UM SISTEMA COMPUTACIONAL PARA ANALISE DE AGRUPAMENTOS

José Eduardo Corrente

Bel. em Matemática

Orientador: Prof. Dr. Cássio Roberto de Melo Godoi

Tese apresentada à Escola Superior de Agricultura "Luiz de Queiroz", da Universidade de São Paulo, para obtenção do título de Doutor em Agronomia, Área de Concentração : Estatística e Experimentação Agronômica.

P I R A C I C A B A

Estado de São Paulo - Brasil

novembro - 1991

Ficha catalográfica preparada pela Seção de Livros da
Divisão de Biblioteca e Documentação - PCAP/USP

C824s Corrente, José Eduardo
 Um sistema computacional para análise de agrupamentos.
 Piracicaba, 1991.
 92p.

Tese - ESALQ
Bibliografia.

1. Análise multivariada - Informática 2. Estatística
aplicada I. Escola Superior de Agricultura Luiz de Quei-
roz, Piracicaba

CDD 519.530285

UM SISTEMA COMPUTACIONAL PARA ANALISE DE AGRUPAMENTOS

JOSÉ EDUARDO CORRENTE

Aprovada em: 02.12.1991

Comissão julgadora:

Prof. Dr. Cássio Roberto de Melo Godoi	ESALQ/USP
Prof ^a . Dr ^a . Maria Cristina Stolf Nogueira	ESALQ/USP
Prof. Dr. Hilton Thadeu Zarate do Couto	ESALQ/USP
Prof. Dr. José Carlos Barbosa	FCAVJ/UNESP
Prof. Dr. Carlos Roberto Padovani	IBBMA/UNESP

Prof. Dr. CÁSSIO ROBERTO DE MELO GODOI
Orientador

À minha família
ofereço

AGRADECIMENTOS

Ao Prof. Dr. Cássio Roberto de Melo Godoi pela orientação deste trabalho.

Ao Fábio Rogero pela elaboração dos programas do sistema.

À Luciane Brajão e Maria Lúcia Coelho pela amizade e incentivo recebidos.

Às amigas Maria del Pilar, Rosemary Ferreira Viégas, Ivani Pozar Otsuk, Licianá V. A. S. Chalita e Roseli Aparecida Leandro pelo companheirismo e apoio constantes.

Aos colegas do Departamento de Matemática e Estatística da ESALQ/USP, professores e funcionários, pela atenção sempre.

A todos que direta ou indiretamente colaboraram para que fosse possível a realização deste trabalho.

SUMÁRIO

RESUMO.....	vi
SUMMARY.....	viii
1. INTRODUÇÃO.....	1
2. REVISÃO DA LITERATURA.....	3
2.1 Medidas de Similaridade e Dissimilaridade.....	5
2.1.1. Medidas de Similaridade e Dissimilaridade para dados quantitativos.....	6
2.1.1.1. Distância Euclideana.....	6
2.1.1.2. Distância Euclideana Média.....	7
2.1.1.3. Distância Generalizada de Mahalanobis.....	7
2.1.1.4. Valor Absoluto.....	9
2.1.1.5. Distância de Minkowski.....	10
2.1.1.6. Coeficiente de Gower.....	10
2.1.1.7. Coeficiente de Similaridade de Cattell.....	10
2.1.1.8. Coeficiente de Cambera.....	11
2.1.1.9. Coeficiente de Correlação.....	11
2.1.2. Medidas de similaridade e dissimilaridade para dados qualitativos.....	12
2.1.2.1. Distância Euclideana Média.....	13
2.1.2.2. Coeficiente de Concordâncias Simples.....	14
2.1.2.3. Coeficiente de Concordâncias Positivas.....	15
2.2. Algoritmos Para Análise de Agrupamento.....	19
2.2.1. Métodos Hierárquicos.....	19
2.2.1.1. Método da Mínima Variância de Ward.....	19

2.2.1.2. Método da Ligação Simples ou do Vizinho Mais Próximo.....	28
2.2.1.3. Método da Ligação Completa ou do Vizinho Mais Distante	34
2.2.1.4. Método da Ligação de Densidades - Método do k-ésimo Vizinho Mais Próximo.....	38
2.2.2. Métodos de Partição.....	44
2.2.2.1. Método "Leading".....	45
2.2.2.2. Método das k-Médias.....	47
3. METODOLOGIA.....	52
3.1. Configuração e Montagem do Sistema.....	52
3.2. Técnicas para Análise de Agrupamentos.....	57
4. RESULTADOS E DISCUSSÃO.....	58
4.1. Descrição do Sistema.....	58
5. CONCLUSÕES.....	89
REFERÊNCIAS BIBLIOGRÁFICAS.....	91

UM SISTEMA COMPUTACIONAL PARA ANÁLISE DE AGRUPAMENTOS

Autor: José Eduardo Corrente

Orientador: Prof. Dr. Cássio R. de M. Godoi

RESUMO

A análise de agrupamentos tem sido utilizada com frequência na experimentação agronômica, principalmente nas áreas de Genética, Melhoramento de Plantas e Animais e Ciências Florestais.

A dificuldade surge, no entanto, da falta de um sistema simples que permita ao pesquisador analisar os dados.

Assim, este trabalho tem o objetivo de apresentar um sistema de análise de agrupamentos, segundo a sugestão de SPÄTH (1980), onde se considera os dados quantitativos, qualitativos nominais e ordinais em separado, construindo uma matriz de distância única.

O sistema, construído em linguagem Pascal, é do estilo interativo de árvores de menus. Possui opções de entrada para os três tipos de dados considerados, que podem ser lidos via teclado ou arquivo texto.

Possui ainda, as opções de medidas, como a Distância Euclideana, a Distância Euclideana Média e a Distância de Minkowsky, para dados quantitativos, e o Coeficiente de Jaccard, o Coeficiente de Concordância Simples, a Distância Euclideana Média e a Distância de Sokal, para dados qualitativos nominais e ordinais.

A matriz de distância única, conforme sugerida por SPÄTH (1980) é construída através da atribuição de um peso para cada matriz de distância parcial, e a seguir, opta-se por um dos métodos de análise, que pode ser o método de partição ou hierárquico.

O método de partição implementado no sistema é chamado método "Leading" e, o hierárquico, é o método do vizinho mais próximo ou da ligação simples.

O sistema também oferece opções de saída dos resultados que podem ser feitas via impressora ou arquivo texto.

A COMPUTACIONAL SYSTEM FOR CLUSTER ANALYSIS

Author: José Eduardo Corrente

Adviser: Cássio R. M. Godoi

SUMMARY

Cluster analysis has been frequently used in Agronomy, mainly in Genetics, Improvement of Plants and Animals and Forest Sciences.

The difficulty arises, however, due to the lack of a simple system that allows the researcher to analyse his data.

Thus, this work intends to introduce a computing system for cluster analysis, according to the suggestion of SPÄTH (1980), where one considers quantitative, nominal and ordinal qualitative data, separately, with single matrix of distances as a result.

The system, constructed in Pascal language, was elaborated in menu tree style. It has options of input for the three kinds of data, that may be read via keyboard or text file.

It still has, the option of measurements as the Euclidean Distance, the Mean Euclidean Distance and the Minkowsky Distance, for the quantitative data, and Jaccard Coeficient, the Simple Concordance Coeficient, the Mean Euclidean Distance and the Sokal Distance, for the nominal and ordinal qualitative data.

The single matrix of distances, according to the sugestion of SPÄTH (1980), is constructed through the atribution of a weight for each distance matrix, and then, choosing one of the analysis methods, that may be the partition or hierarchical method.

The partition method implemented in the system is called "Leading" method and the hierarchical is the nearest neighbour method or single linkage method.

The system also offers output options, that can be done via printer or text file.

1. INTRODUÇÃO

Em muitos casos é comum efetuarem-se medidas de naturezas distintas de um conjunto de unidades amostrais. Isso é frequentemente observado em ensaios biológicos, agrônômicos, florestais, etc, e de posse dessas medidas, objetiva-se encontrar a melhor maneira de descrever os padrões de similaridade ou dissimilaridade mútuas do conjunto de unidades.

Dentre os diferentes métodos que podem ser usados com este propósito, destaca-se a Análise de Agrupamentos.

Assim, dado um conjunto de m observações (objetos, indivíduos, entidades, etc.) medidas segundo n variáveis, o interesse está em obter um esquema que possibilite reunir as unidades em um número de grupos tal que exista homogeneidade dentro e heterogeneidade entre grupos, segundo algum critério conveniente de homogeneidade.

A Análise de Agrupamentos constitui um método multivariado com o objetivo de propor uma estrutura classificatória ou de reconhecimento de grupos.

As técnicas de agrupamentos foram

desenvolvidas por analistas de problemas aplicados, dando ênfase à solução de problemas específicos, ao contrário de outros campos da Análise Multivariada, que se acham teoricamente mais fundamentados. Na análise de agrupamentos, as afirmativas empíricas nem sempre tem respaldo teórico. Muitas técnicas são propostas, mas não há ainda uma teoria generalizada e amplamente aceita.

Para se proceder a uma análise de agrupamentos, existem diversos métodos e diversas propostas de análise, dependendo do tipo de dados considerado.

O uso de técnicas de agrupamento tornou-se freqüente a partir do desenvolvimento da informática que possibilitou o surgimento de algoritmos para agrupar elementos através de critérios de otimização, até então inviáveis, dadas as dificuldades de cálculo.

Assim o presente trabalho tem o objetivo de construir um sistema computacional eficiente, com facilidades de entradas e saídas, com vários tipos de medidas de similaridade ou dissimilaridade entre as observações e alguns dos algoritmos mais usados e aceitos, uma vez que os sistemas disponíveis para esse tipo de análise nem sempre são simples de manipular e também são restritos em termos de opções de métodos de análise.

2. REVISÃO DA LITERATURA

A idéia básica da análise de agrupamentos é que dado o conjunto de observações $x_1, x_2, \dots, x_n$, onde $x_i' = [x_{i1}, x_{i2}, \dots, x_{ip}]'$, devemos adotar algum critério em que se tenha coesão interna dos objetos e isolamento externo entre grupos, isso devido ao fato de não se ter uma definição formal de agrupamentos segundo CORMACK (1971).

Sendo assim, de acordo com BUSSAB *et alii* (1990) pode-se adotar a seguinte classificação para as técnicas de análise de agrupamentos:

1) TÉCNICAS HIERÁRQUICAS - as observações são agrupadas pelas suas distâncias, isto é, tomando-se n agrupamentos como ponto de partida, os indivíduos são agrupados passo a passo até que todos tenham sido agrupados num único grupo. A característica desse processo é que, uma vez alocado um objeto este não poderá ser realocado em outro grupo nos passos seguintes. Os estágios de análise são usualmente representados graficamente por meio de um diagrama de árvore, comumente chamado de dendrograma.

2) TÉCNICAS DE PARTIÇÃO - os agrupamentos obtidos produzem uma partição do conjunto de objetos. Para esta técnica é preciso que se tenha a pré-fixação dos critérios que produzam medidas sobre a qualidade da partição de cada etapa. O uso de métodos de partição pressupõe também o conhecimento do número de componentes da partição desejada.

3) TÉCNICAS DE COBERTURA - os agrupamentos obtidos recobrem o conjunto de objetos, mas podem possuir indivíduos comuns.

As técnicas hierárquicas podem ainda ser subdivididas em dois tipos:

i) Aglomerativas - onde, através de adesões sucessivas dos n objetos, vão sendo obtidos $n-1, n-2, \dots$ grupos, até reunir todos os objetos num único grupo.

ii) Divisivas - partem de um único grupo e, por divisão, vão sendo obtidos 2, 3, etc. grupos.

O primeiro estágio na maioria dos métodos de análise de agrupamentos é, a partir da matriz de dados $n \times p = [X_{ij}]$, a obtenção da matriz S ou D ($n \times n$) de similaridades ou dissimilaridades, medidas das relações entre os pares de indivíduos. Alta similaridade indica que dois indivíduos são próximos em relação ao conjunto de

variáveis, enquanto que alta dissimilaridade indica o contrário.

Temos uma série de medidas de proximidade (similaridade ou dissimilaridade) e a escolha de qual usar dependerá do conjunto de dados disponível e, principalmente, do significado de semelhança que se deseja medir.

2.1. MEDIDAS DE SIMILARIDADE E DISSIMILARIDADE.

Sejam P e Q dois pontos de $\mathbb{R}^r$ do conjunto de objetos de interesse. Uma função real $d(P, Q)$ é uma distância se possui as seguintes propriedades:

- i) $d(P, Q) = d(Q, P)$ (simetria);
- ii) $d(P, Q) \geq 0$;
- iii) $d(P, P) = 0$;
- iv) $d(P, Q) = 0$ se e somente se $P = Q$;
- v) $d(P, Q) \leq d(P, R) + d(R, Q)$ (desigualdade triangular).

Para alguns autores, dentre estes MARDIA et alii (1980), mesmo que (iv) e (v) não estejam satisfeitos, $d(P, Q)$ é ainda chamado de coeficiente de dissimilaridade entre P e Q .

O termo dissimilaridade apareceu em função de que à medida que $d(P, Q)$ cresce, P e Q tornam-se cada vez mais dissimilares.

Desse mesmo modo podemos definir uma medida

de similaridade. Assim, dados dois pontos P e Q , uma medida de similaridade entre P e Q , denotada por $s(P,Q)$, deve satisfazer as seguintes propriedades:

- i) $s(P,Q) = s(Q,P)$;
- ii) $|s(P,Q)| \geq 0$;
- iii) $s(P,Q)$ cresce à medida que a semelhança entre P e Q cresce.

Apresentamos a seguir algumas medidas de similaridades e dissimilaridades, como citado em BUSSAB *et alii* (1990).

2.1.1. MEDIDAS DE SIMILARIDADE E DISSIMILARIDADE PARA DADOS QUANTITATIVOS.

2.1.1.1. DISTÂNCIA EUCLIDEANA.

Considere a matriz X de coordenadas reais, como o descritor dos indivíduos que serão investigados. A medida mais conhecida para indicar a proximidade entre dois indivíduos i e i' é a distância Euclideana, dada por:

$$d_{i,i'} = \left[\sum_{j=1}^p \left(x_{ij} - x_{i'j} \right)^2 \right]^{1/2}$$

ou em linguagem matricial:

$$d_{i,i'} = \left[(x_i - x_{i'})'(x_i - x_{i'}) \right]^{1/2}$$

2.1.1.2. DISTÂNCIA EUCLIDEANA MÉDIA.

A distância Euclideana cresce à medida que cresce o número de caracteres (variáveis). Uma maneira razoável de contornar esse problema é dividir esse valor pela raiz quadrada do número de variáveis, isto é

$$\Delta_{i,i'} = \frac{1}{\sqrt{p}} d_{i,i'}$$

Essa distância é apenas um reescalonamento da distância anterior, possuindo as mesmas propriedades e, portanto, produzindo os mesmos resultados se submetidos às técnicas de Análise de Agrupamentos. Entretanto, esse coeficiente pode ser usado na ausência de dados para algumas variáveis ("missing value").

2.1.1.3. DISTÂNCIA GENERALIZADA DE MAHALANOBIS.

A distância generalizada de Mahalanobis entre as médias amostrais $\underline{i}$ e $\underline{i'}$ é dada matricialmente por

$$D_{i,i'}^2 = \left[\bar{X}_i - \bar{X}_{i'} \right]' S^{-1} \left[\bar{X}_i - \bar{X}_{i'} \right]$$

onde S é a matriz de dispersão amostral comum à todas as unidades, $\bar{X}_i$ e $\bar{X}_{i'}$ são vetores p -dimensionais referentes às médias dos grupos i e i' , $i \neq i'$, $i, i' = 1, 2, \dots, n$. Pode-se também usar a matriz de correlação amostral R no lugar da matriz de dispersão S .

Essa distância, além de ponderar pela variabilidade de cada uma das componentes, leva em conta também o grau de correlação entre elas. Este fato torna muito difícil a interpretação dos resultados obtidos.

Exemplo. Suponha que pontos são medidos segundo o vetor (x_1, x_2) com média $(0, 0)$ e matriz de correlações dada por:

$$R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

e são dados os pontos $x_i = (0, 0)$ e $x_{i'} = (1, 1)$.

A distância de Mahalanobis entre os indivíduos i e i' é:

$$D_{i,i'}^2 = \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right]' R^{-1} \left[\begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right]$$

onde

$$R^{-1} = \frac{1}{1 - r^2} \begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix}$$

Efetuada-se o produto, temos que:

$$D_{i,i'}^2 = \frac{2}{1 + r} \Rightarrow D_{i,i'} = \sqrt{\frac{2}{1 + r}}$$

Analizando o valor de $D_{i,i'}$, observamos que:

$$D_{i,i'} = \begin{cases} 1 & \text{se } r = 1 \\ \sqrt{2} & \text{se } r = 0 \\ +\infty & \text{se } r \rightarrow -1 \end{cases}$$

Isso mostra que contrariamente à distância Euclideana, a distância de Mahalanobis depende das correlações entre as variáveis.

2.1.1.4. VALOR ABSOLUTO.

Em vez dos desvios quadráticos, é muito comum o uso do valor absoluto:

$$d_{i,i'} = \sum_{j=1}^p w_j |x_{ij} - x_{i'j}|$$

onde os w_j 's representam as ponderações para as variáveis. Os valores mais usados são os $w_j = 1$ ou $w_j = \frac{1}{p}$. Essa

medida é conhecida como métrica "city-block".

2.1.1.5. DISTÂNCIA DE MINKOWSKI.

Uma generalização da distância Euclideana é dada por:

$$d_{i,i'} = \left[\sum_{j=1}^p w_j |x_{ij} - x_{i'j}|^k \right]^{1/k}$$

onde os w_j 's representam as ponderações para as variáveis e k é um inteiro qualquer.

2.1.1.6. COEFICIENTE DE GOWER.

Esse coeficiente é baseado na proporção da variação relativa à maior discrepância possível

$$d_{i,i'} = -\log_{10} \left[1 - \frac{1}{p} \sum_{j=1}^p \frac{|x_{ij} - x_{i'j}|}{\max(x_j) - \min(x_j)} \right]$$

onde temos dentro do somatório um tipo de padronização das variáveis.

2.1.1.7. COEFICIENTE DE SIMILARIDADE DE CATTEL

O coeficiente de similaridade de Cattell é dado por

$$c_{i,i'} = \frac{2 \left[p - \frac{2}{3} \right] - d_{i,i'}^2}{2 \left[p - \frac{2}{3} \right] + d_{i,i'}^2}$$

onde $d_{i,i'}^2$ é a distância Euclideana com variáveis padronizadas.

Uma outra forma dessa medida é devida a Cattell e Coutter (CORMACK, 1971), onde é feita uma outra correção

$$c_{i,i'} = \frac{\sqrt{2p} - d_{i,i'}^2}{\sqrt{2p} + d_{i,i'}^2}$$

2.1.1.8. COEFICIENTE DE CAMBERRA.

Fornece a proporção da diferença em relação ao total

$$d_{i,i'} = \frac{1}{p} \sum_{j=1}^p \frac{|x_{ij} - x_{i'j}|}{x_{ij} + x_{i'j}}$$

2.1.1.9. COEFICIENTE DE CORRELAÇÃO

Dentre os coeficientes de similaridade o mais empregado é o coeficiente de correlação de Pearson definido

por:

$$r_{i,i'} = \frac{\sum_{j=1}^P (x_{i,j} - \bar{x}_i)(x_{i',j} - \bar{x}_{i'})}{\sqrt{\left[\sum_{j=1}^P (x_{i,j} - \bar{x}_i)^2 \right] \left[\sum_{j=1}^P (x_{i',j} - \bar{x}_{i'})^2 \right]}}$$

$$\text{onde } \bar{x}_i = \frac{1}{P} \sum_{j=1}^P x_{i,j} \text{ e } \bar{x}_{i'} = \frac{1}{P} \sum_{j=1}^P x_{i',j}.$$

2.1.2. MEDIDAS DE SIMILARIDADE E DISSIMILARIDADE PARA DADOS QUALITATIVOS.

Ainda, de acordo com BUSSAB *et alii* (1990), como é frequente o uso de critérios qualitativos na procura de elementos semelhantes, tem-se a necessidade de obter coeficientes que definam o grau de similaridade ou dissimilaridade entre objetos, segundo variáveis desse tipo.

Por simplicidade, iniciaremos pelo caso onde as variáveis envolvidas são do tipo binário. Neste caso, as propostas de coeficientes de proximidade são muitas. Praticamente, qualquer medida de associação para as chamadas tabelas de contingência pode ser usadas como medida de proximidade.

Consideremos então a seguinte tabela de

contingência 2×2 , com dois indivíduos, i e i' , onde temos variáveis apenas do tipo binário:

		indivíduo i		total
		1	0	
indivíduo i'	1	$a_{i'i}$	$b_{i'i}$	$a_{i'i} + b_{i'i}$
	0	$c_{i'i}$	$d_{i'i}$	$c_{i'i} + d_{i'i}$
		$a_{i'i} + c_{i'i}$	$b_{i'i} + d_{i'i}$	$p = a_{i'i} + b_{i'i} + c_{i'i} + d_{i'i}$

onde,

$a_{i'i}$: número de variáveis com valor 1 para os dois indivíduos (concordantes).

$b_{i'i}$: número de variáveis com valor 1 para o indivíduo i' e 0 para o indivíduo i (discordantes).

$c_{i'i}$: número de variáveis com valor 0 para o indivíduo i' e 1 para o indivíduo i (discordantes).

$d_{i'i}$: número de variáveis com o valor 0 para os dois indivíduos (concordantes).

2.1.2.1. DISTÂNCIA EUCLIDEANA MÉDIA.

Nesta situação particular, pode se calcular a distância Euclideana média entre os indivíduos i e i' , como

sendo:

$$d_{ii} = \left[\frac{1}{p} \sum_{j=1}^p (x_{ij} - x_{ij'})^2 \right]^{1/2} = \left(\frac{b_{i'i} + c_{i'i}}{p} \right)^{1/2} =$$

$$= \left(\frac{b_{i'i} + c_{i'i}}{a_{i'i} + b_{i'i} + c_{i'i} + d_{i'i}} \right)^{1/2}.$$

Desse modo, vemos que essa medida indica a proporção de atributos não coincidentes nos dois indivíduos. Quanto maior esse número, maior a diferença entre os indivíduos, sendo portanto uma medida de dissimilaridade. Pode-se verificar que a sua amplitude de variação é entre 0 e 1.

2.1.2.2. COEFICIENTE DE CONCORDÂNCIA SIMPLES.

Seja

$$s_{ii} = \frac{a_{i'i} + d_{i'i}}{a_{i'i} + b_{i'i} + c_{i'i} + d_{i'i}}$$

então os valores altos de s correspondem a maior similaridade entre os objetos. Este coeficiente também varia entre 0 e 1.

2.1.2.3. COEFICIENTE DE CONCORDÂNCIAS POSITIVAS.

Às vezes deseja-se medir a similaridade baseando-se apenas na presença da característica, e não na ausência. Isso nos leva a:

$$S_{i,i} = \frac{a_{i,i}}{a_{i,i} + b_{i,i} + c_{i,i} + d_{i,i}}$$

ou um outro mais usado, conhecido como COEFICIENTE DE JACCARD:

$$S_{i,i} = \frac{a_{i,i}}{a_{i,i} + b_{i,i} + c_{i,i}}$$

que leva em conta apenas a presença da característica, uma vez que (0,0) seria coincidência, mas na ausência.

ROMESBURG (1984) e BUSSAB *et alii* (1990), fornecem uma tabela de outras medidas que podem ser usadas para medir similaridade ou dissimilaridade entre variáveis binárias.

Quando a variável qualitativa possui mais de dois níveis, o artifício usual é transformação em variáveis binárias através da criação de variáveis "dummy" e aplicar os coeficientes vistos anteriormente.

Segundo BUSSAB *et alii* (1990), podemos supor o vetor de critérios qualitativos

$$y = (y_1, y_2, \dots, y_r)$$

onde o i -ésimo componente assume r_i níveis, codificados de modo que:

$$y_i = j \text{ com } j = 1, 2, \dots, r_i.$$

Suponha também que $\sum r_i = p$. Cada componente irá dar origem a r_i variáveis binárias $x_k(i)$ tal que:

$$x_k(i) = \begin{cases} 1 & \text{se } y_i = k \\ 0 & \text{em caso contrário.} \end{cases}$$

Assim, o vetor y de dimensão $1 \times r$ será transformado no vetor x de dimensão $1 \times p$, formado por componentes binárias. Esquematicamente, tem-se:

$$y = (y_1, y_2, \dots, y_r) \longrightarrow$$

$$+ \quad x = (\underbrace{0, \dots, 1, \dots, 0}_{i_1}^{x_k(i)} ; \dots ; \underbrace{0, \dots, 1, \dots, 0}_{i_r}^{x_k(r)}).$$

Sem perda de generalidade, o vetor x' será indicado por p coordenadas binárias x_i , isto é:

$$x' = (x_1, x_2, \dots, x_p)$$

e tem-se a situação anterior.

MANZATO (1983) fornece uma outra maneira de se analisar dados obtidos em $\underline{s}$ categorias de populações multinomiais, baseado em proporções.

Para o caso de variáveis qualitativas do tipo ordinal, usa-se a mesma estratégia vista anteriormente, porém deve-se levar em conta a ordem.

É bastante comum na experimentação a necessidade de diferentes tipos de critérios na definição da proximidade entre objetos. Sem perda de generalidade, as variáveis podem ser reagrupadas de modo que apareçam primeiro as variáveis qualitativas nominais, depois as variáveis qualitativas ordinais e finalmente as variáveis quantitativas. Esquemáticamente, tem-se:

$$y' = (y_1, \dots, y_p)' = (y_n, y_o, y_q)'$$

SPÄTH (1980) e ROMESBURG (1984) fornecem detalhes sobre como proceder a uma análise para este tipo de situação. Alguns pontos que podemos ressaltar, são:

a) constróem-se coeficientes de proximidade, c_n , c_o e c_q para cada um dos três vetores y_n , y_o e y_q ;

b) em seguida, constróem-se um único coeficiente ponderado, dado por

$$c(i,i') = w_1 c_n(i,i') + w_2 c_o(i,i') + w_3 c_q(i,i')$$

onde os w_i 's são escolhidos convenientemente. A construção deste coeficiente exige alguns cuidados especiais como:

i) os coeficientes de semelhança do mesmo sentido (similaridade ou dissimilaridade);

ii) intervalos de variação dos coeficientes aproximadamente iguais;

iii) conjunto de pesos adequados e justificáveis.

No caso (iii), o que tem sido usado é ponderar pelo número de variáveis envolvidas. Mas se um tipo de variável for mais importante, a ponderação pode melhorar muito os resultados.

Utilizando-se as medidas de semelhança abordadas acima, vários algoritmos para análise de agrupamento tem sido propostos e muitos já estão implementados em pacotes computacionais, como SAS, BMDP, SPSS, etc.

2.2. ALGORÍTMOS PARA ANÁLISE DE AGRUPAMENTO.

Vejamos a seguir alguns dos métodos de análise de agrupamento.

2.2.1. MÉTODOS HIERÁRQUICOS.

2.2.1.1. MÉTODO DA MÍNIMA VARIÂNCIA DE WARD.

Este método, proposto por Ward¹ e citado por SAS INSTITUTE (1987), é um método hierárquico e segue uma série de passos começando com n agrupamentos e terminando com apenas um, contendo todos os objetos. Em cada passo, o método une quaisquer dois agrupamentos que resulta no menor crescimento do valor de um índice E , chamado de índice da soma de quadrados, ou da variância. Isto significa que, para cada passo, verifica-se todos os possíveis pares de agrupamentos, calcula-se o valor de E e seleciona-se aquele par com menor valor de E .

Para cada tentativa de agrupamento, E é calculado como segue:

- 1) Calcula-se a média de cada agrupamento;
- 2) Calcula-se a distância entre cada objeto de um dado agrupamento e a média deste agrupamento;

¹WARD, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, Alexandria, 58: 236-44. 1963.

3) Calcula-se a soma de quadrados das distâncias obtendo a soma de quadrados "dentro" de cada agrupamento;

4) Calcula-se o valor de E adicionando-se as somas de quadrados de cada agrupamento.

Este método é mais simples quando aplicado diretamente à matriz de dados quantitativos.

Para ilustrar o método, consideremos o seguinte exemplo:

Exemplo. Suponha que temos 5 parcelas de terra e dois atributos que descrevem importantes propriedades:

i) capacidade de retenção da água ("water-holding"), expressa em porcentagem;

ii) porcentagem do peso de matéria orgânica.

O objetivo é encontrar as parcelas mais similares, de acordo com esses atributos. Os dados são:

parcelas	atributos	
	1	2
1	10	5
2	20	20
3	30	10
4	30	15
5	5	10

Aplicando o método de Ward para encontrarmos os agrupamentos, comecemos com o passo zero, onde temos 5

agrupamentos e os valores médios para cada agrupamento serão os próprios valores:

$$\begin{aligned}
 E = & \underbrace{(10-10)^2 + (5-5)^2}_{\text{grupo 1}} + \underbrace{(20-20)^2 + (30-30)^2}_{\text{grupo 2}} + \\
 & \underbrace{(30-30)^2 + (10-10)^2}_{\text{grupo 3}} + \underbrace{(30-30)^2 + (15-15)^2}_{\text{grupo 4}} + \\
 & \underbrace{(5-5)^2 + (10-10)^2}_{\text{grupo 5}} = 0
 \end{aligned}$$

Passo 1.

Precisamos calcular E para todas as partições possíveis usando-se dois atributos. As partições possíveis estão no Quadro 1.

Quadro 1 - Partições possíveis considerando-ser agrupamentos para duas parcelas.

Passo	Possíveis partições				E
1	(12)	3	4	5	162.5
	(13)	2	4	5	212.5
	(14)	2	3	5	250.0
	(15)	2	3	4	25.0
	(23)	1	4	5	100.0
	(24)	1	3	5	62.5
	(25)	1	3	4	162.5
	(34)	1	2	5	12.5 *
	(35)	1	2	4	312.5
	(45)	1	2	3	325.0

Para o primeiro caso, onde temos as formações (12), 3, 4 e 5, calculamos a média dos agrupamento (12). Para o 1º atributo, esse valor é $(10+20)/2=15.0$ e para o 2º atributo é $(5+20)/2=12.5$. Assim:

$$\begin{aligned}
 E = & \underbrace{(10-15)^2 + (5-12.5)^2 + (20-15)^2 + (20-12.5)^2}_{\text{grupo (12)}} + \\
 & + \underbrace{(30-30)^2 + (10-10)^2}_{\text{grupo 3}} + \underbrace{(30-30)^2 + (15-15)^2}_{\text{grupo 4}} + \\
 & + \underbrace{(5-5)^2 + (10-10)^2}_{\text{grupo 5}} = 162.5
 \end{aligned}$$

Continuando esse processo, encontramos o valor de E para todas as possíveis partições, cujos valores constam no Quadro 1.

Como o menor valor de E é 12.5, temos como resultado os agrupamentos 1, 2, (34) e 5, unindo-se os grupos 3 e 4.

Passo 2.

Unindo-se 3 e 4, verificamos as outras possíveis duas uniões, considerando os agrupamentos restantes. Neste caso os resultados estão no Quadro 2.

Quadro 2 - Possíveis partições, considerando o grupo (34).

Passo	Possíveis partições			E
2	(34)	(12)	5	175.0
	(34)	(15)	2	37.5 *
	(34)	(25)	1	175.0
	(134)	2	5	316.7
	(234)	1	5	116.7
	(345)	1	2	433.3

Considerando o primeiro caso, onde temos as possibilidades (34), (12) e 5, temos:

$$\text{média de (34)} : \begin{cases} (30+30)/2 = 30 & \text{para o atributo 1} \\ (10+15)/2 = 12.5 & \text{para o atributo 2} \end{cases}$$

$$\text{média de (12)} : \begin{cases} (10+20)/2 = 15 & \text{para o atributo 1} \\ (5+20)/2 = 12.5 & \text{para o atributo 2} \end{cases}$$

Então,

$$E = \underbrace{(30-30)^2 + (10-12.5)^2 + (30-30)^2 + (15-12.5)^2}_{\text{grupo (34)}} +$$

$$\underbrace{(10-15)^2 + (5-12.5)^2 + (20-15)^2 + (20-12.5)^2}_{\text{grupo (12)}} +$$

$$+ \underbrace{(5-5)^2 + (10-10)^2}_{\text{grupo 5}} = 175.0$$

Analogamente calcula-se E para as outras possíveis partições obtendo-se $E = 37.5$ como o menor valor. Temos então os agrupamentos (34), (15) e 2 para esse passo.

Passo 3.

Neste passo as possíveis partições constam no Quadro 3.

Quadro 3 - Possíveis partições considerando-se os grupos (34) e (15).

Passo	Possíveis partições		E
3	(234)	(15)	141.7 *
	(125)	(34)	245.9
	(1345)	2	568.8

Para o primeiro caso, onde temos as partições (234) e (15), temos:

$$\text{média de (234)} : \begin{cases} (20+30+30)/3 = 26.7 & \text{para o atributo 1} \\ (20+10+15)/3 = 15 & \text{para o atributo 2} \end{cases}$$

$$\text{média de (15)} : \begin{cases} (10+5)/2 = 7.5 & \text{para o atributo 1} \\ (5+0)/2 = 7.5 & \text{para o atributo 2} \end{cases}$$

Assim:

$$E = \underbrace{(20-26.7)^2 + (20-15)^2 + (30-26.7)^2 + (10-15)^2 + (30-26.7)^2 + (15-15)^2}_{\text{grupo (234)}} +$$

$$\underbrace{(10-7.5)^2 + (5-7.5)^2 + (5-7.5)^2 + (10-7.5)^2}_{\text{grupo (15)}} +$$

$$= 141.7$$

Logo, unimos 2 com (34), fornecendo os agrupamentos (234) e (15).

Passo 4.

Neste passo só é possível uma união que é (12345). Assim temos:

$$\text{média de (12345): } \begin{cases} (10+20+30+30+5)/5=19 \text{ p/ o atributo 1} \\ (5+20+10+15+10)/5=12 \text{ p/ o atributo 2} \end{cases}$$

e o valor de E é:

$$E = (10-19)^2 + (5-12)^2 + (20-19)^2 + (20-12)^2 + (30-19)^2 + \\ + (10-12)^2 + (30-19)^2 + (15-12)^2 + (5-19)^2 + (10-12)^2 = 650$$

O resultado obtido pode ser representado num dendrograma, conforme a Figura 1.

Figura 1. Árvore produzida pelo método de Ward aos dados de solos.

Podemos observar que E cresce não-linearmente. No exemplo, temos que E cresce em progressão: 0.0, 12.5, 37.5, 141.7, 650. Isto fornece uma árvore bem definida, onde os agrupamentos saltam aos olhos. Contudo isto não é uma significância real pois E é uma função do quadrado de diferenças e tende a crescer não-linearmente quando o número de agrupamentos torna-se menor. Por isso, pode-se construir a árvore considerando a raiz quadrada de E , sugestão que não é muito praticada.

Andenberg¹ citado por ROMESBURG (1984) faz várias considerações sobre o método de Ward. O índice E é um coeficiente de dissimilaridade sensível a translações aditiva e proporcionais aos dados.

O segundo ponto é que o método de Ward não garante uma partição ótima dos objetos em agrupamentos, isto é, podem existir outras partições eventualmente melhores que a obtida por este método.

¹ANDENBERG, M.R. Cluster analysys for aplications.
New York, Academic Press. 1973.

2.2.1.2. MÉTODO DA LIGAÇÃO SIMPLES OU DO VIZINHO MAIS PRÓXIMO.

Neste método, proposto por Florek *et alii*² e mais tarde reutilizado por McQuitty³ e Sneath e Sokal⁴, conforme citado por SAS INSTITUTE (1987), a distância entre dois agrupamentos é a distância mínima entre eles.

MARDIA *et alii* (1979) fornece uma descrição do algoritmo para agrupar indivíduos vizinhos, do seguinte modo:

a) Ordenam-se os $\frac{1}{2} n(n+1)$ pontos da matriz de distâncias em ordem crescente, uma vez que, tendo-se uma matriz de dados $n \times p$, a matriz de distâncias terá dimensão $n \times n$, e simétrica;

²FLOREK, K.; LUKASZEWICZ, J.; PERKAL, J.; ZUBRZYCKI, S. Sur la liasion et la division des point d'un ensemble fini. Colloquium Mathematical, Warsaw, 2: 282-5, 1951a.

²FLOREK, K. ; LUKASZEWICZ, J. ; PERKAL, J.; ZUBRZYCKI, S. Taksonomia Wroclawska. Przegląd Antropologiczny, Warsaw, 17: 193-211, 1951b.

³MCQUITTY, L.L. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevance. Educational and Psychological Measurement, 17: 207-29, 1957

⁴SNEATH, P.H.A. e SOKAL, R.R. Numerical taxonomy; the principles and practice of numerical classification. San Francisco. Freeman, 1973. 573p.

b) Seja $C_1, C_2, \dots, C_m$ o agrupamento inicial, contendo um ponto, isto é $x_i \in C_i$;

c) Sem qualquer perda de generalidade, seja

$$d_{r_1 s_1} = \min \{r, s\}$$

tal que x_{r_1} e x_{s_1} são vizinhos. Então, dois pontos são agrupados e temos $m-1$ agrupamentos, onde $C_{r_1} + C_{s_1}$ é um novo agrupamento;

d) Seja $d_{r_2 s_2}$ a próxima menor distância. Se nem r_1 e s_1 são iguais a r_2 e s_2 , respectivamente, os novos agrupamentos são $C_{r_1} + C_{s_1}$, $C_{r_2} + C_{s_2}$ mais os restantes. Se $r_2 = r_1$ e $s_2 \neq s_1$ os $m-2$ novos agrupamentos são $C_{r_1} + C_{s_1} + C_{s_2}$, mais os restantes.

e) O processo termina quando todos os objetos estiverem agrupados em um único grupo.

Pode-se observar que este algoritmo usa a relação de recorrência:

$$d_{k(i,j)} = \min \{ d_{ik}, d_{jk} \}.$$

Exemplo. Considerando ainda o exemplo inicial, temos que a matriz de distâncias, usando a distância Euclideana é dada por:

$$D_1 = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{ccccc} & 1 & 2 & 3 & 4 & 5 \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \left[\begin{array}{ccccc} 0 & 18.00 & 20.60 & 22.40 & 7.07 \\ & 0 & 14.10 & 11.20 & 18.00 \\ & & 0 & 5.00 & 25.00 \\ & & & 0 & 25.50 \\ & & & & 0 \end{array} \right] \end{array} .$$

Passo 1.

Como $d(3,4) = 5.00$ é a menor distância, unimos 3 e 4.

Passo 2.

Calculemos a nova matriz de distâncias, usando a relação de recorrência:

$$D_2 = \begin{array}{c} 1 \\ 2 \\ (34) \\ 5 \end{array} \begin{array}{ccccc} & 1 & 2 & (34) & 5 \\ \begin{array}{c} 1 \\ 2 \\ (34) \\ 5 \end{array} & \left[\begin{array}{ccccc} 0 & 18.00 & 20.60 & 7.07 \\ & 0 & 11.20 & 18.00 \\ & & 0 & 25.00 \\ & & & 0 \end{array} \right] \end{array} .$$

onde:

$$d_{1(34)} = \min \{d_{13}, d_{14}\} = \min \{20.60, 22.40\} = 20.60$$

$$d_{2(34)} = \min \{d_{23}, d_{24}\} = \min \{14.10, 11.20\} = 11.20$$

$$d_{5(34)} = \min \{d_{53}, d_{54}\} = \min \{25.00, 25.50\} = 25.00.$$

Passo 3

Como $d(1,5) = 7.07$ é a menor distância, unimos 1 e 5.

Passo 4.

Calculemos a nova matriz de distâncias:

$$D_3 = \begin{matrix} & \begin{matrix} (15) & 2 & (34) \end{matrix} \\ \begin{matrix} (15) \\ 2 \\ (34) \end{matrix} & \begin{bmatrix} 0 & 18.00 & 20.60 \\ & 0 & 11.20 \\ & & 0 \end{bmatrix} \end{matrix}$$

onde:

$$d_{2(15)} = \min \{d_{12}, d_{25}\} = \min \{18.00, 18.00\} = 18.00$$

$$d_{(15)(34)} = \min \{d_{1(34)}, d_{5(34)}\} = \min \{20.60, 25.00\} = 20.60$$

Passo 5

Como $d(2,(34)) = 11.20$ é a menor distância unimos 2 a (34).

Passo 6.

Calculando a nova matriz de distâncias, temos:

$$D_4 = \begin{matrix} & \begin{matrix} (15) & (234) \end{matrix} \\ \begin{matrix} (15) \\ (234) \end{matrix} & \begin{bmatrix} 0 & 18.00 \\ & 0 \end{bmatrix} \end{matrix}$$

onde:

$$d_{(15)(234)} = \min \{d_{2(15)}, d_{(15)(34)}\} = \min \{18.00, 20.60\} = 18.00$$

Passo 7.

Unimos (15) com (234) cuja distância é 18.00.

Temos então o dendrograma, que está na Figura 2.

Figura 2. Árvore produzida pelo método do vizinho mais próximo aos dados de solos.

FISHER e VON NESS (1971) e HARTIGAN (1974) citam propriedades teóricas para o método da ligação simples, mas em estudos de simulação, não tem dado bons

resultados. Por não se considerar nenhuma restrição na forma dos agrupamentos, a ligação simples sacrifica sua execução, detectando alongamentos ou agrupamentos irregulares.

2.2.1.3. MÉTODO DA LIGAÇÃO COMPLETA OU DO VIZINHO MAIS DISTANTE.

O método da ligação completa, desenvolvido por Sorensen¹ e citado por SAS INSTITUTE (1987), é semelhante à ligação simples, exceto que a distância entre dois agrupamentos é agora definida como a maior distância entre pares de elementos em cada agrupamento.

MARDIA *et alii* (1979) também fornece um algoritmo para este método, que é análogo ao método da ligação simples. Aqui, usa-se a relação de recorrência:

$$d_{k(ij)} = \max(d_{jk}, d_{ik})$$

para definir a distância entre o grupo k e o grupo (ij) formado pela fusão dos grupos i e j .

Exemplo. Considerando ainda o exemplo anterior, temos que a matriz de distâncias é dada por:

$$D_1 = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \left[\begin{array}{ccccc} 0 & 18.00 & 20.60 & 22.40 & 7.07 \\ & 0 & 14.10 & 11.20 & 18.00 \\ & & 0 & 5.00 & 25.00 \\ & & & 0 & 25.50 \\ & & & & 0 \end{array} \right] \end{matrix}$$

¹SORENSEN, T. A method of establishing groups of equal amplitude in analysis of vegetation on danish commons. Biological Skrifter, 5: 1-34, 1948.

Aplicando o método da ligação completa, temos:

Passo 1.

Sendo $d(3,4) = 5.00$ a menor distância, unimos 3 e 4.

Passo 2.

Calculando a nova matriz de distâncias, temos:

$$D_2 = \begin{matrix} & \begin{matrix} 1 & 2 & (34) & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ (34) \\ 5 \end{matrix} & \begin{bmatrix} 0 & 18.00 & 22.40 & 7.07 \\ & 0 & 14.10 & 18.00 \\ & & 0 & 25.50 \\ & & & \bullet \end{bmatrix} \end{matrix}$$

onde:

$$d_{1(34)} = \max \{d_{13}, d_{14}\} = \max \{20.80, 22.40\} = 22.40$$

$$d_{2(34)} = \max \{d_{23}, d_{24}\} = \max \{14.10, 11.20\} = 14.10$$

$$d_{5(34)} = \max \{d_{53}, d_{54}\} = \max \{25.00, 25.50\} = 25.50.$$

Passo 3.

Sendo $d(1,5) = 7.07$ a menor distância, unimos 1 e 5.

Passo 4.

Calculando a nova matriz de distâncias, temos:

$$D_3 = \begin{matrix} & \begin{matrix} (15) & 2 & (34) \end{matrix} \\ \begin{matrix} (15) \\ 2 \\ (34) \end{matrix} & \begin{bmatrix} 0 & 18.00 & 25.50 \\ & 0 & 14.10 \\ & & 0 \end{bmatrix} \end{matrix}$$

onde:

$$d_{2(15)} = \max \{d_{12}, d_{25}\} = \max \{18.00, 18.00\} = 18.00$$

$$d_{(15)(34)} = \max \{d_{1(34)}, d_{5(34)}\} = \max \{22.40, 25.50\} = 25.50.$$

Passo 5.

Como $d(2, (34)) = 14.10$ é a menor distância, unimos 2 a (34).

Passo 6.

Calculando a nova matriz de distâncias, temos:

$$D_4 = \begin{matrix} & \begin{matrix} (15) & (234) \end{matrix} \\ \begin{matrix} (15) \\ (234) \end{matrix} & \begin{bmatrix} 0 & 25.50 \\ & 0 \end{bmatrix} \end{matrix}$$

onde:

$$d_{(15)(234)} = \max \{d_{2(15)}, d_{(15)(34)}\} = \max \{18.00, 25.50\} = 25.50.$$

Passo 7.

Unimos (15) a (234) formando um único agrupamento, cuja distância é $d((15), (234)) = 25.50$.

O dendrograma está na Figura 3.

Figura 3. Árvore produzida pelo método do vizinho mais distante aos dados de solos dos de solos.

Este método produz agrupamentos de diâmetros aproximadamente iguais, podendo ser mal interpretado na presença de dados discrepantes.

2.2.1.4. MÉTODO DA LIGAÇÃO DE DENSIDADES - MÉTODO DO k-ÉSIMO VIZINHO MAIS PRÓXIMO.

Esse método, desenvolvido por WONG & LANE (1983) é utilizado para formação de agrupamentos de alta densidade, podendo ser aplicado à matriz de dados ou à matriz de dissimilaridades.

Considerando as observações $x_1, \dots, x_n$ como uma amostra aleatória com valores em $\mathbb{R}^p$, $p \geq 1$ com densidade f , a proposta deste método é descobrir agrupamentos de alta densidade, a partir da densidade f , consistindo de dois estágios: no primeiro estágio, o método do k-ésimo vizinho mais próximo é usado para obter um estimador f_n para a densidade subjacente f , onde a densidade estimada no ponto x é definida como

$$f_n(x) = \frac{k}{n V_k(x)} \quad (1)$$

onde $V_k(x)$ é o volume da menor esfera centrada em x contendo k observações amostrais.

Denyone e Wagner¹, citado em WONG & LANE (1983), mostraram que $f_n(x)$ definido como acima é um

¹DENYONE, L.P. e WAGNER, T.J. The strong uniform consistency of nearest neighbours density estimates. *Annals of Mathematical and Statistical*. 5: 536-40. 1977.

estimador uniformemente consistente para f .

Os agrupamentos de alta-densidade definidos sobre esta estimativa uniformemente consistente da densidade f_n , pode ser usada como um estimador amostral dos agrupamentos de alta-densidade definida sobre f . Estes agrupamentos amostrais de alta-densidade são construídos no segundo estágio deste algoritmo.

No segundo estágio, a matriz D de distâncias $d(x_i, x_j)$, $1 \leq i, j \leq n$, para n observações é calculada considerando as seguintes definições:

Def.1. Duas observações x_i e x_j são ditas vizinhas se $d^*(x_i, x_j) \leq d_k(x_i)$ ou $d_k(x_j)$, onde $d^*(\dots)$ é a distância Euclideana e $d_k(x_i)$ é a distância do k -ésimo vizinho mais próximo ao ponto x_i .

Def.2. A distância $d(\dots)$ entre as observações x_i e x_j é dada por:

$$d(x_i, x_j) = \begin{cases} \frac{1}{2} \left[\frac{1}{f_N(x_i)} + \frac{1}{f_N(x_j)} \right] & \text{se } x_i \text{ e } x_j \text{ são vizinhos} \\ +\infty & \text{em c. c.} \end{cases}$$

Usando (1) da página anterior, temos que:

$$d(x_i, x_j) = \begin{cases} \frac{N}{2k} \left[V_k(x_i) + V_k(x_j) \right] & \text{se } x_i \text{ e } x_j \text{ são vizinhos} \\ +\infty & \text{em c. c.} \end{cases}$$

Portanto, distâncias finitas são definidas somente para pares de observações que estão na mesma vizinhança em $\mathbb{R}^p$. Em seguida aplica-se o método da ligação simples à essa matriz de distâncias para se obter os agrupamentos de alta densidade.

Desde que os agrupamentos de alta densidade são invariantes por transformação monótona da função de densidade, as distâncias do k -ésimo vizinho mais próximo $d_k(x_i)$, $i=1, \dots, n$, são usadas ao invés dos $V_k(x_i)$'s no seguinte algoritmo:

Passo 1. Para $i=1, \dots, n$, calcula-se $d_k(x_i)$, a distância do k -ésimo vizinho mais próximo de x_i .

Passo 2. Calcule a matriz de distâncias D como segue:

$$d(x_i, x_j) = \begin{cases} \frac{1}{2} \left[d_k(x_i) + d_k(x_j) \right] & \text{se } d^*(x_i, x_j) \leq d_k(x_i) \\ & \text{ou } d^*(x_i, x_j) \leq d_k(x_j) \\ +\infty & \text{em c. c.} \end{cases}$$

onde d^* é a distância Euclideana.

Passo 3. Aplicar o método da ligação simples à matriz D para obter os agrupamentos.

Este procedimento não é aplicável a grandes conjuntos de dados, mas é melhor aceito para pequenas amostras.

Exemplo. Considerando ainda os dados de solos, temos que a matriz de distâncias, usando-se a distância Euclideana é dada por:

$$D_1 = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0 & 18.00 & 20.60 & 22.40 & 7.07 \\ & 0 & 14.10 & 11.20 & 18.00 \\ & & 0 & 5.00 & 25.00 \\ & & & 0 & 25.50 \\ & & & & 0 \end{bmatrix} \end{matrix} .$$

Utilizando o sistema SAS, executou-se a análise para essa matriz de distâncias, obtendo-se os resultados, que constam das Tabelas 1 e 2 considerando 2 e 3 grupos.

Tabela 1 - Resultado obtido usando o programa SAS, para o método de ligação de densidades, aplicado aos dados de solos, considerando-se 2 agrupamentos.

SAS 19:34 Thursday, September 12, 1991

9

Density Linkage Cluster Analysis

K = 2

Number of Clusters	Clusters Joined		Frequency of New Cluster	Fusion Density	Maximum Density in Each Cluster	
					Lesser	Greater
4	parcela2	parcela5	2	0.0400	0.0400	0.0400
3	parcela3	parcela4	2	0.0400	0.0400	0.0400
2	CL4	parcela1	3	0.0331	0.0223	0.0400

2 model clusters have been formed

Tabela 2. - Resultado obtido usando o programa SAS, para o método de ligação de densidades, aplicado aos dados de solos, considerando-se 3 agrupamentos.

SAS 19:34 Thursday, September 12, 1991

13

Density Linkage Cluster Analysis

K = 3

Number of Clusters	Clusters Joined		Frequency of New Cluster	Fusion Density	Maximum Density in Each Cluster	
					Lesser	Greater
4	parcela2	parcela5	2	0.0328	0.0268	0.0424
3	CL4,	parcela4	2	0.0268	0.0268	0.0424
2	CL3	parcela1	4	0.0239	0.0167	0.0424
1	CL2	parcela3	5	0.0237	0.0213	0.0424

One model cluster has been formed

Os dendrogramas produzidos foram reproduzidos nas figuras 4 e 5.

Figura 4. Dendrograma obtido pelo método da ligação de densidades produzido pela análise de dados solos, para 2 grupos.

Figura 5. Dendrograma obtido pelo método da ligação de densidades produzido pela análise de dados solos, para 3 grupos.

2.2.2) MÉTODOS DE PARTIÇÃO.

As técnicas de partição caracterizam-se pela possibilidade de separar os dados em grupos disjuntos, de modo a se obter maior homogeneidade possível dentro dos grupos e maior heterogeneidade entre grupos.

Assim, uma partição de tamanho m de n objetos, é uma coleção de conjuntos não-vazios C_j , $j=1,2,\dots,n$, tal que:

$$i) C_1 \cup C_2 \cup \dots \cup C_m = \{x_1, x_2, \dots, x_n\}.$$

$$ii) C_j \cap C_k = \emptyset \text{ para } j \neq k.$$

de modo que em cada C_i tenhamos coesão interna e boa separação entre os grupos.

Assim, essas técnicas exigem a pré-fixação de critérios que produzam medidas sobre a qualidade da partição produzida. O uso dos métodos de partição pressupõe também o conhecimento do número de partições desejadas. Daí o problema passa a ser a procura de uma partição dos n objetos em m grupos, de modo a tornar ótimo um critério de otimização da partição.

Isso poderia ser feito construindo todas as partições possíveis, determinar o valor da medida para cada uma e selecionar a melhor partição. Porém, segundo Duran e

Odell¹ citado por BUSSAB *et alii* (1990), o número de possibilidades é assintoticamente da ordem de m^{n-1} . Desse modo fica inviável uma solução através da investigação completa das partições. Logo, os processos de partição tendem a investigar algumas partições, procurando encontrar a partição ótima ou uma alternativa que seja quase ótima.

Vejamos algumas dessas técnicas.

2.2.2.1) MÉTODO "LEADING".

A fim de se obter uma técnica de partição de um conjunto de dados, vamos considerar a representação de uma partição de n indivíduos por um vetor p de m componentes. Se $p_i = j$, então o i -ésimo indivíduo pertence ao j -ésimo agrupamento; se $p_i = 0$, significa que o i -ésimo indivíduo ainda não pertence a qualquer agrupamento. Postulemos que os conjuntos componentes da partição devam ser vazios e vamos supor que, no vetor p , para cada $j=1,2,\dots,m$, existe pelo menos um $i \in \{1,2,\dots,n\}$, com $p_i = j$. Valores de p_i tais que $p_i = 0$ ou $p_i \geq 1$ e $p_i \leq n$ não serão admitidos.

Esse procedimento é chamado Procedimento Heurístico, encontrado em SPÄTH (1980). Podemos observar que, de acordo com esse procedimento, cada objeto é considerado uma vez e imediatamente é alocado a um agrupamento. Isso pode ser

¹DURAN, B.S. & ODELL, P.L. *Cluster Analysis: a survey*.
Berlin, Springer-Verlog, 1974.

feito aplicando-se este método diretamente aos dados originais ou a uma matriz de distâncias.

A descrição do algoritmo é a seguinte:

1. Defina os números mínimo e máximo de grupos. Defina também um valor de DELTA, que é o valor que incrementa uma constante pré-fixada a fim de encontrarmos a partição.

2. Seja um valor da matriz de distâncias o qual comparamos com a constante pré-fixada. Se for menor, as observações são alocadas num primeiro grupo. Caso contrário, as observações não serão alocadas a outro grupo e essa constante será incrementada pelo valor de DELTA. Isso é feito até que o número máximo de grupos seja alcançado.

3. É definido internamente um vetor de zeros, onde a cada passo do algoritmo, esses zeros vão sendo substituídos pelo número de grupos requerido. Em cada alocação feita, testa-se se ainda existe zeros no vetor, cuja dimensão é igual ao do número de observações. Se existirem posições nulas, incrementa-se a constante pré fixada pelo valor de DELTA. Em caso contrário, o número de grupos é incrementado, iniciando-se novamente o processo.

4. Quando o número máximo de grupos é atingido, o processo termina.

HARTIGAN (1974) considera esse procedimento como bastante rápido de se encontrar a partição considerada ótima.

2.2.2.2) MÉTODO DAS k-MÉDIAS.

SPÄTH(1980) descreve que o algoritmo usado no método das k-médias é muito empregado em situações práticas.

Seja C_j um agrupamento com um número de elementos $m_j \geq 2$. Seja $C_k \neq \emptyset$ um subconjunto próprio de C_j e consideremos $C_p = C_j - C_k = C_j \cap C_k^c$ a diferença (não vazia) entre C_j e C_k , isto é, o conjunto de elementos de C_j que não pertencem a C_k .

Seja

$$\bar{x}_q = \frac{1}{m_q} \sum_{i \in C_q} x_i \quad (q = j, k, p)$$

também conhecidos como centróides e

$$e_q = \sum_{i \in C_q} ||x_i - \bar{x}_q||^2 \quad (q = j, k, p) \quad (1)$$

a soma de quadrados das distâncias dos elementos do subconjunto, calculado a partir da centróide. Desenvolvendo (1), temos que:

$$e_q = \sum_{i \in C_q} ||x_i||^2 - m_q ||\bar{x}_q||^2 \quad (q=j, k, p)$$

A centróide $\bar{x}_p$ e a soma de quadrados das distâncias e_p podem ser expressos como função de $\bar{x}_j, \bar{x}_k$, e_j e e_k , e que podem ser facilmente demonstradas:

$$\bar{x}_p = \frac{m_j \bar{x}_j - m_k \bar{x}_k}{m_j - m_k}$$

e

$$e_p = e_j - e_k - \frac{m_j \bar{x}_j - m_k \bar{x}_k}{m_j - m_k} || \bar{x}_j - \bar{x}_k ||^2.$$

Para o caso especial em que $C_k = \{ k \}$, isto é, o agrupamento tem apenas um elemento, as fórmulas acima, tornam-se:

$$\bar{x}_p = \frac{m_j \bar{x}_j - x_k}{m_j - 1}$$

e

$$e_p = e_j - \frac{m_j}{m_j - 1} || \bar{x}_j - x_k ||^2.$$

Assim, o algoritmo das k -médias, iniciando com uma dada partição, cada ponto x_i é transferido experimentalmente do agrupamento C_r para todos os outros agrupamentos C_j , $j=1, \dots, n$ e $j \neq r$. Se para pelo menos um $j \neq r$ tivermos que

$$\frac{m_r}{m_r - 1} || \bar{x}_r - x_i ||^2 > \frac{m_j}{m_j + 1} || \bar{x}_j - x_i ||^2$$

então x_i é alocado ao grupo j para o qual o lado direito da

expressão acima é minimizado, o que significa que

$e_\nu + e_r$ torna-se

$$e_r = \frac{m_r}{m_r - 1} || \bar{x}_r - x_i ||^2 + e_\nu + \frac{m_\nu}{m_\nu + 1} || \bar{x}_\nu - x_k ||^2$$

de modo que $d = \sum_{j=1}^n e_j$ é reduzido ao mínimo. Se este não for o caso, o próximo ponto é considerado. Tantos passos são executados através dos objetos $i=1, \dots, n$ quantos forem necessários, até não ocorrer mais mudanças.

Exemplo: Considerando o mesmo exemplo dado anteriormente, temos os dados:

parcelas	atributos	
	1	2
1	10	5
2	20	20
3	30	10
4	30	15
5	5	10

e queremos encontrar as parcelas mais similares.

Consideremos os agrupamentos iniciais:

$$C_1 = \{ 1, 5 \} \quad e \quad C_2 = \{ 2, 3, 4 \}$$

Temos então:

parcelas	atributos		grupo 1		$d^2(.,c)$	grupo 2		$d^2(.,c)$
	1	2	1	2		1	2	
1	10	5	10	5	12.5			
2	20	20				20	20	69.5
3	30	10				30	10	36.1
4	30	15				30	15	11.1
5	5	10	5	10	12.5			
centróide			7.5	7.5		26.7	15.0	
total			$e_1 = 25.0$			$e_2 = 116.7$		

Vemos que $d=141.7$, e como o atributo (20,20) referente à parcela 2, está com uma distância relativamente alta, tentemos realocá-lo para o grupo 1.

Assim, temos:

parcelas	atributos		grupo 1		$d^2(.,c)$	grupo 2		$d^2(.,c)$
	1	2	1	2		1	2	
1	10	5	10	5	47.3			
2	20	20	20	20	138.8			
3	30	10				30	10	6.3
4	30	15				30	15	6.3
5	5	10	5	10	42.3			
centróide			11.7	11.7		30	12.6	
total			$e_1 = 233.3$			$e_2 = 12.6$		

Com isso, temos que $d=246.3$, que é um valor mais alto que o anterior. Assim, deixamos o atributo (20,20)

no grupo 2 como antes e encerramos o procedimento com os agrupamentos

$$C_1 = \{ 1, 5 \} \quad e \quad C_2 = \{ 2, 3, 4 \}$$

3. METODOLOGIA

3.1 Configuração e Montagem do Sistema

A sugestão proposta por SPÄTH (1980), vista na revisão da literatura, onde se propõe construir uma matriz única de distâncias, reunindo nessa matriz uma ponderação das matrizes parciais de distâncias para dados quantitativos, qualitativos nominais e ordinais, foi utilizada para a construção do sistema.

Para a representação da estrutura lógica do sistema, utilizou-se os Diagramas Hierárquicos de Blocos ("Structure Chart") descrito por YOURDON *et alii* (1979).

Os Diagramas Hierárquicos de Blocos contribuem para evidenciar a hierarquia do sistema, de forma simples, simbólica e muito eficiente na descrição da estrutura lógica.

Esta técnica de documentação de sistemas apresentado na Figura 5, o Diagrama Hierárquico de Blocos mostra a estrutura lógica do procedimento "GRUPAMENTO MULTIDIMENSIONAL (ANÁLISE CLUSTER)".

Figura 6 - Diagrama Hierárquico de Blocos para o sistema de Análise de Agrupamentos.

- ① { Tipo de entrada de dados
- ② { Tipo de dados
- ③ { dimensão da matriz de dados quantitativos
- ③ { dados quantitativos
- ⑤ { matriz de distância para dados quantitativos
- ⑤ { dados qualitativos nominais
- ⑦ { matriz de distância para dados qualitativos nominais
- ⑧ { matriz de dados qualitativos ordinais
- ⑧ { matriz de distância para dados qualitativos ordinais
- ⑩ { matriz única de distância
- ⑩ { grupos formados pelo método de partição
- ⑫ { grupos formados e médias e variâncias dos grupos para variáveis quantitativas
- ⑬ { impressão dos grupos formados pelo método hierárquico

Figura 7. Parâmetros de Entrada (Q) e Saída (D) do Sistema.

Os retângulos que aparecem na Figura 6 denominam-se módulos do sistema, implicando cada um numa atividade ou conjunto de atividades. As descrições das atividades realizadas pelos módulos encontram-se nas legendas que aparecem dentro dos retângulos. Na Figura 2, temos a descrição do símbolo caracterizado por um círculo, uma flecha e um número, que representam os parâmetros de entrada e saída dos módulos.

Os módulos são conectados por segmentos de retas denominados conexões, que mostram as relações entre eles.

A flecha curva representa uma repetição ("loop"). No diagrama, a flecha envolve todas as conexões, assim, a sequência de execução dos módulos poderá ser repetida.

Os losangos são símbolos de decisão. Assim, em função de uma informação, será executado o módulo "Manipule Quantitativos", por exemplo, ou um outro módulo qualquer na hierarquia. Esse tipo de diagrama não mostra quais as informações a serem dadas para a passagem de um módulo para outro.

Os Diagramas Hierárquicos de Blocos devem ser lidos de cima para baixo. Nesse caso, a execução será:

- "Selecione a Entrada de Dados";
- "Selecione o Tipo de Dados";

- "Calcule a Matriz de Distância Única";
- "Analise os Dados".

Escolhendo o fluxo "MANIPULE DADOS QUANTITATIVOS", "MANIPULE DADOS QUALITATIVOS NOMINAIS" ou "MANIPULE DADOS QUALITATIVOS ORDINAIS", a sequência será:

- "Entre com Dados";
- "Padronizar Dados (no caso de dados quantitativos)";
- "Escolher Medidas";
- "Calcular a Matriz de Distâncias Parcial".

Em seguida o controle voltará para "Grupamento Multidimensional (Análise Cluster)", que poderá entrar no módulo "Calcular Matriz de Distância Única" e "Analise os Dados".

Depois dos dados analisados, para o método de partição, o fluxo segue para o cálculo das médias e variâncias para cada grupo formado e tem-se a opção de saída em impressora ou em arquivo texto.

Encerrada a análise, o fluxo retornará ao módulo inicial, caso se deseje fazer uma nova análise.

YOURDON *et alii* (1979) recomendam que esses diagramas devem incluir todos os módulos e conexões, mas comentam que não é necessário mostrar todas as repetições e decisões, exibindo simplesmente as estruturas deste tipo que

sejam importantes para se entender o módulo principal em linhas gerais.

3.2. TÉCNICAS DE ANÁLISE DE AGRUPAMENTOS

As técnicas de Análise de Agrupamentos utilizadas na construção do sistema, foram:

1) uma técnica de partição, que é o método "Leading", por ser um método bastante rápido na formação dos grupos.

2) uma técnica hierárquica, que é o método do vizinho mais próximo, por ser um método clássico neste tipo de análise.

Os programas foram elaborados segundo os algoritmos já descritos e utilizados num conjunto de dados reais, demonstrando a simplicidade no manuseio do sistema. A obtenção dos grupos, utilizando os dois métodos implementados no sistema, conduziram a resultados praticamente análogos, deixando com isto o usuário mais tranquilo em termos de escolha dos métodos, já que não é possível saber qual o melhor procedimento a ser utilizado. Isto acontece devido ao caráter exploratório desse tipo de análise.

4. RESULTADOS E DISCUSSÃO

4.1 DESCRIÇÃO DO SISTEMA

O sistema foi desenvolvido, de acordo com o Diagrama Hierárquico de Blocos proposto e transformado em linhas de programação, utilizando-se a linguagem Pascal 5.0, e pode ser implementado em qualquer micro-computador da linha PC/XT ou AT. O programa deve ser implementado num sub-diretório de disco rígido, pois vários arquivos auxiliares são criados, deixando livre um maior espaço de memória para sua execução, devido ao fato de que esse tipo de análise, em geral é feita com um grande número de dados.

O programa principal foi chamado de CLUSTER.EXE, e, pela extensão .EXE, não é necessário ter o Turbo Pascal instalado.

Dessa forma, para inicializar a execução, basta digitar:

```
C:> CLUST.
```

Iniciada a execução para um dado conjunto de dados, as opções do "menu" de abertura são mostradas, como

na Figura 8.

```
.. Analise Cluster ..  
Memoria inicial: 453232 bytes  
Memoria atual : 453232 bytes livres  
  
      OPCOES  
  
>> Entrada de dados  
>> Mudar disp. saída de dados  
>> Listar dados  
>> Calculo da Matriz de Distancia Unica  
>> Alterar drive/diretorio  
>> Procedimentos de Analise  
>> Retorna ao Sistema Operacional  
  
Opcao
```

Figura 8. "Menu" de abertura do sistema.

A opção "Entrada de dados" permite ao usuário a entrada de dados via arquivo texto ou teclado, conforme mostra a Figura 9. Se a entrada for via arquivo texto (dados gravados em arquivos WS, SideKick, Norton, por exemplo), esse arquivo deverá conter a extensão ".dat".

Figura 9. Opções de entrada de dados.

A Figura 10 mostra o "menu" para escolha dos tipos de dados: quantitativos, qualitativos nominais e ordinais.

Para dados quantitativos, pode-se entrar com os dados colocados na forma matricial, onde para as linhas da matriz, tem-se o número de observações a serem agrupadas e nas colunas o número de variáveis consideradas. O número máximo de observações permitido é 400 e o número máximo de variáveis quantitativas é 25, sendo possível então a entrada de uma matriz 400×30 .

Figura 10. Opções de escolha do tipo de dados.

Para dados qualitativos nominais, temos apenas dados binários, e, como para dados quantitativos, permite-se a entrada de até 400 observações e 50 variáveis, uma vez que para esse tipo de análise, os dados qualitativos nominais são mais frequentes. Já para dados qualitativos ordinais, é permitida a entrada de até 400 observações e 30 variáveis.

Apenas para dados quantitativos, é dada a opção de se fazer uma padronização nos dados, utilizando-se para isso, a média e a variância por colunas, e, evitando de se calcular coeficientes de similaridades ou distâncias

entre observações que não tenham a mesma grandeza.

Após feita a leitura de cada tipo de dados, o fluxo se faz para a escolha das medidas para dados quantitativos, qualitativos nominais e ordinais, como mostram as Figuras 11 e 12.

Figura 11. Opções de medidas para dados quantitativos.

Figura 12. Opções de medidas para dados qualitativos nominais e ordinais.

É importante salientar que para o cálculo da matriz única de distâncias, que é a quarta opção do "menu" de abertura, é necessário pelo menos dois tipos de dados.

No caso em que se deseja analisar apenas dados quantitativos, por exemplo, pode-se construir um arquivo auxiliar para dados qualitativos com dados binários. Em seguida, faz-se a entrada dos dados utilizando-se a opção para dados qualitativos (neste caso tanto faz os dados serem nominais ou ordinais). Assim, tem-se agora, dois conjuntos

de dados diferentes e para obter a matriz de distâncias única, onde a ponderação das matrizes parciais é feita, indica-se com zero o peso para a matriz que é desnecessária, ficando, assim, apenas a matriz que contém as distâncias de interesse.

As matrizes de distâncias parciais, que é como estamos nos referindo para cada tipo de dados, são guardadas num arquivo texto e num arquivo real (reservado para uso do sistema), permitindo assim o acesso ao usuário desses resultados parciais. Apenas o arquivo real com esses dados é usado para o cálculo da matriz de distâncias única.

Uma outra observação a ser feita é que pode-se calcular as matrizes de distâncias parciais utilizando-se um tipo de medida diferente para cada conjunto de dados. Por exemplo, para dados quantitativos, em geral usa-se a Distância Euclideana para a matriz parcial. Já para dados qualitativos, em geral, usa-se o Coeficiente de Jaccard. Isso pode causar problemas na hora de calcular a matriz de distâncias única. É preciso observar que essas duas medidas funcionam em sentidos opostos, isto é, uma mede a dissimilaridade e a outra mede a similaridade entre as observações. Assim, a maneira de se contornar esse problema, é, por exemplo, ponderar a matriz parcial para dados quantitativos com um peso positivo e a matriz parcial para dados qualitativos com um peso negativo. Isso dá um mesmo sentido às medidas consideradas. Se o usuário não achar

conveniente proceder dessa forma, pode-se ponderar a matriz única usando outras transformações de medidas de dissimilaridade em medidas de similaridade, e vice-versa, sugeridas na literatura. O importante para essa matriz única é que se tenha medidas de mesmo sentido.

A ponderação mais usual é aquela em que se leva em conta o número de variáveis em relação ao total. Assim, para um dado conjunto de dados contendo variáveis quantitativas, qualitativas nominais e qualitativas ordinais, podemos usar a seguinte ponderação:

$$\text{- peso para quantitativas} = \frac{\# \text{ var. quantitativas}}{\# \text{ total de variáveis}}$$

$$\text{- peso para qual. nominais} = \frac{\# \text{ var. qual. nominais}}{\# \text{ total de variáveis}}$$

$$\text{- peso para qual. ordinais} = \frac{\# \text{ var. qual. ordinais}}{\# \text{ total de variáveis}}$$

e, obviamente, não esquecendo que o tipo da medida deve ser levado em conta.

A preocupação que surge no caso de termos pesos negativos, é que, se esse número for grande, poderemos ter distâncias negativas. Isso foi solucionado, encontrando-se o menor valor da matriz calculada (ou, o

maior valor absoluto) e adicionando-se esse valor a todos os elementos da matriz. Com isso, a matriz ficará deslocada desse valor, produzindo de imediato um agrupamento, já que teremos elementos nulos fora da diagonal. Isso acontecerá, é claro, se o método utilizado para a formação dos agrupamentos for hierárquico, uma vez que eles sempre procuram o menor valor da matriz de distâncias para começar a executar o algoritmo.

Devemos salientar também, que se o usuário já estiver com a matriz de distâncias calculada por algum outro método qualquer, basta que o arquivo onde está essa matriz tenha o nome de UNICA.111. Para que o programa utilize essa matriz, é necessário fazer uma entrada para dados quantitativos informando o número de variáveis e de observações, e passar aos procedimentos de análise. Isso porque durante a execução, um arquivo com esse nome é criado para guardar a matriz de distâncias calculada durante a execução do programa. Mensagens aparecerão na tela perguntando se essa matriz já existe, conforme indica a Figura 13.

Figura 13. Arquivo UNICA.111 contendo a matriz única de distâncias.

É fornecido também ao usuário optar pela forma de saída dos resultados:

- i) impressora;
- ii) arquivo texto;
- iii) ambos (impressora e arquivo).

Essa opção é feita no instante em que se optar pelo método de análise para se formar os grupos. Um "menu" de opção de saída aparecerá na tela, conforme a Figura 14, para que o usuário opte pelo que melhor lhe convier.

```
.. Analise Cluster ..  
Memoria inicial: 453232 bytes  
Memoria atual : 453232 bytes livres
```

Destino dos resultados obtidos

```
>> 1 Saída em Disco  
>> 2 Saída na Impressora  
>> 3 Em ambos  
  
>> Opcao:
```

Opcao

Figura 14. Menu para escolha da saída dos resultados da análise.

Existe também o inconveniente de se escolher alguma medida e a execução ser abortada. Nestes casos, o usuário não precisará recomeçar a execução do programa novamente. Basta recomeçar pela última entrada de dados e escolher outra medida para o cálculo da distância.

Desta forma, o sistema construído fornece ao usuário um modo de execução bastante simplificada, bastando seguir as observações feitas acima e escolhendo as opções corretas nos "menus" de escolha.

Quanto ao tempo de processamento, podemos afirmar que isso depende do número de variáveis e do número de observações. Quanto maiores esses números, maior será o tempo de execução, devido à natureza das medidas, onde muitas comparações são feitas, principalmente no cálculo da matriz de distâncias para dados qualitativos, e devido à natureza dos algoritmos utilizados.

Como ilustração para uso do sistema, utilizaremos um conjunto de dados fornecido por VETTORAZZI (1988). Esses dados foram coletados a partir de imagens de satélite no mapeamento e caracterização de áreas homogêneas quanto aos critérios de rede de drenagem e relevo (denominadas unidades fisiográficas) para diferentes tipos de solos. Os dados encontram-se na Tabela 3, 4 e 5.

As variáveis consideradas são:

1) Dados Quantitativos, que estão na Tabela 3.

- i) Densidade de Drenagem
- ii) Frequência de Rios
- iii) Comprimento Médio de Rios

2) Dados Qualitativos Nominais, que estão na Tabela 4.

a) Rede de Drenagem

- i) Densidade
 - 0 - insuficiente
 - 1 - muito baixa

- 2 - baixa
- 3 - média
- 4 - alta
- 5 - muito alta

ii) Grau de Uniformidade

- 0 - insuficiente
- 1 - não uniforme
- 2 - pouco uniforme
- 3 - uniforme

iii) Ângulos de Junção

- 0 - insuficiente
- 1 - ângulos agudos
- 2 - ângulos retos
- 3 - ângulos obtusos

iv) Angularidade

- 0 - insuficiente
- 1 - baixa
- 2 - média
- 3 - alta

v) Grau de Controle

- 0 - insuficiente
- 1 - muito baixo

2 - baixo

3 - médio

4 - alto

vi) Tipo de modelo

0 - insuficiente

1 - dendrídrico

2 - paralelo

3 - sub-paralelo

4 - retangular

5 - sub-retangular

b) Relevo

i) Tipo

1 - subnormal

2 - normal

3 - normal e excessivo

4 - excessivo

ii) Classe

0 - insuficiente

1 - plano

2 - plano a suave ondulado

3 - Suave Ondulado a ondulado

4 - ondulado

5 - ondulado a fortemente ondulado

6 - fortemente ondulado

iii) Comprimento de vertentes

- 1 - não aplicável
- 2 - curtas
- 3 - médias
- 4 - longas
- 5 - muito longas

c) Dados Qualitativos Ordinais, que estão na Tabela 5.

i) Posição Topográfica

- 1 - terço inferior da vertente (ti)
- 2 - terço médio da vertente (tm)
- 3 - terço superior da vertente (ts)
- 4 - interflúvio (i)
- 5 - ti + tm
- 6 - ti + tm + ts
- 7 - ti + tm + ts + i
- 8 - tm + ts
- 9 - tm + ts + i
- 10 - ts + i
- 11 - insuficiente

ii) Orientação

- 0 - não orientada
- 1 - pouco orientada
- 2 - orientada

Tabela 3- Características de drenagem superficial e do relevo das unidades fisiográficas - dados quantitativos.

Unidades Fisiográficas	Densidade de Drenagem	Frequência de Rios	Comprimento Médio de Rios
1	0.24	0.80	0.30
2	4.02	11.15	0.25
3	4.17	5.22	0.80
4	2.18	2.15	1.01
5	4.85	12.11	0.35
6	3.23	2.31	1.40
7	1.42	0.55	2.50
8	4.88	5.94	0.82
9	3.09	5.64	0.55
10	3.68	4.21	0.87
11	2.81	5.31	0.53
12	3.16	7.37	0.43
13	3.31	9.79	0.34
14	1.59	3.25	0.44
14	2.50	2.54	0.68
16	4.89	4.44	1.10
17	1.55	1.05	1.47
18	5.85	13.85	0.42
19	5.05	15.71	0.32
20	3.00	7.00	0.43
21	3.54	3.54	1.00
22	2.27	2.16	1.05
23	2.86	4.14	0.69
24	4.50	8.84	0.51
25	0.00	0.00	0.00
26	0.00	0.00	0.00
27	0.21	0.33	0.65
28	4.00	7.50	0.53
29	0.29	0.35	0.84
30	0.00	0.00	0.00
31	0.11	0.28	0.40
32	0.34	0.18	1.90
33	0.00	0.00	0.00
34	0.28	0.65	0.40
34	0.00	0.00	0.00
35	0.00	0.00	0.00
37	0.00	0.00	0.00
38	0.00	0.00	0.00
39	1.65	1.41	1.17
40	0.97	2.93	0.33
41	3.45	1.92	1.60
43	2.52	1.39	1.81
44	4.20	11.24	0.38
45	5.08	10.61	0.48
46	5.00	12.06	0.41
47	2.59	3.19	0.81
48	2.13	1.53	1.36
49	0.25	0.13	1.89
50	2.49	2.10	1.18
51	4.54	8.72	0.52
52	2.94	5.14	0.57
53	3.68	9.38	0.39
54	3.12	4.90	0.64
55	5.71	13.06	0.44
56	2.38	2.90	0.82
57	2.17	2.20	0.68
58	3.30	7.04	0.47
59	1.27	1.71	0.78
60	0.97	1.00	0.97
61	0.82	0.41	2.00
62	0.73	0.36	2.03
63	0.58	0.44	1.32
64	0.58	0.54	1.06
65	0.00	0.00	0.00
66	2.50	5.00	0.50

Tabela 4 - Características de drenagem superficial e do relevo das unidades fisiográficas - dados qualitativos nominais.

Unidades Fisiográficas	Densidade	Estado da Uniformidade	Ângulo de Junção	Angularidade	Estado do Contorno	Forma do Modelo	Tipo do Relevo	Classe	Comprimento de Vertentes
1	1	0	0	0	0	0	2	2	1
2	1	0	1	2	2	1	4	2	2
3	2	2	1	1	2	2	2	4	2
4	2	2	1	1	2	1	4	2	2
5	2	0	1	1	2	0	2	4	4
6	2	2	0	2	2	2	2	4	2
7	4	2	1	1	2	1	2	4	2
8	4	2	1	1	1	1	2	5	2
9	4	2	1	1	2	1	2	5	2
10	2	2	1	1	2	1	2	3	2
11	4	2	1	1	2	1	2	2	2
12	4	2	1	1	2	1	2	2	2
13	4	2	1	1	1	1	2	2	2
14	2	0	1	1	1	2	2	2	2
15	2	1	1	1	1	2	2	2	2
16	2	0	1	0	0	0	2	2	1
17	2	2	1	1	2	2	2	2	2
18	2	2	1	1	2	1	4	2	2
19	2	2	1	1	2	1	4	2	2
20	4	2	0	1	2	0	2	2	2
21	2	2	1	1	1	1	2	2	2
22	2	1	1	1	1	1	2	2	2
23	2	1	1	1	2	1	2	2	2
24	0	0	0	0	0	0	2	2	1
25	0	0	0	0	0	0	2	2	1
26	0	0	0	0	0	0	2	2	1
27	1	0	1	1	0	1	2	2	1
28	4	2	1	1	2	1	2	2	2
29	1	0	0	0	0	0	2	2	1
30	1	0	0	0	0	0	2	2	1
31	1	0	0	0	0	0	2	2	1
32	1	0	0	0	0	0	2	2	1
33	1	0	0	0	0	0	2	2	1
34	1	0	0	0	0	0	2	2	1
35	1	0	0	0	0	0	2	2	1
36	0	0	0	0	0	0	2	2	1
37	0	0	0	0	0	0	2	2	1
38	0	0	0	0	0	0	2	2	1
39	3	1	1	2	2	2	2	2	4
40	2	1	1	1	2	1	4	2	4
41	2	2	1	1	2	1	4	2	4
42	2	2	1	1	2	1	4	2	4
43	2	2	1	1	2	1	4	2	4
44	2	2	1	1	2	1	4	2	4
45	2	2	1	1	2	1	4	2	4
46	2	2	1	1	2	1	4	2	4
47	2	2	1	1	2	1	4	2	4
48	2	2	1	1	2	1	4	2	4
49	2	2	1	1	2	1	4	2	4
50	2	2	1	1	2	1	4	2	4
51	2	2	1	1	2	1	4	2	4
52	2	2	1	1	2	1	4	2	4
53	2	2	1	1	2	1	4	2	4
54	2	2	1	1	2	1	4	2	4
55	2	2	1	1	2	1	4	2	4
56	2	2	1	1	2	1	4	2	4
57	2	2	1	1	2	1	4	2	4
58	2	2	1	1	2	1	4	2	4
59	2	2	1	1	2	1	4	2	4
60	2	2	1	1	2	1	4	2	4
61	2	2	1	1	2	1	4	2	4
62	2	2	1	1	2	1	4	2	4
63	2	2	1	1	2	1	4	2	4
64	2	2	1	1	2	1	4	2	4
65	2	2	1	1	2	1	4	2	4
66	2	2	1	1	2	1	4	2	4
67	2	2	1	1	2	1	4	2	4
68	2	2	1	1	2	1	4	2	4
69	2	2	1	1	2	1	4	2	4
70	2	2	1	1	2	1	4	2	4
71	2	2	1	1	2	1	4	2	4
72	2	2	1	1	2	1	4	2	4
73	2	2	1	1	2	1	4	2	4
74	2	2	1	1	2	1	4	2	4
75	2	2	1	1	2	1	4	2	4
76	2	2	1	1	2	1	4	2	4
77	2	2	1	1	2	1	4	2	4
78	2	2	1	1	2	1	4	2	4
79	2	2	1	1	2	1	4	2	4
80	2	2	1	1	2	1	4	2	4
81	2	2	1	1	2	1	4	2	4
82	2	2	1	1	2	1	4	2	4
83	2	2	1	1	2	1	4	2	4
84	2	2	1	1	2	1	4	2	4
85	2	2	1	1	2	1	4	2	4
86	2	2	1	1	2	1	4	2	4
87	2	2	1	1	2	1	4	2	4
88	2	2	1	1	2	1	4	2	4
89	2	2	1	1	2	1	4	2	4
90	2	2	1	1	2	1	4	2	4
91	2	2	1	1	2	1	4	2	4
92	2	2	1	1	2	1	4	2	4
93	2	2	1	1	2	1	4	2	4
94	2	2	1	1	2	1	4	2	4
95	2	2	1	1	2	1	4	2	4
96	2	2	1	1	2	1	4	2	4
97	2	2	1	1	2	1	4	2	4
98	2	2	1	1	2	1	4	2	4
99	2	2	1	1	2	1	4	2	4
100	2	2	1	1	2	1	4	2	4

Tabela 5 - Características de drenagem superficial e do relevo das unidades fisiográficas - dados qualitativos ordinais.

Unidades Fisiográficas	Posição Topográfica	Orientação
1	10	0
2	6	1
3	1	1
4	7	1
5	7	1
6	1	0
7	5	0
8	5	1
9	6	0
10	5	2
11	8	1
12	8	1
13	6	0
14	7	0
15	7	0
16	11	0
17	6	1
18	6	1
19	6	2
20	6	1
21	1	0
22	6	0
23	6	0
24	7	1
25	10	0
26	10	0
27	10	0
28	6	1
29	10	0
30	10	0
31	10	0
32	10	0
33	9	0
34	9	0
35	10	0
36	10	0
37	9	0
38	9	0
39	5	0
40	8	1
41	1	2
42	1	2
43	6	1
44	6	1
45	6	1
46	6	1
47	6	0
48	1	1
49	1	1
50	6	1
51	6	2
52	7	0
53	8	1
54	8	0
55	9	1
56	5	2
57	6	1
58	8	1
59	6	2
60	6	0
61	7	0
62	7	1
63	7	0
64	7	0
65	11	0
66	6	2

A análise foi feita pelo método de partição, proposto no sistema, e utilizamos inicialmente $\Delta = 0.1$. O algoritmo conduz a análise até encontrar o valor de Δ que fornece a melhor divisão. No caso, foi pedida a divisão em 3, 4, 5, 6, 7 e 8 grupos.

A análise se procedeu da seguinte forma:

1 - com a opção de entrada de dados quantitativos, consideraram-se as 66 observações e 3 variáveis, usando a distância Euclideana para o cálculo da matriz D1.

2 - com a opção de entrada de dados qualitativos nominais, consideraram-se as 66 observações e 46 variáveis do tipo "dummy" e para o cálculo da matriz D2, usou-se o coeficiente de Jaccard.

3 - com a opção de entrada de dados qualitativos ordinais, consideraram-se as 66 observações e 7 variáveis, também do tipo "dummy" e para o cálculo da matriz D3, usou-se o coeficiente de Concordância Simples.

Para construir a matriz de distâncias única, usaram-se os seguintes coeficientes:

$$\text{coeficiente para D1} = \frac{3}{56} = 0.0536$$

$$\text{coeficiente para D2} = - \frac{46}{56} = -0.8214$$

$$\text{coeficiente para D3} = - \frac{7}{56} = -0.1250.$$

Os resultados obtidos estão na Tabela 6, que fornece a formatação de saída do sistema.

De acordo com a saída fornecida ,temos os resultados:

Para 3 grupos:

Grupo 1: formado pelas observações 1, 4, 6, 7, 9, 11, 12, 14, 15, 16, 17, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 47 ,48 , 49, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66.

Grupo 2: formado pelas observações 2, 3, 5, 8, 10, 13, 18, 19, 20, 31, 32, 46, 51.

Grupo 3: formado pelas observações 50 e 52.

Para 4 grupos:

Grupo 1: formado pelas obsevações 1, 6, 14, 16, 21, 25, 26, 27, 28, 29, 30, 33, 34, 35, 36, 37, 38, 43, 44, 45, 48, 49, 53, 54, 56, 58, 59, 60, 61, 62, 63, 64, 65, 66.

Grupo 2: formado pelas observações 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 18, 19, 20, 23, 24, 31, 39, 40, 41, 42, 46, 47, 55, 57.

Grupo 3: formado pelas observações 7, 17 e 32.

Grupo 4: formado pelas observações 15, 22, 50, 51, 52.

Para 5 grupos:

Grupo 1: formado pelas observações 1, 6, 14, 16, 21, 25, 26, 27, 28, 29, 30, 33, 34, 35, 36, 37, 38, 43, 44, 45, 48, 49, 53, 54, 56, 58, 60, 61, 62, 63, 64, 65, 66.

Grupo 2: formado pelas observações 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 18, 19, 20, 23, 24, 31, 32, 39, 40, 41, 42, 46, 47, 55, 57.

Grupo 3: formado pelas observações 7, 17 e 32.

Grupo 4: formado pelas observações 15, 22, 50, 51, 52.

Grupo 5: nenhuma observação alocada.

Para 6 grupos:

Resultado análogo ao de 5 grupos.

Para 7 grupos:

Grupo 1: formado pelas observações 1, 16, 25, 26, 27,

30, 34, 37, 38, 43, 44, 48, 49, 53, 54, 56, 58, 59, 60, 64, 65, 66.

Grupo 2: formado pelas observações 2, 5, 9, 12, 13, 18, 19, 23, 24, 28, 29, 31, 33, 35, 36, 39, 40, 41, 42, 45, 46, 47, 55, 62, 63.

Grupo 3: formado pelas observações 3, 4, 6, 8, 17, 20, 21, 22.

Grupo 4: formado pelas observações 7, 10, 11, 14, 15.

Grupo 5: formado pela observação 32.

Grupo 6: formado pelas observações 50, 57, 61.

Grupo 7: formado pelas observações 51 e 52.

Para a divisão em 8 grupos, temos resultados análogos ao de 7 grupos.

De acordo com VETTORAZZI (1988), as análises foram feitas no sistema SAS, usando o método da ligação média, considerando os dados todos como quantitativos.

A partir do dendograma obtido, o autor conseguiu os seguintes agrupamentos:

Grupo 1: formado pelas observações 1, 16, 25, 26, 27,

29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 65, 66.

Grupo 2: formado pelas observações 2, 5, 13, 18, 19, 24, 43, 44, 45, 51, 53, 55.

Grupo 3: formado pelas observações 17, 61, 62, 63.

Grupo 4: formado pelas observações 3, 4, 6, 7, 8, 9, 10, 11, 12, 14, 15, 20, 21, 22, 23, 28, 40, 41, 42, 46, 47, 48, 49, 52, 57, 54, 58, 64.

Grupo 5: formado pelas observações 39, 50, 56, 59, 60.

Observa-se assim, que os resultados são parecidos, uma vez que o autor considera na análise com o sistema SAS, 5 grupos e com o sistema proposto, foram conseguidos 7 grupos, sendo o grupo 5 formado por uma observação, os grupos 6 e 7 formados por 3 observações e o grupo 4 formado por 5 observações. O autor comenta ainda, que os grupos 3, 4 e 5 são de observações que não se encaixam nos grupos 1 e 2 e que, os grupos 3 e 5 foram formados com um coeficiente muito baixo, através de cortes no dendrograma. Mesmo assim, ressalta-se a existência de três grupos apenas, que foi o resultado obtido considerando este tipo de procedimento. A listagem de saída dos resultados da análise, encontra-se na Tabela 4.

Assim, pode-se notar que o processo de se construir uma matriz de distâncias única, através da separação dos dados em quantitativos, qualitativos nominais e ordinais, conduz a um resultado tão bom quanto ao dos pacotes estatísticos disponíveis. Ainda, a vantagem de se fazer a análise pelo sistema construído, reside no fato do sistema ser composto por "menus", onde o usuário escolhe a opção que melhor lhe convier, o que em geral não acontece com outros pacotes, que em geral exige do usuário um conhecimento inicial da linguagem a ser usada.

Tabela 6 - Formato de saída (em impressora ou arquivo texto) dos resultados obtidos da análise dos dados das unidades fisiográficas, usando o método de partição (continuação).

1	2	3	4	5	6	7	8	0	0	8	0	0	0	0	0	0	5	2	8	0	0	0	0	1	0
1	2	3	4	5	6	7	8	0	7	0	0	0	0	0	0	0	4	1	7	0	0	0	0	0	0
1	2	3	4	5	6	7	8	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M = B Rho = 0.40																									
1	2	3	4	2	5	6	7	7	3	7	8	0	0	0	0	0	0	2	7	0	0	0	0	1	1
1	2	3	1	0	5	6	6	2	6	7	8	0	0	0	0	0	4	1	6	0	0	0	0	0	0
1	2	3	4	0	5	5	1	5	6	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M = B Rho = 0.50																									
1	2	3	4	2	5	6	7	7	3	7	7	8	0	0	0	0	2	2	7	5	4	0	0	1	1
1	2	3	1	0	5	6	6	2	6	6	7	8	8	0	0	1	1	6	0	3	0	4	0	0	0
1	2	3	4	0	5	5	1	5	5	6	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
M = B Rho = 0.60																									
1	2	3	4	2	4	5	4	6	3	4	6	6	6	4	7	4	2	2	6	8	4	0	0	1	1
1	2	3	1	3	0	3	5	2	3	4	5	0	3	0	3	1	1	5	5	3	4	4	0	0	0
1	2	3	2	0	2	5	1	2	3	5	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0
M = B Rho = 0.70																									
1	2	3	4	2	4	5	3	4	3	3	2	6	7	4	8	4	2	2	4	3	4	2	2	1	1
1	2	3	1	3	8	2	3	2	2	1	4	6	3	0	3	1	1	3	2	3	1	1	0	0	0
1	2	3	2	0	1	2	1	1	2	3	5	2	4	0	0	0	0	0	0	0	0	0	0	0	0
M = B Rho = 0.80																									
1	2	3	3	2	3	4	3	2	3	3	2	2	3	3	1	3	2	2	3	3	3	2	2	1	1
1	2	2	1	2	5	2	1	2	2	1	1	2	2	2	2	1	1	2	2	2	1	1	6	7	7
1	1	2	1	6	1	1	1	1	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0

		Caracterizacão dos Grupos	
		Media	Desvio Padrao
-----*			
Grupo 1	Variavel 1	1.970784	1.626875
Grupo 2	Variavel 1	3.545385	1.648471
Grupo 3	Variavel 1	2.715000	0.225000
Grupo 1	Variavel 2	3.133529	3.516685
Grupo 2	Variavel 2	7.565385	4.821356
Grupo 3	Variavel 2	3.620000	1.520000
Grupo 1	Variavel 3	0.767451	0.613996
Grupo 2	Variavel 3	0.641538	0.415726
Grupo 3	Variavel 3	0.875000	0.305000
Grupo 1	Variavel 1	1.634118	1.732100
Grupo 2	Variavel 1	3.271250	1.390996
Grupo 3	Variavel 1	1.103333	0.542361
Grupo 4	Variavel 1	2.930000	0.790519
Grupo 1	Variavel 2	2.811471	3.605390
Grupo 2	Variavel 2	6.140417	4.441854
Grupo 3	Variavel 2	0.593333	0.356495
Grupo 4	Variavel 2	4.132000	2.554137
Grupo 1	Variavel 3	0.630294	0.575262
Grupo 2	Variavel 3	0.730417	0.429859
Grupo 3	Variavel 3	1.990000	0.465689
Grupo 4	Variavel 3	0.860000	0.265556
Grupo 1	Variavel 1	1.634118	1.732100
Grupo 2	Variavel 1	3.271250	1.390996
Grupo 3	Variavel 1	1.103333	0.542361
Grupo 4	Variavel 1	2.930000	0.790519

Tabela 6 - Formato de saída (em impressora ou arquivo texto) dos resultados obtidos da análise dos dados das unidades fisiográficas, usando o método de partição (continuação).

Grupo 1	Variavel 2	2.811471	3.605390
Grupo 2	Variavel 2	6.140417	4.441854
Grupo 3	Variavel 2	0.593333	0.356495
Grupo 4	Variavel 2	4.132000	2.554137
Grupo 1	Variavel 3	0.630294	0.575262
Grupo 2	Variavel 3	0.730417	0.429859
Grupo 3	Variavel 3	1.990000	0.465689
Grupo 4	Variavel 3	0.860000	0.265556
Grupo 1	Variavel 1	1.634118	1.732100
Grupo 2	Variavel 1	3.271250	1.390996
Grupo 3	Variavel 1	1.103333	0.542361
Grupo 4	Variavel 1	2.930000	0.790519
Grupo 1	Variavel 2	2.811471	3.605390
Grupo 2	Variavel 2	6.140417	4.441854
Grupo 3	Variavel 2	0.593333	0.356495
Grupo 4	Variavel 2	4.132000	2.554137
Grupo 1	Variavel 3	0.630294	0.575262
Grupo 2	Variavel 3	0.730417	0.429859
Grupo 3	Variavel 3	1.990000	0.465689
Grupo 4	Variavel 3	0.860000	0.265556
Grupo 1	Variavel 1	1.582609	1.702919
Grupo 2	Variavel 1	2.649600	1.884657
Grupo 3	Variavel 1	2.965000	0.959544
Grupo 4	Variavel 1	1.420000	0.000000
Grupo 5	Variavel 1	0.340000	0.000000
Grupo 6	Variavel 1	2.330000	0.160000
Grupo 7	Variavel 1	3.695000	0.755000
Grupo 1	Variavel 2	2.829565	3.557071
Grupo 2	Variavel 2	5.442000	5.189510
Grupo 3	Variavel 2	3.781667	1.734228
Grupo 4	Variavel 2	0.550000	0.000000
Grupo 5	Variavel 2	0.180000	0.000000
Grupo 6	Variavel 2	2.150000	0.050000
Grupo 7	Variavel 2	6.930000	1.790000
Grupo 1	Variavel 3	0.585217	0.540136
Grupo 2	Variavel 3	0.688800	0.561429
Grupo 3	Variavel 3	0.900000	0.317884
Grupo 4	Variavel 3	2.600000	0.000000
Grupo 5	Variavel 3	1.900000	0.000000
Grupo 6	Variavel 3	1.080000	0.100000
Grupo 7	Variavel 3	0.545000	0.025000
Grupo 1	Variavel 1	1.582609	1.702919
Grupo 2	Variavel 1	2.649600	1.884657
Grupo 3	Variavel 1	2.965000	0.959544
Grupo 4	Variavel 1	1.420000	0.000000
Grupo 5	Variavel 1	0.340000	0.000000
Grupo 6	Variavel 1	2.330000	0.160000
Grupo 7	Variavel 1	3.695000	0.755000
Grupo 1	Variavel 2	2.829565	3.557071
Grupo 2	Variavel 2	5.442000	5.189510
Grupo 3	Variavel 2	3.781667	1.734228
Grupo 4	Variavel 2	0.550000	0.000000
Grupo 5	Variavel 2	0.180000	0.000000

Tabela 6 - Formato de saída (em impressora ou arquivo texto) dos resultados obtidos da análise dos dados das unidades fisiográficas, usando o método de partição (continuação).

Grupo 6	Variavel 2	2.150000	0.050000
Grupo 7	Variavel 2	6.930000	1.790000
Grupo 1	Variavel 3	0.585217	0.540136
Grupo 2	Variavel 3	0.688800	0.561429
Grupo 3	Variavel 3	0.900000	0.317884
Grupo 4	Variavel 3	2.600000	0.000000
Grupo 5	Variavel 3	1.900000	0.000000
Grupo 6	Variavel 3	1.080000	0.100000
Grupo 7	Variavel 3	0.545000	0.025000

5. CONCLUSÕES

Considerando-se o objetivo deste trabalho e com base nos resultados obtidos, conclui-se que:

- O sistema construído para análise de agrupamentos constitui-se num sistema amigável, oferecendo uma simplicidade de análise, em geral não encontrada em outros pacotes.

- Os Diagramas Hierárquicos de Blocos, mostraram-se adequados para a etapa de elaboração do sistema.

- A maneira de tratar os dados, sugerida por SPÄTH (1980), mostrou-se bastante eficiente e prática. Porém, é preciso saber utilizar as medidas de dissimilaridade e coeficientes de similaridade, de modo que a combinação dos mesmos possam levar a bons resultados, mesmo utilizando métodos simples, como o método da partição e o método do vizinho mais próximo, como propostos na construção do sistema.

- É interessante lembrar, também, que o sistema proposto é o início de um sistema que pode ser bem mais completo em termos de medidas e procedimentos de análise, já que os algoritmos para análise de agrupamentos são muitos e

as propostas da literatura também tem sido várias.

REFERÊNCIAS BIBLIOGRÁFICAS

- BUSSAB, W.O.; ANDRADE, D.F. ; MYAZAKY, E.S. Introdução à análise de agrupamentos. s.l., Associação Brasileira de Estatística, 1990. 105p.
- CORMACK, R.M. A review of classification. Journal of the Royal Statistics Society. Ser A. London, 134: 321-67, 1971.
- FISHER, L. & VON NESS, J.W. Admissible clustering algorithms. Biometrika, Alexandria, 58:91-104, 1971.
- HARTINGAN, J. Clustering algorithms. New York, J. Wiley 1974. 347p.
- MANZATO, A.J. Análise hierárquica de agrupamentos para distribuições multinomiais. São Paulo, 1983. 128p. (Mestrado - Instituto de Matemática e Estatística/USP).

MARDIA, K.V.; KENT, J.T. ; BIBBY, D. V. Multivariate analysis. Londron, Academic Press, 1979. 521p.

ROMESBURG, H.C. Cluster analysis for researches. Belmont, Lifetime Learning, 1984. 341p.

SAS INSTITUTE. SAS/STAT guide for personal computers; version 6 edition. Cary, 1987. 1028p.

SPÄTH, H. Cluster analysis algorithms for data reduction and classification of objects. New York, J. Wiley, 1980. 226p.

VETTORAZZI, C.A. Interpretação de imagens TM/LANDSAT-5 em duas escalas, na caracterização fisiográfica para mapeamento de solo. Piracicaba, 1988. 184p (Doutorado - Escola Superior de Agricultura "Luiz de Queiroz"/USP).

WONG, M.A & LANE, T. A k-th nearest neighbor clustering procedure. Journal of the Royal Statistics Society. Ser. B, London; 45:362-8, 1983.

YOURDON, E. ; GANE, C.; SARSON, T.; LISTER, T.R. Learning to programing in structure COBOL. 3.ed. New York, Yourdon, 1979. 478p.