UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA ESTATÍSTICA E

CIÊNCIA DA COMPUTAÇÃO - IMECC

UM ESTUDO SOBRE ALGUNS MÉTODOS HIERÁRQUICOS PARA ANÁLISE DE AGRUPAMENTOS

JOSÉ RAIMUNDO GOMES PEREIRA

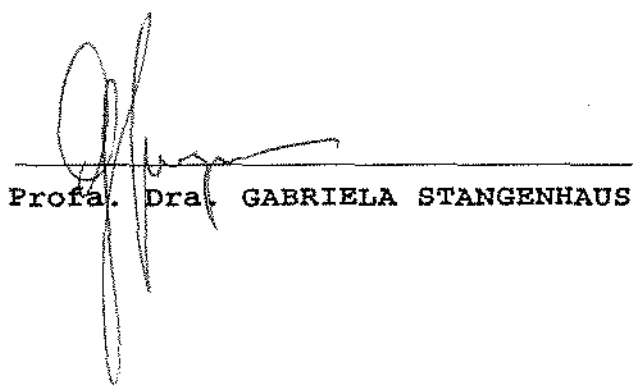
Profa. Dra. GABRIELA STANGENHAUS
Orientadora

CAMPINAS - SÃO PAULO

1993

Este exemplar corresponde à redação final da tese devidamente corrigida e defendida por **JOSÉ RAIMUNDO GOMES PEREIRA** e aprovada pela comissão julgadora.

Campinas, 28 de junho de 1993



Profa. Dra. GABRIELA STANGENHAUS

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação da Universidade Estadual de Campinas - UNICAMP, como requisito parcial para a obtenção do Título de Mestre em Estatística.

Para minha esposa, meus filhos e minha mãe.

AGRADECIMENTOS

Em primeiro lugar a Deus por sua luz em meus caminhos.

À minha família pelo apoio, paciência e tolerância.

À profa. Dra. Gabriela Stangenhaus pelos ensinamentos e pela firmeza da orientação.

Aos meus colegas pelo companheirismo e solidariedade.

À família Falcão e à família Carneiro pelo apoio e amizade.

À "turma do futebol" pelos bons momentos de descontração.

Aos colegas do Departamento de Estatística da U.A. pelo apoio e compreensão.

A todos que de alguma forma contribuiram para a realização deste trabalho.

RESUMO

Dez métodos hierárquicos aglomerativos para Análise de Agrupamentos tiveram seus desempenhos comparados ante a diferentes estruturas de dados. Foi construido um experimento com estrutura fatorial, onde os fatores eram diferentes aspectos de estrutura de dados. A presença de grupos com sobreposição, a matriz de dispersão dentro dos grupos e a correlação entre as variáveis foram alguns dos fatores considerados. Amostras foram simuladas sob as diferentes condições determinadas pelos cruzamentos dos níveis dos fatores. Os métodos foram aplicados à essas amostras e seus desempenhos mensurados quanto a recuperação das estruturas de grupos embutidas nas amostras. Dentre os fatores estudados, a sobreposição dos grupos foi o que mais afetou o desempenho dos métodos. São feitas sugestões para o emprego de alguns dos métodos. Não sendo detectada a presença de observações com valores discrepantes nos dados é sugerido o emprego dos métodos da Média das Ligações e o Centróide. Ante a presença de observações com essa característica é sugerido o emprego dos métodos da Ligação de Densidades em Dois Estágios e do Beta-Flexível. Nos dois casos ou na falta de informações sobre os dados, é sugerido que o método de Ward deve sempre ser empregado.

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

1.1 O PROBLEMA DA ANÁLISE DE AGRUPAMENTOS.

De maneira geral, o objetivo dos métodos de Análise de Agrupamentos é separar um conjunto de objetos em grupos, onde os objetos dentro dos grupos apresentem características homogêneas e que estas sejam heterogêneas entre objetos em grupos distintos. Os objetos podem ser pessoas, animais, plantas, empresas, fábricas, peças arqueológicas, sinais emitidos por satélites, etc.. Esses objetos são descritos por um conjunto de variáveis e, com base nas observações obtidas sob estas variáveis, são determinados o número de grupos, as características dos grupos e os membros desses grupos.

Os métodos empregados na Análise de Agrupamentos são técnicas de análise de dados multivariados, usadas de forma

exploratória quando for apropriado considerar, ou averiguar, a existência de uma estrutura de grupos embutida nos dados. Eles são parte de um processo científico mais geral, cujo objetivo é identificar estruturas nos dados e tentar construir "leis" (modelos) que expliquem estas estruturas.

A aplicação da Análise de Agrupamentos tem finalidades bastante diferenciadas, tais como, a determinação de objetos semelhantes num primeiro estágio de um esquema de amostragem estratificada, formulação de hipóteses sobre a estrutura dos dados e a determinação de esquemas de classificação. Dentro desta última, ocorrem aplicações muito conhecidas, tais como, classificação de plantas, de animais e a classificação de doenças. Outra aplicação importante ocorre quando se tem por objetivo modelar dados ante a presença de estruturas de grupos. Resultados mais eficientes podem ser obtidos tomando-se esta estrutura em consideração, antes de tentar estimar qualquer relação que possa existir nos dados. Evidentemente, a Análise de Agrupamentos pode ser empregado para identificar essa estrutura de grupos.

Os métodos de Análise de Agrupamentos aparecem na literatura com diferentes denominações: métodos para Taxonomia, Q-Análise, Tipologia e Reconhecimento de Padrões não-supervionado ("*unsupervised pattern recognition*") são algumas dessas denominações. Provavelmente, essa variedade de nomes é devida a importância e a intensiva aplicação da Análise de Agrupamentos em

diversas áreas de estudos, tais como, biologia, botânica, medicina, psicologia, sociologia, geografia, arqueologia e inteligência artificial. Essa diversidade de áreas explica também, como foi citado, as diferentes finalidades das aplicações. Em HARTIGAN (1975), EVERITT (1980) e em GNANADESIKAN e KETTENRING (1989) são dadas algumas aplicações em áreas bastante diferenciadas.

Em geral, os dados a serem submetidos à Análise de Agrupamentos podem ser apresentados em uma matriz (N x p), **X**, a qual pode ser escrita como

$$\mathbf{X} = [(x_{ij})] = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix},$$

onde x_{ij} é a obsevação da j-ésima variável para o i-ésimo objeto. Então, pelo que foi descrito, o objetivo da Análise de Agrupamentos é estabelecer um esquema para agrupar os vetores $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, os vetores p-dimensionais caracterizadores dos objetos, em G grupos com $G < N$.

Como um comentário a respeito dos vetores que caracterizam os objetos, é interessante observar que as variáveis x_{ij} dentro de cada $\mathbf{x}_i$ geralmente são correlacionadas e podem ser de diferentes tipos. Essas variáveis podem ser quantitativas (discretas ou contínuas), qualitativas (ordinais ou nominais) ou, como ocorre em

muitas situações, os dados podem envolver uma mistura de diferentes tipos de variáveis.

Outro comentário diz respeito à forma de apresentação dos dados para a análise. Em determinadas circunstâncias, os dados podem ser apresentados através de uma matriz (N x N) de medidas que, de alguma forma, mensuram o relacionamento entre os pares de objetos. Por exemplo, em experimentos em psicologia sobre preferências de um grupo de indivíduos, em vez das respostas para cada indivíduo, os dados são apresentados na forma de uma "medida de concordância" para cada par de indivíduos. Na seção 1.4 será abordada a questão dessas medidas.

1.2 DUAS TÉCNICAS RELACIONADAS COM A ANÁLISE DE AGRUPAMENTOS.

As razões para a utilização da Análise de Agrupamentos são diversas e pode ser incluída entre elas o emprego para redução de dados. Em certas situações, onde o volume de observações é muito grande, não é possível um tratamento mais adequado do problema em mãos, a menos que essas obsevações sejam agrupadas e que os grupos, em algum sentido, possam ser tratados como unidades observadas. Neste caso, os métodos de Análise de Agrupamentos podem ser usados com o objetivo de condensar a informação disponível sobre os N objetos em G grupos, onde G seria muito menor que N. Assim,

considerada como uma técnica de redução de dados, a Análise de Agrupamentos está relacionada com outra técnica de análise de dados multivariados, a Análise de Componentes Principais.

Em Análise de Componentes Principais, um dos principais objetivos é reduzir o conjunto original de variáveis para um conjunto menor de composições lineares ortogonais dessas variáveis. Neste sentido, a Análise de Componentes Principais é uma técnica de agrupamento que tem por objetivo agrupar as variáveis caracterizadoras dos objetos. Embora na maioria dos casos o emprego da Análise de Agrupamentos seja para agrupar os objetos, muitos dos seus métodos também podem ser empregados para agrupar variáveis. De certa forma, ambas são técnicas de redução de dados, com a Análise de Componentes Principais atuando sobre as colunas da matriz **X** e a Análise de Agrupamentos sobre as linhas dessa matriz.

Outra técnica relacionada com a Análise de Agrupamentos é a Análise Discriminante. Nesta última, o objetivo é classificar os objetos em grupos mutuamente exclusivos e exaustivos, usando para isto uma regra de classificação determinada através de uma função das observações. Embora aparentemente muito similares, as duas técnicas têm uma diferença fundamental. Na Análise Discriminante existe a tácita pressuposição de que os grupos são conhecidos "*a priori*" e o problema consiste em associar os objetos a estes grupos previamente estabelecidos. Entretanto, em muitas situações não há certeza da existência de agrupamento nos dados e, desta forma, a

saída é confiar nos dados para decidir sobre a presença do agrupamento e, se for o caso, delinear os grupos. A conjuntura desta última situação, que é o caso da aplicação da Análise de Agrupamentos, diferencia as características de emprego das duas técnicas.

Evidentemente, podem ocorrer situações onde as técnicas de Análise de Agrupamentos e de Análise Discriminante possam ser empregadas conjuntamente. Em um problema prático pode acontecer que haja uma idéia preliminar, ou imprecisa, a respeito dos grupos e o objetivo seja verificar a existência desses grupos e, além disso, estabelecer uma regra para classificar os objetos nos grupos. Em uma situação como esta o emprego combinado das duas técnicas pode ser apropriado. Talvez em decorrência de situações como esta e da semelhança dos objetivos dessas técnicas, alguns autores empregam o termo classificação para designar tanto o problema da Análise de Agrupamentos como o da Análise Discriminante.

1.3 AS TÉCNICAS DE AGRUPAMENTO.

Segundo a literatura, nas últimas décadas o número de métodos de agrupamento propostos tem crescido consideravelmente. Os primeiros métodos propostos foram construídos para a resolução de problemas específicos em certas áreas de estudo, como a biologia e

a psicologia, sem objetivar uma aplicaçào mais geral em estatística. Com o desenvolvimento acentuado dos computadores ocorreu uma "explosão" de métodos mas, ainda assim, muitas propostas "*ad hoc*". Recentemente, o problema tem atraído a atenção de estatísticos e matemáticos e alguns trabalhos foram desenvolvidos numa tentativa de estudar o problema de Análise de Agrupamentos num contexto mais formal (vide Capítulo 2). Esses profissionais, inclusive, contribuiram de forma decisiva para melhorar a eficiência dos algoritmos para a implementação computacional dos métodos considerados na literatura como os mais convenientes.

Os métodos para a Análise de Agrupamentos podem ser classificados em três tipos principais de técnicas: técnicas Hierárquicas, técnicas de Partição e técnicas de Sobreposição. Com as técnicas Hierárquicas os grupos são formados hierarquicamente, podendo o agrupamento ser obtido de duas formas. Uma forma é considerar inicialmente cada um dos N objetos constituindo um grupo e, por meio de sucessivas uniões, chegar a um único grupo contendo todos os objetos. A outra forma é considerar no início um único grupo contendo todos os objetos e, através de sucessivas divisões, terminar com N grupos contendo um único objeto. Estas técnicas produzem uma sequência de agrupamentos, isto é, são obtidas partições do conjunto de objetos em 1, 2, ..., N-1 e N grupos. Dentro dessas técnicas estão alguns dos métodos mais usados e mais discutidos na literatura.

As técnicas de Partição e as de Sobreposição determinam um único agrupamento, porém, os agrupamentos gerados por cada uma delas apresentam características diferentes. Com as técnicas de Partição os grupos são mutuamente exclusivos formando uma partição do conjunto dos objetos. De forma diferente, com as técnicas de Sobreposição o agrupamento obtido permite uma interseção entre os grupos, isto é, um objeto pode estar associado a mais de um grupo.

Na literatura observa-se uma predominância do emprego das técnicas Hierárquicas e é possível identificar alguns motivos para isto. Enquanto ferramentas exploratórias, as técnicas Hierárquicas apresentam uma maior versatilidade, além da simplicidade e da variedade de métodos disponíveis. Pode ser acrescentado também, o fato de que em muitas áreas de estudo o pesquisador tem interesse nas hierarquias dos agrupamentos geradas por esses métodos.

As técnicas de Partição apresentam métodos que por um lado são mais flexíveis e por outro são mais difíceis de usar. Muitos desses métodos permitem alterar uma partição que tenha sido considerada inadequada, o que não é possível com os métodos Hierárquicos. A dificuldade consiste no fato de que, para o emprego desses métodos, são necessárias informações adicionais como, por exemplo, informação à respeito do número de grupos ou sobre os "centros" dos grupos a serem considerados na partição inicial. É evidente que em muitas situações práticas não há informações deste tipo e isto dificulta o emprego dessas técnicas.

Com relação às técnicas de Sobreposição, suas aplicações têm sido limitadas e, talvez, isto seja devido ao fato de que, na maioria das aplicações de Análise de Agrupamentos a procura seja por grupos mutuamente exclusivos. Argumentando que é possível imaginar muitas situações onde a interseção entre os grupos seja apropriada, GNANADESIKAN e KETTENRING (1989) comentam que a existência de poucos trabalhos com as técnicas de sobreposição, talvez, seja devido mais ao fato de existirem poucos métodos eficientes dentro dessas técnicas.

Uma questão envolvida nos problemas de Análise de Agrupamentos é o emprego de outras técnicas em conjunto com os métodos de agrupamento, onde alguns cuidados devem ser observados. Por exemplo, é muito comum o emprego de Análise de Componentes Principais, onde as observaçòes são projetadas sobre as duas ou três primeiras componentes principais e estas são apresentadas graficamente para que os grupos sejam identificados visualmente. Uma outra forma do emprego desta técnica é usar essas componentes, ou mais delas, como as variáveis caracterizadoras dos objetos nos métodos de agrupamento. GNANADESIKAN e KETTENRING (1989) alertam para os riscos envolvidos nestes procedimentos devido a questão da adequabilidade dessas projeções, uma vez que elas podem simplesmente encobrir os grupos existentes. SEBER (1984) ilustra esta questão ao considerar um conjunto de dados bidimensionais contendo três grupos bastante distintos apresentados na Figura 1.1. A projeção das observações sobre a primeira componente principal

indica a presença de um único grupo, os grupos existentes somente são visualizados na projeção sobre a segunda componente principal. Neste caso, por exemplo, considerar somente a primeira componete na análise seria completamente inútil para a determinação dos grupos. É evidente que podem ser imaginadas situações semelhantes para dados com mais dimensões.

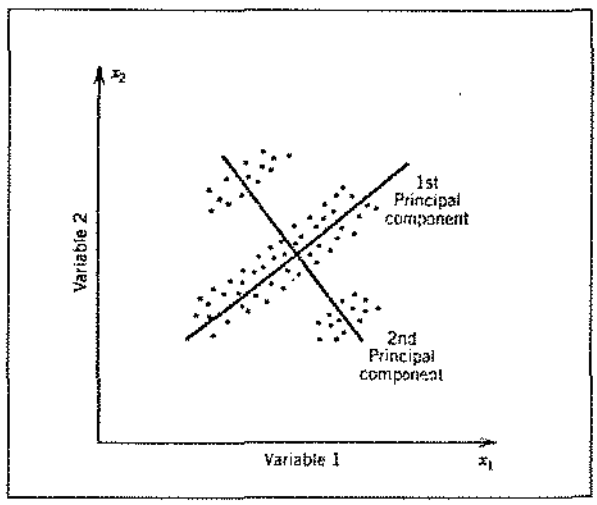


Figura 1.1 Três grupos de observações bidimensionais e o eixo das componentes principais.

É claro que o emprego de outras técnicas como um auxílio aos métodos de agrupamentos são procedimentos válidos. Em SEBER (1984), seção 4.1, são descritas várias técnicas gráficas que podem ser empregadas com esse objetivo. O emprego delas pode sugerir a presença de grupos nos dados. A questão é que não é aconselhável confiar exclusivamente nestas técnicas, considerando principalmente que as interpretações tornam-se mais difíceis à medida que aumenta a dimensão dos dados. Como colocado em GNANADESIKAN e KETTENRING

(1989): "se a análise tem um objetivo sério, então é melhor confiar nos métodos de Análise de Agrupamentos apesar de suas limitações e imperfeições".

Na seção seguinte serão discutidas as etapas envolvidas na aplicação da Análise de Agrupamentos. No Capítulo 2 as técnicas Hierárquicas serão novamente abordadas com uma descrição mais pormenorizada. Para uma discusão mais completa das técnicas de Partição e de Sobreposição, indica-se EVERITT (1980) e SEBER (1984).

1.4 AS ETAPAS PARA APLICAÇÃO DA ANÁLISE DE AGRUPAMENTOS.

Três etapas podem ser identificadas na aplicação de técnicas de Análise de Agrupamentos: (i) a seleção das variáveis e uma posterior preparação dos dados para a análise; (ii) a escolha do método a ser aplicado; (iii) a análise dos agrupamentos obtidos.

Na primeira etapa estão a questão da escolha das variáveis mais adequadas para a caracterização dos objetos e, após a coleta dos dados, as questões da padronização e/ou ponderação das variáveis. Aqui ocorrem questões polêmicas como, por exemplo, a questão da padronização das variáveis. Em muitas aplicações de

Análise de Agrupamento as variáveis são padronizadas empregando o desvio padrão das variáveis tomado sobre o conjunto completo dos objetos. Esse procedimento, entretanto, pode afetar seriamente a análise por diluir as diferenças entre os grupos para aquelas variáveis que melhor separam estes grupos. Uma alternativa seria empregar o desvio padrão dentro dos prórpios grupos, porém, como os grupos não são conhecidos "*a priori*" este procedimento não pode ser adotado e, desta forma, esta questão fica em aberto.

Ainda na primeira etapa, quando do emprego das técnicas Hierárquicas, ocorre uma questão adicional que é a escolha da medida que melhor expresse as diferenças entre os objetos. Isto é de fundamental importância, tendo em vista que tais métodos atuam diretamente sobre a matriz com essa medida para cada par de objetos. A medida escolhida tem por finalidade quantificar a distância entre cada dois objetos ou o quanto eles são "parecidos". O sentido de "parecido" depende dos objetivos da análise e, de certa forma, orienta a escolha da medida. Existe uma variedade de medidas que podem ser usadas com esse objetivo e elas são classificadas em medidas de similaridade e medidas de dissimilaridade. Com as medidas de similaridade quanto maior o valor observado mais "parecidos" são os objetos. De forma contrária, com as medidas de dissimilaridade quanto maior for o valor, menos "parecidos" são os objetos. Nas medidas de dissimilaridade estão incluídas as medidas de distância como, por exemplo, a distância Euclidiana e a de Mahalanobis. A escolha da

medida também depende dos tipos das variáveis componentes dos vetores $\mathbf{x}_i$. No caso de mistura de diferentes tipos de variáveis, por exemplo, GOWER (1971) sugere um coeficiente de similaridade que pode ser empregado nesta situação. Em ANDERBERG (1973), em EVERITT (1980) e em KAUFMAN e ROUSSEEUW (1990) muitas das medidas propostas na literatura são descritas e discutidas suas utilizações.

A segunda etapa consiste na escolha do método de agrupamento a ser empregado. Existe uma vasta literatura propondo diferentes métodos e, devido a essa abundância de métodos, o usuário encontra uma certa dificuldade para a escolha do método mais adequado. Essa tarefa pode ser facilitada se o usuário conhecer determinadas características dos métodos disponíveis e associar esse conhecimento aos propósitos de sua investigação. Como será descrito à frente, o objetivo deste trabalho assenta-se exatamente na caracterização de alguns métodos visando a questão colocada acima.

Aceitar os resultados da aplicação de um método de agrupamento sem uma investigação adicional sobre os grupos obtidos, é um procedimento inadequado adotado por alguns usuários. Como colocado em RAND (1971), os métodos produzem agrupamentos independentemente da existência, ou não, de estrutura de grupos nos dados. Em decorrência disto, se faz necessário avaliar os grupos obtidos buscando averiguar se o agrupamento não é apenas uma imposição do método e isto constitui a terceira etapa na aplicação de Análise de Agrupamentos. Embora alguns trabalhos tenham sido

desenvolvidos para uma abordagem mais formal da avaliação dos agrupamentos, por exemplo, BAKER e HUBER (1975), nesta etapa estão envolvidas questões muito subjetivas e o conhecimento do investigador é de fundamental importância para essa avaliação. Alguns autores, inclusive, enfatizam essa subjetividade. CORMACK (1971) afirma que "a classificação (o agrupamento)... não pode ser verdadeira ou falsa, provável ou improvável, somente proveitosa ou não".

Outra questão crucial em Análise de Agrupamentos é decidir sobre o número de grupos, G, existentes nos dados. Com as técnicas de Partição esta questão está embutida na segunda etapa, uma vez que para a aplicação dos métodos o número de grupos precisa ser especificado. Com as técnicas Hierárquicas esta questão faz parte da terceira etapa e o problema consiste em determinar, com base na hierarquia, o valor mais adequado para G. Várias propostas de procedimentos para determinar o número de grupos foram apresentados na literatura. Em EVERITT (1980) e em SEBER (1984) muitas dessas propostas são apresentadas e discutidas. Outra fonte bastante completa é dada em MILLIGAN e COOPER (1985) onde são discutidos e avaliados, através de um estudo de Monte Carlo, trinta procedimentos para determinação do apropriado número de grupos a partir de agrupamentos obtidos por técnicas hierárquicas. Embora de muita importância, esta questão não será discutida neste trabalho.

Evidentemente, todas as etapas descritas acima são de

importância fundamental. Por exemplo, os agrupamentos são gerados com base nos dados obtidos sob as variáveis selecionadas e, de certa forma, a escolha destas variáveis e o tratamento estatístico dado às suas respostas é que determinam os grupos. KAUFMAN e ROUSSEEUW (1990) comentam e ilustram os efeitos da seleção, padronização e ponderação de variáveis sobre a determinação dos grupos. EVERITT (1980) e SEBER (1984) também discutem de forma bastante esclarecedora as questões envolvidas nas etapas aqui consideradas.

1.5 OBJETIVOS DO TRABALHO.

Escolher entre técnicas hierárquicas e técnicas não-hierárquicas depende, principalmente, dos propósitos da investigação, as vezes há interesse num agrupamento com um fixado número de grupos e em outras o interesse pode ser por uma classificação hierárquica para os objetos. Neste ponto, o usuário tem a sua disposição inúmeros métodos e diferentes métodos podem produzir diferentes agrupamentos quando aplicados a um mesmo conjunto de objetos. Uma explicação para este fenômeno é que diferentes métodos são afetados por diferentes aspectos, ou a falta deles, na estrutura dos dados. Desta forma, é importante serem estabelecidas as características dos diversos métodos para servir como orientação ao usuário. A compreensão das características intrínsecas dos métodos é essencial para a escolha do método mais

adequado aos propósitos de uma dada investigação.

O objetivo deste trabalho é comparar alguns métodos de agrupamento sob determinadas estruturas de dados e, paralelamente, determinar aspectos dessas estruturas que afetem o desempenho dos métodos considerados. Serão estudados dez métodos hierárquicos, sendo oito deles considerados os mais conhecidos e mais usados em Análise de Agrupamentos, a saber, o Método da Ligação Simples (MLS), o Método da Ligação Completa (MLC), o Método da Média das Ligações (MML), o Método de McQuitty (MMcQ), o Método Centróide (MCEN), o Método da Mediana (MMED), o Método de Ward (MWARD) e o Método Beta-Flexível (MFLE). Os outros dois métodos, o Método do K-ésimo Vizinho Mais Próximo (MkVP) e o Método da Ligação de Densidades em Dois Estágios (MLDE), são métodos mais recentes para os quais, inclusive, não está reportado na literatura nenhum estudo comparativo.

A escolha de métodos hierárquicos para serem considerados neste trabalho é justificada por serem esses métodos os mais empregados nas aplicações de Análise de Agrupamentos. Um critério também adotado para a escolha desses métodos, foi a existência de uma literatura descrevendo e discutindo suas propriedades e, acrescentado a isso, o fato de estarem estes métodos implementados em um único "*software*". Os métodos estão implementados no Sistema de Programas SAS.

Como uma observação, evidentemente não se pode pensar em comparar métodos de agrupamento sem considerar a questão computacional. Um aspecto fundamental, por exemplo, é a questão da eficiência dos algoritmos implementando os métodos. Esta questão, entretanto, não será abordada aqui, tendo em vista que o "*software*" empregado tem a reconhecida eficiência dos seus programas.

A questão de comparar métodos de agrupamento pode ser abordada de diferentes formas. DUBES e JAIN (1976) discutem algumas formas que podem ser considerados para estas comparações. Os métodos podem ser comparados sob critérios matemáticos formais, porém, segundo os autores, as questões subjetivas inerentes ao problema de Análise de Agrupamentos não permitem que esta abordagem seja adotada de forma realista.

Outra abordagem seria selecionar um conjunto de dados com propriedades conhecidas, aplicar vários métodos e, com base em um crtério especificado, conceituar os métodos segundo seus desempenhos. Aqueles autores consideram esta abordagem inadequada, argumentando que dados com propriedades conhecidas, usualmente, significa que os dados são "bem comportados" em algum sentido e, assim, essa característica pode encobrir as habilidades dos métodos. Outro aspecto é que os resultados obtidos são para um específico conjunto de dados, isto é, somente para a estrutùra contida nestes dados.

Uma outra abordagem é empregar dados simulados, planejados de forma a conterem as características desejadas na investigação. Esta abordagem é considerada bastante apropriada pois permite comparar os desempenhos dos métodos sob diversas estruturas de dados. A maioria dos trabalhos sobre comparações de métodos tem empregado dados simulados recorrendo às técnicas de estudo de Monte Carlo.

Vários estudos de Monte Carlo aparecem na literatura comparando métodos de agrupamento sob diferentes aspectos. Os efeitos da introdução de perturbações aleatórias nos dados e os efeitos da estrutura dos dados, tais como, o tamanho dos grupos, a dimensão das observações e a correlação entre as variáveis sobre o desempenho dos métodos, foram alguns dos aspectos investigados. MILLIGAN (1981) fez uma revisão de muitos dos trabalhos realizados, expondo os aspectos investigados nestes trabalhos e fazendo sugestões para outros estudos.

Um dos primeiros trabalhos foi desenvolvido por RAND (1971). Com dados simulados segundo uma mistura de distribuições normais multivariadas, os métodos foram comparados quanto a capacidade de recuperar os grupos sob as seguintes condições: dados sem erros e dados com perturbações aleatórias (uma variável aleatória N(0;0,01) foi adicionada a cada componente das observações). Os métodos comparados, denominados T/N e AA, não fazem parte daqueles aqui considerados.

CUNNINGHAM e OLGIVIE (1972) e BAKER (1974), introduzindo perturbações na matriz de distâncias entre os objetos, compararam alguns métodos com respeito a recuperação de uma estrutura hierárquica de grupos simulada. CUNNINGHAM e OLGIVIE (1972) estudaram sete métodos e os resultados indicaram que o MML e o MLC apresentaram os melhores desempenhos. BAKER (1974) comparou apenas o MLS e o MLC e os resultados indicaram que o MLS foi o mais afetado com a introdução das perturbações.

KUIPER e FISHER (1975) compararam seis dos mais conhecidos métodos hierárquicos. Com dados simulados sob uma mistura de normais multivariadas, os autores investigaram os efeitos do tamanho dos grupos, do número de grupos e do aumento da variância das observações sobre o desempenho dos métodos. Também foi investigado o efeito da presença de valores discrepantes ("*outliers*") nos dados, usando para isto dados simulados segundo uma mistura de distribuições de Cauchy. Na maioria dos casos o MWARD, seguido do MLC, apresentou os melhores resultados. Uma exceção ocorreu quando foram considerados grupos com tamanhos diferentes, neste caso, o MWARD apresentou desempenho inferior ao dos outros métodos.

No trabalho de EDELBROCK e McLAUGLIN (1980), onde foram considerados dados simulados sob duas diferentes distribuições, uma mistura de distribuições normais e uma mistura de distribuições gamas, e empregando quatro diferentes medidas de similaridade,

cinco métodos hierárquicos foram comparados. Um aspecto também investigado foi a presença de grupos com sobreposição, isto é, no espaço das variáveis os grupos apresentavam regiões superpostas. Quatro dos métodos, o MLS, o MLC, o MML e o MCEN, tiveram seus resultados comparados com os do MWARD, este empregando a distância Euclidiana como a dissimilaridade entre os objetos. Para a recuperação dos grupos sem sobreposição, o MWARD apresentou resultados significativamente superiores e, para grupos com sobreposição, os cinco métodos apresentaram resultados sem diferenças significativas.

MILLIGAN (1980) comparou onze métodos hierárquicos e quatro não-hierárquicos. O autor considerou dados simulados segundo uma mistura de distribuições normais multivariadas, truncando as distribuições para assegurar que os grupos não apresentassem sobreposições. Fixado o número de observações nas amostras simuladas, o autor variou a dimensão dos dados, o número de grupos e o número de objetos por grupo e, ainda, introduziu seis condições para os dados: os dados sem erros, os dados com valores discrepantes, a introdução de uma perturbação aleatória na matriz de distâncias, a adição de uma dimensão aleatória nos dados, empregar uma medida de dissimilaridade diferente da distância Euclidiana e a padronização das variáveis. Dentro dos métodos hierárquicos, o MML apresentou os melhores resultados, porém, foi verificado que o método teve seu desempenho fortemente afetado pela presença de valores discrepantes. O MCEN apresentou resultados

inferiores aos demais sob a condição das distâncias perturbadas e, de maneira geral, os piores resultados foram para o MLS.

Treze métodos foram comparados no trabalho de BAYNE et al. (1980), sendo nove deles hierárquicos. Os dados foram simulados segundo duas distribuições normais bivariadas, representando dois grupos, e foram variados o vetor de médias de um dos grupos, as matrizes de dispersão e o número de objetos por grupo. O MWARD, o MLC e o MML apresentaram os melhores resultados enquanto que os piores foram para o MLS. Diferentes de outros trabalhos (KUIPER e FISHER, 1975; MILLIGAN, 1980), os resultados indicaram que MWARD também obteve o melhor desempenho quando os grupos apresentavam tamanhos diferentes.

Um dos trabalhos mais recentes foi desenvolvido por DUBIEN e WARDE (1987), onde o interesse era investigar o efeito da correlação entre as variáveis. Usando dados simulados segundo distribuições normais bivariadas com o tamanho das amostras e o número de grupos fixados, foram variados o número de objetos por grupo e a correlação entre as variáveis. Como as variáveis foram simuladas com variância unitária, os diferentes valores do coeficiente de correlação, ρ, determinavam diferentes formas geométricas para os grupos no espaço das variáveis e, assim, as conclusões foram apresentadas em termos da performance dos métodos sob as formas de grupos ali consideradas. Os melhores resultados para o MLS foram com grupos de forma alongada ($\rho = 0,9$) e para o

MLC com grupos aproximadamente circulares ($\rho = 0,0$). Os autores também concluiram que o MML não é muito afetado pela forma dos grupos e que o MFLE, com o parâmetro $\beta = -0,25$ (ver Capítulo 2), apresentou resultados melhores que os outros métodos considerados, independentemente do valor de ρ.

Dos trabalhos descritos algumas observações podem ser feitas. A maioria deles emprega dados simulados segundo distribuições normais multivariadas, sendo isto justificado pela importância teórica e prática desta distribuição. Entretanto, para a aplicaçao dos métodos de Análise de Agrupamentos não é necessário normalidade para os dados.

Outra observação diz respeito aos métodos hierárquicos comparados, onde parece que alguns deles se destacaram entre os demais. O MWARD, o MML, o MLC e o MFLE apresentaram resultados superiores na maioria dos trabalhos. Já o MLS, na maioria das vezes, não apresentou resultados satisfatórios.

Um outro aspecto observado nos trabalhos desenvolvidos é a forma como foram conduzidas as comparações. Nos trabalhos mais recentes foram construídos experimentos com uma estrutura fatorial, onde os fatores eram aspectos da estrutura dos dados de interesse na investigação. Por exemplo, os fatores foram o número de grupos, a dimensão dos dados, o número de objetos por grupo e o coeficiente de correlação entre outros. Esses fatores foram cruzados e sob as

combinações de seus níveis as amostras foram simuladas obtendo, assim, os dados com as estruturas desejadas.

Visando atingir os objetivos propostos neste trabalho, foi montado um experimento com estrutura fatorial, onde os fatores foram escolhidos para simular dados com estruturas consideradas não muito bem exploradas em trabalhos anteriores. Dois aspectos, pelo menos, parecem não ter sido convenientemente explorados, a questão da matriz de dispersão dos grupos e a da sobreposição ("*overlapping*") dos grupos. O emprego da norma L_1 como dissimilaridades entre os objetos, também não foi considerada em trabalhos anteriores. Embora também estejam incluídos fatores considerados em trabalhos anteriores, neste trabalho foram selecionados fatores para simular estes aspectos mencionados e, desta forma, os fatores inicialmente escolhidos foram:

1) o número de variáveis caracterizando as observações
2) a matriz de dispersão dentro dos grupos
3) a correlação entre as variáveis
4) a presença e ausência de grupos sobrepostos
5) a presença de um erro aleatório nas componentes das observações
6) a medida de dissimilaridade entre os objetos.

Anteriormente foi comentado que as variáveis dentro dos vetores caracterizadores dos objetos podem ser de diferentes tipos, entretanto, neste trabalho esta questão está limitada ao caso onde

todas essas variáveis são contínuas. Em virtude dessa limitação e por razões já discutidas, os dados foram simulados segundo distribuições normais e, para o caso da sobreposição dos grupos, foram empregadas distribuições normais "contaminadas".

No Capítulo 2 serão descritos detalhadamente cada um dos métodos de agrupamento escolhido. A descrição dos fatores, os níveis considerados e a completa descrição do experimento serão dadas no Capítulo 3 e, além disso, serão descritos os aspectos computacionais envolvidos na simulação dos dados.

CAPÍTULO 2

REVISÃO DOS MÉTODOS DE AGRUPAMENTO

2.1 INTRODUÇÃO.

Uma das maiores dificuldades encontradas para um estudo mais aprofundado dos métodos de agrupamento é a falta de definições formais dos termos envolvidos no problema da Análise de Agrupamentos, a começar pela definição de grupo, o termo "*cluster*" na literatura, a qual tem tido apenas um sentido intuitivo sem qualquer definição formal (EVERITT, 1980). Segundo alguns autores, muitas definições propostas empregam termos os quais são vagos não permitindo, assim, uma completa caracterização da definição. Termos como similaridade, semelhança e grupos naturais as vezes empregados, carecem eles mesmos de uma definição formal.

Uma definição de grupo foi dada por EVERITT (1980). Dados os objetos mensurados sob p variáveis, ele considera esses objetos como pontos no espaço p-dimensional, onde cada variável representa um dos eixos deste espaço, ou seja, o espaço das variáveis. Um grupo é definido como uma região contínua deste espaço contendo uma densidade de pontos relativamente alta e separada de outras regiões semelhantes por regiões com uma densidade de pontos relativamente baixa.

Considerada esta definição dada por EVERITT, os grupos podem ser, simplesmente, definidos como regiões de alta densidade de pontos. Ainda assim, permanecem questões em aberto que, em muitas situações, levam a tomadas de decisões subjetivas. Dependendo do problema, pode ser necessário ter que decidir, como um primeiro critério na determinação dos grupos, entre grupos que apresentem uma acentuada separação ou por grupos mais compactos, no sentido de maior homogeneidade dentro dos grupos, ou seja, em grupos com menor dispersão. SEBER (1984) ilustra de uma forma bastante simples essa subjetividade em decidir como considerar os grupos. Ele apresenta o gráfico da Figura 2.1 e comenta: "se a compacticidade é o aspecto mais importante, podemos decidir pela presença de dois ou mais grupos, se a clareza da separação é mais importante nós, provavelmente, decidiremos pela presença de dois grupos".

O que fica claro é que não há uma concordância universal do que constitui um grupo e, provavelmente, uma única definição não

seja suficiente (EVERITT, 1980). WILLIAMS et al. (1971) argumentam que, talvez, fosse melhor evitar completamente o emprego da definição de grupos, dada a quantidade de diferentes considerações sobre eles. Isto é coerente dado o caráter exploratório do emprego dos métodos de Análise de Agrupamentos.

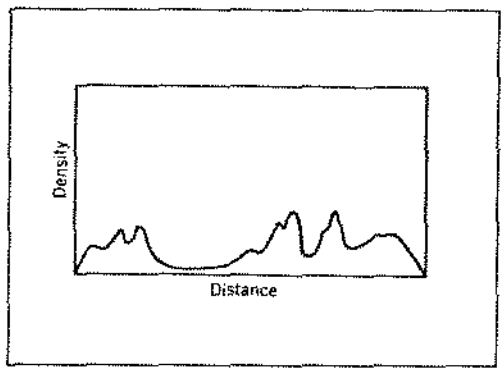


Figura 2.1 Densidade dos pontos.

Um outro agravante às dificuldades expostas é que a maioria dos métodos de agrupamento apresentam tendências a determinar grupos com uma forma particular. Entretanto, na maioria das vezes, em uma análise visando detectar a existência de grupos nos dados não há motivos "*a priori*" para acreditar que os grupos tenham alguma forma específica. Assim, se um determinado método que não seja o mais "adequado" for empregado na análise, o resultado pode ser a imposição de uma estrutura de grupos em vez de realmente recuperar a verdadeira estrutura de grupos embutida nos dados (KALKSTEIN et al., 1987).

Apesar dos aspectos subjetivos envolvidos na aplicaçao de

Análise de Agrupamentos, alguns autores desenvolveram trabalhos estabelecendo definições e propriedades em um contexto matemático mais formal, dentro das quais estudaram e compararam muitos dos métodos mais comuns. JOHNSON (1967), LANCE e WILLIAMS (1968), JARDINE e SIBSON (1968), FISHER e VAN NESS (1971), MILLIGAN (1979), DuBIEN e WARDE (1979), WONG (1982) e WONG e LANE (1983) são alguns exemplos desses trabalhos.

DuBIEN e WARDE (1979) argumentam que os problemas associados com as definições em Análise de Agrupamentos podem ser resolvidos, pelo menos parcialmente, com uma abordagem matemática para os mesmos. Eles consideram os objetos a serem agrupados como um vetor

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}), \; i = 1, 2, \ldots, N,$$

onde as componentes x_{ij} são as respostas sob as variáveis mensuradas. O conjunto de todos os objetos,

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

ao qual eles denominam espaço dos objetos, é considerado embutidos na espaço Euclidiano p-dimensional. Com essa abordagem, os autores dão as seguintes definições:

DEFINIÇÃO 2.1. Um grupo $\mathbf{Y}_K$ é qualquer subconjunto não

vazio do espaço dos objetos.

DEFINIÇÃO 2.2. Um agrupamento **Y** é qualquer partição do espaço dos objetos. Também, $\mathbf{Y} = \{ \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_K \}$ é uma partição de **X** se as condições a seguir se verificam:

(i) para cada $\mathbf{Y}_L \in \mathbf{Y}$, $\mathbf{Y}_L \neq \emptyset$ (conjuto vazio)

(ii) se $\mathbf{Y}_L \in \mathbf{Y}$, $\mathbf{Y}_M \in \mathbf{Y}$ e $\mathbf{Y}_L \neq \mathbf{Y}_M$, então $\mathbf{Y}_L \cap \mathbf{Y}_M = \emptyset$

(iii) $\bigcup_{k=L}^{k} \mathbf{Y}_k = \mathbf{Y}$

Assim, pelas definições um grupo é simplesmente uma coleção de pontos e um agrupamento é um tipo especial de coleção de grupos.

Na seção seguinte serão feitas considerações sobre as técnicas Hierárquicas de agrupamentos e os métodos empregados neste trabalho serão descritos, buscando suas caracterizações através de suas propriedades.

2.2 TÉCNICAS HIERÁRQUICAS DE AGRUPAMENTO.

O interesse aqui são pelas técnicas hierárquicas, uma vez que os métodos de agrupamento considerados neste trabalho pertencem a esta classe de métodos. As técnicas hierárquicas podem ser subdivididas em aglomerativas e divisivas. As técnicas hierárquicas

aglomerativas procedem por considerar, no início do processo de agrupamento, os N objetos a serem agrupados como N grupos distintos e, através de sucessivas fusões, determinam um único grupo de N objetos ao final do processo. As técnicas divisivas, de maneira oposta, consideram inicialmente um único conjunto de N objetos e, através de sucessivas divisões, chegam a N grupos distintos. Os métodos a serem estudados aqui são todos aglomerativos.

Os métodos hierárquicos aglomerativos operam basicamente sobre a matriz de similaridades ou de dissimilaridades entre os objetos, a qual, por conveniência de exposição, será referenciada como matriz de dissimilaridades. Como posto em ANDERBERG (1973), esses métodos seguem um procedimento mais ou menos geral, com os passos dados a seguir:

1. Iniciam com N grupos cada um consistindo de um único objeto. Os grupos serão denotados por $\mathbf{Y}_1$, $\mathbf{Y}_2$, ..., $\mathbf{Y}_N$.
2. Determinam na matriz de dissimilaridades o par de grupos com menor dissimilaridade. Os grupos selecionados serão denotados por $\mathbf{Y}_I$ e $\mathbf{Y}_J$ enquanto d_{IJ} denotará a dissimilaridade entre eles.
3. Unem os dois grupos selecionados formando um novo grupo $\mathbf{Y}_L = \mathbf{Y}_I \cup \mathbf{Y}_J$. Calculam as novas medidas de dissimilaridades entre o novo grupo $\mathbf{Y}_L$ e os outros grupos restantes. O número total de grupos é diminuido de

um.

4. Executam os passos 2 e 3 (N - 1) vezes, isto é, até que todos os objetos estejam em um único grupo.

A diferença entre os distintos métodos consiste na forma pela qual definem o par com "menor dissimilaridade" (passo 2) e na maneira pela qual determinam as dissimilaridades entre o novo grupo formado e os grupos restantes (passo 3).

Os métodos hierárquicos geram partições do conjunto dos objetos de tal forma que, em cada estágio da análise, os grupos são disjuntos e contém os grupos do estágio anterior ou são disjuntos deles. Isto, portanto, cria uma hierarquia entre as partições geradas nas distintas fases. Dado isto, é comum representar a hierarquia obtida através de um dendograma, o qual é um diagrama bidimensional ilustrando as fusões, ou divisões, realizadas em cada sucessivo estágio da análise. Os grupos formados em cada nível do dendograma constituem as partições do conjunto de observações.

DUBIEN e WARDE (1979) também apresentam definições para hierarquia e métodos hierárquicos de agrupamento. Com $\mathbf{Y}^K$ denotando um agrupamento de tamanho K (com K grupos), eles dão as seguintes definições:

DEFINIÇÃO 2.3. Uma hierarquia, *H*, sobre o espaço dos

objetos é uma sequência ordenada de agrupamentos, onde cada agrupamento está contido no anterior. Simbolicamente,

H: $\mathbf{Y}^N$, $\mathbf{Y}^{N-1}$, ..., $\mathbf{Y}^2$, $\mathbf{Y}^1$ onde $\mathbf{Y}^N \subsetneq \mathbf{Y}^{N-1} \subsetneq \ldots \subsetneq \mathbf{Y}^2 \subsetneq \mathbf{Y}^1$

DEFINIÇÃO 2.4. Um método aglomerativo de agrupamento é qualquer método de agrupamento, m, o qual produz uma hierarquia sobre o espaço dos objetos sujeito às seguintes restrições:

(i) $\mathbf{Y}^N$ é o agrupamento inicial

(ii) O agrupamento $\mathbf{Y}^{K-1}$, $K \leq N$, é obtido a partir do agrupamento $\mathbf{Y}^K$ por unir os dois grupos mais "próximos" em $\mathbf{Y}^K$, isto é, se $\mathbf{Y}_I$ e $\mathbf{Y}_J \in \mathbf{Y}^K$ são considerados os grupos mais "próximos", então $\mathbf{Y}_I \cup \mathbf{Y}_J \in \mathbf{Y}^{K-1}$.

Desta forma, segundo os autores, a aplicação de um método aglomerativo sobre o conjunto dos objetos resulta num tipo especial de hierarquia e, através disto, impõe uma estrutura hierárquica sobre o espaço dos objetos.

Como um cometário inicial, nota-se que as técnicas aglomerativas tem a desvantagem de não permitir a mudança de um objeto de um grupo para outro. Dois objetos que em algum estágio da análise são unidos para formar um novo grupo, permanecem juntos em um mesmo grupo até o final da análise. Em outra palavras, se uma

partição inicial for inadequada não será possível alterá-la.

Como citado anteriormente, em cada estágio ao formar um novo grupo pela união de grupos do estágio anterior, $\mathbf{Y}_L = \mathbf{Y}_I \cup \mathbf{Y}_J$, as dissimilaridades d_{KL}, onde $\mathbf{Y}_K$ é um outro grupo qualquer, são determinados de forma diferenciada em cada método. LANCE e WILLIAMS (1967) introduziram uma fórmula de recorrência generalizada que permite a determinação de d_{KL}, para muitos dos métodos hierárquicos aglomerativos mais conhecidos. A fórmula é dada por

$$d_{KL} = \alpha_I d_{KI} + \alpha_J d_{JK} + \beta d_{IJ} + \gamma |d_{IK} - d_{JK}|. \qquad (2\text{-}1)$$

Os parâmetros α_I, α_J, γ e β em (2-1), definem cada um dos métodos. Em termos computacionais, um método que possa ser enquadrado nesta fórmula tem a vantagem de que na sua aplicação, em cada estágio da análise, necessita apenas das informações da matriz de dissimilaridades do estágio anterior. Em particular, a maioria dos métodos considerados neste trabalho satisfazem a esta fórmula de recorrência.

Nas subseções a seguir os métodos serão descritos de uma forma pormenorizada.

2.2.1 Método da Ligação Simples (MLS).

Este método é um dos mais antigos e mais simples. A dissimilaridade entre dois grupos, $\mathbf{Y}_K$ e $\mathbf{Y}_L$, é definido como a menor das dissimilaridade dentro de cada par de objetos formado por um objeto pertencente a $\mathbf{Y}_K$ e o outro pertencente a $\mathbf{Y}_L$. Desta forma, a dissimilaridade é dada por

$$d_{KL} = \min_{\substack{k \in Y_K \\ l \in Y_L}} d_{kl} \qquad (2\text{-}2)$$

Observa-se que no estágio inicial, onde cada grupo é formado por um único objeto, d_{KL} é a própria dissimilaridade entre o objeto em $\mathbf{Y}_K$ e o objeto em $\mathbf{Y}_L$.

Em virtude da definição de d_{KL}, o método é as vezes denominado o Método do Vizinho Mais Próximo.

Empregando a fórmula de LANCE e WILLIAMS e considerando como anteriormente $\mathbf{Y}_L = \mathbf{Y}_I \cup \mathbf{Y}_J$ e $\mathbf{Y}_K$ um outro grupo qualquer, as dissimilaridades d_{KL} são determinadas com $\alpha_I = \alpha_J = \frac{1}{2}$, $\beta = 0$ e $\gamma = -\frac{1}{4}$. Portanto,

$$d_{KL} = \frac{1}{2} d_{IK} + \frac{1}{2} d_{KJ} - \frac{1}{4} |d_{IK} - d_{KJ}|. \qquad (2\text{-}3)$$

O MLS é também um dos métodos mais discutidos na literatura gerando, inclusive, algumas controvérsias entre autores . JOHNSON (1967) mostrou algumas propriedades, tais como, a monotonicidade crescente da sequência formada pelas dissimilaridades que geram as uniões dos grupos e a invariância dos agrupamentos obtidos sob transformações monótonas da matriz de dissimilaridades. Quando as dissimilaridades originais são transformadas por uma função estritamente monótona, a hierarquia gerada pelo MLS não é alterada.

LANCE e WILLIAMS (1967) estabeleceram três propriedades dentro das quais, segundo os autores, os métodos deveriam ser considerados. Com base em uma delas, eles definem um método como "contrator de espaço" se ele se comporta como se houvesse uma contração do espaço na vizinhança dos grupos, apresentando uma tendência maior a aproximar os objetos aos grupos já existentes em vez de formar novos grupos. Os autores atribuem esta propriedade ao MLS e o consideram obsoleto argumentando que, com este comportamento, o método encobre os limites dos grupos existentes. Esta observação também foi comentada por CORMACK (1971) e demonstrada de uma maneira mais formal por DUBIEN e WARDE (1979).

A tendência comentada acima, em termos práticos, significa que o método é incapaz de lidar com objetos intermediários aos grupos. Quando entre dois grupos, mesmo grandes e claramente distintos, houver um objeto em uma posição intermediária o método não é capaz de manter estes grupos separados. Um único objeto entre

dois grupos é suficiente para conectá-los. Esta tendência é denominada de "encadeamento" ("*chaining*") na literatura e sendo relevante observar que esta tendência leva a formação de grupos com forma alongada e, por isso, pode ocorrer que objetos em um mesmo grupo, situados em extremos opostos, sejam bastante dissimilares.

Outro aspecto abordado por LANCE e WILLIAMS (1967) foi considerar um método quanto a ser "compatível" ou "incompatível". Um método é "compatível" se as medidas de dissimilaridades entre os grupos, calculadas nos diversos estágios da análise, são do mesmo tipo daquelas do estágio inicial. Será "incompatível" se esta propriedade não se verifica. A importância dessa característica está no fato de que, se as medidas iniciais entre os objetos tiverem algum significado, isto é, forem interpretáveis, as medidas entre os grupos formados também será interpretável. Em particular, o MLS é um método "compatível".

JARDINE e SIBSON (1968) introduziram algumas condições matemáticas numa estrutura axiomática as quais, na visão desses autores, deveriam ser satisfeitas pelos métodos de agrupamento. Ainda segundo os autores, essas condições tomadas conjuntamente, são satisfeitas somente pelo MLS. Como exemplos dessas condições, os autores requerem que: (i) o método deve ser uma "transformação contínua" dos dados, isto é, pequenas mudanças nos dados devem produzir pequenas mudanças nos resultados; (ii) o método deve ser "bem definido", isto é, o resultado é o mesmo para qualquer

permutação do conjunto dos objetos (em alguns métodos a presença de empates entre as dissimilaridades pode levar a diferentes resultados quando alterada a ordem dos objetos); (iii) o método é "invariante" ao efeito de escala, isto é, a multiplicação das dissimilaridades entre os objetos por um fator de escala, $c > 0$, não altera o agrupamento. Outros autores consideram as condições exigidas por JARDINE e SIBSON como muito severas e sem muito sentido prático (WILLIAMS et al., 1971; CORMACK, 1971; GOWER, 1971).

Das condições propostas por JARDINE e SIBSON (1968), uma das mais criticadas, é a condição de "continuidade". Segundo CORMACK (1971), da forma proposta, a continuidade é uma propriedade dos dados e não uma propriedade analítica dos métodos. GOWER (1971) argumenta que seria mais razoável, em vez de rejeitar um método sob essa condição, verificar o quanto os dados poderiam ser perturbados até a continuidade ser violada. Se fossem necessárias somente pequenas perturbações, os dados não seriam adequados para um agrupamento hierárquico.

No processo de agrupamento é necessário apenas a escolha de um mínimo, por isso, a invariância sob transformações que preservem a ordem das dissimilaridades. Do ponto de vista prático, isto significa que algumas medidas são equivalentes quando empregadas como dissimilaridades neste método.

O MLS apresenta uma vantagem relativa sobre outros métodos, ele é um dos poucos que pode gerar grupos com formas alongadas ("serpentinas alongadas multidimensionais"; ANDERBERG, 1973), diferentes das formas usuais como elipsóides ou hiperesferas. Em áreas de estudo onde grupos com estas formas possam ocorrer, o MLS pode ser muito conveniente (KOPP, 1978a; KAUFMAN e ROUSSEEUW, 1990).

2.2.2 Método da Ligação Completa (MLC).

Este método é oposto ao MLS na definição de dissimilaridade entre dois grupos. Aqui, a dissimilaridade entre $\mathbf{Y}_K$ e $\mathbf{Y}_L$ é definida como a maior das dissimilaridades dentro de cada par de objetos formado por um elemento de $\mathbf{Y}_K$ e o outro pertencente a $\mathbf{Y}_L$, isto é,

$$d_{KL} = \max_{\substack{k \in \mathbf{Y}_K \\ l \in \mathbf{Y}_L}} d_{kl} \qquad (2\text{-}4)$$

Pela forma como é definido a dissimilaridade d_{KL}, o método também recebe a denominação de Método do Vizinho Mais Distante.

Como uma observação inicial sobre o método, verifica-se que d_{KL} é a dissimilaridade entre os membros de $\mathbf{Y}_K$ e $\mathbf{Y}_L$ que estão mais afastados. Se $\mathbf{Y}_K$ e $\mathbf{Y}_L$ forem unidos, então quaisquer dois objetos no

grupo resultante terão dissimilaridade no máximo igual a d_{KL}. Se a dissimilaridade for uma distância, d_{KL} será o diâmetro da menor bola que contém os pontos resultantes da união de $\mathbf{Y}_K$ e $\mathbf{Y}_L$.

Este método satisfaz a fórmula recursiva de LANCE e WILLIAMS, sendo definido com $\alpha_I = \alpha_J = \frac{1}{2}$, $\beta = 0$ e $\gamma = \frac{1}{4}$, isto é,

$$d_{KL} = \frac{1}{2}d_{IK} + \frac{1}{2}d_{JK} + \frac{1}{4}\,|d_{IK} - d_{JK}|, \qquad (2\text{-}5)$$

onde, como anteriormente, $\mathbf{Y}_L = \mathbf{Y}_I \cup \mathbf{Y}_J$ e $\mathbf{Y}_K$ é outro grupo qualquer.

Tomando por base a definição original dada em (2-4), o método envolve somente a escolha de um máximo sendo, portanto, invariânte sob transformações monótonas das dissimilaridades. Desta forma, valem aqui os comentários feitos a respeito do MLS para esta propriedade. Com relação a outras propriedades, MLC apresenta características opostas ao MLS.

Segundo a proposta de LANCE e WILLIAMS (1967) este método é classificado como "dilatador de espaço", significando que o mesmo se comporta como se o espaço na vizinhança dos grupos fosse expandido. Neste contexto, isto significa que o método apresenta uma tendência maior a levar os objetos que ainda não tenham sido agrupados a atuarem como núcleos para formação de novos grupos em

vez de conectá-los aos grupos já existentes. Na visão de JOHNSON (1967), é a tendência do método em minimizar o diâmetro dos grupos em cada estágio da análise.

A tendência comentada no parágrafo anterior é exatamente oposta àquela do MLS. Do ponto de vista prático, enquanto o MLS tende a formar poucos grupos com pouca homogeneidade, a tendência do MLC é formar muitos grupos de tamanhos pequenos e mais compactos, isto é, os grupos apresentam uma maior homogeneidade (CORMACK, 1971; KOPP, 1978b). Esta homogeneidade é no sentido de que, dentro dos grupos, as dissimilaridades entre os objetos são pequenas. Uma consequência disto é que objetos relativamente similares permanecerão em grupos diferentes em grande parte da análise, sendo unidos somente nos estágios finais (KAUFFMAN e ROUSSEEUW, 1990).

Ainda dentro da abordagem proposta por LANCE e WILLIAMS (1967), o MLC é um método "compatível", no sentido descrito anteriormente.

Como observado por KOPP (1978b), o MLC tem o inconveniente de que, quando a dissimilaridade mínima ocorre para mais de um par de objetos (grupos), o método pode produzir diferentes agrupamentos dependendo do par de objetos (grupos) selecionados para serem unidos primeiro. Dentro da estrutura proposta por JARDINE e SIBSON (1968), esta característica leva o método a ser considerado como

"mal definido".

Outro incoveniente, abordado por LANCE et al. (1971), é que o MLC é extremamente sensível à presença de valores aberrantes nos dados. Isto também foi verificado empiricamente no trabalho de MILLIGAN (1980).

2.2.3 Método da Média das Ligações (MML).

Para definir a dissimilaridade entre os grupos, em vez de considerar valores extremos como MLS e MLC, este método toma a média das dissimilaridades entre todos os pares de objetos, com cada par formado por um objeto de cada grupo envolvido. Desta forma,

$$d_{KL} = \frac{1}{N_K N_L} \sum_{\substack{k \in Y_K \\ l \in Y_L}} d_{kl}, \qquad (2\text{-}6)$$

onde N_K e N_L são os números de objetos em $\mathbf{Y}_K$ e $\mathbf{Y}_L$, respectivamente.

Pela definição de d_{KL}, o método situa-se entre MLS e o MLC e, segundo KOPP (1978c), tira proveito da estabilidade de um e da propriedade de homogeneidade do outro. Os grupos são caracterizados pela média de todas as dissimilaridades entre seus membros e,

assim, não dependendo dos valores extremos das dissimilaridades, não é possível fazer uma caracterização dos grupos formados em termos de máxima ou mínima associação entre seus membros.

Como observado por CORMACK (1971), este método será usado somente com medidas de dissimilaridades para as quais tenha sentido tomar suas médias.

Pela fórmula recursiva de LANCE e WILLIAMS, o método é definido com $\alpha_I = \frac{N_I}{N_L}$, $\alpha_J = \frac{N_J}{N_L}$, $\beta = 0$ e $\gamma = 0$. Portanto,

$$d_{KL}=\frac{N_I}{N_L}d_{IK}+\frac{N_J}{N_L}d_{JK}, \qquad (2\text{-}7)$$

onde, novamente $\mathbf{Y}_L = \mathbf{Y}_I \cup \mathbf{Y}_J$, $\mathbf{Y}_K$ um outro grupo qualquer e $N(\cdot)$ é o número de objetos de cada grupo ($N_L = N_I + N_J$).

Na abordagem de LANCE e WILLIAMS (1967) algumas considerações interessantes podem ser feitas sobre MML. Ele é "compatível", desde que faça sentido a média das dissimilaridades (restrição comentada anteriormente). Os autores consideram que o método não apresenta tendências marcantes de "contrator de espaço" ou de "dilatador de espaço" e, por isso, o consideram "conservador de espaço". O significado disto é que o MML não tem a tendência do MLS para formar poucos grupos e nem a do MLC para formar muitos grupos.

Neste sentido, o MML seria mais dependente da estrutura existente nos dados que os outros dois métodos.

Dentro das condições propostas por JARDINE e SIBSON (1968), o MML não satisfaz a condição de ser uma "transformação contínua" dos dados. Entretanto, como já foi comentado, do ponto de vista prático, outros autores não consideram isto como um inconveniente para um método de agrupamento.

FISHER e VAN NESS (1971) propuseram uma abordagem buscando estabelecer, para alguns métodos, a admissibilidade sob uma dada propriedade. Os autores formularam propriedades e verificaram se um dado método era admissível, ou não, sob cada uma das propriedades, admissível no sentido de satisfazer a propriedade. No trabalho de KOPP (1978c) é relatado que o MML satisfaz algumas dessas propriedades. Uma delas, denominada "admissibilidade sob a omissão de grupo", diz que, sendo $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_M\}$ uma partição obtida pelo método e sendo $\mathbf{X}$ o conjunto de todos os objetos, se o método for novamente aplicado sobre $\mathbf{X} - \mathbf{Y}_I$, o conjunto $\mathbf{X}$ sem os objetos do grupo $\mathbf{Y}_I$, resultará numa partição com M - 1 grupos que será exatamente $\mathbf{Y} - \mathbf{Y}_I$, a partição $\mathbf{Y}$ reduzida do grupo $\mathbf{Y}_I$.

Uma discusão interessante foi dada por KALKSTEIN et al. (1987). Eles adotaram uma abordagem em termos do espaço Euclidiano; a dissimilaridade entre os objetos era dada pela distância Euclidiana. Considerando d_{kl} em (2-6) como a distância Euclidiana ao

quadrado, os autores mostram que d_{KL} pode ser escrita como

$$d_{KL} = \frac{W_K}{N_K} + \frac{W_L}{N_L} + D_{KL}^2 \ , \qquad (2\text{-}8)$$

onde $\mathbf{W}_K$ e $\mathbf{W}_L$ são as somas de quadrados dentro do grupo $\mathbf{Y}_K$ e $\mathbf{Y}_L$, respectivamente, e D'_{KL} é a distância Euclidiana ao quadrado entre os centróides de $\mathbf{Y}_K$ e $\mathbf{Y}_L$. Nestas condições, verifica-se que o MML, de certa forma, apresenta um vício em unir grupos com menor variância entre seus membros. Para três grupos que tenham seus centróides equidistantes, o método unirá aqueles dois para os quais a soma das variâncias dentro dos grupos for menor. Os autores comentam que apesar da variância ser um critério forte para a formação dos grupos, a distância entre os grupos também o é, por isso, a capacidade do método em minimizar a variância dentro dos grupos e maximizar a variância entre os grupos.

KAUFMAN e ROUSSEEUW (1990) comentam que, quanto a forma, o MML tem uma tendência a determinar grupos com forma grosseiramente esférica mas, apesar disto, o método é relativamente robusto para lidar com grupos com outras formas. Comentário também feito por DUBIEN e WARDE (1987) a partir de investigações empíricas.

2.2.4 Método de McQuitty (MMcQ).

Neste método a dissimilaridade entre os grupos não é explicitamente definida. Diferente dos outros métodos já descritos, aqui é definida apenas a aquação recursiva para determinar as dissimilaridades entre os grupos formados e os grupos restantes.

O método inicia por unir os dois objetos com menor dissimilaridade. Em cada estágio ao unir dois grupos, $\mathbf{Y}_I$ e $\mathbf{Y}_J$, para formar um novo grupo, $\mathbf{Y}_L$, as dissimilaridades são determinadas através da fórmula

$$d_{KL} = \frac{1}{2}d_{IK} + \frac{1}{2}d_{JK}. \qquad (2\text{-}9)$$

Segundo KAUFMAN e ROUSSEEUW (1990), esta fórmula, apesar de muito simples, não corresponde a qualquer definição de dissimilaridade entre grupos. Os autores também ilustram o fato de que resultados contraditórios podem ocorrer com o emprego de (2-9). Eles consideram um exemplo simples, onde um objeto $\mathbf{x}_i$ tem a mesma dissimilaridade com outros dois objetos $\mathbf{x}_j$ e $\mathbf{x}_k$. É mostrado que, se forem unidos primeiramente $\mathbf{x}_i$ e $\mathbf{x}_j$ serão obtidas dissimilaridades diferentes daquelas obtidas com a união inicial de $\mathbf{x}_i$ e $\mathbf{x}_k$. É evidente que os resultados, nos dois casos, deveriam ser iguais. Na abordagem de JARDINE E SIBSON (1968), com a característica descrita acima, o método é considerardo "mal definido".

Apesar deste método não ter sido discutido no trabalho original de LANCE e WILLIAMS (1967), outros autores (SNEATH e SOKAL, 1973; MILLIGAN, 1979) associam ao MMcQ as propriedades propostas naquele trabalho. O método é considerado "compatível" e "conservador de espaço". O método também satisfaz a equação recursiva de LANCE e WILLIAMS, uma vez que em (2-9) verifica-se $\alpha_I = \alpha_J = \frac{1}{2}$, $\beta = 0$ e $\gamma = 0$.

Alguns autores consideram o MMcQ como uma variante do MML, por isso, as vezes aparece na literatura como o método da Média Ponderada das Ligações. KAUFMAN e ROUSSEEUW (1990) justificam o emprego deste nome considerado uma situação onde um grupo $\mathbf{Y}_I$ com muitos objetos é unido a um grupo $\mathbf{Y}_J$ com poucos objetos. Os autores argumentam que ao usar (2-9), os objetos do grupo $\mathbf{Y}_J$ terão um peso muito maior que os objetos do grupo $\mathbf{Y}_I$. Por outro lado, o MML associa o mesmo peso a cada par de objetos, vide equação (2-6), apesar dos aparentes pesos na equação recursiva (2-7).

De maneira geral, o MMcQ é muito pouco discutido na literatura.

2.2.5 Método Centróide (MCEN).

Este método foi inicialmente proposto para lidar com dados representando observações de variáveis contínuas. No desenvolvimento do processo, os grupos são definidos no espaço

Euclidiano e são representados pelas coordenadas dos seus centróides. Em cada estágio, são unidos aqueles dois grupos com a menor distância entre seus centróides, os dois grupos mais similares neste sentido.

Sendo cada objeto descrito por um vetor $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, i = 1, 2, ..., N, o centróide do grupo $\mathbf{Y}_K$ é dado por

$$\bar{\boldsymbol{x}}_K = (\bar{x}_{K1}, \bar{x}_{K2}, \ldots, \bar{x}_{Kp}),$$

onde

$$\bar{x}_{Kf} = \frac{1}{N_K} \sum_{g \in Y_K} x_{gf} \qquad (2\text{-}10)$$

e N_K é o número de objetos em $\mathbf{Y}_K$, para f = 1, 2, ..., p.

No MCEN a dissimilaridade entre dois grupos, $\mathbf{Y}_K$ e $\mathbf{Y}_L$, é dada pela distância Euclidiana entre seus centróides, isto é,

$$d_{KL} = \|\bar{\boldsymbol{x}}_K - \bar{\boldsymbol{x}}_L\|. \qquad (2\text{-}11)$$

Se na matriz de dissimilaridade original a dissimilaridade

entre os objetos é dada pela distância Euclidiana ao quadrado, pode ser mostrado (KAUFMAN e ROUSSEUW, 1990) que o método é definido através da fórmula recursiva de LANCE e WILLIAMS com $\alpha_I = \frac{N_I}{N_L}$, $\alpha_J = \frac{N_J}{N_L}$, $\beta = \alpha_I\alpha_J$ e $\gamma = 0$, portanto,

$$d_{KL} = \frac{N_I}{N_L} d_{IK} + \frac{N_J}{N_L} d_{JK} - \frac{N_I N_J}{(N_L)^2} d_{IJ} \qquad (2\text{-}12)$$

Dentro da abordagem proposta por LANCE e WILLIAMS (1967), o MCEN é considerado "conservador de espaço" e é "compatível" para medidas de dissmilaridades que satisfaçam as propriedades de distância. A equação dada em (2-12), vale somente para a medida de dissimilaridade ali considerada. A respeito disto, KAUFMAN e ROUSSEUW (1990) observam que a equação (2-12) pode ser usada com outras medidas de dissimilaridade, porém, isto não é recomendado pois pode levar a resultados que não possam ser interpretados.

Em relação a questão da medida de dissimilaridade empregada, ANDERBERG (1973) foi mais radical. Ele afirma que o método pode ser usado com qualquer medida de dissimilaridade, porém, os resultados somente terão uma interpretação clara se a dissimilaridade for a distância Euclidiana, mesmo sem empregar a fórmula recursiva (2-12). Por esta observação, o método tem o incoveniente de ter seu emprego restrito a uma dada medida de dissimilaridade.

ANDERBERG (1973) também comenta outra desvantagem do MCEN. Se os tamanhos dos dois grupos unidos forem muito diferentes, o centróide do grupo resultante estaria muito próximo do grupo maior e, podendo mesmo, ficar dentro deste grupo. Se isto ocorre, as características próprias do grupo menor, no sentido aqui considerado, estarão virtualmente perdidas.

A observação comentada acima pode ser vista escrevendo o centróide do grupo resultante como função dos centróides dos grupos unidos. Unindo dois grupos, $\mathbf{Y}_I$ e $\mathbf{Y}_J$, o centróide do grupo resultante, $\mathbf{Y}_L$, é dado por

$$\bar{\boldsymbol{x}}_L = \frac{N_I}{N_L}\bar{\boldsymbol{x}}_I + \frac{N_I}{N_L}\bar{\boldsymbol{x}}_J \qquad (2\text{-}13)$$

Por (2-13) observa-se que o centróide de $\mathbf{Y}_L$ está sobre um segmento de reta, entre $\bar{\boldsymbol{x}}_I$ e $\bar{\boldsymbol{x}}_J$, e mais próximo do centróide do grupo com maior número de objetos. Isto torna os grupos maiores mais influentes que os de menor tamanho.

KALKSTEIN et al. (1987) comentaram uma tendência do MCEN que eles denomimam de "bola de neve". Os autores ilustram a tendência do método em formar um ou dois grupos grandes, contendo a maioria dos objetos e que crescem por absorver grupos pequenos, e numerosos grupos de um único objeto.

Outro aspecto considerado inconvenientemente neste método é o fato de que os níveis de dissimilaridade sob os quais os grupos são unidos, não formam uma sequência monótona. Pode ocorrer que a dissimilaridade de uma dada união seja menor que a dissimilaridade de uma união em um estágio anterior. MILLIGAN (1979), empregando a fórmula recursiva de LANCE e WILLIAMS, mostrou que as condições, necessárias e suficientes, para a monotonicidade das dissimilaridades de união são:

$$\begin{aligned} &\text{(i)} \quad \alpha_I+\alpha_J+\beta \geq 1 \ ; \\ &\text{(ii)} \quad \alpha_I+\alpha_J \geq 0 \ ; \\ &\text{(iii)} \quad \gamma \geq -\min(\alpha_I, \alpha_J) \ . \end{aligned} \tag{2-14}$$

Em (2-12) a condição (i) não está satisfeita. O inconveniente da falta de monotonicidade está no fato de dificultar a apresentação dos resultados de uma forma conveniente, por exemplo, em um dendograma (KAUFMAN e ROUSSEEUW, 1990).

Segundo KUIPER e FISHER (1975), uma possível situação onde o MCEN é bastante apropriado, é quando for aceitável supor que os grupos embutidos na amostra sejam de tamanhos bastante diferenciados.

2.2.6 Método da Mediana (MMED).

Foi visto que no MCEN os grupos maiores exercem uma influência maior na determinação dos centróides dos grupos

resultantes das uniões. GOWER (1967) argumentou que nem sempre isto é conveniente e propôs um método onde os grupos sejam ponderados igualmente, independente do número de objetos. Na proposta, os grupos não são representados pelos seus centróides e sim por uma quantidade denominada "centro" do grupo.

Dentro da proposta de GOWER, no estágio inicial onde os grupos são formados por um único objeto, o centro do grupo $\mathbf{Y}_I$, $\mathbf{c}_I$, são as próprias mensurações do objeto neste grupo (as coordenadas do ponto representando este objeto). Ao unir dois grupos $\mathbf{Y}_I$ e $\mathbf{Y}_J$, o centro do grupo resultante $\mathbf{Y}_L$, é dado por

$$\boldsymbol{c}_L = \frac{1}{2}\boldsymbol{c}_I + \frac{1}{2}\boldsymbol{c}_J. \qquad (2\text{-}15)$$

A dissimilaridade entre um grupo $\mathbf{Y}_K$ e um grupo $\mathbf{Y}_L$ é dada pela distância Euclidiana entre seus centros, então,

$$d_{KL}=\|\boldsymbol{c}_K-\boldsymbol{c}_L\| \qquad (2\text{-}16)$$

GOWER justificou esta formulação, que é uma variante do MCEN, considerando uma situação onde vários objetos idênticos estejam incluidos na amostra. Estes objetos seriam representados por pontos idênticos e, de certa forma, viciariam o grupo formado por eles através do número de objetos no grupo. Esta seria uma situação onde a formulação do MMED seria mais conveniente que a do MCEN.

Segundo LANCE e WILLIAMS (1967), o MMED é considerado "conservador de espaço" e completamente "compatível" para dissimilaridades dadas pela distância Euclidiana, medida para a qual GOWER definiu o método. Para obter a equação recursiva, basta fazer $\alpha_I = \alpha_J = \frac{1}{2}$ e $\beta = -\frac{1}{4}$ em (2-12). A equação é então

$$d_{KL} = \frac{1}{2}d_{IK} + \frac{1}{2}d_{JK} - \frac{1}{4}d_{IK} \qquad (2\text{-}17)$$

LANCE E WILLIAMS (1967) observam que ao tomar um triângulo no espaço Euclidiano com os vértices dados pelos centros dos grupos $\mathbf{Y}_I$, $\mathbf{Y}_J$ e $\mathbf{Y}_K$, o centro do grupo $\mathbf{Y}_L$ está situado no ponto médio do menor lado deste triângulo [vide (2-15)]. A distância d_{KL} é o comprimento da mediana que tem um dos extremos no vértice dado pelo centro de $\mathbf{Y}_K$. Isto sugere o nome dado ao métodos, embora, também seja denominado Método de GOWER.

KAUFMAN e ROUSSEEUW (1990) comentam que a equação (2-17) não dá uma definição explícita de dissimilaridade entre grupos, devido ao fato de que os centros dos grupos não são bem definidos. Os autores argumentam que os centros dos grupos não são definidos de uma forma única, uma vez que eles dependem da ordem em que os grupos são formados e, com isso, a equação (2-17) pode levar a resultados contraditórios como aqueles comentados com respeito ao MMcQ [Equação (2-9)].

De maneira geral, o MMED tem propriedades análogas àquelas do MCEN, além disso, carece das mesmas propriedades como, a invariância dos resultados sob transformações das dissimilaridades e da monotonicidade na sequência formada pelas dissimilaridades que geram as uniões. KOPP (1978c) comenta que a escolha entre um método e outro é uma decisão que depende de objetivos práticos. O autor exemplifica com uma situação onde há informação "a priori" da existência de grupos pequenos, porém, importantes na amostra e considera que neste caso o MMED deve ser preferido em relação ao MCEN.

Ainda sobre a semelhança entre o MCEN e o MMED, um comentário interessante foi feito por CUNNINGHAM e OLGILVIE (1972). Eles simularam dados com uma estrutura hierárquica e submeteram a vários métodos de agrupamentos. Foi observado que o MCEN e o MMED introduziram distorções sobre os dados e, desta forma, os autores sugerem que estes métodos não são apropriados quando o objetivo da análise for a obtenção de estruturas hierárquicas.

2.2.7 Método de Ward (MWARD).

O procedimento empregado neste método é proveniente de uma proposta de método de agrupamento dada por WARD (1963). Na proposta o autor descreve um método muito geral, onde em cada estágio do agrupamento os grupos são formados de maneira que um dado critério,

denominado função objetivo, seja otimizado. O autor ilustrou a aplicação do método usando como função objetivo a soma de quadrados dos desvios (SQ) em relação a média dentro dos grupos, em cada estágio eram unidos aqueles grupos para os quais ocorria um acréscimo mínimo na SQ.

A SQ dentro de um grupo $\mathbf{Y}_K$ é definida como a soma das distâncias Euclidianas ao quadrado entre as coordenadas dos objetos neste grupo e seu centróide, isto é,

$$SQ_K = \sum_{i \in \mathbf{Y}_K} \|\mathbf{x}_i - \bar{\mathbf{x}}_K\|^2 \qquad (2\text{-}18)$$

Em cada estágio é considerada a SQ total entre os grupos, dada por

$$SQ_{TOTAL} = \sum_{i=1}^{G} SQ_i,$$

onde G é o número de grupos no estágio.

No estágio inicial, onde cada grupo é composto de um único objeto, é evidente que $SQ_{TOTAL} = 0$. Nos estágios posteriores, ao ser efetuada uma união $\mathbf{Y}_L = \mathbf{Y}_I \cup \mathbf{Y}_J$, o acréscimo na SQ_{TOTAL} é dado por

$$\Delta SQ_{TOTAL} = SQ_L - SQ_I - SQ_J, \qquad (2\text{-}20)$$

uma vez que as SQ nos outros grupos não se alteram. Então, pelo exposto, o objetivo é unir em cada estágio os dois grupos para os quais ΔSQ_{TOTAL} seja mínimo.

WISHART (1969) mostrou que a equação (2-20) pode ser escrita como

$$\Delta SQ_{TOTAL} = \frac{N_I N_J}{N_L} \|\bar{\mathbf{x}}_I - \bar{\mathbf{x}}_J\|^2 \qquad (2\text{-}21)$$

Por (2-21), verifica-se que o acréscimo na SQ_{TOTAL} é proporcional a distância Euclidiana ao quadrado entre os centróides dos grupos unidos.

O MWARD usa a formulação dada em (2-21) como a dissimilaridade entre os grupos. Em um dado estágio os grupos $\mathbf{Y}_I$ e $\mathbf{Y}_J$ serão unidos se eles têm a menor dissimilaridade, onde esta é dissimilaridade é dada por

$$d_{IJ} = \frac{N_I N_J}{N_I + N_J} \|\bar{\mathbf{x}}_I - \bar{\mathbf{x}}_J\|^2 \qquad (2\text{-}22)$$

Este método satisfaz a equação recursiva de LANCE e WILLIAMS (KAUFMAN e ROUSSEEUW, 1990), a qual é dada por

$$d_{KL} = \frac{N_I N_K}{N_L + N_K} d_{IK} + \frac{N_J N_K}{N_L + N_K} d_{JK} - \frac{N_K}{N_L + N_K} d_{IJ}. \qquad (2\text{-}23)$$

Da equação (2-23), verifica-se que a sequência de dissimilaridades de união é monótona, uma vez que os coeficientes satisfazem as condições dadas em (2-14). Isto também pode ser verificado ao relacionar-se (2-21) com (2-22), pois o valor de ΔSQ_{TOTAL} em cada estágio é , no mínimo, igual ao do estágio anterior.

Em princípio, o MWARD é indicado para agrupar objetos mensurados sob variáveis contínuas, empregando a distância Euclidiana como dissimilaridade entre os objetos. Alguns autores consideram esta questão de uma forma mais amena. ANDERBERG (1973) argumenta que o método pode ser empregado com qualquer medida de dissimilaridade para agrupar objetos ou variáveis, embora isso gere dificuldades para analisar os grupos obtidos. KOPP (1978c) afirma que o emprego da distância Euclidiana não é tão restritivo como no caso do MCEN e do MMED.

O MWARD apresenta uma tendência a formar grupos com o mesmo número de objetos (KALKSTEIN et al., 1987). Da equação (2-22), verifica-se que, se um grupo está situado à mesma distância com relação a outros dois grupos com diferentes número de objetos, ele será unido àquele grupo com menor número de objetos, isto é, o método dá prioridade a unir grupos com poucos objetos. Este comportamento, de certa forma, gera uma tendência a formar grupos

de mesmo tamanho.

De trabalhos de investigações empíricas, algumas considerações foram feitas sobre o MWARD. Numa situação onde os dados foram simulados segundo uma distribuição normal multivariada, com a mesma matriz de dispersão e igual número de objetos por grupo, KUIPER e FISHER (1975) relatam que o método classificou os objetos tão bem quanto uma função linear discriminante com parâmetros conhecidos. EVERITT (1980) relatou que o método recuperou corretamente grupos simulados com distribuição normal bivariada, com as matrizes de dispersão iguais e as variáveis independentes ($\rho = 0$), porém, foi ineficiente em recuperar os grupos quando introduzida uma correlação entre as variáveis ($\rho = 0,75$). No trabalho de MILLIGAN (1980), o autor concluí que o desempenho do método é muito afetado pela presença de observações com valores aberrantes nos dados.

2.2.8 Método Beta-Flexível (MFLE).

Este método foi proposto por LANCE e WILLIAMS (1967). Os autores derivaram o método a partir da fórmula recursiva que eles introduziram, equação (2-1), impondo as seguintes retrições sobre os parâmetros:

$$\text{i)} \quad \alpha_I + \alpha_J + \beta = 1;$$

$$\text{ii)} \quad \alpha_I = \alpha_J; \qquad (2\text{-}24)$$

iii) $\beta < 1$;

iv) $\gamma = 0$.

Considerando as restrições acima, o método tem a equação recursiva dada por

$$d_{KL} = \frac{1}{2}(1-\beta)\,d_{IK} + \frac{1}{2}(1-\beta)\,d_{JK} + \beta d_{IJ}. \qquad (2\text{-}25)$$

Da equação (2-25), verifica-se que o método depende do valor de β, o que sugere a sua denominação. Na prática, o valor do parâmentro β é especificado pelo usuário.

Embora a equação (2-25) não represente qualquer definição de dissimilaridade entre os grupos, o processo de agrupamento é semelhante aos métodos já descritos. Em cada estágio são unidos os grupos com menor dissimilaridade e, após essas uniões, as dissimilaridades são determinadas empregando a equação (2-25).

Quanto às características do método, evidentemente que elas dependem do valor especificado para o parâmetro β. No trabalho de LANCE E WILLIAMS (1967) é comentado que o método pode ser extremamente "contrator de espaço" especificando o valor β suficientemente próximo a 1 e, com isso, levar a formação de grupos com formas alongadas devido a presença do "encadeamento". Decrescendo β para valores negativos, o método torna-se gradativamente "dilatador de espaço" e, desta forma, podem ser

obtidos grupos mais homogêneos. Os autores argumentam que se consideram incapazes de definir rigorosamente o valor de β para o qual o método possa ser considerado "conservador de espaço", entretanto, sugerem que este valor deve ser um número negativo, pequeno em valor absoluto, e propõem usar $\beta = -0,25$.

O método é "compatível" para a distância Euclidiana. As condições dadas em (2-14) são satisfeitas pelos coeficientes da equação (2-25), portanto, as dissimilaridades de união de grupos formam uma sequência não decrescente.

Na visão de KOPP (1978c), a dependência sob o valor de β pode ser vista como uma vantagem do MFLE com relação a outros métodos. Segundo o autor, empregando diferentes valores de β, várias análises podem ser executadas sobre os dados e, desta forma, uma possível estrutura de grupos estável pode ser identificada.

2.2.9 Método do k-ésimo Vizinho Mais Próximo (MkVP).

WONG e LANE (1983) propuseram este método, considerando a situação onde os objetos formam uma amostra de uma população, de tal forma a permitir fazer-se inferência sobre os grupos existentes nesta população. No desenvolvimento do método, os autores empregaram o conceito de grupos de alta-densidade. Considerando a

população dos objetos tendo densidade **f** e os objetos como pontos no espaço p-dimensinal, para todo $\mathbf{f}_0 > 0$, um grupo de alta-densidade no nível $\mathbf{f}_0$ é um subconjunto C do espaço p-dimensional, tal que seus pontos **x** satisfazem a $\mathbf{f}(\mathbf{x}) > \mathbf{f}_0$, isto é,

$$C = \{ \mathbf{x} \mid \mathbf{f}(\mathbf{x}) > \mathbf{f}_0 \}$$

Outro conceito utilizado é o de árvore de grupos. Uma árvore é uma família **T** de grupos onde se $\mathbf{Y}_I$ e $\mathbf{Y}_J \in \mathbf{T}$, uma das três condições a seguir é satisfeita:

$$\begin{aligned} &\text{(i)} \quad \boldsymbol{Y}_I \subset \boldsymbol{Y}_J \\ &\text{(ii)} \quad \boldsymbol{Y}_J \subset \boldsymbol{Y}_I \\ &\text{(iii)} \quad \boldsymbol{Y}_I \cap \boldsymbol{Y}_J = 0 \ . \end{aligned} \tag{2-26}$$

É relevante observar que os grupos de alta-densidade formam uma árvore (HARTINGAN, 1975).

Considerando uma amostra de observações, $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_N$, da população com densidade **f** e sendo **T** uma árvore de grupos de alta-densidade definidos sobre **f**, a idéia do método é usar esta amostra para gerar uma estimativa $\mathbf{T}_N$ de **T**. A estimativa T_N é uma árvore de grupos de alta-densidade definidos sobre $\mathbf{f}_N$, onde $\mathbf{f}_N$ é uma estimativa da densidade **f**. Para obter $\mathbf{f}_N$ é empregado o estimador de densidade do k-ésimo vizinho mais próximo, o que sugere a denominação do método.

O estimador do k-ésimo vizinho mais próximo é definido como segue: a estimativa da densidade no ponto $\mathbf{x}_i$ é dada por

$$f_N(\boldsymbol{x}_i) = \frac{k}{NV_K(\boldsymbol{x}_i)} ,$$

onde $V_k(\mathbf{x}_i)$ é o volume da menor esfera, centrada em $\mathbf{x}_i$, que contém k observações de uma amostra de tamanho N.

Por conveniência de exposição, os objetos serão referenciados como observações da amostra nas definições a seguir dadas por WONG e LANE (1983).

DEFINIÇÃO 2.5. Duas observações, $\mathbf{x}_i$ e $\mathbf{x}_j$, são ditas vizinhas se

$$d_{\boldsymbol{x}_i\boldsymbol{x}_j} \leq \max[d_k(\boldsymbol{x}_i), d_k(\boldsymbol{x}_j)] ,$$

onde $d_{\boldsymbol{x}_i\boldsymbol{x}_j}$ é a distância entre $\mathbf{x}_i$ e $\mathbf{x}_j$, enquanto que $d_k(\mathbf{x}_i)$ é a distância do k-ésimo vizinho mais próximo ao ponto $\mathbf{x}_i$, ambas sob a métrica Euclidiana.

DEFINIÇÃO 2.6. A distância d' entre duas observações $\mathbf{x}_i$ e $\mathbf{x}_j$ é dada por

$$d^*_{x_i x_j} = \frac{1}{2}\left[\frac{1}{f_N(x_i)} + \frac{1}{f_N(x_j)}\right] = \frac{n}{2k}\left[V_k(x_i) + V_k(x_j)\right], \quad \text{se } \mathbf{x}_i \text{ e } \mathbf{x}_j \text{ são vizinhos}$$

$d^*_{x_i x_j} = \infty$, caso contrário.

O método emprega d^* como a dissimilaridade entre as observações e realiza o agrupamento em dois passos:

1. Determina a matriz $\mathbf{D}^*$, a matriz de distâncias entre as observações, isto é,

$$D^* = [d^*_{x_i x_j}] \ , \ i, j = 1, 2, \ldots, N.$$

2. Aplica o método da Ligação Simples (MLS) sobre a matriz $\mathbf{D}^*$ para obter a árvore amostral de grupos.

Os autores provam que a árvore de grupos produzida por este método é fortemente consistente para a árvore populacional $\mathbf{T}$. Um procedimento de agrupamento, ou equivalentemente a árvore $\mathbf{T}_N$, é dito ser fortemente consistente para grupos de alta-densidade, ou para $\mathbf{T}$, se para quaisquer dois grupos $\mathbf{Y}_I$ e $\mathbf{Y}_J \in \mathbf{T}$, satisfazendo a $\mathbf{Y}_I \cap \mathbf{Y}_J = \emptyset$, verifica-se

$$p(\mathbf{Y}_{I(N)} \cap \mathbf{Y}_{J(N)} = \emptyset, \textit{ quando } N \to \infty) = 1 \qquad (2\text{-}27)$$

onde $\mathbf{Y}_{I(N)}$ e $\mathbf{Y}_{J(N)}$ são os menores grupos em $\mathbf{T}_N$ contendo os pontos amostrais pertencentes, respectivamente, a $\mathbf{Y}_I$ e $\mathbf{Y}_J$.

O resultado limite dado em (2-27), significa que a relação de grupos em $\mathbf{T}_N$ converge fortemente para a relação de grupos em $\mathbf{T}$. Para provar este resultado, é usado o fato de que o estimador de densidade empregado é uniformemente consistente para $\mathbf{f}$, com probabilidade um, desde que $\mathbf{f}$ seja uniformemente contínua e que $k = k(N)$, quando $N \to -$, satisfaça a

$$\text{(i)} \quad \frac{k(N)}{N} \to 0$$

$$\text{(ii)} \quad \frac{k(N)}{\log N} \to \infty.$$

Os autores do método afirmam, ainda, que um método de agrupamento consistente não impõe qualquer estrutura geométrica sobre os grupos produzidos.

Fica claro que o método depende do valor de k, com $1 \leq k \leq N$. Sobre o valor deste parâmetro, os autores afirmam não haver uma recomendação única para a escolha desse valor, entretanto, sugerem uma regra prática escolhendo $k = 2 \log_2 N$ para N variando de 50 a 500. Em uma aplicação ilustrativa, eles usam $k = 5$ para agrupar 52 observações e relatam a eficiência do método em recuperar os grupos simulados.

2.2.10. Método da Ligação de Densidades em Dois Estágios (MLDE).

Este método foi desenvolvido por W. S. SARLE do SAS INSTITUTE INC. Na versão aqui considerada é uma modificação do MkVP.

Na descrição do método é definido um grupo modal como sendo um grupo com pelo menos k objetos, onde k é parâmetro do MkVP. Segundo o manual do SAS, no seu procedimento CLUSTER, empregando o MkVP é possivel que os grupos modais sejam unidos antes que todas as observações tenham sido agrupados, por exemplo, as observações com valorese extremos. A modificação introduzida é para garantir que todas as observações estejam em um grupo modal, antes que estes sejam unidos.

O método se desenvolve em dois estágios. No primeiro estágio os grupos modais são formados. Neste estágio é imposta uma restrição: dois grupos só podem ser unidos se, pelo menos um deles, tem menos membros do que o valor especificado pelo parâmetro k. No final deste estágio cada objeto pertence a um grupo modal.

No segundo estágio o Método da Ligação Simples (MLS) é empregado para agrupar hierarquicamente os grupos modais. Também, segundo o manual citado, ao final desse segundo estágio o número de grupos pode ser maior que um, caso haja grandes intervalos entre os grupos modais ou quando o valor de k for pequeno.

As únicas informações disponíveis na literatura sobre esse método são aquelas contidas no manual supracitado.

CAPÍTULO 3

O EXPERIMENTO PARA COMPARAÇÃO DOS MÉTODOS

3.1 INTRODUÇÃO.

Dado o objetivo deste trabalho, serão apresentadas as condições em que os métodos foram comparados e como foram mensurados seus desempenhos. No que diz respeito a primeira questão, no Capítulo 1 foi comentado o experimento adotado, um experimento com estrutura fatorial, onde os fatores determinam os aspectos da estrutura dos dados relevantes nas comparações. Portanto, a idéia foi simular conjuntos de dados com uma especificada estrutura de grupos, estes grupos com especificadas características, e então submeter estes conjuntos aos métodos aqui considerados. A capacidade do método em recuperar os grupos simulados foi observada, ou seja, o grau de concordância dos grupos

obtidos pelos métodos com aqueles embutidos nos dados simulados. Esse grau de concordância foi mensurado por um critério denominado estatística de RAND, o qual será descrito na próxima seção.

3.2 UMA MEDIDA DE RECUPERAÇÃO DE GRUPOS: A ESTATÍSTICA DE RAND.

Um primeiro passo para conduzir as comparações entre os métodos é estabelecer um critério objetivo para mensurar os resultados obtidos para cada um dos métodos. No problema aqui considerado, o critério deve quantificar objetivamente o grau de concordância entre os grupos embutidos nos dados simulados com os grupos obtidos pela aplicação de um dado método. RAND (1971) propôs uma medida, aqui denominada estatística de RAND (ESTRAND), a qual pode ser empregada como um critério para comparar os métodos. Neste critério o objetivo é verificar como os objetos, considerados aos pares, estão associados em dois dados agrupamentos.

Para definir a estatística de RAND considere a definição a seguir.

DEFINIÇÃO 3.1 Considere um conjunto de objetos $\mathbf{X}$ e dois agrupamentos, $\mathbf{Y}$ e $\mathbf{Y}'$, desse conjunto. Se um par de objetos, $(\mathbf{x}_i, \mathbf{x}_j)$, estão em um mesmo grupo em $\mathbf{Y}$ e em $\mathbf{Y}'$ ou se eles estão associados a grupos diferentes tanto

em **Y** como em **Y'**, então, é dito que ocorreu uma associação similar para o par de objetos $(\mathbf{x}_i, \mathbf{x}_j)$.

A definição da estatística de RAND é dada a seguir.

DEFINIÇÃO 3.2 Dado um conjunto de objetos $\mathbf{X}' = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ e dois agrupamentos de **X**, $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_K\}$ e $\mathbf{Y}' = \{\mathbf{Y}'_1, \mathbf{Y}'_2, \ldots, \mathbf{Y}'_{K'}\}$, a estatística de RAND, ESTRAND, entre **Y** e **Y'** é dada por

$$\text{ESTRAND}(\mathbf{y}, \mathbf{y}') = \frac{\sum_{i<j}^{N} \gamma_{ij}}{\binom{N}{2}}, \qquad (3\text{-}1)$$

onde $\gamma_{ij} = \begin{cases} 1, & \textit{se ocorre uma associação similar} \\ & \textit{para o par } (\mathbf{x}_i, \mathbf{x}_j) \textit{ em } \mathbf{Y} \textit{ e } \mathbf{Y}'. \\ 0, & \textit{caso contrário.} \end{cases}$

Por (3-1), verifica-se que a estatística assume valores no intervalo [0,1], uma vez que dá o número de pares similarmente associados, dividido pelo número total de pares de pontos.

Com base na definição, o problema pode ser abordado considerando-se **Y** como o agrupamento constituído pelos grupos simulados nas amostras de tamanho N e **Y'** como o agrupamento obtido com a aplicação de um dado método, para o qual é fixado k' = k. Com

esta abordagem, a estatística é uma medida da capacidade do método em recuperar os grupos simulados.

Esta estatística é uma das mais empregadas em investigações sobre métodos de agrupamento. Sua utilização é considerada não muito adequada, quando o número de grupos nos agrupamentos são diferentes e, para este caso, MOREY e AGRESTI (1984) apresentam uma correção para a estatística. No caso onde o número de grupos são iguais, a abordagem proposta por RAND (1971) é considerada bastante apropriada.

3.3 A DESCRIÇÃO DO EXPERIMENTO.

3.3.1 Os Fatores.

Como foi comentado, os fatores tem a finalidade de determinar os aspectos na estrutura dos dados que sejam de interesse na investigação. Os aspectos de interesse neste trabalho estão relacionados com a matriz de dispersão dos grupos, a correlação entre as variáveis e a questão da sobreposição dos grupos. Parece que estes aspectos não foram explorados de uma forma mais abrangente em trabalhos anteriores. Alguns deles foram considerados isoladamente e, ainda assim, as conclusões não foram muito esclarecedoras quanto à performance dos métodos ante a presença de

dados com características envolvendo esses aspectos. Uma exceção ocorre para o trabalho de DUBIEN e WARDE (1987), comentado na seção 1.5, onde foi explorado a questão da correlação entre as variáveis.

Um problema de ordem prática que ocorre nas aplicações de Análise de Agrupamentos, é que há uma tendência por parte dos pesquisadores a incluir muito mais variáveis do que seriam necessárias. Este procedimento leva a inclusão de variáveis que em nada contribuem para a separação dos grupos. Esta questão, de certa forma,está relacionada com a escolha do número de variáveis para caracterizar as observações simuladas para este trabalho. Aqui não havia interesse em considerar dados com muitas dimensões, uma vez que isto dificultaria o entendimento do efeito das características dos dados e, em particular, a forma geométrica dos grupos. Com muitas dimensões não seria possível uma visualização desse aspecto. Assim, decidiu-se considerar observações com duas e três dimensões. Considerando o objetivo do trabalho, pode também ser acrescentado que, se um dado método não tem um bom desempenho nessas situações simples, não espera-se que ele tenha um bom desempenho em situações mais complexas.

Com relação à matriz de dispersão dentro dos grupos, o objetivo era verificar o efeito dos grupos terem essas matrizes diferentes comparado com o caso das matrizes serem iguais. Trabalhos anteriores deram indicações de que alguns métodos pareciam ser afetados por esta questão, como por exemplo, o MWARD.

Adicionado à questão acima, variar o grau de correlação entre as variáveis permitiria criar diferentes formas para os grupos no espaço das variáveis. Assim, decidiu-se incluir o coeficiente de correlação, ρ, como um fator no experimento. Como comentado em DUBIEN e WARD (1987), aumentando o valor de ρ, com $\rho \geq 0$, é possível variar a configuração dos pontos de uma forma aproximadamente circular para uma mais elíptica. Com isto, para aqueles métodos mais apropriados à determinação de grupos com forma circular, seria possível verificar a robustez desses métodos ao afastamento desta condição.

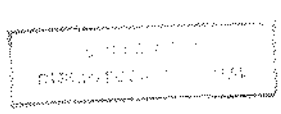
Outro aspecto de interesse, foi verificar o desempenho dos métodos ante a situação onde os grupos apresentassem regiões superpostas no espaço das variáveis. MILLIGAN e ISAAC (1980) argumentam que não é adequado considerar este aspecto, uma vez que os métodos hierárquicos de agrupamento não foram desenvolvidos para detectar estruturas com grupos sobrepostos ("*overlapping*"). Neste trabalho, entretanto, esta questão é considerada relevante, dado que na prática esse aspecto pode ser configurado por meio de observações com valores discrepantes ("*outliers*") ou mesmo se a real distribuição das observações apresenta valores nos extremos com probabilidades relativamente altas, por exemplo, maiores que as da distribuição normal (distribuições com as "caudas" mais pesadas).

Outro aspecto considerado de interesse prático foi a

introdução de um erro aleatório nas coordenadas das observações. O objetivo era reproduzir o efeito de um erro de mensuração sobre os dados. Com esta finalidade foram criadas duas situações, observações "sem erro" e observações "com erro". Para simular a situação "com erro", uma variável aleatória $N(0; \sigma^2)$ foi adicionada, de forma independente, a cada coordenada das observações, onde σ^2 era igual a 10% da variância original da respectiva variável componente.

Na maioria dos trabalhos desenvolvidos, a distância Euclidiana foi empregada como dissimilaridade entre os objetos. EDELBROCK e McLUGHLIN (1980) e MILLIGAN (1980) foram uns dos poucos trabalhos onde foram incluídas outra medidas de dissimilaridade (ou de similaridade). Sendo os dados construídos no espaço Euclidiano, empregar uma medida de dissimilaridade diferente da distância Euclidiana, pode ser vista como uma forma de erro na utilização dos métodos. Entretanto, os resultados indicaram que os métodos são um tanto robusto com respeito a este tipo de erro (MILLIGAN, 1980). Apesar disto, aqui havia interesse em verificar o efeito de empregar a norma L_1 como dissimilaridade, principalmente, no caso da sobreposição dos grupos, onde pela forma empregada na simulação, seriam geradas observações com valores bastante discrepantes.

Todos esses fatores foram submetidos a investigações preliminares as quais serão descritas a seguir.



3.3.2 Investigações Iniciais.

Algumas invesigações preliminares foram desenvolvidas buscando verificar, de uma forma apenas exploratória, a influência dos fatores escolhidos sobre o desempenho dos métodos. Estas investigações consistiram de pequenos experimentos de simulação, onde os fatores eram sempre considerados em dois níveis, isto é, foram montados experimentos com estrutura fatorial 2^v.

Dentro dos fatores considerados alguns não apresentaram evidências de afetarem os resultados dos métodos. O fator número de variáveis, a dimensionalidade das observações, foi um desses fatores. Como comentado, os níveis para este fator foram "duas variáveis" e "três variáveis". É possível que o resultado não significativo deste fator seja devido a diferença mínima entre as dimensionalidades consideradas, entretanto, não havia interesse em considerar dimensões para os dados maiores que estas, como já foi justificado. Por este resultado decidiu-se empregar somente as observações bidimensionais no experimento final.

Para o fator introduzindo um erro aleatório sobre as variáveis componentes das observações, os resultados não apresentaram evidências de que afetava a performance dos métodos. Neste fator um nível era "sem erro" e o outro era "com erro", sendo neste último introduzindo o erro aleatório como descrito anteriormente. Dado este resultado, decidiu-se não incluí-lo no

experimento final.

O fator dissimilaridade entre os objetos, que tinha em um nível a norma L_1 e no outro a norma L_2, não apresentou efeito significativo sobre as respostas. Isto indicava que usar a norma L_1 ou a norma L_2 não afetava o desempenho dos métodos. Decidiu-se então considerar somente a norma L_2, a distância Euclidiana, como dissimilaridade entre os objetos.

Alguns desses experimentos foram repetidos com diferentes tamanhos de amostra. Em todos os casos, o número de grupos foi sempre igual a dois, porém, três diferentes situações com respeito ao tamanho da amostra, N, e o número de objetos por grupos, N_1 e N_2, foram investigadas: (i) $N = 40$, $N_1 = 28$, $N_2 = 12$; (ii) $N = 100$, $N_1 = 70$, $N_2 = 30$; (iii) $N = 100$, $N_1 = 50$, $N_2 = 50$. Os resultados obtidos para estes diferentes casos foram todos muito semelhantes e decidiu-se considerar apenas um deles. Optou-se por $N = 40$, $N_1 = 28$ e $N_2 = 12$.

De maneira geral, estas investigações foram bastante esclarecedoras para a escolha dos fatores e a forma como conduzir o experimento. A seguir as condições fixadas para a simulação dos dados, como também a descrição dos níveis dos fatores incluídos no experimento final são apresentadas.

3.3.3 As Estruturas Selecionadas.

A partir dos resultados das investigações iniciais, três fatores foram considerados para criar as estruturas para simular as observações no experimento final: a matriz de dispersão, a correlação entre as variáveis e a sobreposição dos grupos.

Das investigações iniciais, também ficou decidido considerar amostras contendo observações de dois grupos. As amostras tinham tamanho N = 40, com N_1 = 28 observações em um grupo e N_2 = 12 observações no outro.

As observações foram simuladas segundo uma distribuição normal bivariada, onde o vetor de médias e a matriz de dispresão eram fixados segundo os grupos a que pertenciam. Assim

$$\mathbf{x}_i \sim N_2(\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K), \quad i = 1, 2, \ldots, N,$$

onde $\boldsymbol{\mu}_K$ e $\boldsymbol{\Sigma}_K$ representam, respectivamente, o vetor de médias e a matriz de dispersão do k-ésimo grupo, k = 1, 2. Pelo que foi descrito, $\boldsymbol{\Sigma}_K$ também depende dos fatores considerados.

Será descrito a seguir como foram considerados os fatores selecionados e os seus níveis para determinar as diferentes estruturas nos dados simulados.

1. MATRIZ DE DISPERSÃO DENTRO DOS GRUPOS.

Dado o objetivo para o qual considerou-se este fator, foi adotado que em um nível os grupos teriam matrizes de dispersão diferentes e no outro nível essas matrizes seriam iguais. Assim, fixou-se as seguintes matrizes.

NÍVEL 1

$$\Sigma_1 = \begin{bmatrix} 1 & cov_{11} \\ cov_{11} & 16 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 25 & cov_{12} \\ cov_{21} & 4 \end{bmatrix}$$

NÍVEL 2

$$\Sigma_1 = \begin{bmatrix} 9 & cov_2 \\ cov_2 & 9 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 9 & cov_2 \\ cov_2 & 9 \end{bmatrix}$$

Dentro de cada matriz o valor das covariâncias foram determinadas segundo os níveis do fator coeficiente de correlação, os quais são descritos a seguir.

2. COEFICIENTE DE CORRELAÇÃO.

Este fator estabelecendo o coeficiente de correlação, ρ,

entre as variáveis foi considerado em três níveis. O objetivo era estabelecer três níveis de correlação: variáveis não correlacionadas, baixa correlação e alta correlação e, assim, os valores de ρ foram fixados em 0,0, 0,4 e 0,8.

3. A SOBREPOSIÇÃO DOS GRUPOS.

Como já foi discutido, este fator tinha por finalidade criar uma situação onde os grupos apresentassem uma região de sobreposição. Para criar essa situação foi introduzida uma "contaminação" sobre as observações simuladas. Introduzir essa "contaminação" consistiu em gerar as observações segundo uma distribuição "normal contaminada", com um percentual α de contaminação. Fixado o valor de α em 20%, as observações foram geradas com

$$\mathbf{x}_i \sim 0.80\ N_2\ (\mu_K,\ \Sigma_K) + 0,20\ N_2\ (\mu_K^*,\ \Sigma^*_K),$$

onde $\mu_K^* = \mu_K$ e $\Sigma_K^* = 25 \times \Sigma_K$.

Pelo exposto acima, verifica-se que as observações com contaminação em cada grupo, além de não terem distribuição normal, esperava-se que 20% delas apresentem valores discrepantes sobrepondo-se a região do outro grupo. As variáveis nessas observações tinham um desvio padrão cinco vezes maior que o das observações sem contaminação.

3.3.4 Vetores de Médias.

Os vetores de médias foram considerados fixos em cada grupo, independentemente das mudanças nos níveis dos fatores. O critério empregado para a escolha destes vetores foi que as elipses com 95% de confiança, centradas neles e construídas com base na matriz de dispersão dos grupos com $\rho = 0.0$, não apresentassem sobreposição, embora com pontos nos contornos bastante próximos.

Para dar uma idéia da distância com relação a primeira dimensão das observações (eixo horizontal) entre as elipses representativas dos grupos, a distância Euclidiana entre os dois pontos mais próximos, um em cada elipse, foi de aproximadamente 10% da média aritmética entre os desvios padrão das variáveis representando aquela dimensão em cada grupo. Isto se verificando nos dois níveis do fator matriz de dispersão.

Empregando as propriedades das distribuições normais bivariadas (JOHNSON e WICHERN, 1988), os vetores de médias foram

$$\mu_1 = \begin{bmatrix} 3 \\ 5 \end{bmatrix} \qquad \text{e} \qquad \mu_2 = \begin{bmatrix} 18 \\ 5 \end{bmatrix}.$$

Nas Figuras 3.1 e 3.2 são apresentados esboços das elipses

com 95% de confiança dos grupos nas combinações dos níveis dos fatores Matriz de Dispersão e Coeficiente de Correlação, para o caso de grupos sem sobreposição. Três casos típicos da distribuição dos pontos para o caso de grupos com sobreposição, são apresentados na Figura 3.2.

3.3.5 A Estrutura do Experimento.

Resumindo o exposto até aqui, o experimento consistiu em empregar os dez métodos descritos no Capítulo 2 sobre os dados simulados com estruturas determinadas pelo cruzamentos dos níveis dos fatores. Ao todo, foram doze diferentes estruturas de dados

As aplicações dos métodos sobre as amostras simuladas dentro das estruturas, foram feitas de duas formas diferentes. No primeiro caso, que será denominado Experimento 1, cada amostra simulada foi submetida a apenas um dos métodos. No outro caso, o Experimento 2, todos os métodos foram aplicados a cada uma das amostras simuladas.

A sequência de passos empregados, dentro de cada combinação dos fatores, para a execução do Experimento 1 foi a seguinte:

P1. Foi gerada uma amostra de 40 observações, com 28 observações no grupo 1 e 12 observações no grupo 2.

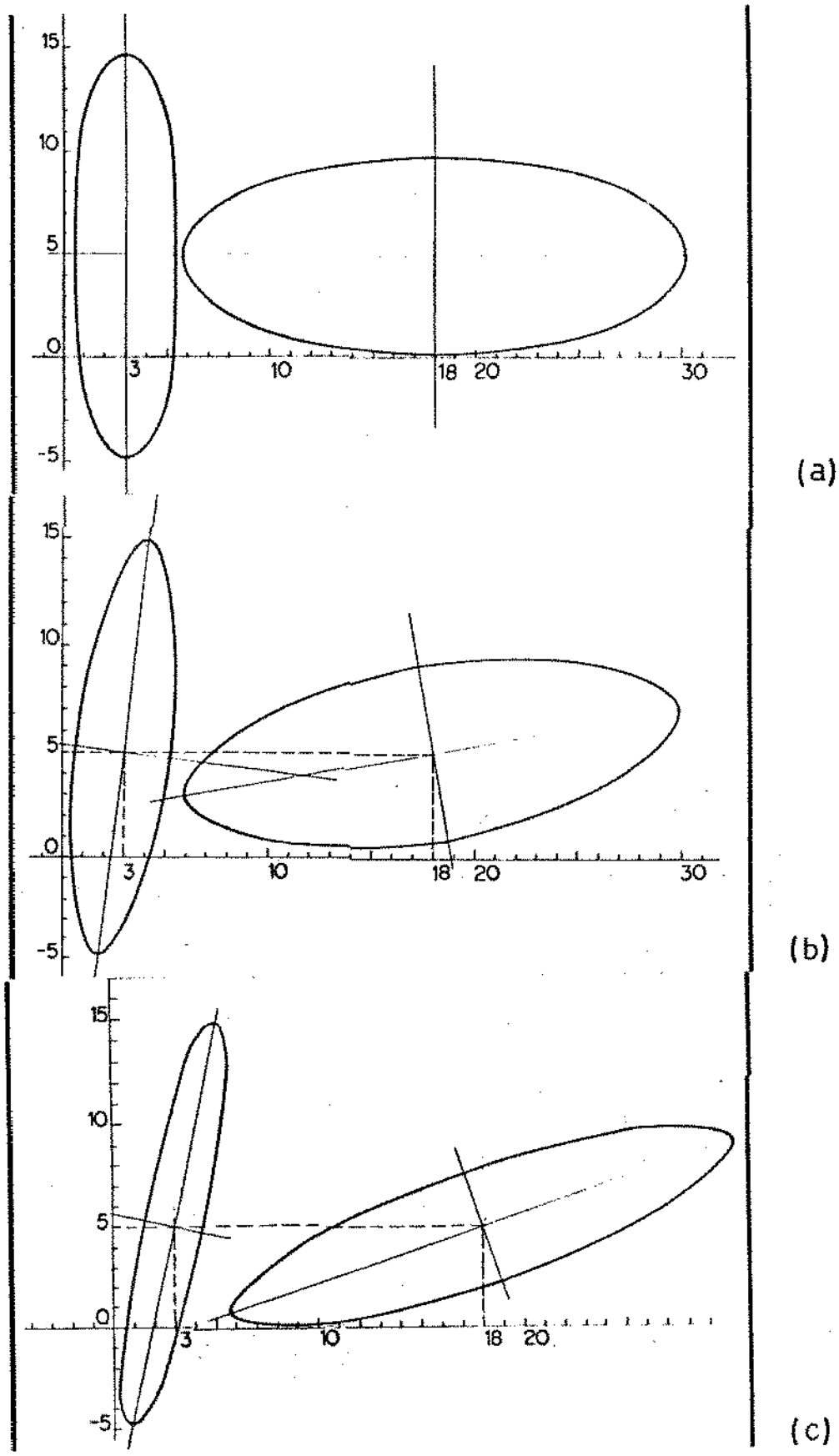


Figura 3.1 Esboço das elipses com 95% de confiança para grupos sem sobreposiçao e matrizes diferentes. (a) $\rho = 0,0$. (b) $\rho = 0,4$. (c) $\rho = 0,8$.

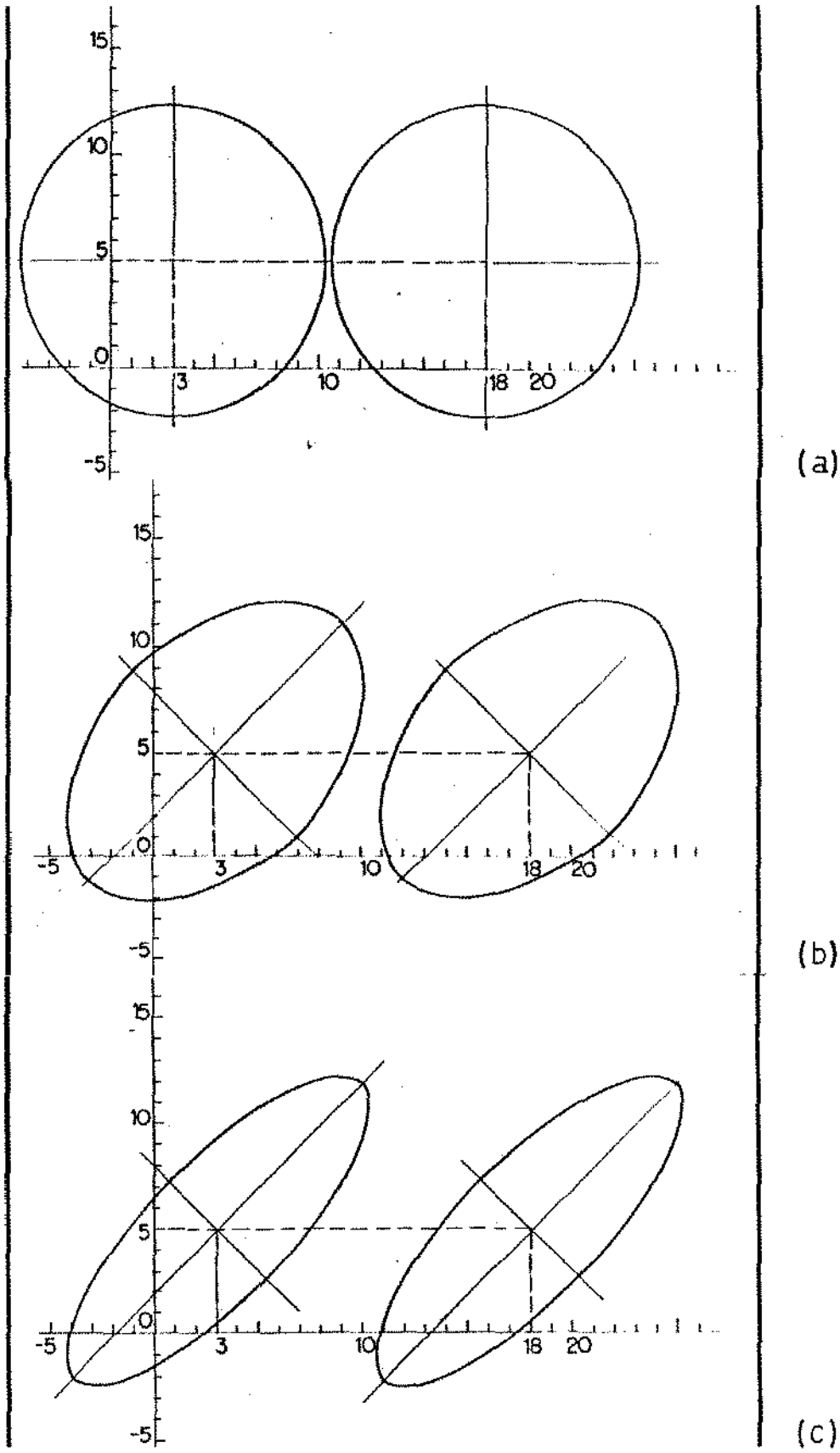
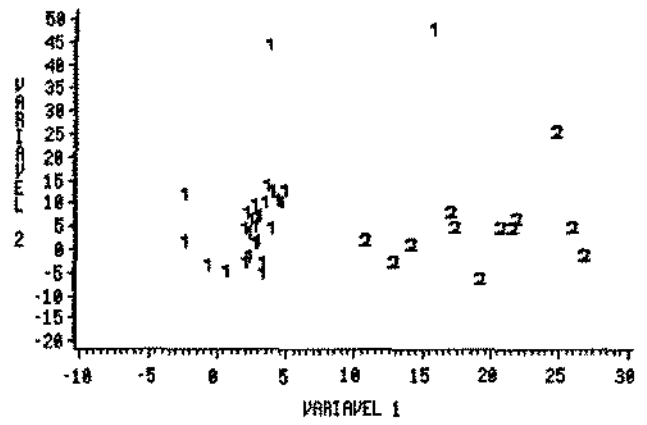
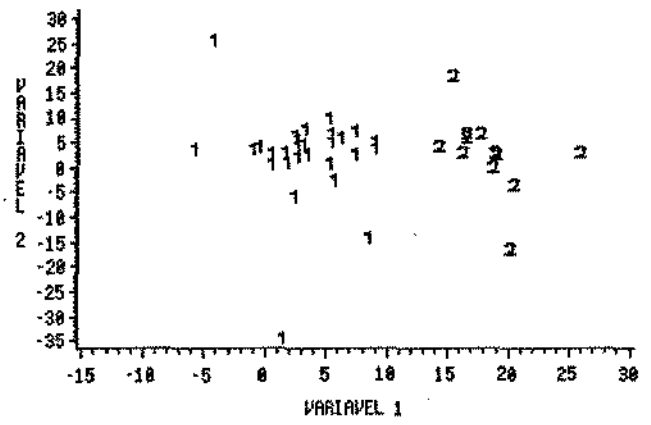


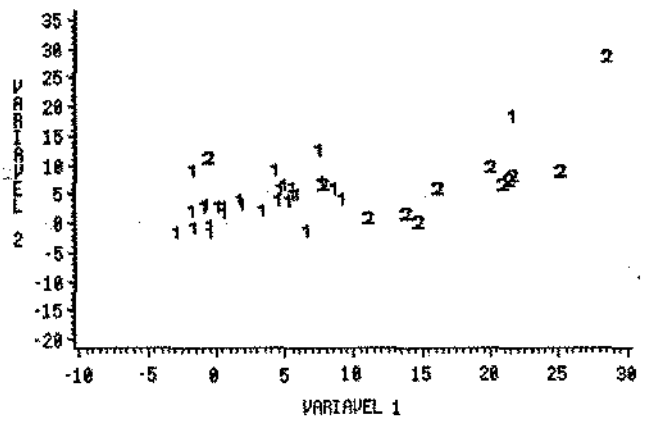
Figura 3.2 Esboço das elipses com 95% de confiança para os grupos sem sobreposição e matrizes iguais. (a) $\rho = 0,0$. (b) $\rho = 0,4$. (c) $\rho = 0,8$.



(a)



(b)



(c)

Figura 3.3 Três casos típicos dos pontos com grupos com sobreposição. (a) matrizes diferentes e $\rho = 0,4$. (b) matrizes iguais e $\rho = 0,0$. (c) matrizes iguais e $\rho = 0,8$.

P2. Sobre a amostra, empregando o Procedimento CLUSTER do SAS, era aplicado um dos métodos de agrupamento com um corte em dois grupos.

P3. O valor da ESTRAND era calculada com base no agrupamento simulado e o agrupamento obtido pelo método.

Esta sequência de passos foi repetida 50 vezes para cada um dos métodos. Com 12 caselas de combinações de níveis, perfaz um total de 600 observações da ESTRAND por método.

O Experimento 2 se diferenciava do Experimento 1 nos passos P2 e P3 da sequência acima. No Experimento 2 esses passos foram dados por:

P2'. Empregando o Procedimento CLUSTER do SAS, cada um dos dez métodos era aplicado sobre a amostra, todos com um corte em dois grupos.

P3'. Para cada um dos métodos, era calculado o valor da ESTRAND com base no grupamento simulado e o obtido pelo método.

Neste segundo experimento, a sequência de passos foi repetida 20 vezes, gerando 20 observações de ESTRAND por método em cada uma das estruturas. Obteve-se, então, 240 (12 x 20) observações para

cada um dos métodos.

Dado que as amostras simuladas eram independentes, a diferença entre esses experimentos diz respeito a independência das observações da ESTRAND. No Experimento 1 as respostas foram independentes dentro dos métodos e, também, entre os métodos. No Experimento 2, essas respostas são independentes dentro dos métodos, porém, entre eles não havia independência.

O Experimento 1 foi montado com o propósito de submeter as observações da ESTRAND a uma Análise de Variância, onde o modelo empregado incluía o método de agrupamento com um fator. Como será descrito no Capítulo 4, uma análise para as respostas sob cada método também foi efetuada. Com as respostas do Experimento 2, o objetivo era analisar o perfil do comportamento dos métodos em cada uma das estruturas, uma vez que todos os métodos analisavam os mesmos dados.

A seguir serão descritos os procedimentos empregados para gerar as observações nas amostras simuladas.

3.4 SIMULAÇÃO DAS OBSERVAÇÕES.

Como foi descrito, cada observação foi gerada segundo uma distribuição normal bivariada, com parâmetros fixados segundo o grupo a que pertencia e a estrutura desejada. Supondo esses parâmetros $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, cada observação $\mathbf{X}' = [x_1, x_2]$ ($\mathbf{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$) foi

gerada empregando a transformação

$$\mathbf{X} = \mathbf{cZ} + \boldsymbol{\mu},$$

onde $\mathbf{Z}' = [Z_1, Z_2]$ é um vetor onde as componentes são variáveis aleatórias independentes com distribuição N(0, 1), $\mathbf{Z} \sim N_2(\mathbf{0}, \mathbf{I}_2)$, e $\mathbf{c}$ é tal que $\mathbf{c}'\mathbf{c} = \boldsymbol{\Sigma}$.

Para obter $\mathbf{c}$ foi empregada a decomposição de Cholesky (ver JOHNSON, 1987). Os vetores $\mathbf{Z}$ e a decomposição de Cholesky foram obtidos empregando o Procedimento IML do SAS. Para obter as componentes Z_i de $\mathbf{Z}$, este procedimento emprega um algoritmo baseado no método de geração de variáveis aleatórias normais de BOX-MÜLLER (BUSTOS e FRERY, 1992).

Um comentário sobre a eficiência do gerador empregado: o método de BOX-MÜLLER envolve a geração de variáveis independentes com distribuição Uniforme (0, 1) e, evidentemente, a qualidade do método depende da qualidade do gerador de observações da distribuição Uniforme empregado. O gerador de observações da distribuição Uniforme (0, 1) empregado pelo SAS, do tipo congruencial multiplicativo com módulo primo, possui boas propriedades estatísticas (FISHMAN e MOORE, 1982). Desta forma, o gerador de observações da N(0,1) empregado nesta simulação gera adequadamente esta distribuição.

Para gerar as observações contaminadas, adotou-se o seguinte algoritmo (JOHNSON, 1987):

1. Gerar uma variável U, onde U ~ Uniforme (0, 1)

   se $U \leq (1-\alpha)$ executar o passo 2

   se $U > (1-\alpha)$ executar o passo 3

2. Gerar $\mathbf{X}$ segundo uma N_2 $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

3. Gerar $\mathbf{X}$ segundo uma N_2 $(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$.

Todas as observações foram geradas dentro do Procedimento IML do SAS. As simulações foram feitas empregando um computador IBM 3090 do Centro de Computação da UNICAMP.

CAPÍTULO 4

RESULTADOS E CONCLUSÕES

4.1 INTRODUÇÃO.

No capítulo anterior foi descrito o experimento realizado, sendo especificados os fatores considerados para a geração dos dados, a medida empregada para mensurar o desempenho dos métodos e a forma como os métodos foram aplicados às amostras simuladas. Pelo que foi descrito, foram adotados dois procedimentos distintos, denominados Experimento 1 e Experimento 2, para aplicar os métodos de agrupamentos. No Experimento 1 as respostas obtidas foram independentes tanto dentro dos métodos como, também, entre os métodos. Para as respostas do Experimento 2 não houve independência entre os métodos.

Com as respostas do Experimento 1 os métodos tiveram seus desempenhos comparados através das médias dentro das estruturas

simuladas. A Análise de Variância e as técnicas de comparações múltiplas, foram aplicadas sobre as respostas desse experimento.

Os dados do Experimento 2 foram analisados apenas de forma descritiva, objetivando identificar métodos com perfis semelhantes entre as diversas estruturas de dados consideradas.

Na exposição das análises serão adotadas algumas codificações. A estatística de RAND será denotada por ESTRAND, os fatores Sobreposição de Grupos, Matriz de Dispersão e Coeficiente de Correlação serão denotadas, respectivamente, por SOBREPOS, MATRDISP E COEFCORR. Os níveis desses fatores serão considerados como dados na Tabela 4.1.

TABELA 4.1 DESCRIÇÃO DOS NÍVEIS DOS FATORES

FATOR	NÍVEL		
	1	2	3
SOBREPOS	Grupos sem sobreposição (SOBREPOS 1)	Grupos com sobreposição (SOBREPOS 2)	-
MATRDISP	Matrizes de Dispersão diferentes (MATRDISP 1)	Matriz de dispersão iguais (MATRDISP 2)	-
COEFCORR	Coeficiente de correlação $\rho = 0,0$ (COEFCOR 1)	Coeficiente de correlação $\rho = 0,4$ (COEFCOR 2)	Coeficiente de Correlação $\rho = 0,8$ (COEFCOR 3)

Outra codificação adotada diz respeito as estruturas simuladas nos dados através da combinação dos níveis dos fatores. Para facilitar a exposição dos resultados, o termo ESTRUTURA foi empregado para designar essas combinações de níveis. Na Tabela 4.2 são especificadas essas denominações.

TABELA 4.2 DESCRIÇÃO DAS ESTRUTURAS

ESTRUTURA	NÍVEIS DOS FATORES		
	SOBREPOS	MATRDISP	COEFCORR
1	1	1	1
2	1	1	2
3	1	1	3
4	1	2	1
5	1	2	2
6	1	2	3
7	2	1	1
8	2	1	2
9	2	1	3
10	2	2	1
11	2	2	2
12	2	2	3

A seguir serão apresentados os resultados das análises dos experimentos considerados.

4.2 ANÁLISE DO EXPERIMENTO 1.

4.2.1 Modelo com o Método como Fator.

A primeira análise conduzida sobre as respostas desse

experimento foi uma Análise de Variância, cujo modelo subjacente incluía o método de agrupamento como um fator, os outros fatores descritos e todas as interações entre eles. Desta forma, considerando o método de agrupamento como um fator com dez níveis, estava sendo considerado um experimento com estrutura fatorial com 10x2x2x3 =120 tratamentos. Para a aplicação desta metodologia fez-se necessário proceder a uma transformação das respostas.

As respostas dadas pela ESTRAND expressam proporções, a proporção de pares de objetos considerados corretamente classificados. Segundo a literatura (BOX et ali., 1978; MONTGOMERY, 1991), uma transformação adequada para este tipo de resposta é dada por

$$y = \text{arcsen } \sqrt{ESTRAND} \; . \qquad (4\text{-}1)$$

As análises foram executadas sobre os valores de y, obtidos pela transformação dada em (4-1).

Ajustando o modelo considerado, a análise dos resíduos indicava que mais fatores poderiam ser considerados no modelo para melhor explicar as respostas. Evidentemente, como já foi comentado, inúmeros outros fatores poderiam ser considerados neste contexto mas, dentro dos objetivos deste trabalho, tomou-se apenas os fatores de maior interesse. De maneira geral, entretanto, o comportamento dos resíduos não apresentaram tendências grosseiras que invalidassem o emprego da Análise de Variância e, assim, foi

dada continuidade a essa análise paramétrica.

A tabela de ANOVA é apresentada na Tabela 4.2. Dos resultados apresentados, verificou-se que quase todas as interações envolvendo o fator MÉTODO apresentaram efeitos significativos ao nível de significância (n.s.) $p < 0,01$. Somente as interações MÉTODO*MATRDISP*COEFCORR e MÉTODO*SOBREPOS*MATRDISP*COEFCORR não apresentaram efeitos significativos.

No trabalho de CUNNINGHAM e OLGIVIE (1971) também foi reportado o resultado comentado acima. Os autores verificaram uma interação muito forte entre os métodos e as estruturas dos dados.

TABELA 4.2 TABELA DE ANOVA PARA O MODELO COM QUATRO FATORES

FATORES	GL	SQ	QM	F	Pr >F
METODO (A)	9	9.146	1.016	27.45	< 0,01
SOBREPOS (B)	1	162.861	162.861	4399.79	< 0,01
MATRDISP (C)	1	27.821	27.821	751.61	< 0,01
COEFCORR (D)	2	0.532	0.266	7.19	< 0,01
A*B	9	9.898	1.099	29.71	< 0,01
A*C	9	3.112	0.346	9.34	< 0,01
A*D	18	2.104	0.117	3.16	< 0,01
B*C	1	11.657	11.657	314.94	< 0,01
B*D	2	0.247	0.123	3.34	0.0354
C*D	2	0.409	0.204	5.52	< 0,01
A*B*C	9	2.575	0.286	7.73	< 0,01
A*B*D	18	1.7	0.094	2.55	< 0,01
A*C*D	18	0.816	0.045	1.22	0.2309
B*C*D	2	0.458	0.229	6.18	< 0,01
A*B*C*D	18	0.394	0.022	0.59	0.9090
RESIDUO	5880	217.652	0.037		
TOTAL	5999	451384			

Considerando a presença da interação entre o MÉTODO e os outros fatores e com a finalidade de facilitar a compreeeenção dos efeitos dos fatores sobre os métodos, decidiu-se analisar as respostas para cada um dos métodos separadamente.

4.2.2 A Análise de cada Método.

A segunda parte da análise consistiu em empregar, para cada um dos métodos de agrupamento, uma Análse de Variância sobre as respostas cujo modelo subjacente incluía oa fatores SOBREPOS, MATRDISP e COEFCORR com suas respectivas interações. Quando evidenciado a significância do efeito de fatores ou de suas interações, procedimentos de comparações múltiplas foram conduzidos para os níveis dos fatores ou para combinações de níveis, no caso das interações. Para essas comparações empregou-se o teste "*multiple range*" de RYAN-EINOT-GABRIEL-WELSCH (Procedimento GLM do Sistema de Programas SAS), sendo este teste escolhido por ser considerado na literatura como um dos mais poderosos.

A seguir, serão expostos os resultados das análises segundo cada um dos métodos de agrupamento. Nestas exposições serão apresentadas tabelas de médias contendo as médias das respostas transformadas, seus desvios padrão (entre parênteses) e, para ajudar a compreensão dos resultados, as médias das respostas sem transformação, apresentadas entre colchetes.

1) MÉTODO DA LIGAÇÃO SIMPLES (MLS).

Para este método, a Análise de Variância indicou efeito significativo para todos os fatores e para as interações entre eles. A interação SOBREPOS*MATRDISP*COEFCORR foi considerada significativa ao n.s. p = 0,0473 e, desta forma, o efeito de cada fator foi analisado considerando os níveis dos outros fatores.

Os gráficos mostrando o comportamento das médias das respostas para a interação dos fatores são apresentados na Figura 4.1. Comparando-se os gráficos (a) e (b), verifica-se um decréscimo nas médias dentro da condição de grupos com sobreposição e, além disso, os gráficos sugerem um comportamento diferenciado da interação MATRDISP*COEFCORR entre os níveis de SOBRPOS. Também, pelos gráficos (c), (d) e (e), observa-se um comportamento similar da interação SOBRPOS*MATRDISP entre os níveis de COEFCORR, onde é sugerido médias miores para a condição de matrizes iguais (MATRDIS2).

As médias dentro das combinações níveis dos três fatores são apresentados na Tabela 4.4. Aplicando testes comparativos entre essas médias verificou-se que:

(1) fixados os níveis de MATRDISP e COEFCORR, havia um decréscimo significativo nas médias das respostas ao mudar de grupos sem sobreposição para grupos com sobreposição, isto dentro de cada combinação dos níveis dos outros fatores (p = 0,01);

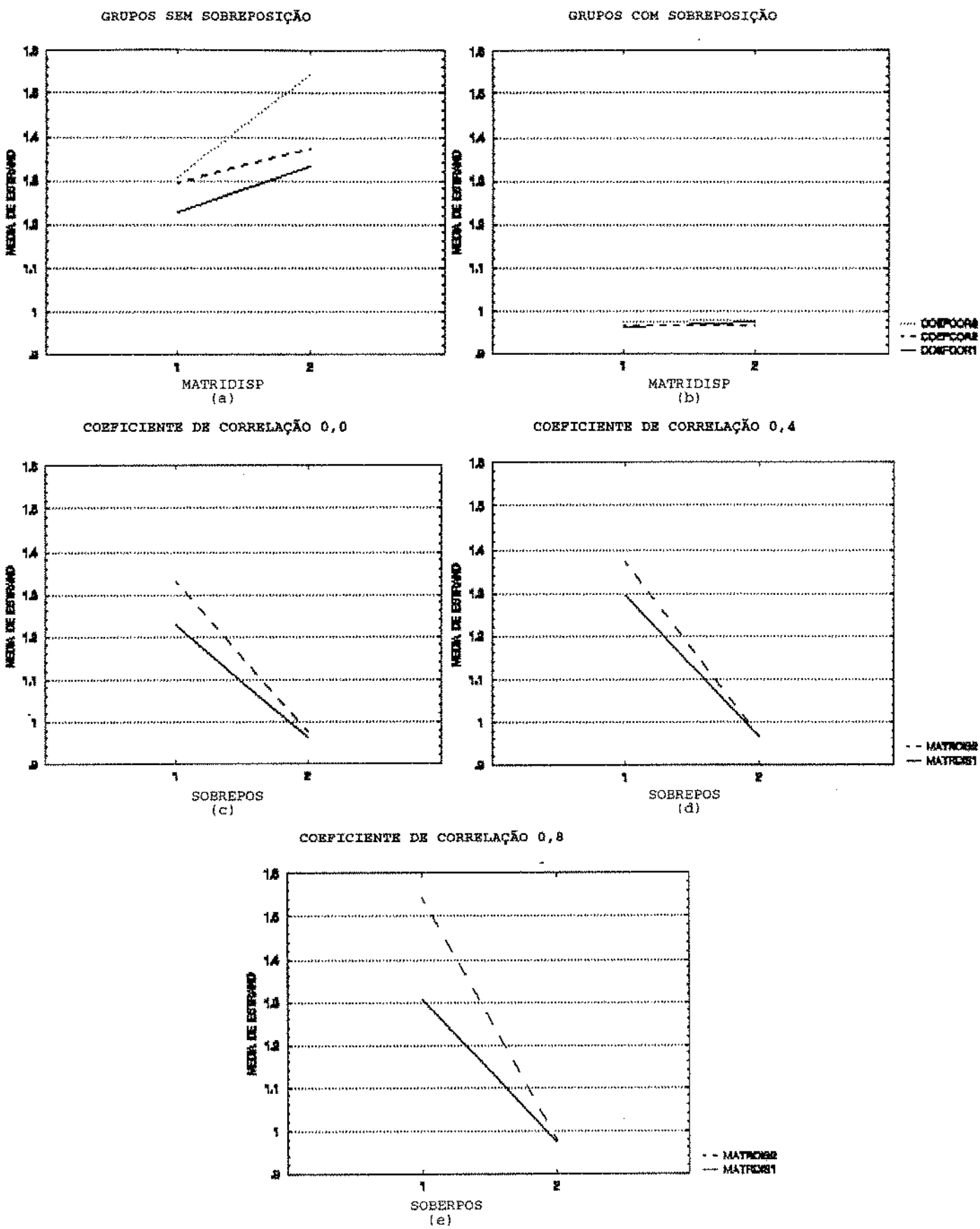


Figura 4.1 MLS. Interação MATRDISP*COEFCORR: (a) grupos sem sobreposição; (b) grupos com sobreposição. Interação SOBREPOS*MATRDISP: (c) $\rho = 0,0$; (d) $\rho = 0,4$; (e) $\rho = 0,8$.

TABELA 4.4 MÉDIAS NOS NÍVEIS DE SOBREPOS*MATRDISP* COEFCORR PARA O MLS.

SOBREPOS	MATRDISP	COEFCORR 1	COEFCORR 2	COEFCORR 3
1	1	1.229 (0,036) [0,836]	1,294 (0,036) [0,874]	1,306 (0,035) [0,879]
	2	1,332 (0,035) [0,893]	1,374 (0,034) [0,911]	1,545 (0,015) [0,988]
2	1	0,963 (0,005) [0,673]	0,966 (0,005) [0,677]	0,972 (0,006) [0,681]
	2	0,974 (0,005) [0,684]	0,967 (0,004) [0,677]	0,980 (0,010) [0,688]

NOTA: 1) 50 observações de ESTRAND por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

(2) dentro da condição de grupos sem sobreposição e dentro de cada nível de COEFCORR, as médias nos níveis de MATRDISP apresentaram diferenças significativas somente no caso de $\rho = 0,8$, onde a média com matrizes iguais foi significativamente superior ($p = 0,01$);

(3) sob a condição de grupos sem sobreposição e dentro de cada nível de MATRDISP, foi verificada diferença significativa entre os níveis de COEFCORR somente no caso de matrizes iguais, onde a média com $\rho = 0,8$ foi significativamente maior que nos outros níveis deste fator ($p = 0,01$);

(4) dentro da condição de grupos com sobreposição, não foram detectadas diferenças significativas entre as médias nas

combinaçoes dos nívies de MATRDISP e COEFCORR.

O resultado obtido com grupos sem sobreposição, matrizes iguais e $\rho = 0,8$ é condizente com as características deste método, pois neste caso, os grupos apresentavam uma forma mais alongada. O método também foi muito fortemente afetado pela sobreposição dos grupos.

2) MÉTODO DA LIGAÇÃO COMPLETA (MLC).

Os três fatores e as interações de dois fatores SOBREPOS*MATRDISP e SOBREPOS*COEFCORR, apresentaram efeitos significativos ao n.s. $p < 0,02$.

As interações de dois fatores são apresentadas graficamente na Figura 4.2. O gráfico (a), apresentando a interação SOBREPOS*MATRDISP, sugere um decréscimo nas médias sob as duas condições de MATRDISP ao considerar-se grupos com sobreposição e sendo este decrésimo mais acentuado sob a condição de matrizes iguais (MATRDIS2).

No gráfico (b) da Figura 4.2, onde é apresentada a interação SOBREPOS*COEFCORR, também verifica-se um decréscimo nas médias sob os três níveis de COEFCORR ao passar para a condição de grupos com sobreposição. Entretanto, o comportamento das médias sob $\rho = 0,0$ e $\rho = 0,4$ apresentam retas praticamente paralelas e, desta forma, é

sugerido que a interação seja devido ao comportamento das médias com $\rho = 0,8$.

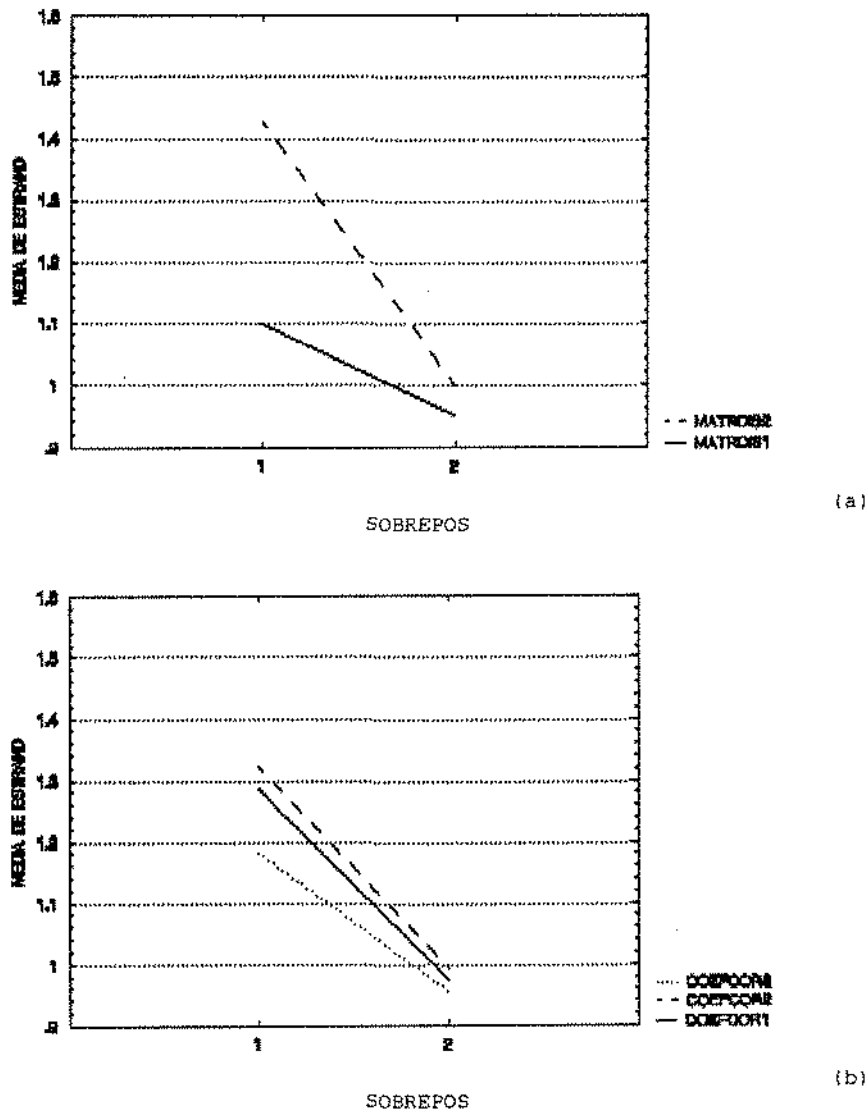


Figura 4.2 MLC. (a) interação SOBREPOS*MATRDISP. (b) interação SOBREPOS*COEFCORR.

As Tabelas 4.5 e 4.6 apresentam, respectivamente, as médias para as interações SOBREPOS*MATRDISP e SOBREPOS*COEFCORR.

Da comparação das médias apresentadas na tabelas 4.5 e 4.6

verificou-se que:

(1) dentro de cada nível de SOBREPOS (Tabela 4.5), as médias com matrizes iguais foram significativamente maiores que no outro nível de MATRDISP ($p = 0,01$);

(2) sob a condição de grupos sem sobreposição (Tabela 4.6), a média sob a condição de $\rho = 0,8$ foi significativamente menor que nos outros níveis de COEFCORR, porém, no outro nível de SOBREPOS não foram detectadas diferenças significativas entre as medidas nos níveis de COEFCORR ($p = 0,01$).

Para comparar os níveis fator SOBREPOS devem ser considerados os níveis de MATRDISP e COEFCORR. Comparando-se as médias apresentadas na Tabela 4.7 verificou-se que:

TABELA 4.5 MÉDIAS NOS NÍVEIS DE SOBREPOS* COEFCORR PARA O MLC.

SOBREPOS	MATRDISP	
	1	2
1	1,099 (0,022) [0,746]	1,429 (0,018) [0,940]
2	0,949 (0,010) [0,654]	1,003 (0,011) [0,701]

NOTA: 1) 150 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

(1) dentro das combinações dos níveis de MATRDISP e COEFCORR, quase todas as médias sob a condição de grupos sem sobreposição foram signifcativamente maiores que as médias no outro nível de

SOBREPOS, a exceção ocorrendo para a condição de matrizes iguais com $\rho = 0,08$. ($p = 0,01$);

TABELA 4.6 MÉDIAS NOS NÍVEIS DE SOBREPOS*COEFCORR PARA O MLC.

SOBREPOS	COEFCORR 1	COEFCORR 2	COEFCORR 3
1	1,285 (0,027) [0,865]	1,325 (0,028) [0,878]	1,182 (0,032) [0,785]
2	0,975 (0,014) [0,675]	0,995 (0,013) [0,696]	0,958 (0,012) [0,662]

NOTA: 1) 100 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

TABELA 4.7 MÉDIAS NOS NÍVEIS DE SOBREPOS*MATRDISP* COEFCORR PARA O MLC.

SOBREPOS	MATRDISP	COEFCORR 1	COEFCORR 2	COEFCORR 3
1	1	1,155 (0,040) [0,783]	1,160 (0,040) [0,787]	0,981 (0,031) [0,668]
	2	1,415 (0,025) [0,948]	1,490 (0,024) [0,969]	1,383 (0,039) [0,903]
2	1	0,960 (0,021) [0,660]	0,965 (0,016) [0,671]	0,922 (0,012) [0,633]
	2	0,990 (0,018) [0,691]	1,025 (0,019) [0,721]	0,993 (0,021) [0,692]

NOTA: 1) 100 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

As características deste método sugerem que ele seja mais apropriado para lidar com grupos com forma circular. O resultado para a condição de matrizes diferentes, com $\rho = 0,8$ e sem sobreposição, o caso onde os grupos apresentavam a forma mais alongada, é coerente com essas características. O método também foi muito afetado pela sobreposção dos grupos.

3) MÉTODO DA MÉDIA DAS LIGAÇÕES (MML).

Os fatores SOBREPOS e MATRDISP, ao n.s. $p < 0,01$, e as interações SOBREPOS*MATRDISP e MATRDISP*COEFCORR com $p < 0,01$ e $p = 0,029$, respectivamente, apresentaram efeitos significativos para as respostas sob esse método.

Os gráficos das interações significativas são apresentadas na Figura 4.3. Do gráfico (a) nesta figura, apresentando a interação SOBREPOS*MATRDISP, observa-se um decréscimo nas médias para os grupos com sobreposição e, além disso, indica um decréscimo mais acentuado para condição de matrizes iguais. Também é sugerido por este gráfico que as médias sob a condição de matrizes iguais são superiores às médias no outro nível de MATRDISP.

O gráfico (b) da Figura 4.3, o qual apresenta a interação MATRDISP*COEFCORR, indica um acréscimo nas médias ao passar da

condição de matrizes diferentes para matrizes iguais, isto dentro de cada nível de COEFCORR. Este gráfico sugere que a interação é devido ao comportamento das médias sob o coeficiente de correlaçao $\rho = 0,8$ (COEFCOR3), uma vez que o comportamento das médias nos outros níveis de COEFCORR são muito semelhantes (retas paralelas).

A Tabela 4.8 e a Tabela 4.9 apresentam as médias para as interações significativas. Na Tabela 4.10 são apresentadas as médias nos níveis de SOBREPOS*MATRDISP*COEFCORR. Comparando-se as médias das Tabelas 4.8 e 4.9 verificou-se que:

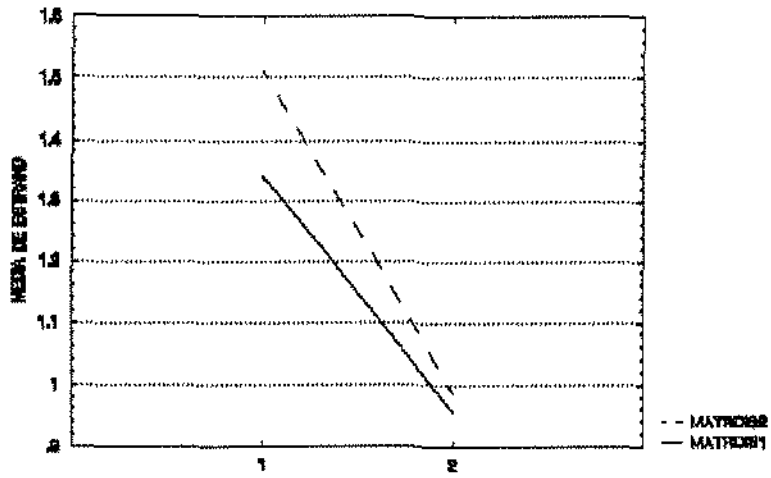
(1) as médias para grupos com sobreposição (Tabela 4.8) foram significativamente menores que as médias no outro nível de SOBREPOS, isto dentro de cada nível da MATRDISP ($p = 0,01$);

(2) dentro de cada nível de MATRDISP (Tabela 4.9) não havia difernças significativas entre as médias nos níveis de COEFCORR.

Através das médias apresentadas na Tabela 4.10, comparando-se os níveis de MATRDISP dentro das combinações dos nívies de SOBREPOS e COEFCORR verificou-se que:

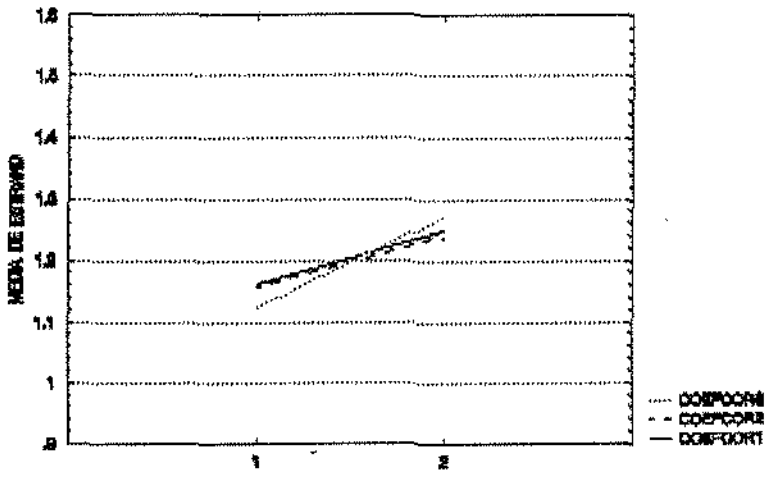
(1) para grupos sem sobreposição e dentro de cada nível de COEFCORR, as médias com matrizes iguais foram significativamente maiores que no outro nível de MATRDISP ($p = 0,01$);

(2) com grupos superpostos e dentro de cada nível de COEFCORR, não foram detectadas diferenças significativas entre as médias nos níveis de MATRDISP, ($p = 0,01$).



(a)

SOBREPOS



(b)

MATRDISP

Figura 4.3 MML. (a) interação SOBREPOS*MATRDISP. (b) interação SOBREPOS*COEFCORR.

Os resultados indicam que o método não foi afetado pela forma dos grupos, entretanto, foi significativamente afetado pela matriz de dispersão no caso de grupos sem sobreposção. Neste caso, o método apresentou um melhor desempenho sob a condição de matrizes iguais. Observa-se também que o método foi significativamente

afetado pela presença de grupos com sobreposição.

TABELA 4.8 MÉDIAS NOS NÍVEIS DE SOBREPOS* MATRDISP PARA O MML.

SOBREPOS	MATRDISP	
	1	2
1	1,340 (0,018) [0,906]	1,512 (0,009) [0,984]
2	0,955 (0,005) [0,665]	0,988 (0,009) [0,691]

NOTA: 1) 150 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

TABELA 4.9 MEDIAS NOS NIVEIS DE MATRDISP * COEFCORR PARA O MML.

MATRDISP	COEFCORR		
	1	2	3
1	1,161 (0,025) [0,795]	1,157 (0,025) [0,792]	1,124 (0,025) [0,770]
2	1,245 (0,028) [0,838]	1,235 (0,029) [0,828]	1,269 (0,029) [0,846]

NOTA: 1) 100 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

Os resultados verificados para este método estão de acordo com aqueles reportados em trabalhos anteriores (MILLIGAN, 1980; DUBIEN e WARD, 1987).

TABELA 4.10 MÉDIAS NOS NÍVEIS DE SOBREPOS*MATRDISP* COEFCORR PARA O MML.

SOBREPOS	MATRDISP	COEFCORR 1	COEFCORR 2	COEFCORR 3
1	1	1,367 (0,029) [0,922]	1,352 (0,030) [0,914]	1,300 (0,035) [0,882]
	2	1,494 (0,019) [0,982]	1,506 (0,017) [0,982]	1,535 (0,013) [0,991]
2	1	0,956 (0,005) [0,667]	0,962 (0,009) [0,670]	0,948 (0,010) [0,658]
	2	0,996 (0,015) [0,697]	0,965 (0,008) [0,674]	1,003 (0,019) [0,701]

NOTA: 1) 50 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

4) MÉTODO DE McQUITTY (MMcQ).

Para este método a interação de três fatores apresentou efeito significativo sobre as respostas ao n.s. p = 0,0279. Cada fator for analisado considerando os níveis dos outros fatores.

Para ajudar a interpretação da interação SOBREPOS*MATRDISP*COEFCORR, a Figura 4.4 apresenta os gráficos das interações entre os fatores. Os gráficos (a) e (b) nesta figura sugerem um decréscimo nas médias para a condição de grupos com sobreposição e, além disso, indica um comportamento diferenciado das médias nos níveis de COEFCORR entre os níveis de SOBREPOS.

Os gráficos (c), (d) e (e) da Figura 4.4 confirmam os comentários feitos acima e indicam valores maiores para as médias sob a condição de matrizes iguais (MATRDIS2). Destes gráficos também é observado que a interação SOBREPOS*MATRDISP é muito semelhante nos níveis de $\rho = 0,0$ e $\rho = 0,4$, entretanto, com $\rho = 0,8$ as médias sob a condição de matrizes diferentes apresenta um comportamento diferenciado o que, provavelmente, configura a interação entre os fatores.

As médias para interação dos três fatores são apresentadas na Tabela 4.11. Comparando-se as médias verificou-se que:

(1) para cada combinação dos níveis de MATRDISP e COEFCORR, havia um decréscimo significativo ao passar para condição de grupos com sobreposição ($p = 0,01$);

(2) para a condição de grupos sem sobreposição e dentro de cada nível de COEFCORR, as médias sob a condição de matrizes iguais foram significativamente maiores que no outro nível de MATRDISP ($p = 0,01$);

(3) com grupos superpostos não haviam diferenças significativas entre as média em cada combinação dos níveis de MATRDISP e COEFCORR ($p = 0,01$);

(4) fixados os níveis de SOBREPOS e MATRDISP, não haviam diferenças significativas entre as médias nos níveis de COEFCORR.

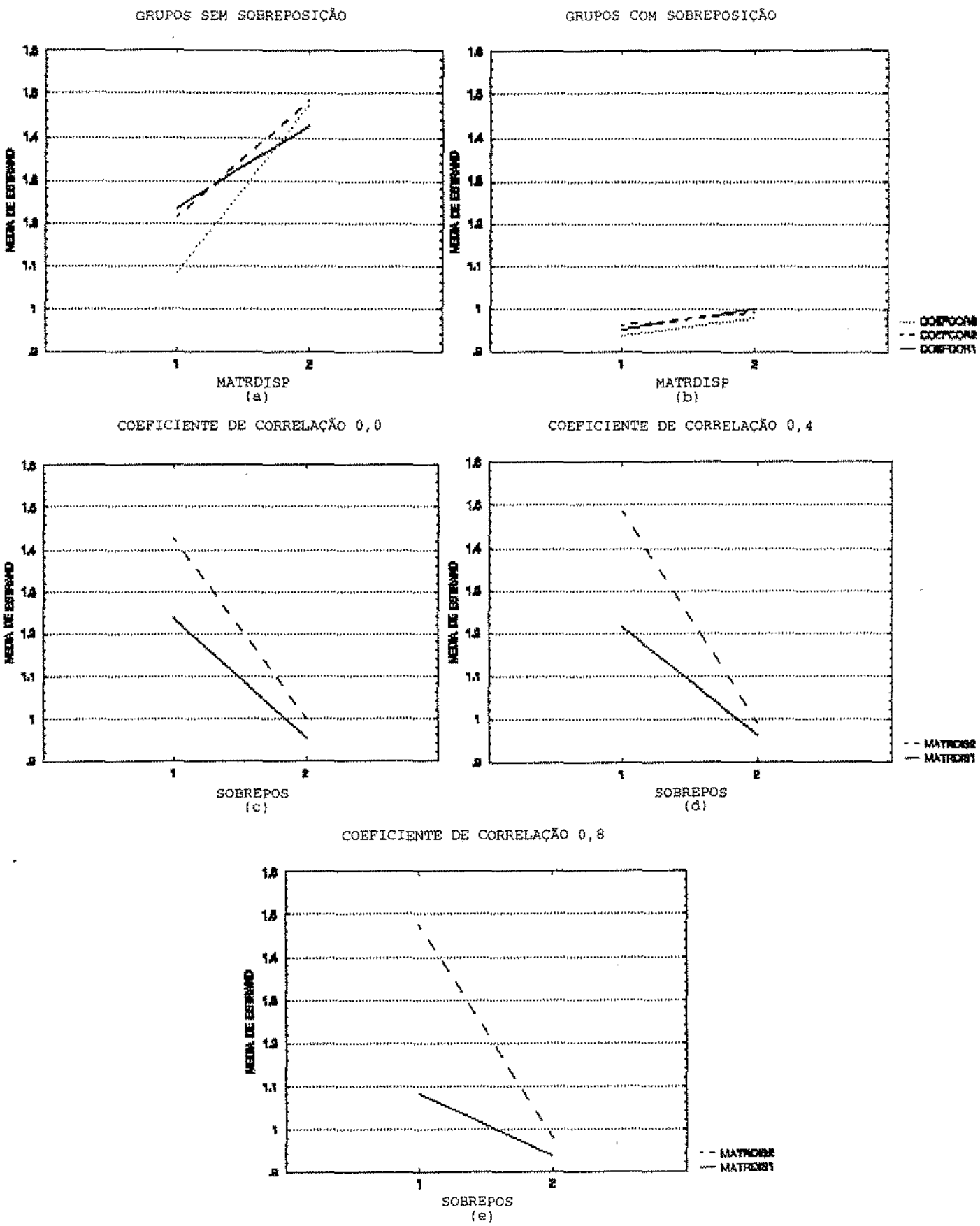


Figura 4.4 MMcQ. Interação MATRDISP*COEFCORR: (a) dentro de grupos sem sobreposição; (b) dentro de grupos com sobreposição. Interação SOBREPOS*MATRDISP: (c) $\rho = 0,0$; (d) $\rho = 0,4$; (e) $\rho = 0,$

TABELA 4.11 MEDIAS NOS NIVEIS DE SOBREPOS * MATRDISP * COEFCORR PARA O MMcQ.

SOBREPOS	MATRDISP	COEFCORR 1	COEFCORR 2	COEFCORR 3
1	1	1,236 (0,043) [0,826]	1,213 (0,046) [0,803]	1,080 (0,039) [0,732]
	2	1,428 (0,030) [0,942]	1,485 (0,020) [0,974]	1,477 (0,029) [0,956]
2	1	0,955 (0,014) [0,662]	0,962 (0,014) [0,669]	0,937 (0,012) [0,647]
	2	1,001 (0,014) [0,704]	0,989 (0,015) [0,691]	0,980 (0,018) [0,681]

NOTA: 1) 50 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

Os resultados indicam que o método foi muito afetado pela presença de grupos com sobreposição. Para a condição de grupos sem sobreposição, as médias com matrizes iguais apresentaram melhores resultados e a forma dos grupos não afetou o método.

Dada sua definição, as características deste método não são claras e, assim, impossibilita associar estes resultados a essas características.

5) MÉTODO CENTRÓIDE (MCEN).

Somente os fatores SOBREPOS, MATRDISP e a interação entre

eles, apresentaram efeitos significativos, todos ao n.s. $p < 0,01$.

O gráfico da interação SOBREPOS*MATRDISP, apresentado na Figura 4.5, sugere que as médias sob a condição de matrizes iguais são superiores às médias no outro nível de MATRDISP e, além disso, que as médias decrescem sob a condição de grupos com sobreposição.

A Tabela 4.12 apresenta as médias nos níveis de SOBREPOS*MATRDISP. Da análise dessas médias verificou-se que:

(1) havia um decréscimo significativo nas médias das

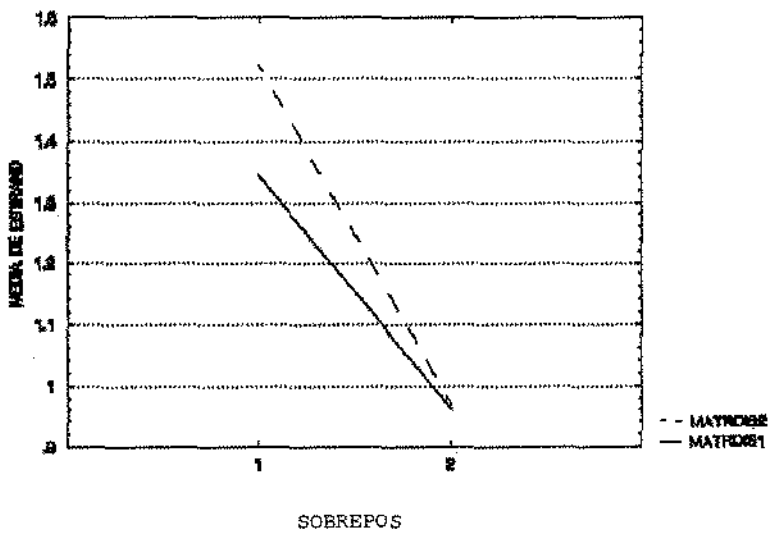


Figura 4.5 MCEN. Interação SOBREPOS*MATRDISP.

respostas ao considerar os grupos com sobreposição, dentro de cada nível de MATRDISP ($p = 0,01$);

(2) sob a condição de grupos sem sobreposição, as médias para grupos com matrizes iguais foram significativamente maiores que no

outro nível de MATRDISP, entretanto, isto não se verificou no outro nível de SOBREPOS (p = 0,01).

Os resultados indicam que o método não foi afetado pela forma dos grupos, entretanto, a sobreposição de grupos afetou significativamente o método. Estes resultados são coerentes com a definição do método, uma vez que ele emprega a distância Euclidiana entre centróides dos grupos como a dissimilaridade entre eles. Essa medida é muito afetada pela presença de observações com valores discrepantes.

TABELA 4.12 MÉDIAS NOS NÍVEIS DE SOBREPOS* MATRDISP PARA O MCEN.

SOBREPOS	MATRDISP	
	1	2
1	1,343 (0,018) [0,908]	1,522 (0,008) [0,987]
2	0,961 (0,006) [0,670]	0,970 (0,004) [0,679]

NOTA: 1) 150 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

6) MÉTODO DA MEDIANA (MMED).

Os três fatores e a interação SOBREPOS*MATRDISP apresentam efeitos significativos, todos ao n.s. p < 0,01.

As médias nos níveis de COEFCORR são apresentadas na Tabela 4.13. Comparando-se essas médias, com n.s. p = 0,01, verificou-se que a média no nível 3, foi significativamente menor que as outras.

O gráfico da interação considerada significativa é apresentado na Figura 4.6. O gráfico indica um decréscimo nas médias para o caso de grupos com sobreposição, sendo este decréscimo mais acentuado para a condição de matrizes iguais, condição que o gráfico sugere ter médias maiores que no outro nível de MATRDISP.

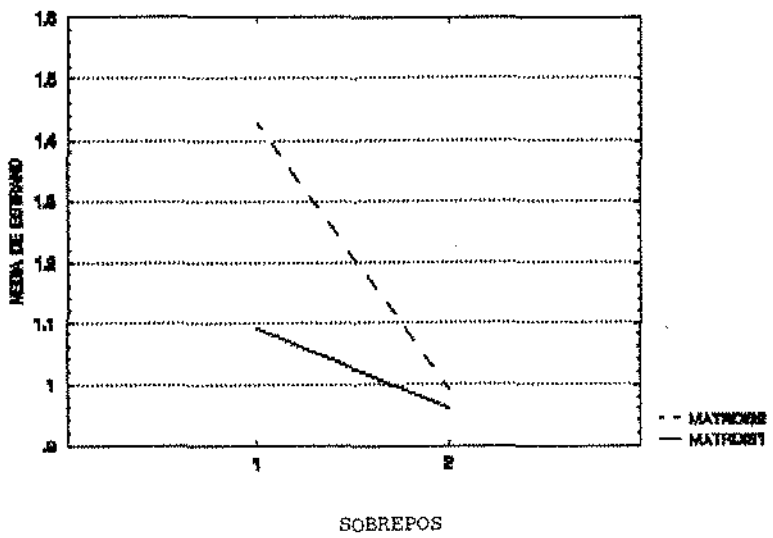


Figura 4.6 MMED. Interação SOBREPOS*MATRDISP.

As médias nos níveis de SOBREPOS * MATRDISP são apresentadas na Tabela 4.14.

Comparando-se as médias apresentadas nas Tabela 4.14 verificou-se que:

TABELA 4.13 AS MÉDIAS NOS NÍVEIS DE COEFCORR PARA O MMED.

NÍVEL	MÉDIA
1	1,149 (0,019) [0,781]
2	1,128 (0,019) [0,765]
3	1,072 (0,019) [0,725]

NOTA: 1) 200 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

TABELA 4.14 MÉDIAS NOS NÍVEIS DE SOBREPOS* MATRDISP PARA O MMED.

SOBREPOS	MATRDISP	
	1	2
1	1,090 (0,022) [0,739]	1,425 (0,020) [0,928]
2	0,960 (0,006) [0,668]	0,991 (0,009) [0,693]

NOTA: 1) 150 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

(1) dentro de cada nível de MATRDISP , havia um decréscimo significativo nas médias das respostas ao passar da condição de grupos sem sobreposição para o outro nível de SOBREPOS ($p = 0,01$);

(2) dentro dos níveis de SOBREPOS, as médias sob a condição de matrizes iguais foram significativamente maiores que as médias

no outro nível de MATRDISP (p = 0,01);

Apesar de ter características semelhantes as do MCEN, os resultados endicaram que este método foi mais sensível às modificações nas estruturas, sendo afetado por todos os fatores considerados. Considera-se estes resultados condizentes com os dos trabalhos anteriores, uma vez que não foram reportados bons desempenhos para este método.

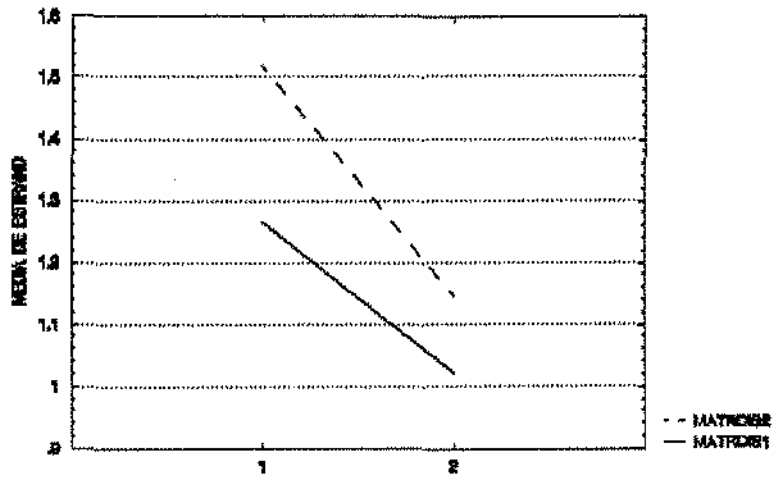
7) MÉTODO DE WARD (MWARD).

Os fatores SOBREPOS, MATRDISP e a interação entre eles apresentam efeitos significativos com p < 0,01. A interação SOBREPOS*COEFCORR também apresentou efeito significativo com p = 0,0105.

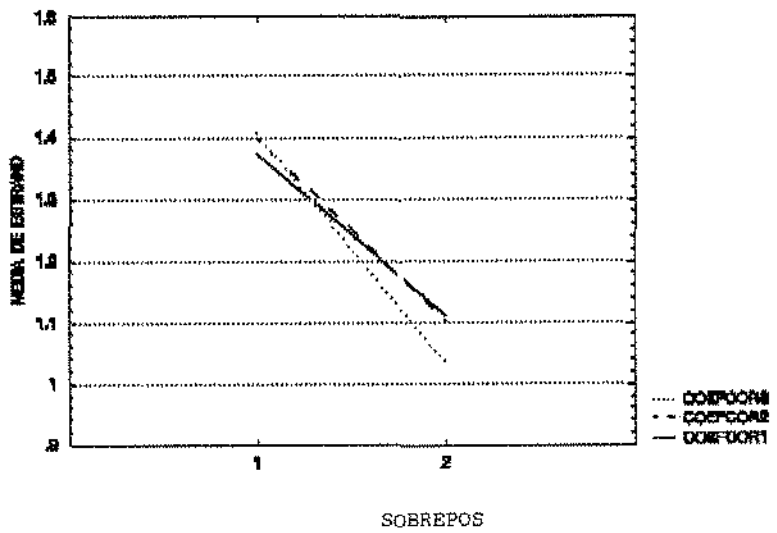
Para ajudar a interpretação das interações significativas, os gráficos dessas interações são apresentadas na Figura 4.7. O gráfico (a) dessa figura, apresentando a interação SOBREPOS*MATRDISP, indica um decréscimo nas médias sob a condição de grupos com sobreposição e, além disso, indica que as médias sob a condição de matrizes iguais são maiores que no outro nível de MATRDISP.

A presença da interação entre SOBREPOS e COEFCORR é evidenciada no gráfico (b) da Figura 4.7. Este gráfico indica que

as médias em todos os níveis de COEFCORR decrescem sob a condição de grupos com sobreposição, entretanto, parece haver indicação de um comportamento não muito diferenciado entre as médias nos níveis de COEFCORR.



(a)



(b)

Figura 4.7 MWARD. (a) interação SOBREPOS*MATRDISP. (b) interção SOBREPOS*COEFCORR.

Da comparação das médias nos cruzamentos dos níveis dos

fatores envolvidos nas interações significativas, médias estas apresentadas na Tabela 4.15 e na Tabela 4.16, verificou-se que:

(1) dentro de cada nível de SOBREPOS (Tabela 4.15), as médias sob a condição de matrizes iguais foram significativamente maiores que as médias no outro nível de MATRDISP ($p = 0,01$);

(2) comparando as médias entre os níveis de COEFCORR e dentro de cada nível de SOBREPOS (Tabela 4.16), foi detectada diferença significativa somente no caso de grupos com sobreposição, onde a média com $\rho = 0.8$ foi significativamente menor que a média com $\rho = 0,0$ ($p = 0,01$);

TABELA 4.15 MEDIAS NOS NIVEIS DE SOBREPOS * MATRDISP PARA O MWARD.

SOBREPOS	MATRDISP	
	1	2
1	1,265 (0,022) [0,854]	1,519 (0,009) [0,985]
2	1,021 (0,013) [0,713]	1,144 (0,015) [0,804]

NOTA: 1) 150 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

As médias para o fator SOBREPOS dentro das combinações dos níveis de MATRDISP e COEFCORR, são apresentadas na Tabela 4.17. Comparando-se essas médias verificou-se que:

(1) fixando os níveis de MATRDISP e COEFCORR, havia um decréscimo significativo nas médias ao passar para a condição de grupos com sobreposição ($p = 0,01$);

TABELA 4.16 MEDIAS NOS NIVEIS DE SOBREPOS* COEFCORR PARA O MWARD.

SOBREPOS	COEFCORR		
	1	2	3
1	1,372 (0,024) [0,913]	1,397 (0,024) [0,925]	1,407 (0,025) [0,921]
2	1,113 (0,020) [0,779]	1,101 (0,016) [0,778]	1,033 (0,019) [0,718]

NOTA: 1) 100 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

TABELA 4.17 MÉDIAS NOS NÍVEIS DE SOBREPOS*MATRDISP* COEFCORR PARA O MWARD.

SOBREPOS	MATRDISP	COEFCORR		
		1	2	3
1	1	1.268 (0,039) [0,855]	1,260 (0,037) [0,858]	1,266 (0,041) [0,850]
	2	1,475 (0,021) [0,970]	1,535 (0,012) [0,992]	1,548 (0,011) [0,993]
2	1	1,046 (0,024) [0,734]	1,047 (0,019) [0,741]	0,969 (0,025) [0,665]
	2	1,181 (0,030) [0,824]	1,154 (0,024) [0,817]	1,097 (0,026) [0,771]

NOTA: 1) 50 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

(2) dentro de cada nível de COEFCORR, as médias sob a condição de matrizes diferentes e grupos sem sobreposição, não apresentaram diferenças significativas com relação as médias sob a

condição de matrizes iguais e grupos superpostos (p = 0,01).

Este método também foi afetado pela sobreposição dos grupos, entretanto, verifica-se que algumas médias dentro da condição de grupos superpostos não apresentaram diferenças significativas para médias com grupos sem sobreposição. Ter apresentado um melhor desempenho sob a condição de matrizes iguais, é um resultado condizente com as características desse método.

8) MÉTODO BETA-FLEXÍVEL (MFLE).

Para este método, os fatores SOBREPOS, MATRDISP e a interação entre eles apresentaram efeitos significativos, todos com o n.s. p < 0,01.

O gráfico da interação SOBREPOS*MATRDISP, apresentado na Figura 4.8, sugere que as médias sob a condição de matrizes iguais são maiores que as médias na outra condição de MATRDISP e, nos dois níveis deste fator, há um decréscimo nas médias sob a condição de grupos com sobreposição.

A Tabela 4.18 apresenta as médias nas combinações de níveis de SOBREPOS e MATRDISP. Comparando-se essas médias verificou-se que:

(1) havia um decréscimo significativo nas médias ao passar

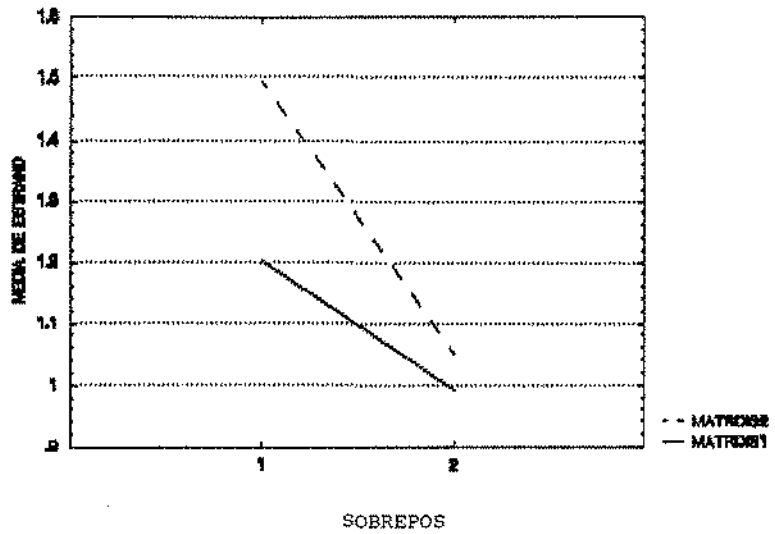


Figura 4.8 MFLE. Interação SOBREPOS*MATRDISP.

para a condição de grupos com sobreposição, isto dentro de cada nível de MATRDISP (p = 0,01);

(2) dentro de cada nível de SOBREPOS, as médias sob a condição de matrizes iguais são significativamente maiores que as médias com matrizes diferentes (p = 0,01).

TABELA 4.18 MEDIAS NOS NIVEIS DE SOBREPOS * COEFCORR PARA O MFLE.

SOBREPOS	MATRDISP	
	1	2
1	1,202 (0,024) [0,808]	1,493 (0,014) [0,969]
2	0,992 (0,012) [0,690]	1,047 (0,012) [0,737]

NOTA: 1) 150 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

Os resultados indicam que o método não foi afetado pela forma dos grupos. Este resultado está de acordo com aqueles reportados por DUBIEN e WARDE (1987).

9) MÉTODO DO k-ÉSIMO VIZINHO MAIS PRÓXIMO (MkVP).

Da Análise de Variância foram detectados efeitos significativos para os fatores SOBREPOS e MATRDISP, ao n.s. $p < 0,01$, e para a interação entre eles com $p = 0,0103$.

O gráfico da interação SOBREPOS*MATRDISP, apresentado na Figura 4.9, indica um decréscimo nas médias sob a condição dos grupos com sobreposição e que as médias com matrizes iguais são maiores que as médias no outro nível de MATRDISP.

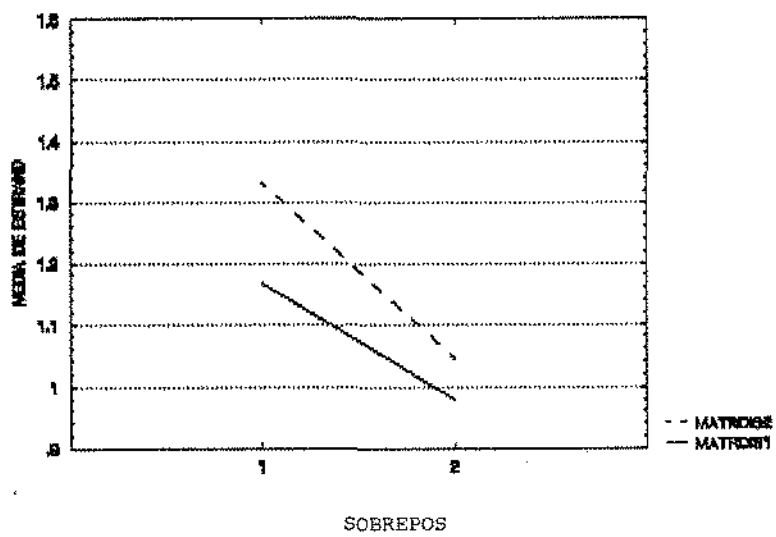


Figura 4.9 MkVP. Interação SOBREPOS*MATRDISP.

As médias nos cruzamentos dos níveis de SOBREPOS e MATRDISP são dadas na Tabela 4.19. Comparando-se essas médias verificou-se que:

(1) as médias sob a condição de grupos com sobreposição foram significativamente menores que as médias no outro nível de SOBREPOS, isto dentro de cada nível de MATRDISP (p = 0,01);

(2) dentro de cada nível de SOBREPOS, as médias sob a condição de matrizes iguais foram significativamente maiores que no outro nível de MATRDISP (p = 0,01).

TABELA 4.19 MEDIAS NOS NIVEIS DE SOBREPOS * MATRDISP PARA O MkVP.

SOBREPOS	SOBREPOS	
	1	2
1	1,172 (0,023) [0,791]	1,331 (0,023) [0,882]
2	0,980 (0,010) [0,683]	1,045 (0,012) [0,734]

NOTA: 1) 150 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

A forma dos grupos não afetou o desempenho do método. Este resultado é coerente com a definição do método pois, segundo os seus autores, ele não depende da estrutura geométrica dos grupos

10) MÉTODO DA LIGAÇÃO DE DENSIDADES EM DOIS ESTÁGIOS (MLDE).

Os fatores SOBREPOS e MATRDISP, ao n.s. $p < 0,01$, e a interação SOBREPOS * COEFCORR, com $p = 0,0407$, apresentaram efeitos significativos para as respostas sob esse método.

As médias nos níveis de MATRDISP são apresentadas na Tabela 4.20. Comparando-se essas médias verificou-se que a média com matrizes iguais foi significativamente maior que sob a condição de matrizes diferentes.

TABELA 4.20 MEDIAS NOS NIVEIS DE MATRDISP PARA O MLDE.

NIVEL	MÉDIA
1	1,166 (0,015) [0,800]
2	1,246 (0,017) [0,838]

NOTA: 1) 300 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

Na Figura 4.10 é apresentado o gráfico da interação SOBREPOS*COEFCORR. O gráfico sugere um decréscimo das médias em todos os níveis de COEFCORR para a condição de grupo com sobreposição, embora, sem diferenças acentuadas entre as médias.

As médias nos cruzamentos dos níveis de SOBREPOS e COEFCORR

são dadas na Tabela 4.21. Comparando-se essas médias, verificou-se que:

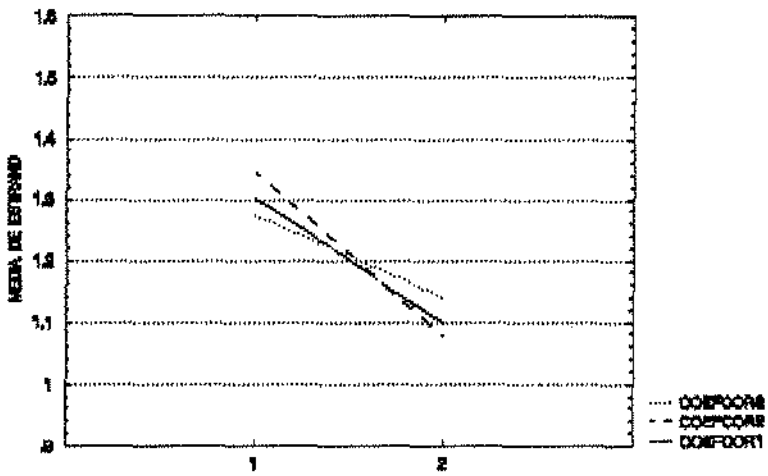


Figura 4.10 MLDE. Interação SOBREPOS*COEFCORR.

TABELA 4.21 MEDIAS NOS NIVEIS DE SOBREPOS * COEFCORR PARA O MLDE.

SOBREPOS	COEFCORR		
	1	2	3
1	1,303	1,344	1,274
	(0,029)	(0,027)	(0,032)
	[0,867]	[0,892]	[0,838]
2	1,099	1,079	1,138
	(0,022)	(0,020)	(0,021)
	[0,764]	[0,756]	[0,797]

NOTA: 1) 100 observações por casela
2) (.) desvio-padrão
3) [.] média das observações não transformadas

(1) dentro de cada nível de COEFCORR havia um decréscimo significativo nas médias para a condição de grupos sobrepostos ($p = 0,01$);

(2) dentro de cada nível de SOBREPOS não foram detectadas diferenças significativas entre as médias nos níveis de COEFCORR;

Pelos resultados, o método não foi afetado pela forma dos grupos o que é condizente com a sua definição, uma vez que este método é uma modificação do MkVP.

4.2.3. Resumo dos Resultados.

Dos fatores considerados, o fator SOBREPOS afetou significativamente todos os métodos. Todos os métodos apresentaram um decréscimo significativo em suas médias da ESTRAND ao considerar a condição de grupos com sobreposiçao. Para o MWARD, entretanto, verificou-se um resultado diferente dos demais métodos. Para este método, as médias sob a condição de grupos sem sobreposição e matrizes iguais, não apresentaram diferenças significativas com relação às médias sob a condição de grupos com sobreposição e matrizes iguais, isto dentro de cada nível de COEFCORR.

Para o fator MATRDISP, todos os métodos apresentaram médias significativamente maiores no nível 2 deste fator, a condição de matrizes de dispersão dentro dos grupos iguais.

Somente alguns métodos foram afetados pelo fator COEFCORR. Para o MLS a média com $\rho = 0,8$ foi significativamente maior que as

médias com outros valores de ρ, isto dentro da condição de matrizes de dispersão dentro dos grupos iguais e grupos sem sobreposição. Para o MLC a média com ρ = 0,8 foi significativamente menor dentro da condição de grupos sem sobreposição. Dentro da condição de grupos com sobreposição, para o MWARD a média com ρ = 0,8 foi significativamente menor que a média com ρ = 0,0. Para os outros métodos não foram detectados efeitos significativos das mudanças nos níveis de COEFCORR.

4.3 ANÁLISE DO EXPERIMENTO 2.

Como já foi comentado, as respostas para esse experimento foram independentes somente se consideradas sob cada método separadamente. Uma Análise de Variância foi conduzida sobre as respostas para cada um dos métodos e os resultados, de maneira geral, ratificaram aqueles descritos para o Experimento 1: os métodos foram todos afetados significativamente pela condição de grupos com sobreposição, os métodos apresentaram melhores desempenhos sob a condição de matrizes iguais e, quanto ao fator COEFCORR, alguns métodos foram mais afetados que outros pelas mudanças nos níveis deste fator, ressaltando que esses métodos são aqueles identificados no outro experimento.

As médias dentro das estruturas segundo cada um dos métodos são apresentados na Tabela 4.22 e a Figura 4.11 apresenta o gráfico dessas médias através das estruturas.

Com base nas médias apresentadas na Tabela 4.22 e com ajuda do gráfico, observa-se que alguns métodos sofreram decréscimos nas médias muito mais acentuados que outros, ao passar para as estruturas com sobreposição de grupos (Estruturas 7 a 12). O gráfico sugere que os métodos que apresentam estes decréscimos mais acentuados são o MLS, o MML e o MCEN. Com decréscimos menos acentuados o gráfico sugere o MWARD, o MFLE e o MLDE. O gráfico também parece sugerir a presença de grupos de métodos com comportamento semelhante entre as estruturas com sobreposição. Um grupo parece bem definido, o dos métodos com resultados "superiores" formado pelo MWARD e o MLDE. Poderia ser considerado um grupo com os métodos MLS, MLC, MML, MMcQ, MCEN e MMED, os métodos com resultados "inferiores". Os métodos MFLE e MkVP poderiam, talvez, constituir um terceiro grupo de métodos com resultados "intermediários".

Dentro das estruturas sem sobreposição não é muito clara a separação dos métodos. Entretanto, o gráfico permite visualizar que o MWARD, o MML, o MFLE e o MCEN apresentam médias entre as "superiores" em todas essas estruturas. O gráfico também sugere que o MkVP é o método com resultados "inferiores".

Para complementar a análise das respostas desse experimento, foi executada uma Análise de Agrupamentos sobre os métodos com o objetivo de verificar possíveis grupos de métodos com desempenhos semelhantes. Cada método foi descrito pelas suas médias dentro das

TABELA 4.22 MÉDIAS DA ESTRAND DENTRO DAS ESTRUTURAS SEGUNDO OS MÉTODOS.

METODO	ESTRUTURA										
	1	2	3	4	5	6	7	8	9	10	11	12
1. MLS	0,908 (0,035)	0,874 (0,037)	0,906 (0,037)	0,792 (0,045)	0,843 (0,042)	0,979 (0,020)	0,578 (0,007)	0,565 (0,005)	0,575 (0,006)	0,556 (0,005)	0,563 (0,005)	0,553 (0,003)
2.MLC	0,965 (0,014)	0,925 (0,021)	0,884 (0,030)	0,932 (0,017)	0,960 (0,015)	0,937 (0,024)	0,615 (0,024)	0,618 (0,032)	0,605 (0,029)	0,639 (0,032)	0,636 (0,035)	0,660 (0,035)
3.MML	0,976 (0,010)	0,925 (0,021)	0,985 (0,006)	0,949 (0,011)	0,985 (0,008)	0,997 (0,002)	0,575 (0,009)	0,561 (0,006)	0,579 (0,006)	0,560 (0,006)	0,575 (0,016)	0,590 (0,024)
4.MMCQ	0,953 (0,025)	0,902 (0,028)	0,883 (0,036)	0,949 (0,013)	0,937 (0,031)	0,967 (0,018)	0,570 (0,009)	0,598 (0,024)	0,587 (0,020)	0,603 (0,027)	0,613 (0,025)	0,576 (0,023)
5.MCEN	0,976 (0,010)	0,929 (0,020)	0,953 (0,022)	0,963 (0,010)	0,990 (0,004)	0,977 (0,020)	0,575 (0,009)	0,561 (0,006)	0,591 (0,013)	0,562 (0,006)	0,575 (0,018)	0,587 (0,024)
6.MMED	0,911 (0,030)	0,900 (0,026)	0,865 (0,036)	0,930 (0,018)	0,918 (0,032)	0,924 (0,031)	0,590 (0,021)	0,595 (0,024)	0,626 (0,026)	0,591 (0,023)	0,653 (0,035)	0,611 (0,030)
7.MWARD	0,978 (0,010)	0,948 (0,015)	0,978 (0,009)	0,961 (0,010)	0,985 (0,005)	0,997 (0,002)	0,837 (0,035)	0,794 (0,035)	0,830 (0,038)	0,808 (0,025)	0,858 (0,025)	0,850 (0,022)
8.MFLE	0,976 (0,010)	0,948 (0,015)	0,964 (0,013)	0,961 (0,009)	0,985 (0,005)	0,990 (0,007)	0,774 (0,037)	0,697 (0,040)	0,740 (0,042)	0,733 (0,035)	0,731 (0,040)	0,771 (0,034)
9.MKVP	0,885 (0,042)	0,762 (0,045)	0,787 (0,048)	0,753 (0,044)	0,852 (0,041)	0,876 (0,033)	0,715 (0,044)	0,774 (0,046)	0,743 (0,043)	0,622 (0,027)	0,686 (0,041)	0,680 (0,041)
10.MLDE	0,921 (0,035)	0,816 (0,044)	0,787 (0,048)	0,942 (0,022)	0,917 (0,029)	0,876 (0,033)	0,802 (0,040)	0,825 (0,042)	0,823 (0,041)	0,840 (0,015)	0,809 (0,035)	0,762 (0,041)

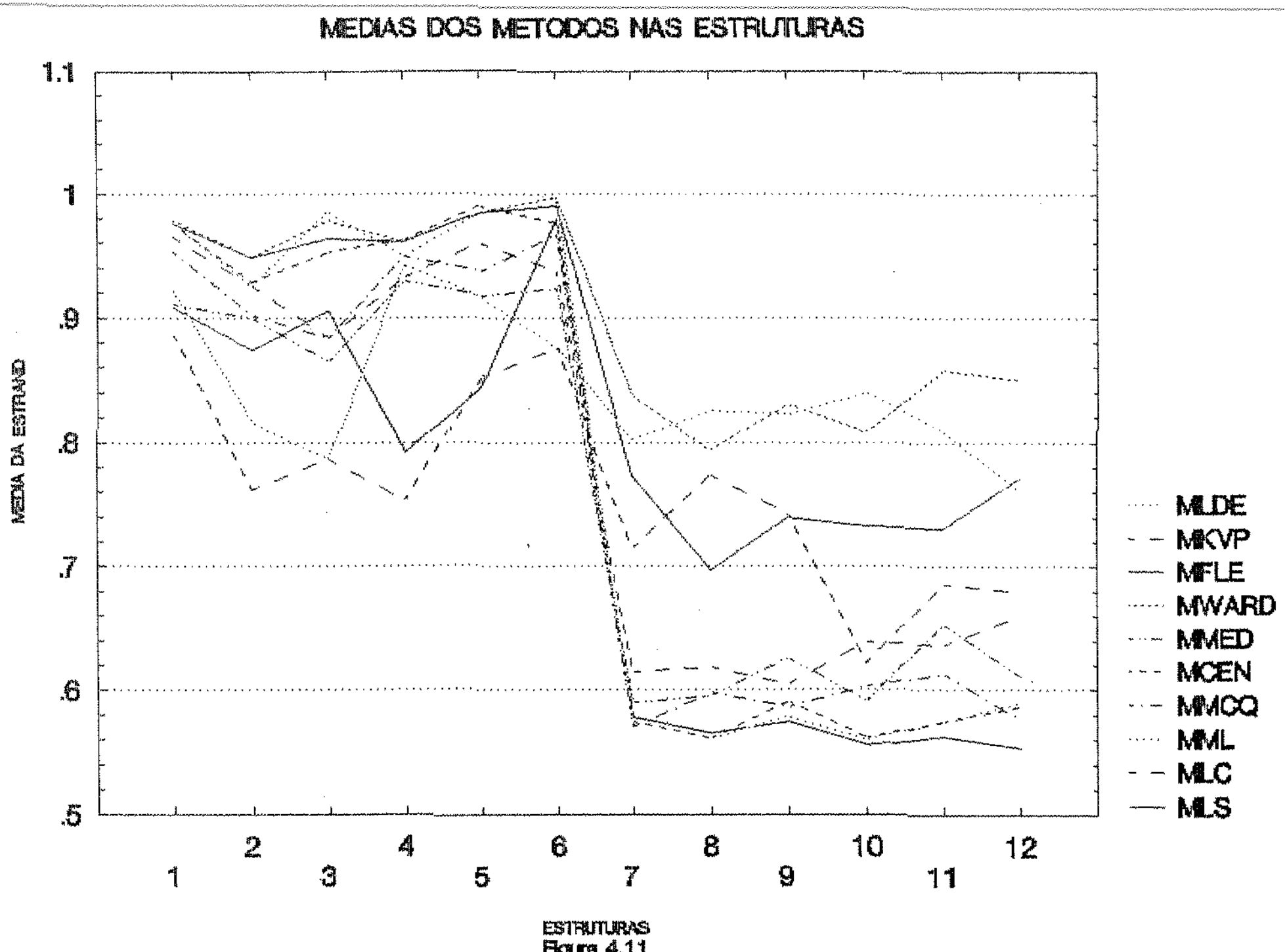

MEDIAS DOS METODOS NAS ESTRUTURAS
MEDIA DA ESTRUTURA
1.1
1
.9
.8
.7
.6
.5
1
2
3
4
5
6
7
8
9
10
11
12
ESTRUTURAS
MLDE
MKVP
MFLE
MWARD
MMED
MCEN
MMCQ
MML
MLC
MLS

Figura 4.11

estruturas, isto é, cada método foi descrito por

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i12}),\ i = 1,2, \ldots, 10,$$

onde x_{ij} corresponde a média do i-ésimo método para a j-ésima estrutura. Para efetuar os agrupamentos foram empregados o MWARD, o MML e o MLDE, sendo estes métodos escolhidos considerando que o MML apresentou um dos melhores desempenhos nas estruturas sem sobreposição, o MLDE para as estruturas com sobreposição e o MWARD obteve este comportamento para todas as estruturas. Os dendogramas para os métodos empregados são apresentados na Figura 4.12.

Através de seus dendogramas, o MWARD e o MML parecem sugerir a existência de três grupos de métodos e os grupos são os mesmos nos dois métodos. Os grupos são:

G_1 = { MLS, MML,MCEN},

G_2 = { MLC, MMcQ, MMED, MFLE, MkVP},

G_3 = {MWARD, MLDE}.

Estes grupos apresentam uma certa coerência. O grupo G_1 contém os métodos com bons desempenhos nas estruturas sem sobreposição e, ao mesmo tempo, os que apresentam decréscimos mais acentuados ao passar para as estruturas com sobreposição. No grupo G_2 estão os métodos com desempenhos de intermediário a inferior entre as estruturas. Os métodos do grupo G_3 parecem ser os métodos com resultados mais "estáveis", o MWARD obteve sempre um dos melhores desempenhos através de todas as estruturas e o MLDE, sendo um dos

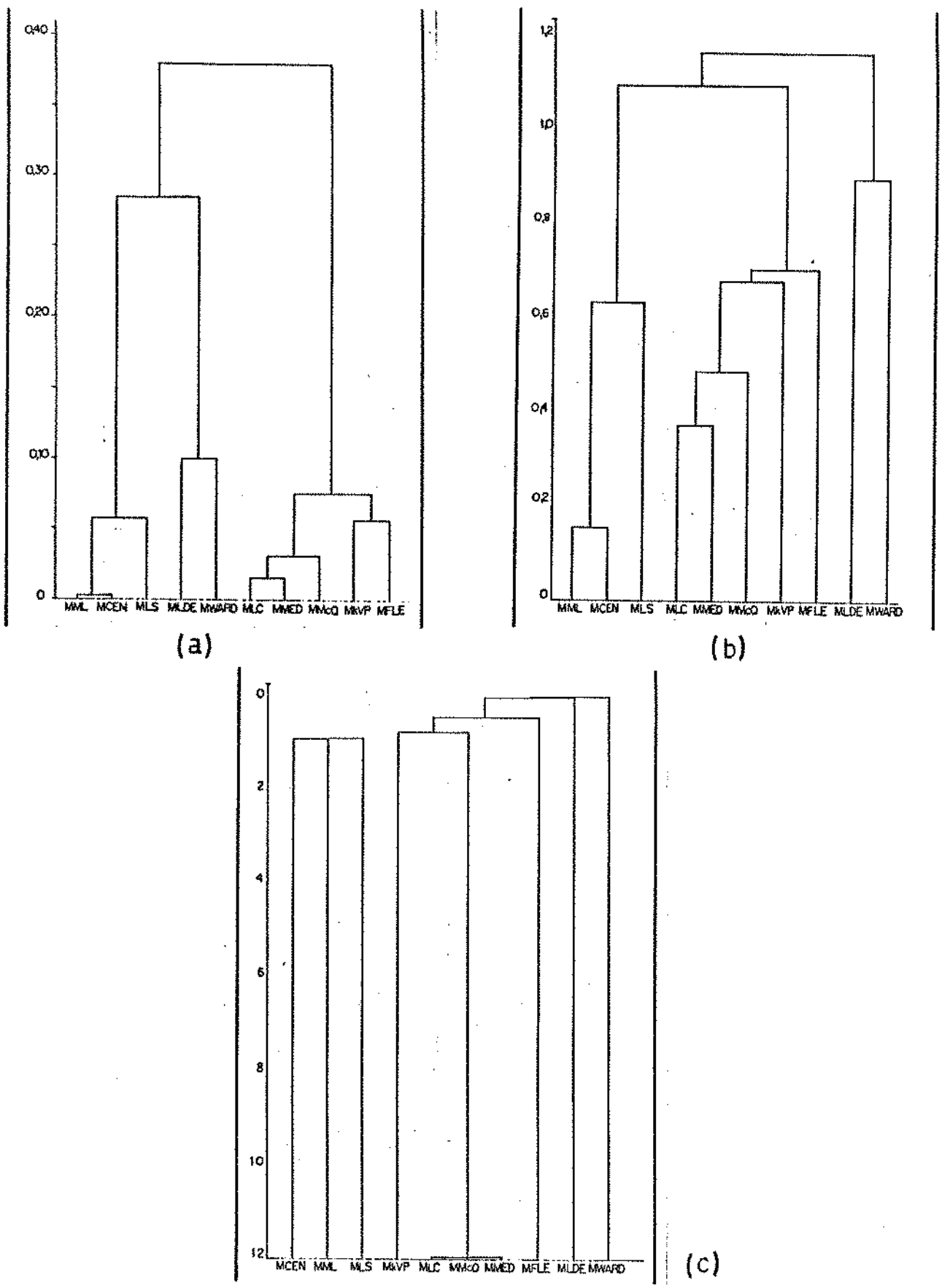


Figura 4.12 Dendograma dos métodos considerando todas as estruturas. (a) MWARD. (b) MML. (c) MLDE.

menos afetados com os grupos com sobreposição, teve médias muito "próximas" em todas as estruturas.

Para o MLDE, o dendograma não é bem definido em termos do número de grupos a considerar, entretanto, se for considerado o número de grupos sugeridos pelos outros dois métodos, os grupos são

G_1 = { MLS, MML, MCEN},

G_2 = { MLC, MMcQ, MMED, MkVP, MFLE, MLDE} e

G_3 = { MWARD}.

Este agrupamento também pode ser considerado apropriado, uma vez que colocar o MWARD isoladamente em um grupo é justificado pelo fato de que os resultados deste método podem ser considerados como os "melhores" através de todas as estruturas.

A Análise de Agrupamentos também foi aplicada considerando separadamente as estruturas sem sobreposição de grupos, Estruturas 1 a 6, e as estruturas com sobreposição de grupos, Estruturas de 7 a 12. Desta forma, em cada um dos casos os métodos foram descritos pelas seis médias nas respectivas estruturas. Os dendogramas para a análise nas estruturas sem sobreposição são apresentados na Figura 4.13 e para as estruturas com sobreposição na Figura 4.14.

Para as estruturas sem sobreposição, os dendogramas do MWARD e do MML sugerem que devem ser considerados dois grupos de métodos.

A composição dos grupos são as mesmas, sendo

G_1 = { MLS, MML, MCEN, MWARD},

G_2 = { MLC, MMcQ, MMED, MFLE, MkVP, MLDE}.

Como já foi comentado, os métodos do G_1 são os métodos com os resultados "superiores" para essas estruturas. O G_2 contém os métodos com as médias menos "estáveis" através dessas estruturas.

No dendograma do MLDE não parece bem definido o número de grupos a considerar, através dele poderiam ser consideradas mais de dois grupos. Entretanto, considerando-se dois grupos, os grupos montados por este método são os mesmos do MWARD e do MML.

Dentro das estruturas com sobreposição, os resultados apresentados pelo MWARD e pelo MML sugerem a presença de dois grupos de métodos. Os dois grupos montados por esses métodos são os mesmos e exibem aspectos evidenciados no gráfico da Figura 4.11. Os grupos foram

G_1 = { MWARD, MLDE},

G_2 = {MLS, MLC, MML, MMcQ, MCEN, MMED, MFLE, MkVP}.

Sendo considerado três grupos de métodos para as estruturas com sobreposição, os grupos apresentados pelo MWARD e pelo MML também são os mesmos. Eles colocam os "objetos" MFLE e MkVP juntos em um terceiro grupo

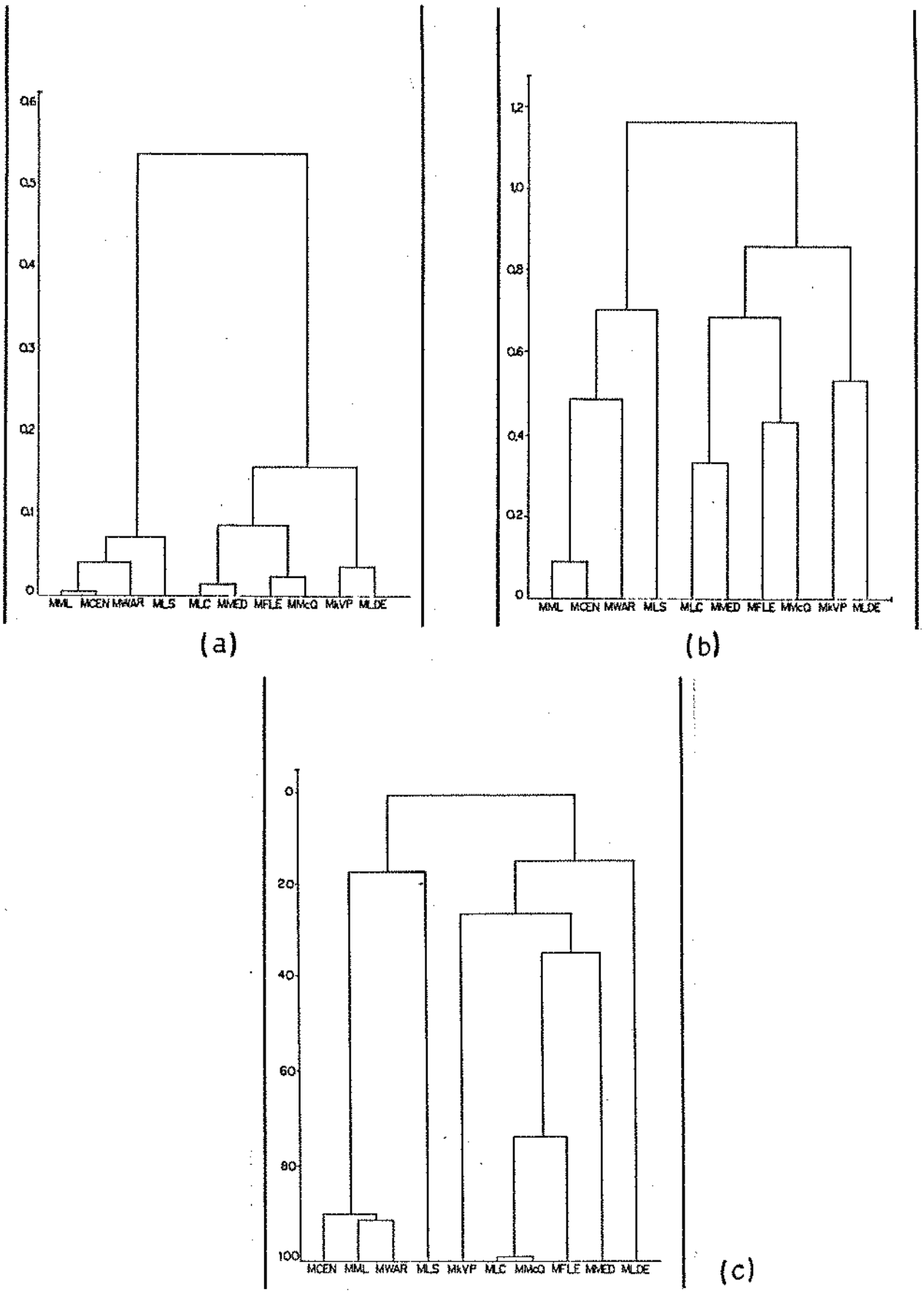


Figura 4.13 Dendogramas dos métodos para as estruturas sem sobreposição. (a) MWARD. (b) MML. (c) MLDE.

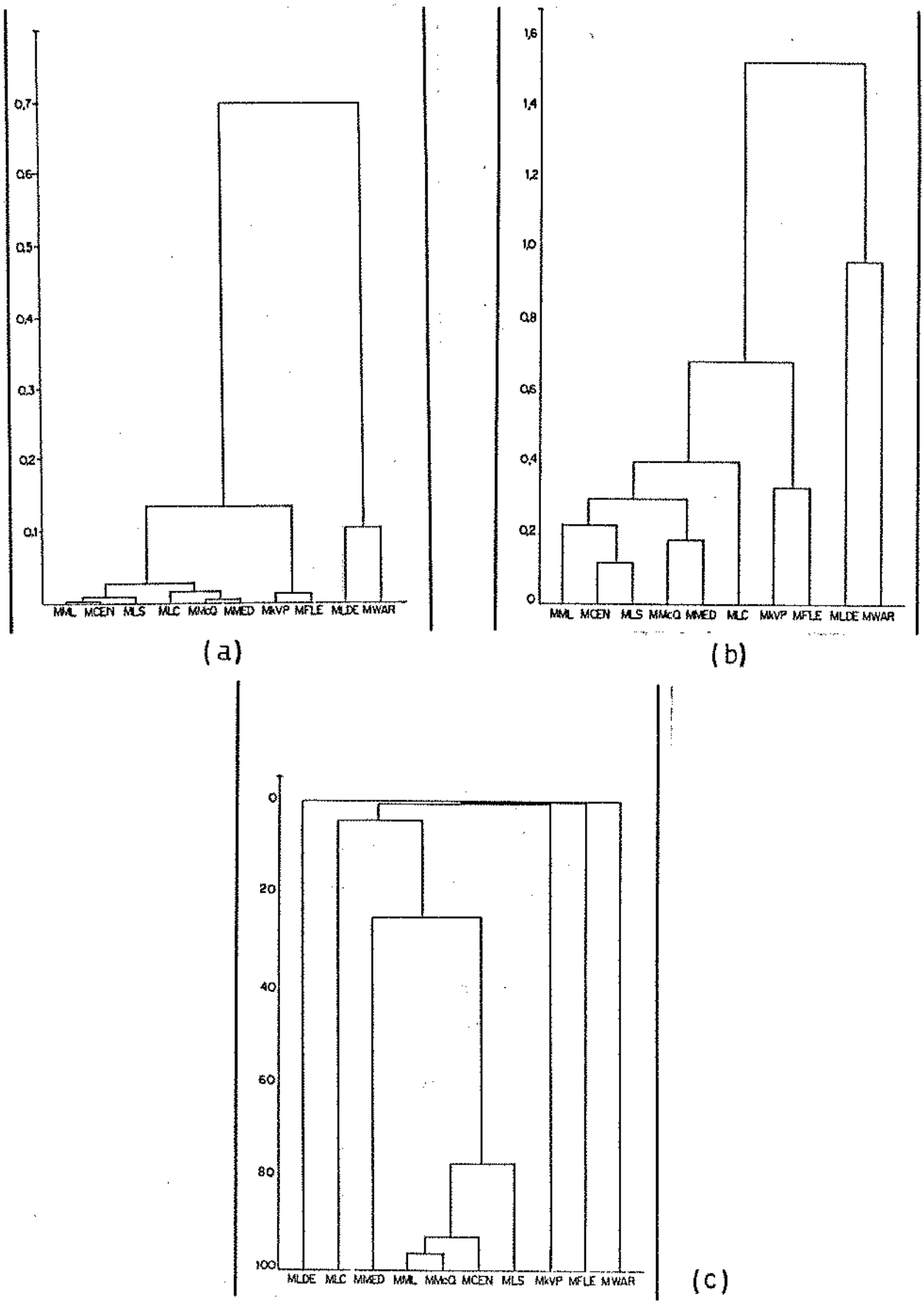


Figura 4.14 Dendogramas dos métodos para as estruturas com sobreposição. (a) MWARD. (b) MML. (c) MLDE.

Para as estruturas com sobreposição o dendograma do MLDE, também não sugere claramente o número de grupos. Considerando o número de grupos sugerido pelos outros métodos, o MLDE fornece os seguintes grupos.

G_1 = {MLS, MLC, MML, MMcQ, MCEN, MED, MWARD, MFLE, MkVP}

G_2 = {MLDE}

Ainda sob a condição acima e tomando três grupos, o MLDE coloca o "objeto" MWARD em um terceiro grupo.

Em todos os casos o MWARD e MML apresentaram resultados bastante apropriados aos aspectos evidenciados no gráfico das médias. O agrupamento montado pelo MLDE para o caso das estruturas com sobreposição, não foi o mais apropriado, entretanto, para estruturas todas juntas considera-se o agrupamento montado por este método como o mais apropriado.

Na seção a seguir serão apresentadas as conclusões das análises e sugestões para aplicações dos métodos de agrupamento.

4.4 CONCLUSÕES E SUGESTÕES

O objetivo deste trabalho era identificar aspectos das estruturas de dados, dentre algumas estruturas especificadas, que afetassem os dez métodos de agrupamentos considerados.

Paralelamente, pretendia-se comparar os desempenhos desses métodos dentro das estruturas consideradas como também, se possível, identificar o método, ou os métodos, com desempenhos que pudessem ser considerados como superiores e, além disso, em quais das condições isso se verificava. As conclusões das análises realizadas sobre as respostas dos experimentos montados com a finalidade de atingir esse objetivos, são apresentadas a seguir.

Para os aspectos da estrutura de dados considerados (os fatores), foi verificado que todos os métodos tiveram seus desempenhos afetados quando considerada a condição de grupos com sobreposição, condição em que os métodos apresentaram um decréscimo nas suas médias. Ficou evidente, entretanto, que alguns métodos foram mais afetados que outros sob essa condição. Os métodos considerados mais afetados foram: o da Ligação Simples, o da Média das Ligações e o Centróide. Esses métodos que para a condição de grupos sem sobreposiao apresentaram resultados entre os "melhores", ao considerar a sobreposição de grupos ficaram entre os métodos com resultados "inferiores". Os métodos de Ward, o Beta-Flexível, o do k-ésimo Vizinho Mais Próximo e o da Ligação de Densidades em Dois Estágios são os métodos considerados menos afetados pela sobreposição dos grupos. Os outros podem ser considerados em uma posição intermediária.

Outro aspecto observado com respeito a questão da sobreposição, é que os resultados indicam que a presença de grupos

com sobreposição da forma considerada aqui, observações com valores discrepantes, obscurecem o efeito dos outros fatores.

Os resultados considerados acima estão de acordo com alguns trabalhos anteriores (EDELBROCK e McLAUGHLIN, 1980; MILLIGAN, 1980) e sugerem que alguns métodos não são os mais adequados se, por exemplo, existirem observações com valores discrepantes ("*outliers*") nos dados. Sob essa condição, os métodos da Ligação Simples, da Média das Ligações e o Centróide parecem não muitos apropriados e, por outro lado, o método de Ward e o da Ligações de Densidades em Dois Estágios poderiam ser empregados.

Um outro aspecto investigado foi o efeito das matrizes de dispersão dentro dos grupos, sob as condições de matrizes diferentes e matrizes iguais. Todos os métodos apresentaram melhor desempenho sob a condição de matrizes iguais no caso de grupos sem sobreposição. Os métodos da Ligação Completa, da Mediana, de Ward o Beta-Flexível, do k-ésimo Vizinho Mais Próximo e o da Ligação de Densidades em Dois Estágios apresentaram melhor desempenho com matrizes iguais nas duas condições de sobreposição. É possível que a disposição dos grupos no espaço das variáveis tenha contribuido para esses resultados, uma vez que ao considerar $\rho = 0,4$ ou $\rho = 0,8$ era esperado que sob a condição de matrizes iguais, os grupos ficassem mais afastados que sob a condição de matrizes diferentes (vide Figura 3.1) e, evidentemente, os métodos deveriam apresentar melhor desempenho sob aquela condição.

O outro aspecto investigado foi o efeito da correlação entre as variáveis. Fixada a condição das matrizes de dispersão como diferentes ou iguais, o valor do coeficiente de correlação (ρ) entre as variáveis foi tomado como 0,0, 0,4 ou 0,8. Os resultados indicaram que somente alguns métodos apresentaram evidências de serem afetados pelas mudanças em ρ. Somente os métodos da Ligação Simples, da Ligação Completa, da Mediana e o de Ward apresentarm resultados significativamente afetados pela mudança no valor de ρ. Dentro da condição de grupos sem sobreposição e matrizes de dispersão iguais, a média sob $\rho = 0,8$ para o método da Ligação Simples foi significativamente superior às médias sob os outros valores de ρ, enquanto que para o método da Ligação Completa esta média foi significativamente inferior. Este resultado é bastante coerente com as características desses dois métodos, uma vez que o MLS é indicado para lidar com grupos com forma alongada e MLC para aqueles mais esféricos. Para o método de Ward, dentro da condição de grupos com sobreposição, a média com $\rho = 0,8$ foi significativamente menor que com $\rho = 0,0$. O método da Mediana apresentou uma média significativamente inferior com $\rho = 0,8$.

Os comentários colocados no parágrafo anterior, sobre o método da ligação Simples e o da Ligação Completa, estão de acordo com aqueles feitos em DUBIEM e WARDE (1987), entretanto, sobre o método Mediana e o método de WARD não foram reportado em trabalhos anteriores. Os resultados sugerem que alguns métodos são mais robustos às mudanças nas formas dos grupos e, como numa situação

prática usualmente a forma dos grupos não pode ser especificada "*a priori*", isto dá uma certa vantagem a esses métodos.

Da comparação entre os métodos não é possível estabelecer o melhor método, entretanto, foi verificado que alguns métodos apresentaram desempenhos que, de alguma forma, se diferenciaram dos outros. Quando considerados as estruturas sem sobreposição de grupos, os métodos da Média da Ligações, Centróide e o de Ward formaram um grupo de métodos cujos resultados foram considerados os melhores. Para a estrutura envolvendo os grupos com sobreposição, onde pela forma empregada na simulação ocorreram observações com valores discrepantes, os métodos de Ward e da Ligação de Densidades em Dois Estágios apresentaram os melhores resultados.

Os resultados observados neste trabalho permitem fazer alguns comentários. Na maioria dos trabalhos anteriores o método de Ward foi sempre um dos métodos apresentando um dos melhores desempenhos para as diversas condições investigadas (vide seção 1.5). Aqui neste trabalho este método também foi o que apresentou os melhores resultados. Duas condições foram reportados em trabalhos anteriores sob as quais este método apresentou desempenhos inferiores aos demais, considerar grupos com tamanhos diferentes (KUIPER e FISHER, 1975) e considerar as variáveis com uma correlação relativamente alta (EVERITT, 1980). Essas condições foram consideradas aqui e, ainda assim, o método teve os melhores resultados.

Os métodos da Média das Ligações e o Centróide, também são métodos considerados com bons desempenhos em trabalhos anteriores. Estes métodos apresentam a vantagem de serem pouco afetados pela forma dos grupos, entretanto, têm a desvantagem de serem muito afetados por observações com valores discrepantes. Como já foi citado, para as estruturas sem sobreposição de grupos, neste trabalho estes métodos apresentaram um bom desempenho.

Para o método Beta-Flexível também foram observados "bons" resultados em todas as estruturas, em particular, para as estruturas com sobreposição de grupos. Este método apresenta ainda, como já foi comentado, a possibilidade de tomar diferentes valores para o seu parâmetro β e tornar o método adequado às diferentes formas de grupos.

O método da Ligação de Densidades em Dois Estágios não foi considerado em trabalhos anteriores mas, sob as condições investigadas aqui, este método apresentou resultados bastante interessantes. Ele pode ser considerado o método menos sensível à condição de grupos com sobreposição aqui considerada, isto sugere ser este método uma boa alternativa para lidar com conjuntos de dados apresentando observações discrepantes.

Como já foi comentado, não é possível indicar o melhor método de agrupamento. Com base nos resultados deste trabalho e dos trabalhos anteriores, considera-se conveniente empregar mais de um

método de agrupamento nas análises. Não sendo verificadas observações com valores discrepantes nos dados sugere-se empregar o método da Média das Ligações e o método Centróide. Havendo informações sobre a presença de observações com valores discrepantes sugere-se empregar o método da Ligação de Densidades em Dois Estágios e o método Beta-Flexível. Nos dois casos e na falta de informação sobre os dados sugere-se que o método de Ward deva sempre ser empregado.

Uma outra sugestão diz respeito a uma proposta de método de agrupamento. Amenizar os efeitos da presença de observações com valores discrepantes nos dados, é sempre uma preocupação quando se emprega qualquer técnica estatística e, dados os resultados deste trabalho, esta questão é muito relevante para os métodos de agrupamento. Nas primeiras investigações neste trabalho foi considerada uma medida mais robusta como dissimilaridade entre os objetos, a norma L_1, com objetivo de verificar se isto contribuiria para um melhor desempenho dos métodos. Entretanto, parece não ser isto suficiente, uma vez que resta ainda a questão da dissimilaridade entre os grupos. A idéia, portanto, é empregar uma medida robusta como dissimilaridade entre os grupos. A sugestão seria uma modificação nos métodos que empregam média na determinação de dissimilaridades entre grupos, onde seria empregada a mediana em vez da média. Dois exemplos seriam o método da Média das Ligações e o Centróide e, com isto, teríamos mais duas alternativas de métodos para lidar com observações discrepantes nos dados.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ANDERBERG, M. R. (1973) *Cluster Analysis for Applications*. Academic Press: NY.

2. BACKER, F. B. (1974). Stability of Two Hierarchical Grouping Techniques. Case I: Sensitivity to data errors. *J. Amer. Stat. Assoc.*, 346, 440-445.

3. BACKER, F. B. e HUBERT, L. J. (1975). Measuring the Power os Hierarchical Cluster Analysis. *J. Amer. Stat. Assoc.*, 349, 31-38.

4. BAYNE, C. K.; BEAUCHAMP, J. J.; BEGOVICH, C. L. e KANE, V. E. (1980). Monte Carlo Comparasions of Selected Clustering Procedures. *Pattern Recognition*, 12, 51-62.

5. BOX, G. E. P. e MÜLLER, M. E. (1958). A Note on the Generation

of Random Normal Deviates. *Ann. Math. Statis.*, 29, 610-611.

6. BOX, G. E. P.; HUNTER, W. G. e HUNTER, J. S. (1978). *Statistics for Experimenters*. John Wiley & Sons: NY.

7. BUSTOS, O. H. e FRERY, A. C. (1992). *Simulação Estocástica: Teoria e Algoritmos (Versão Completa)*. Monografias de Matemática, nº 49, CNPq - IMPA.

8. CORMACK, R. M. (1971). A Review of Classification. *Journal of the Royal Statistical Society (Series A)*, 134, 321-353.

9. CUNNINGHAM, K. M. e OLGIVIE, J. C. (1972). Evaluation of Hierarchical Grouping Tecnniques: A Preliminary Study. *The Computer Journal*, 15, 209-213.

10. DAGPUNAR, J. (1988). *Principles of Randon Variate Generation*. Claredon Press: Oxford.

11. DUBES, R. e JAIN, A. K. (1976). Clustering Techniques: The user's Dilemma. *Pattern Recognition*, 8, 247-260.

12. DUBIEN, J. L. e WARDE, W. D. (1979). A Mathematical Comparision of the Members of Infinity Family of Aglomerative Clustering Algorithms. *The Canadian Journal of Statistics*, 7, 29-38.

13. DUBIEN, J. L. e WARDE, W. D. (1987). A Comparision of Aglomerative Clustering Methods whit Respect to Noise. *Communications in Statistics - Theory and Methods*, 16, 1433-1460.

14. EDELBROCK, C. e McLAUGHLIN, B. (1980). Hierarchical Cluster Analysis Using Interclass Correlations: A Mixture Model Study. *Multivariate Behavorial Research*, 15, 299-318.

15. EVERITT, B. (1980). *Cluster Analysis. Second Edition*. Halted Press: London.

16. FISHER, L. e VAN NESS, J. W. (1971). Admissible Clustering Procedures. *Biometrika*, 58, 91-104.

17. FISHMAN, G. S. e MOORE, L. R. (1982). A Statistical Evaluation of Multiplicative Congruencial Generators with Modulus 2^{31} - 1. *J. Amer. Stat. Assoc*, 77, 129-136.

18. GNANADESIKAN, R. e KETTENRING, J. R. (1989). Discriminant Analysis and Clustering. Panel on Discriminant Analysis, Classification and Clustering. *Statistical Science*, 4, 34-69.

19. GOWER, J. C. (1967). A Comparasion of Some Methods of Cluster Analysis. *Biometrics*, 23, 623-628.

20. GOWER, J. C. (1971). Discussion of a Paper by R. M. Comarck. *Journal of the Royal Statistics Society (Series A)*, 134, 360-365.

21. HOAGLIN, D. C. e ANDREWS, D. F. (1975). The Reporting of Comparation-Based Results in Statistics. *The American Statistician, Statistical Computing*. 29, 122-126.

22. JARDINE, N. e SIBSON, R. (1968). The Construction of Hierarchic and Non-Hierarchc Classification. *Computer Journal*, 11, 177-183.

23. JOHNSON, M. E. (1987). *Multivariate Statistical Simulation*. John Wiley & Sons: NY.

24. JOHNSON, S. C. (1967). Hierarchical Clustering Schems. *Psychometrika*, 32, 241-254.

25. JOHNSON, R. A. e WICHERN, D. W. (1988). *Applied Multivariate Statistical Analysis. Second Edition.* Prentice Hall International, INC.

26. KALKSTEIN, L. S.; TAN, G. e SKINDLOV, J. A. (1987). An Evaluation of Three Clustering Procedures for Use in Synoptic Climatological Classification. *Journal of Climate and Applied Metereology*, 26, 716-730.

27. KAUFMAN, L. e ROUSSEEUW, P. J. (1990). *Finding Groups in Data. An Introduction*. John Wiley & Sons, INC.

28. KOPP, B. (1978a). Hierarchical Classification I: Single Linkage Method. *Biometrical Journal*, 20, 495-501.

29. KOPP, B. (1978b). Hierarchical Classification II: Complete Linkage Method. *Biometrical Journal*, 20, 597-602.

30. KOPP, B. (1978c). Hierarchical Classification III: Average Linkage, Median, Centroid, Ward, Flexible - Strategy. *Biometrical Journal*, 20, 703-711.

31. KUIPER, F. K. e FISHER, L. (1975). A Monte Carlo Comparasion of Clustering Procedures. *Biometrics*, 31, 777-783.

32. LANCE, G. N. e WILLIAMS, W. T. (1967). A General Theory of Classificatory Sorting Strategies I. Hierarchical Systems. *Computer Journal*, 9, 373-380.

33. McQUITTY, L. L. (1966). Similarity Analysis by Reciprocal Pairs for Discrete and Continous Data. *Educational and Psychological Meansurement*, 26, 825-831.

34. MILLIGAN, G. W. (1979). Utrametric Hierarchical Clustering Algorithms. *Psychometrika*, 44, 343-346.

35. MILLIGAN, G. W. (1980). An Examination of the Effect of Six Types of Error Pertubation on Fiftten Clustering Algorithms. *Psychometrika*, 45, 325-342.

36. MILLIGAN, G. W. (1981). A Review of Monte Carlo Tests of Cluster Analysis. *Multivariate Behavorial Reserch*, 16, 379-407.

37. MILLIGAN, G. W. e COOPER, M. C. (1985). An Examination of Procedures for Determining the Number of Clusters in Data Set. *Psychometrika*, 50, 159-179.

38. MILLIGAN, G. W. e ISAAC, P. D. (1980). The Validation of Four Utrametric Clustering Algorithms. *Pattern Recognition*, 12, 41-50.

39. MONTGOMERY, D. C. (1991). *Design and Analysis of Experiments. Third Edition*. John Wiley & Sons: NY.

40. MOREY, L. C. e AGRESTI, A. (1984). The Mensurement of Classification Agreement: An Adjustment to the Rand Statistics for Chance Agreement. *Educational and Psychological Mensurement*, 44, 33-37.

41. RAND, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *J. Amer. Stat. Assoc.*, 66, 846-850.

42. SAS INSTITUTE INC (1988). *SAS/IML User's Guide: Release 6.03 Edition*. Cary, NC: SAS Institute INC.

43. SAS INSTITUTE INC (1989). *SAS/STAT User's Guide: Version 6, Fourth Edition. Volume 1 e 2* . Cary, NC: SAS Institute INC.

44. SEBER, G. A. F. (1984). *Multivariate Observations*. John Wiley & Sons: NY.

45. SIBSON, R. (1971). Some Observations on a Paper by Lance e Williams. *The Computer Journal*, 14, 156-157.

46. SNEATH, P. H. D. e SOKAL, R. R. (1973). *Numerical Taxonomy. The Principles and Pratice of Numerical Classifications*. W. H. Freeman and Company: San Francisco.

47. WARD, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *J. Amer. Stat. Assoc.* 58, 236-244.

48. WILLIAMS, W. T.; LANCE, G. N.; DALE, M. B. e CLIFFORD, H. T. (1971). Controversy Concerning The Criteria for Taxonometric Strategies. *The Computer Journal*, 14, 162-164.

49. WISHART, D. (1969). An Algorithm for Hierarchical Classification, *Biometrics*, 25, 165-170.

50. WONG, M. A. e LANE, T. (1983). A kth Nearest Neighbour Clustering Procedure. *Journal of the Royal Statistics Society (Series B)*, 45, 362-368.