UNIVERSIDADE ESTADUAL DE CAMPINAS
DEPARTAMENTO DE ESTATÍSTICA - IMECC

A Distância de Mahalanobis para Misturas de Variáveis Categóricas e Contínuas;
Aplicação na Análise de Agrupamento

Pledson Guedes de Medeiros

Orientação
Profa. Dra. Regina Célia Carvalho Pinto Moran

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação da Universidade Estadual de Campinas para obtenção de título de Mestre em Estatística
Campinas - S.P.
1995

FICHA CATALOGRAFICA ELABORADA PELA
BIBLIOTECA DO IMECC DA UNICAMP

Medeiros, Pledson Guedes de
M467d A distancia de Mahalanobis para misturas de variaveis categoricas e continuas; aplicacao na analise de agrupamentos / Pledson Guedes de Medeiros. -- Campinas, ISP : s.n.], 1995.

Orientadora : Regina Celia Carvalho Pinto Moran
Dissertacao (mestrado) - Universidade Estadual de Campinas, Instituto de Matematica, Estatistica e Ciencia da Computacao.

1. Analise multivariada. 2. Distancias - Medicao. 3. Variaveis aleatorias. I. Moran, Regina Celia Carvalho Pinto. II. Universidade Estadual de Campinas. Instituto de Matematica, Estatistica e Ciencia da Computacao. III. Titulo.

A Distância de Mahalanobis para Misturas de Variáveis Categóricas e Contínuas; Aplicação na Análise de Agrupamento

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pelo Sr. Pledson Guedes de Medeiros e aprovada pela comissão julgadora.

Campinas, 22 de maio de 1995

Profa. Dra. Regina Célia Carvalho Pinto Moran

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do título de mestre em Estatística.

Dedico este trabalho a minha mãe
Ilda Guedes de Medeiros e a
Dione Maria Valença

Gostaria de expressar meus agradecimentos

- À Deus por ter me dado a chance de crescer profissionalmente e de conhecer pessoas maravilhosas durante o mestrado, especialmente a Dione;

- À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, pelo apoio financeiro;

- À Profa. Dra. Regina C. C. P. Moran pelo apoio em todos os momentos e pela sugestão e orientação em um assunto tão interessante;

- Ao Prof. Dr. Paulo César Formiga Ramos pelo incentivo durante a graduação;

- Aos professores da UFRN pela compreenssão e liberação para que eu pudesse concluir este trabalho;

- Aos amigos que fiz durante o mestrado que, para não pecar por esquecimento, prefiro não citar nomes;

- Aos funcionários da secretaria do Departamento de Estatística da Unicamp, especialmente ao Marcelo e ao Ermerson que, sempre que requisitados, não mediram esforços para ajudar;

- Aos funcionários da secretaria de Pós Graduação e da Biblioteca, que também foram muito prestativos;

Sumário

Resumo

Este trabalho aborda o problema da quantificação da distância existente entre indivíduos mensurados sob o contexto de misturas. Inicialmente é feito um estudo com relação a caracterização de misturas, por meio da obtenção da expressão para a matriz de variâncias e covariâncias na forma de blocos. A seguir, como contribuição deste trabalho à literatura, é determinada a expressão em blocos para a inversa desta matriz e, utilizando a mesma como matriz de ponderação, tem-se uma extensão da Distância de Mahalanobis para este contexto. Além da implementação computacional desta extensão, também é feita uma aplicação desta distância utilizando Técnicas Hierárquicas de Agrupamento.

Capítulo 1

Introdução

1.1 Considerações Preliminares

1.1.1 Dados com Misturas de Variáveis

É muito comum, ao se trabalhar com conjuntos de dados reais, depararmo-nos com a existência de variáveis de diferentes tipos pois várias áreas do conhecimento tratam, frequentemente, problemas nos quais a escolha das variáveis focalizadas recai tanto sobre categóricas, quanto sobre contínuas. Por exemplo, na área de educação uma população de alunos pode ser estudada focalizando variáveis sócio-econômicas de forma categorizada e variáveis de desempenho acadêmico de forma contínua. Em um outro exemplo, na área de estudos populacionais, uma região pode ser tipificada pelo estudo de seus municípios em variáveis categorizadas e contínuas. As primeiras podem referir-se a presença/ausência ou intensidade de determinados indicadores de desenvolvimento tais como, escolas, hospitais, bancos, etc. Já as variáveis do segundo tipo podem conter informações como km de ruas asfaltadas, consumo de energia elétrica, entre outras. Quando isto acontece, dizemos que estamos trabalhando com um conjunto de dados com *misturas de variáveis categóricas e contínuas* que, de agora em diante, chamaremos apenas de *misturas*. O problema de *misturas* de variáveis contínuas e categóricas vem sendo, ultimamente, bastante estudado no contexto de *Análise Discriminante*, visto que a existência de variáveis

categóricas pode gerar sub-populações de indivíduos. Neste campo podemos destacar a grande contribuição que vem sendo dada, ao longo destas duas últimas décadas, por Olkin e Tate e, principalmente, por Krzanowski. Mais recentemente, Vlachonikolis tem publicado vários trabalhos sobre este assunto.

1.1.2 A Estrutura de Inter-relação no Contexto de Misturas

Quando estamos trabalhando com um conjunto de dados onde há *misturas*, é importante levar em consideração que as variáveis contínuas podem possuir um comportamento diferente se observadas com relação as combinações dos valores assumidos pelas variáveis categóricas, ou seja, pode existir uma estrutura de inter-relação multivariada entre as contínuas e estas combinações, o que caracterizaria a existência de sub-populações de indivíduos. Este fato é de grande importância quando desejamos, por exemplo, obter um agrupamento desses indivíduos, pois os grupos a serem formados podem ser muito influenciados, e até mesmo caracterizados por esta estrutura de inter-relação. Para esclarecer melhor o que foi dito, consideremos a seguinte situação fictícia:

Exemplo 1.1: Em um determinado estudo foram mensuradas 3 variáveis categóricas: X_1 = Sexo (m e f), X_2 = Renda Familiar (1 : mais de 20 salários, 2 : entre 5 e 20 salários e 3 : menos de 5 salários) e X_3 = Cor (1 : preto, 2 : moreno, 3 : amarelo e 4 : branco). Neste mesmo estudo também foram mensuradas 3 variáveis contínuas X_4, X_5 e X_6, sobre cada indivíduo. Esquematicamente, considerando o vetor $\underset{\sim}{X} = (X_1, X_2, X_3)'$, podemos vislumbrar abaixo todas as possíveis combinações dos valores assumidos por cada variável categórica:

(X_1, X_2, X_3)	$\#combin.$	(X_1, X_2, X_3)	$\#combin.$
$(m, 1, 1)$	1	$(f, 1, 1)$	13
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(m, 1, 4)$	4	$(f, 1, 4)$	16
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(m, 3, 1)$	9	$(f, 3, 1)$	21
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(m, 3, 4)$	12	$(f, 3, 4)$	24

Pode acontecer, por exemplo, das mensurações das variáveis contínuas X_4, X_5 e X_6, nos indivíduos da combinação 4 (homem, de cor branca com renda superior a 20 salários), terem um comportamento bem diferente das feitas sobre os indivíduos da combinação 21 (mulher, de cor negra com renda inferior a 5 salários). Esta dependência induz a uma estrutura de inter-relação entre as combinações das variáveis categóricas e as contínuas. Com isso, se tentarmos agrupar os indivíduos objetos deste estudo, intuitivamente somos levados a pensar que esta estrutura de inter-relação terá uma grande influência nos grupos a serem formados. Por isto, são necessários alguns cuidados especiais quando estamos lidando com dados com *misturas*.

1.1.3 Procedimentos Iniciais para o Trato com misturas

Existem, na literatura, vários procedimentos criados para se trabalhar com dados na presença de *misturas*. Segundo Cochran e Hopkins (1961), podemos categorizar as variáveis contínuas e proceder como se todas fossem categóricas. Já segundo Krzanowski (1980), se as variáveis categóricas estão ordenadas, o procedimento mais simples é designar um escore numérico arbitrário para cada categoria possível destas variáveis, e proceder como se elas fossem contínuas. Outra solução possível, também citada em Krzanowski (1980), seria analisar os dois tipos de variáveis separadamente. Apesar destas alternativas serem válidas, elas apresentam sérios problemas, tais como:

- Atribuir às variáveis uma mensuração maior que a observada, no caso de tratar variáveis categóricas como contínuas;
- Perda de informação, no caso de tratar variáveis contínuas como categóricas;
- Não considerar qualquer associação existente entre as variáveis contínuas e categóricas, no caso de tratá-las separadamente.

Em uma abordagem estritamente paramétrica, antes de fazer qualquer análise em dados com *misturas*, precisamos formular um modelo que sirva de base para esta análise. Como observa Krzanowski (1988), a distribuição normal multivariada tem sido bastante usada como base para a análise em dados contínuos, enquanto a distribuição multinomial é, segundo ele, a base natural para dados categóricos. Intuitivamente, uma combinação dessas distribuições poderia ser um modelo apropriado para *misturas*, ou seja, um modelo satisfatório seria aquele que especificasse a distribuição conjunta de todas as variáveis de modo que tanto a associação entre categóricas e contínuas, quanto as características marginais de cada tipo de variável, possam ser modeladas. Partindo do princípio de que a distribuição conjunta de um grupo de variáveis pode ser expressa como o produto da distribuição marginal de um grupo delas, pela distribuição condicional das restantes dado os valores das primeiras, o primeiro passo para se encontrar um modelo adequado é observar uma divisão das variáveis dentro de dois grupos. No caso de *misturas*, a divisão mais natural é formar um grupo com as contínuas e o outro com as categóricas. Com isso, se pretendemos especificar a distribuição conjunta de q variáveis categóricas e p contínuas, são dois os caminhos possíveis a se seguir: expressá-la como a distribuição condicional das categóricas dado os valores das variáveis contínuas, multiplicada pela distribuição conjunta dessas últimas, ou expressá-la como a distribuição condicional das contínuas dado os valores das variáveis categóricas, multiplicada pela distribuição conjunta dessas últimas.

A primeira alternativa foi defendida por Cox (1972), que sugere que a distribuição conjunta de *misturas* de binárias e contínuas pode ser escrita como uma distribuição

condicional logística das variáveis binárias para dados valores das variáveis contínuas, multiplicada por uma distribuição normal multivariada marginal para as últimas.

Com relação a segunda possibilidade, Tate (1954) propõe um modelo para o caso de *misturas* envolvendo apenas uma variável contínua Y e uma variável categórica binária X (assume apenas valores 0 e 1), chamado *modelo bisserial pontual*, no qual a distribuição marginal da variável categórica binária X é multiplicada pela distribuição condicional de Y dado $X = x$ fixo. Já Olkin e Tate (1961), apresentam uma extensão multivariada para o *modelo bisserial pontual*, chamado *modelo de posição* (*"location model"*). Neste modelo, as variáveis categóricas $(X_1, X_2, \ldots, X_q)$ são arranjadas em forma de uma tabela de contingência, de modo que se X_i possui s_i categorias, a tabela de contingência terá $s = s_1 \times s_2 \times \ldots \times s_q$ posições, onde cada arranjo de $\underset{\sim}{X} = (X_1, X_2, \ldots, X_q)$ define de forma única uma posição nesta tabela e, estas posições, são identificadas como categorias de um vetor multinomial $\underset{\sim}{Z} = (Z_1, Z_2, \ldots, Z_s)'$, também de forma única. Com isso, no *modelo de posição*, a distribuição marginal da posição Z_m observada é multiplicada pela distribuição condicional das contínuas $\underset{\sim}{Y} = (Y_1, Y_2, \ldots, Y_c)'$ dado Z_m, determinando assim a distribuição conjunta entre a posição observada e o vetor contínuo. Vale salientar que no *modelo de posição*, a distribuição condicional das contínuas $\underset{\sim}{Y} = (Y_1, Y_2, \ldots, Y_c)'$ dado Z_m varia apenas com relação ao vetor de médias porém, Krzanowski (1983) sugere uma generalização deste modelo, assumindo que a matriz de variâncias e covariâncias Σ da distribuição condicional $\underset{\sim}{Y} \backslash Z_m$ também varia de acordo com a categoria Z_m observada.

Segundo Krzanowski (1988), a razão para a segunda alternativa ser a mais utilizada é que na primeira, defendida por Cox (1972), a estimação dos parâmetros inevitavelmente acaba em esquemas computacionalmente iterativos. Enquanto que na segunda, o modelo pode ser manuseado de forma mais simples computacionalmente e, por isto, tem sido mais utilizada em problemas com *misturas*. Krzanowski (1993) volta a discutir estas opções e reintera que a escolha da segunda é mais adequada, citando também que os trabalhos mais recentes desenvolvidos nesta área adotam a mesma.

Todavia, quando o interesse em estudo recai sobre uma quantificação da dissimilaridade (distância) existente entre indivíduos mensurados sob *misturas*, que é o tema abordado nesta dissertação, os trabalhos existentes se mostram em uma fase bastante embrionária. Isto decorre, basicamente, da dificuldade existente para a obtenção desta quantificação, pois quando queremos estudar indivíduos com relação a proximidade ou distância existente entre os mesmos, a lógica nos manda observá-los segundo as variáveis porém, se o número de variáveis mensuradas não for muito pequeno, fica difícil perceber visualmente o quanto eles são parecidos ou distintos.

1.2 O Problema Abordado e Como é Tratado

Formalmente, podemos dizer que o nosso problema é o seguinte: Dada uma amotra de n objetos (indivíduos) de um vetor misto $\underset{\sim}{W}$, queremos quantificar as dissimilaridades existentes entre os indivíduos $\underset{\sim}{w_i}$'s, $i = 1, 2, \ldots, n$, mensurados, onde,

$$\underset{\sim}{W} = (\underset{\sim}{X} \mid \underset{\sim}{Y}).$$

$\underset{\sim}{X} = (X_1, X_2, \ldots, X_c)$:Vetor linha de variáveis categóricas e,
$\underset{\sim}{Y} = (Y_1, Y_2, \ldots, Y_p)$:Vetor linha de variáveis contínuas.

Para tratar este problema, os trabalhos existentes na literatura utilizam os populares coeficientes de dissimilaridade (onde quanto maior o valor observado mais distantes estão os indivíduos) para que possamos obter a quantificação desejada. Dessa forma, a pergunta que passamos a nos fazer é a seguinte: Qual o coeficiente de similaridade ou dissimilaridade (distância) mais adequado ao problema em questão? Esta indagação é muito comum entre os estudiosos do assunto, dado que não existe um coeficiente melhor que os outros para qualquer situação mas, sim, o mais adequado a um determinado caso. Isto decorre, obviamente, do alto grau de subjetividade inerente a estes coeficientes.

Segundo Gower (1970), diferentes situações podem levar à definição de diferentes

distâncias pois, em determinados casos, é possível montar um coeficiente que dê melhores resultados através do conhecimento prévio do problema. Isto nos explica, ao menos intuitivamente, o motivo da existência de tantos coeficientes para o caso em que todas as variáveis mensuradas são de um único tipo. Um dos exemplos mais recentes, é a distância de Shen, Bie e Chiu (ver Shen, Bie e Chiu, 1993), proposta com a finalidade de classificar texturas de imagens. Vale salientar que a textura é um aspecto de superfície muito importante para a percepção visual, pois dá informações essenciais para reconhecimento e interpretação de imagens.

Quando pretendemos quantificar a distância existente entre dois indivíduos, no contexto de *misturas*, a escolha de um coeficiente de dissimilaridade é bastante discutível devido a natureza dos dados. Esta discussão está relacionada principalmente as seguintes questões:

a) Devemos ponderar ou não a contribuição dos diferentes tipos de variáveis (contínuas, nominais e ordinais)? Se sim, como deve ser feita esta ponderação?

b) Devemos incorporar ou não a estrutura de inter-relação existente entre os tipos de variáveis?

Podemos observar que estes problemas introduzem um auto grau de subjetividade à escolha e, por isto, torna-se imprescindível a opinião do pesquisador responsável pelo estudo em questão. Também para o caso de misturas, tratado neste trabalho, mesmo existindo um número menor de coeficientes, não é concensual escolher a distância mais apropriada. Dentre os existentes para o contexto de misturas, podemos destacar os seguintes:

- Distância combinada,
- Distância a partir do Coeficiente geral de similaridade de Gower,
- Distância obtida por transformação das variáveis,
- Distância de Romesburg.

Além deles, vale ressaltar que Krzanowski (1983) propõe uma distância, baseada no modelo de posição introduzido por Olkin e Tate (1961) e discutido no *Capítulo 2*, que incorpora a estrutura de inter-relação existente entre as variáveis, através da utilização da matriz de variâncias e covariâncias no coeficiente de distância. Porém, este coeficiente não será abordado nesta dissertação, pois apresenta as seguintes restrições à sua utilização para um caso de *misturas*:

a) A distância é construída sob a suposição de que as variáveis contínuas possuem distribuição normal e exige que, caso o mesmo não seja verificado, as variáveis sejam transformadas. Apesar da multinormalidade ser desejada, visto que as distâncias passam a ter um maior significado, esta exigência não é muito interessante pois restringe o nosso estudo às *misturas* de variáveis normais e categóricas.

b) As expressões das covariâncias entre contínuas e categóricas não são apresentadas, o que dificulta a implementação computacional desta distância.

Comumente, o que se faz é utilizar uma distância combinada por coeficientes propostos para cada um dos tipos de variáveis. Outra alternativa, que também é bastante utilizada, é a transformação dos dados de forma que passemos a trabalhar com um único tipo de variável.

1.3 Objetivos

Constatada a limitação existente na literatura, esta dissertação foi confeccionada com os seguintes objetivos:

- Determinar a estrutura de inter-relação existente entre as variáveis para o caso de *misturas*. Pois como pretendemos quantificar a distância existente entre os indivíduos, neste contexto, é importante conhecer esta estrutura;
- Apresentar uma opção à quantificação das distâncias entre indivíduos, no contexto de *misturas*, que incorpore a estrutura de inter-relação existente entre as variáveis.

Pois são poucos os coeficientes existentes e, além disso, os mais conhecidos e utilizados não incorporam esta estrutura;

- Implementar computacionalmente este novo coeficiente de dissimilaridade, para que o mesmo possa ser incorporado à literatura no assunto;

- Apresentar uma aplicação deste coeficiente na *Análise de Agrupamento*, visto que esta técnica exploratória multivariada se constitui no maior campo para a utilização dos coeficientes de dissimilaridade.

1.4 A Estrutura do Trabalho

Em função dos objetivos descritos acima, esta dissertação está estruturada de forma a responder as seguintes questões:

1) Qual a expressão para a matriz de variâncias e covariâncias para *misturas*?

Para respondermos a esta questão, ver *Capítulo 2*, achamos conveniente partir do princípio proposto por Olkin e Tate (1961) para o trato com *misturas*. Princípio este que se baseia na manipulação de cada combinação das variáveis categóricas como uma categoria multinomial, visto que isto permite estudar o comportamento das variáveis contínuas dentro de cada uma das categorias geradas pelos distintos arranjos do vetor de variáveis categóricas. Para tanto, primeiro utilizamos uma transformação para passarmos de um vetor categórico a um multinomial. A partir desta transformação, apresentamos as espressões para as inter-relações existentes entre as categorias multinomiais e as variáveis contínuas. Vale salientar que estas expressões analíticas para a matriz de variâncias e covariâncias foram apresentadas inicialmente por Olkin e Tate (1961) e posteriormente melhor caracterizadas por León (1993), sem a necessidade de suposição de normalidade às variáveis contínuas. Como contribuição deste trabalho, tivemos o cuidado de dar uma maior interpretabilidade aos resultados obtidos. As expressões são apresentadas em

blocos nas formas algébrica e matricial, tanto para o caso populacional quanto para o amostral.

2) Após a determinação da forma analítica para a estrutura de inter-relação entre as variáveis, no contexto de *misturas*, surge a seguinte questão: qual entre os coeficiente que incorporam a estrutura de inter-relação entre as variáveis, é o mais indicado para o contexto de misturas?

A *distância de Mahalanobis*, proposta por Mahalanobis (1936), tem se tornado um dos coeficientes padrões para o caso em que todas as variáveis envolvidas são quantitativas (ver Kotz e Johnson, 1985). Esta distância, além de ponderar pelo inverso dos desvios padrões de cada uma das variáveis, também leva em conta o grau de inter-relação existente entre elas (ver Bolshev, 1969). Com o objetivo de incorporar estas inter-relações quando as variáveis envolvidas são qualitativas, Balakrishnan e Sangghvi (1968) e Kurczynski (1970) adaptaram a *distância de Mahalanobis* para este caso. Contudo, esta distância não tem sido utilizada no contexto de *misturas* provavelmente, por não se ter a expressão analítica da matriz de variâncias e covariâncias. Como contribuição deste trabalho à literatura vamos fazer uma extensão desta distância para este caso, utilizando a matriz de variâncias e covariâncias determinada no *Capítulo 2*.

3) Para chegarmos a uma *Extensão da Distância de Mahalanobis para misturas*, não basta apenas a determinação da matriz de variâncias e covariâncias em blocos. Precisamos ainda de respostas às seguintes questões: qual é a forma da inversa desta matriz? e, finalmente, qual a expressão analítica para esta *extensão*?

As respostas a estas duas questões, dadas no *Capítulo 3*, é uma contribuição deste trabalho à literatura pois, por definição, a *Distância de Mahalanobis* é escrita em função da inversa da matriz de variâncias e covariâncias e, para chegarmos a *Extensão da Distância de Mahalanobis para misturas*, é necesário a determinação desta inversa. É importante ressaltar que esta inversa é expressa na forma de blocos, o que permite que a forma

analítica da *Extensão da Distância de Mahalanobis para misturas* também o seja. Além desta extensão, no *Capítulo 3* também são apresentados outros *Coeficientes de Distância* existentes para o trato com *misturas*.

4) Após a determinação da forma analítica para a *Extensão da Distância de Mahalanobis para misturas*, a indagação que se faz é: Como utilizar este novo coeficiente na análise de um conjunto de dados reais?

Para que possamos utilizar a *Extensão da Distância de Mahalanobis* em qualquer problema de quantificação da distância entre indivíduos sob *misturas*, e também fazer uma aplicação deste coeficiente de distância para o caso de agrupamento de dados, fazemos a implementação computacional deste coeficiente. Esta implementação, ver *Apêndice D*, é feita através de programação em *SAS*. Para tornar mais fácil o entendimento desta implementação, bem como o programa computacional desenvolvido, é feita uma aplicação em um conjunto de dados com *misturas* e apresentados os resultados, ver *Capítulo 5*. Para facilitar o nosso trabalho, resolvemos utilizar apenas os *Métodos Hierárquicos*, ver *Capítulo 4*, que se mostrassem compatíveis à utilização desta extensão, visto que os mesmos são os mais difundidos na literatura.

Como complementação aos cinco Capítulos que compõem esta dissertação, são apresentados, na forma de apêndices, alguns resultados importantes à melhor compreensão do trabalho.

Capítulo 2

Caracterização de Misturas

2.1 Resumo do Capítulo

Inicialmente, na *seção 2.2*, definimos algumas transformações que podem ser utilizadas para se trabalhar com vetores aleatórios mistos, ilustrando as suas aplicações através de exemplos. Na *seção 2.3* é apresentada a expressão, determinada por Olkin e Tate (1961) sob o modelo de posição, para a matriz de variâncias e covariâncias de um vetor misto. Finalmente, na *seção 2.4*, é feita a dedução das variâncias e covariâncias populacionais e amostrais para *misturas* de um vetor multinomial e um vetor contínuo. Vale salientar que as expressões obtidas na *seção 2.4* são apresentadas tanto na forma algébrica quanto na forma matricial, para facilitar a sua posterior implementação computacional.

2.2 Transformações de Variáveis

Para determinarmos a estrutura de inter-relação existente para o caso de misturas, associando a cada combinação dos valores assumidos pelo vetor categórico uma única categoria multinomial (princípio adotado por Olkin e Tate, 1961), precisamos de uma transformação que possibilite tratar um vetor de dados categóricos como um multinomial. Para tanto, apresentamos a seguir algumas transformações, encontradas na literatura, que podem ser utilizadas para este propósito.

Inicialmente, vamos considerar o caso particular em que todas as variáveis categóricas

são binárias (assumem apenas valores 0 e 1). Para este caso, apresentamos abaixo a transformação usada por Krzanowski (1975).

Definição 2.1: Seja $\underset{\sim}{X} = (X_1, X_2, \ldots, X_q)'$ um vetor aleatório binário, ou seja, as variáveis X_i são todas binárias, $i = 1, 2, \ldots, q$. O vetor $\underset{\sim}{X}$ pode ser expresso como um vetor multinomial $\underset{\sim}{Z} = (Z_1, \ldots, Z_k)'$, onde $k = 2^q$, de modo que cada combinação do vetor $\underset{\sim}{X}$ define uma categoria multinomial, de forma única, dada por $m = 1 + \sum_{i=1}^{q} x_i 2^{(i-1)}$, ou seja, se $\underset{\sim}{x} = (x_1, x_2, \ldots, x_q)'$ for observado, então $z_m = 1$ e $z_i = 0$, $\forall i \neq m$; $i, m = 1, 2, \ldots, k$.

Quando da existência de variáveis categóricas com mais de duas categorias, Krzanowski (1980) usa o artifício de transformar primeiro cada uma destas em um vetor binário, usando a *Definição 2.2*, dada abaixo.

Definição 2.2: Seja X uma variável aleatória categórica com c categorias. X pode ser substituída por $k = c - 1$ variáveis binárias X_i, $i = 1, 2, \ldots, k$, tal que se a i-ésima categoria de X for observada, então $X_i = 1$ e $X_j = 0$, $\forall i \neq j$, $i, j = 1, 2, \ldots, k$. Todas as variáveis binárias são iguais a zero quando a última categoria c for observada.

Após a utilização da transformação da *Definição 2.2*, Krzanowski transforma o vetor formado por todos os vetores binários gerados em um multinomial, usando a *Definição 2.1*. Para ilustrar a aplicação sequencial da *Definição 2.2* e da *Definição 2.1*, e para vislumbrar melhor as mudanças provocadas por estas transformações, vamos considerar o seguinte exemplo:

Exemplo 2.1: Consideremos que em uma pesquisa realizada junto à eleitores brasileiros, com relação ao segundo turno da última eleição presidencial, foram mensuradas as seguintes variáveis categóricas:

X_1 : *Voto para presidente, tal que* $x_{i1} \in \{FHC, Lula\}$.

X_2 : *Escolaridade, tal que* $x_{i2} \in \{Analfab., Prim.\ ou\ Secund., Universit.\}$.

Assumindo que k_i representa o número de categorias da variável X_i, $i = 1, 2$, temos $k = k_1 \times k_2 = 3 \times 2 = 6$ combinações possíveis das categorias das variáveis X_1 e X_2. Aplicando primeiramente a *Definição 2.2*, nas variáveis X_1 e X_2, temos que X_1 será substituída por uma variável binária Y_1, dado que possui apenas duas categorias, enquanto X_2 será substituída por duas variáveis binárias Y_2 e Y_3, pois possui três categorias, ou seja.

$$Y_1 = \begin{cases} 1, \ se \ X_1 = FHC \\ 0, \ se \ X_1 = Lula, \end{cases}$$

e,

$$(Y_2, Y_3) = \begin{cases} (1,0), \ se \ X_2 = Analfab. \\ (0,1), \ se \ X_2 = Prim. \ ou \ Secund. \\ (0,0), \ se \ X_2 = Universit. \end{cases}$$

Com isso, temos a seguinte relação entre as combinações de $\underset{\sim}{X} = (X_1, X_2)'$ e o vetor binário $\underset{\sim}{Y} = (Y_1, Y_2, Y_3)'$:

(X_1, X_2)	(Y_1, Y_2, Y_3)
$(FHC, Analfab.)$	$(1,1,0)$
$(FHC, Prim.ouSecund.)$	$(1,0,1)$
$(FHC, Universit.)$	$(1,0,0)$
$(Lula, Analfab.)$	$(0,1,0)$
$(Lula, Prim.ouSecund.)$	$(0,0,1)$
$(Lula, Universit.)$	$(0,0,0)$

Podemos observar, porém, que os valores $(1,1,1)$ e $(0,1,1)$ do vetor binário $\underset{\sim}{Y} = (Y_1, Y_2, Y_3)'$ gerado, não correspondem a nenhuma combinação do vetor categórico $\underset{\sim}{X}$

$= (X_1, X_2)'$. Com isso, aplicando a *Definição 2.1* no vetor binário $\underset{\sim}{Y} = (Y_1, Y_2, Y_3)'$, temos que este será transformado em um vetor multinomial $\underset{\sim}{Z}$ com 8 categorias, tal que,

(X_1, X_2)	(Y_1, Y_2, Y_3)	m	$(Z_1, \ldots, Z_8)$
$(FHC, Analfab.)$	$(1,1,0)$	4	$(0,0,0,1,0,0,0,0)$
$(FHC, Prim.ouSecund.)$	$(1,0,1)$	6	$(0,0,0,0,0,1,0,0)$
$(FHC, Universit.)$	$(1,0,0)$	2	$(0,1,0,0,0,0,0,0)$
$(Lula, Analfab.)$	$(0,1,0)$	3	$(0,0,1,0,0,0,0,0)$
$(Lula, Prim.ouSecund.)$	$(0,0,1)$	5	$(0,0,0,0,1,0,0,0)$
$(Lula, Universit.)$	$(0,0,0)$	1	$(1,0,0,0,0,0,0,0)$

e,

$$\underset{\sim}{Z} \sim Multinomial(1, p_1, \ldots, p_8). \tag{2.1}$$

onde, $p_i \in (0,1), \forall i = 1, \ldots, 8$, $\sum_{i=1}^{8} p_i = 1$ e,

$$p_7 = P(Z_7 = 1) = P(\underset{\sim}{Y} = (0,1,1)') = 0.$$

$$p_8 = P(Z_8 = 1) = P(\underset{\sim}{Y} = (1,1,1)') = 0.$$

Vale ressaltar todavia, que a aplicação sequencial da *Definição 2.2* e da *Definição 2.1* possue a desvantagem de construir categorias multinomiais com probabilidade de ocorrência igual a zero, dado que as mesmas não representam nenhuma combinação das variáveis categóricas originais, ou seja, passamos a trabalhar com um vetor multinomial que possui um número de categorias maior que o necessário. Ainda com respeito a transformação da *Definição 2.2*, podemos observar que ela não considera a diferença de natureza entre categóricas nominais e ordinais. Uma forma alternativa de resolver este problema, sugerida por Bussab, Miazaki e Andrade (1990) é transformar cada variável categórica nominal ou ordinal em binária, através de transformações existentes para cada tipo. Depois disto, podemos transformar o vetor binário completo gerado em um

multinomial, usando a *Definição 2.1*. Segundo Talkington (1967), para transformar uma variável categórica nominal em um vetor binário, podemos usar a seguinte definição:

Definição 2.3: Sejam $X_1, X_2, \ldots, X_c$ variáveis aleatórias categóricas nominais onde X_i possui k_i categorias, $i = 1, \ldots, c$. Cada componente X_i do vetor $\underset{\sim}{X} = (X_1, X_2, \ldots, X_c)'$ dará origem a k_i variáveis binárias $Y_{j(i)}$ tal que, $l = 1, \ldots, k_i; i = 1, \ldots, c$ e,

$$Y_{l(i)} = \begin{cases} 1, & se\ X_i = l \\ 0, & c.c. \end{cases} \tag{2.2}$$

Já para transformar uma variável categórica ordinal em um vetor binário, uma alternativa é desconsiderar a ordem existente e utilizar ou a transformação da *Definição 2.2* ou a da *Definição 2.3*. Para que esta ordem seja considerada, segundo Sokal e Sneath (1963), devemos utilizar a transformação definida abaixo:

Definição 2.4: Sejam $X_1, X_2, \ldots, X_c$ variáveis aleatórias categóricas ordinais onde X_i possui k_i categorias ordenadas, $i = 1, \ldots, c$. Cada componente X_i do vetor $\underset{\sim}{X} = (X_1, X_2, \ldots, X_c)'$ dará origem a k_i variáveis binárias $Y_{j(i)}$, $j = 1, \ldots, k_i$, $i = 1, \ldots, c$, tal que se $X_i = l_i$,

$$Y_{j(i)} = \begin{cases} 1, & \forall j = 1, \ldots, l_i \\ 0, & \forall j = l_i + 1, \ldots, k_i. \end{cases} \tag{2.3}$$

Para ilustrar a aplicação sequencial das *Definições 2.3* e *2.4* e da *Definição 2.1*, vamos utilizar o seguinte exemplo:

Exemplo 2.2: Consideremos a situação descrita no *Exemplo 2.1*.

Aplicando a *Definição 2.3* na variável categórica nominal X_1, que possui duas categorias, ela será substituída por duas variáveis binárias Y_1 e Y_2, tal que,

$$(Y_1, Y_2) = \begin{cases} (1,0), & se\ X_1 = FHC \\ (0,1), & se\ X_1 = Lula \end{cases}$$

Aplicando a *Definição 2.4* na variável categórica ordinal X_2, que possui três categorias, ela será substituída por três variáveis binárias Y_3, Y_4 e Y_5, tal que,

$$(Y_3, Y_4, Y_5) = \begin{cases} (1,0,0), \ se \ X_2 = Analfab. \\ (1,1,0), \ se \ X_2 = Prim. \ ou \ Secund. \\ (1,1,1), \ se \ X_2 = Universit. \end{cases}$$

Com isso, temos a seguinte relação entre as combinações de $\underset{\sim}{X} = (X_1, X_2)'$ e o vetor binário $\underset{\sim}{Y} = (Y_1, \ldots, Y_5)'$:

(X_1, X_2)	$(Y_1, \ldots, Y_5)$
$(FHC, Analfab.)$	$(1,0,1,0,0)$
$(FHC, Prim.ouSecund.)$	$(1,0,1,1,0)$
$(FHC, Universit.)$	$(1,0,1,1,1)$
$(Lula, Analfab.)$	$(0,1,1,0,0)$
$(Lula, Prim.ouSecund.)$	$(0,1,1,1,0)$
$(Lula, Universit.)$	$(0,1,1,1,1)$

Podemos observar, porém, que vários valores do vetor binário $\underset{\sim}{Y} = (Y_1, \ldots, Y_5)'$ gerado, entre eles $(0, \ldots, 0)$ e $(1, \ldots, 1)$ por exemplo, não correspondem a nenhuma combinação do vetor categórico $\underset{\sim}{X} = (X_1, X_2)'$. Com isso, aplicando a *Definição 2.1* no vetor binário $\underset{\sim}{Y} = (Y_1, \ldots, Y_5)'$, temos que este será transformado em um vetor multinomial $\underset{\sim}{Z}$ com 32 categorias, tal que,

(X_1, X_2)	$(Y_1, \ldots, Y_5)$	m	$(Z_1, \ldots, Z_{32})$
$(FHC, Analfab.)$	$(1,0,1,0,0)$	6	$(0_1, \ldots, 0_5, 1, 0_7, \ldots, 0_{32})$
$(FHC, Prim.ouSecund.)$	$(1,0,1,1,0)$	14	$(0_1, \ldots, 0_{13}, 1, 0_{15}, \ldots, 0_{32})$
$(FHC, Universit.)$	$(1,0,1,1,1)$	30	$(0_1, 0_2, \ldots, 0_{29}, 1, 0_{31}, 0_{32})$
$(Lula, Analfab.)$	$(0,1,1,0,0)$	7	$(0_1, \ldots, 0_6, 1, 0_8, \ldots, 0_{32})$
$(Lula, Prim.ouSecund.)$	$(0,1,1,1,0)$	15	$(0_1, \ldots, 0_{14}, 1, 0_{16}, \ldots, 0_{32})$
$(Lula, Universit.)$	$(0,1,1,1,1)$	31	$(0_1, 0_2, \ldots, 0_{29}, 0_{30}, 1, 0_{32})$

onde, $0_i = 0, \forall i \in \{1, 2, \ldots, 32\}$ e,

$$\underset{\sim}{Z} \sim Multinomial(1, p_1, \ldots, p_{32}). \tag{2.4}$$

e, onde, $p_i \in (0,1), \forall i = 1, \ldots, 32, \sum_{i=1}^{32} p_i = 1$ e, $p_j = 0, \forall j \notin \{6, 7, 14, 15, 30, 31\}$.

Assim como o observado na aplicação sequencial das *Definições 2.2 e 2.1*, a aplicação sequencial das *Definições 2.3, 2.4 e 2.1* também apresenta a desvantagem de construir categorias multinomiais que não representam nenhuma combinação das variáveis categóricas originais.

Krzanowski (1980) também discute uma transformação mais direta, definida abaixo, que na verdade é uma generalização da *Definição 2.1* para o caso de um vetor categórico cujas variáveis possuem mais de duas categorias. Porém, esta transformação não possui a desvantagem de construir categorias multinomiais com probabilidade de ocorrência igual a zero (multinomial com um número de parâmetros maior que o necessário).

Definição 2.5: Sejam $X_1, X_2, \ldots, X_c$ variáveis aleatórias categóricas onde X_i possui k_i categorias, $\forall i = 1, \ldots, c$ tal que o vetor aleatório $\underset{\sim}{X} = (X_1, X_2, \ldots, X_c)'$ pode ser combinado de $k = \prod_{i=1}^{c} k_i$ formas diferentes. Podemos transformar este vetor $\underset{\sim}{X}$ em um vetor multinomial $\underset{\sim}{Z} = (Z_1, \ldots, Z_k)'$ tal que, se a j-ésima combinação de $\underset{\sim}{X}$ for observada, teremos $Z_j = 1$ e $Z_l = 0, \forall l \neq j$ onde $l, j = 1, \ldots, k$.

Para melhor entendimento do efeito desta transformação no estudo individual de cada uma das combinações das variáveis categóricas, consideremos o exemplo a seguir:

Exemplo 2.3: Consideremos também a situação descrita no *Exemplo 2.1*.

Utilizando a *Definição 2.5*, podemos representar $\underset{\sim}{X} = (X_1, X_2)'$ como $\underset{\sim}{Z} = (Z_1, Z_2, \ldots, Z_6)'$ onde:

$$Z_1 = \begin{cases} 1, \ se \ \underset{\sim}{X} = (FHC, \ Analfab.) \\ 0, \ Caso \ Contrário \end{cases} \qquad Z_4 = \begin{cases} 1, \ se \ \underset{\sim}{X} = (Lula, Analfab.) \\ 0, \ Caso \ Contrário \end{cases}$$

$$Z_2 = \begin{cases} 1, \ se \ \underset{\sim}{X} = (FHC, \ Prim. \ ou \ Sec.) \\ 0, \ Caso \ Contrário \end{cases} \qquad Z_5 = \begin{cases} 1, \ se \ \underset{\sim}{X} = (Lula, Prim. \ ou \ Sec.) \\ 0, \ Caso \ Contrário \end{cases}$$

$$Z_3 = \begin{cases} 1, \ se \ \underset{\sim}{X} = (FHC, Universit.) \\ 0, \ Caso \ Contrário \end{cases} \qquad Z_6 = \begin{cases} 1, \ se \ \underset{\sim}{X} = (Lula, Universit.) \\ 0, \ Caso \ Contrário \end{cases}$$

Com isso, temos a seguinte relação entre as combinações de $\underset{\sim}{X} = (X_1, X_2)'$ e o vetor multinomial $\underset{\sim}{Z} = (Z_1, \ldots, Z_6)'$,

(X_1, X_2)	$(Z_1, \ldots, Z_6)$
$(FHC, \ Analfab.)$	$(1,0,0,0,0,0)$
$(FHC, \ Prim. \ ou \ Secund.)$	$(0,1,0,0,0,0)$
$(FHC, Universit.)$	$(0,0,1,0,0,0)$
$(Lula, Analfab.)$	$(0,0,0,1,0,0)$
$(Lula, \ Prim. \ ou \ Secund.)$	$(0,0,0,0,1,0)$
$(Lula, Universit.)$	$(0,0,0,0,0,1)$

e,

$$\underset{\sim}{Z} \sim Multinomial(1, p_1, \ldots, p_6). \tag{2.5}$$

onde,

$$\begin{cases} p_1 = P(\underset{\sim}{X} = (FHC,\ Analfab.)), & p_4 = P(\underset{\sim}{X} = (Lula, Analfab.)) \\ p_2 = P(\underset{\sim}{X} = (FHC,\ Prim.\ ou\ Sec.)), & p_5 = P(\underset{\sim}{X} = (Lula,\ Prim.\ ou\ Sec.)) \\ p_3 = P(\underset{\sim}{X} = (FHC, Universit.)), & p_6 = P(\underset{\sim}{X} = (Lula, Universit.)) \end{cases}$$

tal que, $p_i \in (0,1), \forall i = 1, \ldots, 6$ e $\sum_{i=1}^{6} p_i = 1$.

Podemos observar que ao aplicar a *Definição 2.5* em um vetor categórico, não geramos mais uma multinomial com um número de parâmetros maior que o realmente necessário, o que analogamente acontece quando aplicamos a Definição 2.1 em um vetor originalmente binário. Todavia, tanto a *Definição 2.1* (usada para transformar um vetor binário em multinomial) quanto a *Definição 2.5* (usada para transformar um vetor categórico em multinomial) geram um vetor multinomial $\underset{\sim}{Z}$ completo que, por sua vez, possui a incoveniente propriedade de dependência linear entre as categorias Z_i 's (ver *Proposição B.1,a*). Para solucionar este problema, para o caso de um vetor binário, Vlachonikolis e Marriott (1982) utilizam a seguinte transformação:

Definição 2.6: Seja $\underset{\sim}{X} = (X_1, X_2, \ldots, X_q)'$ um vetor aleatório binário. O vetor $\underset{\sim}{X}$ pode ser substituído por $k = 2^q - 1$ variáveis indicadoras $Z_1, \ldots, Z_k$, de modo que a combinação do vetor $\underset{\sim}{X}$ define a ocorrência de uma variável indicadora Z_i de forma única se $m = \sum_{i=1}^{q} x_i 2^{(i-1)} \in \{1, 2, \ldots, k\}$, onde $z_m = 1$ e $z_j = 0$, $\forall j \neq m$; $j, m = 1, 2, \ldots, k$. Se $m = 0$, então $z_i = 0$, $\forall i = 1, 2, \ldots, k$.

Podemos observar que esta transformação retira uma componente do vetor multinomial $\underset{\sim}{Z}$ completo para evitar dependência linear entre as categorias Z_i 's geradas. Isto corresponde a retirar o parâmetro do modelo multinomial completo, que está associado a uma determinada combinação das variáveis categóricas. Para o caso em que $\underset{\sim}{X}$ é um vetor categórico, vamos fazer uma extensão da *Definição 2.6*:

Definição 2.7: Sejam $X_1, X_2, \ldots, X_c$ variáveis aleatórias categóricas onde X_i possui k_i categorias, $\forall i = 1, \ldots, c$ tal que o vetor $\underset{\sim}{X} = (X_1, X_2, \ldots, X_c)'$ pode ser combinado

de $k = \prod_{i=1}^{c} k_i$ formas diferentes. Vamos omitir uma dessas k combinações e transformar este vetor $\underset{\sim}{X}$ no vetor $(Z_1, \ldots, Z_r)'$ onde $r = k - 1$, tal que se a j-ésima combinação de $\underset{\sim}{X}$ for observada, $Z_j = 1$ e $Z_l = 0, \forall l \neq j$ onde $l, j = 1, \ldots, r$ e, se a k-ésima combinação de $\underset{\sim}{X}$ for observada, então $Z_j = 0, \forall j = 1, \ldots, r$.

Para melhor entendimento do efeito desta transformação no estudo individual de cada uma das combinações das variáveis categóricas, consideremos o exemplo a seguir:

Exemplo 2.4: Para tornar mais fácil a comparação com as outras transformações, vamos considerar também a situação descrita no exemplo 2.1.

Utilizando a *Definição 2.7*, podemos representar $\underset{\sim}{X} = (X_1, X_2)'$ como $\underset{\sim}{Z} = (Z_1, Z_2, \ldots, Z_5)'$ onde:

$$Z_1 = \begin{cases} 1, \ se \ \underset{\sim}{X} = (FHC, \ Analfab.) \\ 0, \ Caso \ Contrário \end{cases} \qquad Z_4 = \begin{cases} 1, \ se \ \underset{\sim}{X} = (Lula, Analfab.) \\ 0, \ Caso \ Contrário \end{cases}$$

$$Z_2 = \begin{cases} 1, \ se \ \underset{\sim}{X} = (FHC, \ Prim. \ ou \ Sec.) \\ 0, \ Caso \ Contrário \end{cases} \qquad Z_5 = \begin{cases} 1, \ se \ \underset{\sim}{X} = (Lula, Prim. \ ou \ Sec.) \\ 0, \ Caso \ Contrário \end{cases}$$

$$Z_3 = \begin{cases} 1, \ se \ \underset{\sim}{X} = (FHC, Universit.) \\ 0, \ Caso \ Contrário \end{cases}$$

Com isso, temos a seguinte relação entre as combinações de $\underset{\sim}{X} = (X_1, X_2)'$ e o vetor multinomial $\underset{\sim}{Z} = (Z_1, \ldots, Z_5)'$:

(X_1, X_2)	$(Z_1, \ldots, Z_5)$
$(FHC, \ Analfab.)$	$(1, 0, 0, 0, 0)$
$(FHC, \ Prim. \ ou \ Secund.)$	$(0, 1, 0, 0, 0)$
$(FHC, Universit.)$	$(0, 0, 1, 0, 0)$
$(Lula, Analfab.)$	$(0, 0, 0, 1, 0)$
$(Lula, \ Prim. \ ou \ Secund.)$	$(0, 0, 0, 0, 1)$

onde,

$$\begin{cases} p_1 = P(\underset{\sim}{X} = (FHC,\ Analfab.)), & p_4 = P(\underset{\sim}{X} = (Lula, Analfab.)) \\ p_2 = P(\underset{\sim}{X} = (FHC,\ Prim.\ ou\ Sec.)), & p_5 = P(\underset{\sim}{X} = (Lula,\ Prim.\ ou\ Sec.)) \\ p_3 = P(\underset{\sim}{X} = (FHC, Universit.)), & 1 - \sum_{i=1}^{5} p_i = P(\underset{\sim}{X} = (Lula, Universit.)) \end{cases}$$

tal que, $p_i \in (0,1), \forall i = 1, \ldots, 5$.

Vale salientar que $(Z_1, \ldots, Z_r)'$ é um vetor multinomial completo sem uma de suas categorias. Categoria esta que foi retirada para evitar a dependência linear entre as categorias Z_i's. A grande vantagem de utilizarmos a transformação da *Definição 2.7* é que a matriz de variâncias e covariâncias do vetor multinomial gerado passa a ser não singular, o que não acontece se utilizarmos o vetor multinomial completo.

2.3 A Matriz de Variâncias e Covariâncias sob o Modelo de Posição

Antes de determinarmos a expressão da matriz de variâncias e covariâncias para o caso em que a estrutura de inter-relação entre as variáveis contínuas varia de acordo com a categoria multinomial observada, vamos definir os modelos e apresentar as expressões encontradas por Tate (1954) e Olkin e Tate (1961) para o caso em que a variação é apenas do vetor de médias das contínuas. Vale salientar que eles apresentam expressões particionadas para a matriz de variâncias e covariâncias, o que permite interpretar os blocos correspondentes as variáveis categóricas e contínuas separadamente.

2.3.1 A Definição dos Modelos

O *modelo de posição* proposto por Olkin e Tate (1961), é uma extensão multivariada do *modelo bisserial pontual* proposto por Tate (1954). O *modelo bisserial pontual* especifica a distribuição de um vetor aleatório misto $\underset{\sim}{W} = (X, Y)$ composto por uma variável aleatória

X binária e uma variável aleatória Y contínua, como sendo o produto da distribuição marginal de X pela distribuição condicional de Y dado $X = x$ fixo, ou seja.

$$f(w) = f(x, y) = f(x)f(y \backslash x). \tag{2.6}$$

onde,

$$X \sim binomial(1, p), \text{ e } p_0 = 1 - p_1. \tag{2.7}$$

e,

$$Y \backslash X = x \sim normal(\mu^{(x)}, \sigma^2), \ x \in \{0.1\}. \tag{2.8}$$

Já o modelo de posição especifica a distribuição de um vetor aleatório $\underset{\sim}{W} = (\underset{\sim}{X}', \underset{\sim}{Y}')'$, composto por um vetor multinomial $\underset{\sim}{X} = (X_1, X_2, \ldots, X_k)$ com k categorias, e por um vetor contínuo $\underset{\sim}{Y} = (Y_1, Y_2, \ldots, Y_p)$, como sendo o produto da distribuição condicional de $\underset{\sim}{Y}$ dado a categoria multinomial x_m observada ($x_m = 1$), pela distribuição marginal de X_m, ou seja,

$$f(\underset{\sim}{w}) = f(x_m = 1, \underset{\sim}{y}) = f(x_m = 1)f(\underset{\sim}{y} \backslash x_m = 1), \tag{2.9}$$

onde,

$$X_m \sim binomial(1, p_m), \ com \ \sum_{m=1}^{k} p_m = 1. \tag{2.10}$$

e,

$$\underset{\sim}{Y} \backslash x_m = 1 \sim normal(\mu^{(m)}, \sigma^2). \tag{2.11}$$

2.3.2 A Expressão da Matriz de Variâncias e Covariâncias

Vamos apresentar agora a expressão da matriz de variâncias e covariâncias populacionais, determinadas por Tate (1954) e Olkin e Tate (1961), para o *modelo bisserial pontual* e para o *modelo de posição*, respectivamente.

Modelo bisserial pontual

Para este modelo, segundo Tate (1954), temos que:

$$\underset{\sim}{\Sigma_W} = \begin{pmatrix} \Psi & \Delta \\ \Delta' & \Gamma \end{pmatrix}, \tag{2.12}$$

onde,

$$\Psi = p_1(1 - p_1), \tag{2.13}$$

$$\Delta = p_i(\mu^{(i)} - \bar{\mu}), i = 0, 1 \text{ e } p_0 = 1 - p_1, \tag{2.14}$$

$$\Gamma = \sigma + \sum_{m=0}^{1} (\mu^{(m)} - \bar{\mu})^2 p_m. \tag{2.15}$$

Modelo de Posição

Para este modelo, segundo Olkin e Tate (1961), temos que:

$$\underset{\sim}{\Sigma_W} = \begin{pmatrix} \Psi & \Delta \\ \Delta' & \Gamma \end{pmatrix}, \tag{2.16}$$

onde, de forma algébrica, temos,

$$\Psi = (\psi_{ij}) = \begin{cases} p_i(1 - p_i), & i = j \\ -p_i p_j, & i \neq j \end{cases} \quad i, j = 1, \ldots, k. \tag{2.17}$$

$$\Delta = (\delta_{ij}) = p_i(\mu_j^{(i)} - \bar{\mu}_j), i = 1, \ldots, k \text{ e } j = 1, \ldots, p \,, \tag{2.18}$$

$$\Gamma = (\gamma_{ij}) = \sigma_{ij} + \sum_{m=1}^{k} (\mu_i^{(m)} - \bar{\mu}_i)(\mu_j^{(m)} - \bar{\mu}_j)p_m; \; i, j = 1, \ldots, p. \tag{2.19}$$

Se considerarmos,

$$U = (u_{ij}) = (\mu_j^{(i)} - \bar{\mu}_j), i = 1, \ldots, k \text{ e } j = 1, 2, \ldots, p, \tag{2.20}$$

$$D = diag(p), \text{ onde } p = (p_1, \ldots, p_k), \tag{2.21}$$

e,

$$\Upsilon = (\sigma_{ij}), \; i, j = 1, \ldots, p, \tag{2.22}$$

podemos reescrever as expressões (2.17) à (2.19) de forma matricial, ou seja,

$$\Psi = D - p'p, \tag{2.23}$$

$$\Delta = DU, \tag{2.24}$$

$$\Gamma = \Upsilon + U'DU. \tag{2.25}$$

2.4 A Determinação da Matriz de Variâncias e Covariâncias para um Vetor Misto

2.4.1 Considerações Iniciais

Como podemos observar nas equações (2.8) e (2.11), os modelos *bisserial pontual* e de *posição* possuem a pré-suposição de que a estrutura de inter-relação entre as contínuas permanece constante, independentemente da categoria multinomial observada. Já um *modelo de posição mais geral*, sugerido por Krzanowski (1983), admite que estas inter-relações podem variar de acordo com a categoria multinomial observada, ou seja,

$$\underset{\sim}{X} \sim multinomial(1, p_1, p_2, \ldots, p_c), \tag{2.26}$$

$$\underset{\sim}{Y} \setminus x_m = 1 \sim normal(\mu^{(m)}, \sigma^2_{(m)}). \tag{2.27}$$

Porém, Krzanowski não apresenta a expressão da matriz de variâncias e covariâncias para esta generalização do modelo de posição. Complementando a proposta de Krzanowski (1983), León (1993) determina as expressões do vetor de médias e da matriz de variâncias e covariâncias para esta situação.

Mesmo não sendo necessário supor normalidade para determinarmos a matriz de variâncias e covariâncias para um vetor misto, apesar de haver ganhos quando a mesma é verificada, a associação de categorias multinomiais às combinações das categóricas (princípio adotado por Olkin e Tate, 1961) facilita em muito a obtenção dos momentos para o caso de misturas. Faremos primeiro a determinação destas covariâncias para o caso populacional para, em seguida, obter os seus estimadores usuais de momentos. As expressões encontradas para estes dois casos serão apresentadas tanto na forma algébrica quanto na forma matricial, com o objetivo de facilitar a interpretação e a implementação computacional das mesmas. Vale salientar que quando se trabalha com um vetor misto de variáveis categóricas e contínuas esta determinação não é tão trivial, visto que as

categorias podem caracterizar sub-populações que precisam ser consideradas quando da obtenção dos momentos.

2.4.2 Determinação para o Caso Populacional

Vamos assumir que $\underset{\sim}{X} = (X_1, \ldots, X_c)'$ é um vetor categórico que pode ser combinado de $k = \prod_{i=1}^{c} k_i$ formas diferentes, dado que X_i possui k_i categorias, $i = 1, \ldots, c$ e $\underset{\sim}{Y} = (Y_1, \ldots, Y_p)'$ é um vetor aleatório contínuo. De acordo com a *Definição B.2*, o vetor misto $\underset{\sim}{W}$ é denotado por:

$$\underset{\sim}{W} = (\underset{\sim}{X}', \underset{\sim}{Y}')'. \tag{2.28}$$

Para podermos estudar o comportamento das variáveis contínuas $Y_1, \ldots, Y_p$ com relação aos resultados observados nas variáveis categóricas, $X_1, \ldots, X_c$, vamos transformar estas últimas em um vetor multinomial como discutido anteriormente. Porém, para evitar dependência linear entre as categorias multinomiais Z_i's geradas (ver *Proposição B.1.a*), vamos gerar uma multinomial com uma categoria a menos utilizando, para isto, a transformação dada na *Definição 2.7*. Evitar esta dependência linear é de grande importância, pois assim evitamos também a singularidade da matriz de variâncias e covariâncias do vetor multinomial (ver *Proposição B.1.c*). Após a aplicação da *Definição 2.7* no vetor $\underset{\sim}{X}$ de (2.28), passamos a trabalhar com o vetor:

$$\underset{\sim}{W}^{*} = (\underset{\sim}{Z}', \underset{\sim}{Y}')', \tag{2.29}$$

onde, $\underset{\sim}{Z} = (Z_1, \ldots, Z_r)'$, tal que $r = k - 1$, $Z_i \in \{0, 1\}$, $\forall i = 1, \ldots, r$.

Como a transformação utilizada gera um vetor multinomial $\underset{\sim}{Z}$ com uma categoria a menos, para que $\underset{\sim}{Z}$ torne-se um vetor multinomial completo basta considerar a variável $Z_0 = 1 - \sum_{i=1}^{r} Z_i$ faltante. Porém, como não nos interessa a singularidade da matriz de variâncias e covariâncias do vetor multinomial, é imprescindível trabalhar com uma multinomial com uma categoria a menos.

Determinação Algébrica

Para simplificar as determinações que virão a seguir, vamos considerar as seguintes equivalências de eventos:

$$[\underset{\sim}{Z} = (Z_1, \ldots, Z_r) : Z_i = 0, \forall i = 1, \ldots, r] = [Z_0 = 1], \tag{2.30}$$

e,

$$[\underset{\sim}{Z} = (Z_1, \ldots, Z_{i-1}, Z_i, Z_{i+1}, \ldots, Z_r) : Z_i = 1 \text{ e } Z_j = 0, \forall j \neq i; j = 1, \ldots, r] = [Z_i = 1]. \tag{2.31}$$

Denotaremos a probabilidade da i-ésima categoria multinomial ser observada, $\forall i = 1, \ldots, r$, por,

$$p_i = P(Z_i = 1) = P(\underset{\sim}{Z} = (Z_1, \ldots, Z_{i-1}, Z_i, Z_{i+1}, \ldots, Z_r) : Z_i = 1 \text{ e } Z_j = 0, \forall j \neq i), \tag{2.32}$$

e, a probabilidade de ocorrência da categoria que foi retirada do vetor multinomial, por,

$$p_0 = P(Z_0 = 1) = 1 - P(\bigcup_{i=1}^{r} [Z_i = 1]) = 1 - \sum_{i=1}^{r} p_i. \tag{2.33}$$

Agora, passaremos à determinação da matriz de variâncias e covariâncias que é obtida em função dos momentos de 1^a ordem não centrados e do produto cruzado com relação as variáveis envolvidas.

Vamos denotar os momentos de 1^a ordem não centrados das variáveis contínuas, condicionados a categoria multinomial observada, como:

$$\mu_i^{(m)} = E(Y_i \backslash Z_m = 1), \ m = 0, \ldots, r \ e \ i = 1, \ldots, p. \tag{2.34}$$

Também para facilitar a notação, vamos denotar:

$$\tilde{\mu}_i = \sum_{m=0}^{r} \mu_i^{(m)} p_m = \sum_{m=0}^{r} E(Y_i \backslash Z_m = 1) P(Z_m = 1). \tag{2.35}$$

E ainda, denotar as covariâncias entre as variáveis contínuas, condicionadas a categoria multinomial observada, como:

$$\sigma_{ij}^{(m)} = Cov(Y_i, Y_j \backslash Z_m = 1),\ m = 0, \ldots, r \in i, j = 1, \ldots, p. \tag{2.36}$$

Para tornar mais fácil a obtenção da matríz de variâncias e covariâncias do vetor populacional $\underset{\sim}{W}^*$, denotada por $\underset{\sim}{\Sigma}_{\underset{\sim}{W}^*}$, vamos particioná-la de forma idêntica a feita por Olkin e Tate (1961), dado em (2.16), ou seja,

$$\underset{\sim}{\Sigma}_{\underset{\sim}{W}^*} = \begin{pmatrix} \Psi & \Delta \\ \Delta' & \Gamma \end{pmatrix}, \tag{2.37}$$

onde,

$$\Psi = (\psi_{ij}) = (Cov(Z_i, Z_j)), i, j = 1, \ldots, r, \tag{2.38}$$

$$\Delta = (\delta_{ij}) = (Cov(Z_i, Y_j)), i = 1, \ldots, r, \text{ e } j = 1, \ldots, p. \tag{2.39}$$

$$\Gamma = (\sigma_{ij}) = (Cov(Y_i, Y_j)), i, j = 1, \ldots, p, \tag{2.40}$$

tal que,

$$\begin{cases} \psi_{ij} : \text{É a covariância entre as categorias multinomiais } Z_i \text{ e } Z_j, \\ \delta_{ij} : \text{É a covariância entre a variável contínua } Y_i \text{ e a categoria multinomial } Z_j, \\ \sigma_{ij} : \text{É a covariância entre as variáveis contínuas } Y_i \text{ e } Y_j. \end{cases}$$

Inicialmente, vamos obter os momentos de 1^a ordem não centrados das variáveis

contínuas e das categorias Z_i's do vetor $\underset{\sim}{Z}$. Como $\underset{\sim}{Z}$ é um vetor multinomial, a distribuição marginal das categorias Z_i's é *binomial*$(1, p_i)$, ver *Proposição B.1.b*, e,

$$EZ_i = p_i, \ \forall i = 1, \ldots, r \ . \tag{2.41}$$

Com relação as variáveis contínuas, considerando todos os resultados possíveis do vetor $\underset{\sim}{Z}$ temos, pela *Proposição B.2*, que:

$$EY_i \ = \ E\{E(Y_i \backslash \ \underset{\sim}{Z})\} \ = \ \sum_{m=0}^{r} E(Y_i \backslash Z_m = 1) P(Z_m = 1), \tag{2.42}$$

e, usando (2.35), temos:

$$EY_i \ = \ \sum_{m=0}^{r} E(Y_i \backslash Z_m = 1) P(Z_m = 1) \ = \ \sum_{m=0}^{r} \mu_i^{(m)} p_m \ = \ \bar{\mu}_i \ . \tag{2.43}$$

Agora, para que possamos determinar completamente (2.37) precisamos obter os momentos conjuntos de 1^a ordem.

Com respeito as categorias Z_i's do vetor multinomial $\underset{\sim}{Z}$ temos por (2.32) que:

$$EZ_i Z_j = \begin{cases} 0, & se \ i \neq j \\ P(Z_i = 1) = p_i, & se \ i = j \end{cases} \quad \forall i, j = 1, \ldots, r \ . \tag{2.44}$$

Com isso, substituindo (2.41) e (2.44) em (2.38), temos que a covariância populacional entre as categorias Z_i e Z_j do vetor multinomial $\underset{\sim}{Z}, \forall i, j = 1, \ldots, r$, é dada por:

$$\psi_{ij} = EZ_i Z_j - EZ_i EZ_j = \begin{cases} -p_i p_j, & se \ i \neq j \\ p_i(1 - p_j), & se \ i = j \end{cases} \tag{2.45}$$

Com respeito aos momentos conjuntos entre as variáveis contínuas e as categorias multinomiais, usando a *Proposição B.2*, (2.32) e (2.34), temos que:

$$\begin{aligned} EZ_iY_j &= E\{E(Z_iY_j \backslash \underset{\sim}{Z})\} = \sum_{m=0}^{r} E(Z_iY_j \backslash Z_m = 1)P(Z_m = 1) \\ &= E(Y_j \backslash Z_i = 1)p_i = \mu_j^{(i)} p_i, \ \forall i = 1, \ldots, r \text{ e } j = 1, \ldots, p \ . \end{aligned} \tag{2.46}$$

Dessa forma, substituindo (2.41), (2.43) e (2.46) em (2.39) temos,

$$\begin{aligned} \delta_{ij} &= EZ_iY_j - EZ_iEY_j = \mu_j^{(i)} p_i - p_i \, \tilde{\mu}_j \\ &= p_i(\mu_j^{(i)} - \tilde{\mu}_j), \qquad \forall i = 1, \ldots, r \text{ e } j = 1, \ldots, p \ . \end{aligned} \tag{2.47}$$

Podemos observar que a covariância entre uma variável contínua Y_j e uma categoria multinomial Z_i é expressa como a diferença entre a esperança condicional de Y_j, dado que a i-ésima categoria multinomial foi observada, e a esperança de Y_j, ponderada pela probabilidade da i-ésima categoria multinomial ser observada.

Com respeito aos momentos conjuntos entre as variáveis contínuas Y_j's, considerando a presença de *misturas* e usando a *Proposição B.2* e (2.32). temos que:

$$\begin{aligned} EY_iY_j &= E\{E(Y_iY_j \backslash \underset{\sim}{Z})\} \\ &= \sum_{m=0}^{r} E(Y_iY_j \backslash Z_m = 1)p_m, \ \forall i, j = 1, \ldots, p \ . \end{aligned} \tag{2.48}$$

Utilizando a *Definição B.5*, podemos escrever os momentos conjuntos das variáveis contínuas Y_j's, condicionados a categoria multinomial observada, como:

$$E(Y_iY_j \backslash Z_m = 1) = Cov(Y_i, Y_j \backslash Z_m = 1) + E(Y_i \backslash Z_m = 1)E(Y_j \backslash Z_m = 1) \ . \tag{2.49}$$

Com isso, usando (2.49), (2.34) e (2.36) em (2.48), temos que:

$$
\begin{aligned}
EY_iY_j &= \sum_{m=0}^{r} E(Y_iY_j \backslash Z_m = 1)p_m \\
&= \sum_{m=0}^{r} [Cov(Y_i, Y_j \backslash Z_m = 1) + E(Y_i \backslash Z_m = 1)E(Y_j \backslash Z_m = 1)]p_m \qquad (2.50) \\
&= \sum_{m=0}^{r} (\sigma_{ij}^{(m)} + \mu_i^{(m)}\mu_j^{(m)})p_m \ ,
\end{aligned}
$$

dessa forma, substituindo (2.43) e (2.50) em (2.40), temos que:

$$
\begin{aligned}
\gamma_{ij} &= EY_iY_j - EY_iEY_j \\
&= \sum_{m=0}^{r} (\sigma_{ij}^{(m)} + \mu_i^{(m)}\mu_j^{(m)})p_m - \bar{\mu}_i\bar{\mu}_j \\
&= \sum_{m=0}^{r} \sigma_{ij}^{(m)}p_m + \sum_{m=0}^{r} \mu_i^{(m)}\mu_j^{(m)}p_m - \bar{\mu}_i\bar{\mu}_j \qquad (2.51) \\
&= \sum_{m=0}^{r} \sigma_{ij}^{(m)}p_m + \sum_{m=0}^{r} (\mu_i^{(m)} - \bar{\mu}_i)(\mu_j^{(m)} - \bar{\mu}_j)p_m \ .
\end{aligned}
$$

Podemos observar que as variâncias e covariâncias de cada variável contínua ficam expressas por duas parcelas: a primeira, é uma média ponderada das covariâncias *dentro* das categorias multinomiais e, a segunda, é uma média ponderada das covariâncias *entre* categorias, já que pondera produtos cruzados dos desvios das médias por categoria, em relação a média geral.

Com isso, voltando a (2.37), temos que a matriz de variâncias e covariâncias populacionais para o vetor misto $\underset{\sim}{W}^* = (\underset{\sim}{Z}', \underset{\sim}{Y}')'$ é dada por:

$$
\underset{\sim}{\Sigma}_{W^*} = \begin{pmatrix} \Psi & \Delta \\ \Delta' & \Gamma \end{pmatrix},
$$

onde, por (2.45), (2.47) e (2.51).

$$
\Psi = (\psi_{ij}) = \begin{cases} -p_ip_j, \ se \ i \neq j \\ p_i(1-p_i), \ se \ i = j \end{cases} \quad i, j = 1, \ldots, r, \qquad (2.52)
$$

$$
\Delta = (\delta_{ij}) = p_i(\mu_j^{(i)} - \bar{\mu}_j), \ i = 1, \ldots, r \ \text{e} \ j = 1, \ldots, p, \qquad (2.53)
$$

$$\Gamma = (\gamma_{ij}) = \sum_{m=0}^{r} \sigma_{ij}^{(m)} p_m \; + \; \sum_{m=0}^{r} (\mu_i^{(m)} - \; \bar{\mu}_i)(\mu_j^{(m)} - \; \bar{\mu}_j) p_m \; . \qquad (2.54)$$

Comparando as expressões (2.52), (2.53) e (2.54) com as obtidas por Olkin e Tate (1961) para o modelo de posição, (2.17), (2.18) e (2.19) respectivamente, podemos observar que aqui além do vetor $\underset{\sim}{Z}$ possuir uma categoria a menos, $i = 1, \ldots, r$ $(r = k - 1)$ enquanto no modelo de posição temos $i = 1, \ldots, k$, a única diferença está na 1^a parcela de Γ, ou seja, aqui temos que σ_{ij} varia de acordo com a categoria de $\underset{\sim}{Z}$ que foi observada, enquanto no modelo de posição σ_{ij} permanece constante. Neste caso fica caracterizado que a diferença de estrutura de sub-populações devido à presença das categorias Z_i's é tanto de locação como de escala.

Examinemos, agora, algumas das expressões derivadas acima:

i) Δ é um bloco no qual há interesse em examinar as contribuições dos fatores às covariâncias resultantes. De fato, cada elemento δ_{ij} resulta da magnitude da diferença entre a média da variável Y_j, no grupo gerado pela categoria Z_i, e a média geral, ponderada pelo tamanho relativo deste grupo, p_i. Percorrendo uma coluna desta matriz, e isolados seus fatores, as magnitudes das diferenças observadas entre a média de uma dada variável, ao longo das categorias multinomiais, podem ser avaliadas tendo em conta o tamanho destas. Percorrendo cada linha, e como é mantida a proporção, aparece a contribuição das p variáveis ao longo das categorias.

ii) O exame das parcelas de Γ ajudará, certamente, a avaliar as diferenças entre grupos, para cada variável, relativisando tanto quanto à variância total γ_{ij}, como quanto à contribuição da variação dentro dos grupos, vinda da primeira parcela. Por estes motivos, é importante obter os resultados das duas parcelas que compõe os γ_{ij}, de modo a não perder informação das contribuições *dentro* e *entre* categorias.

Determinação Matricial

Seguindo uma notação análoga a utilizada por Olkin e Tate (1961), e expressa nas equações (2.20) a (2.22) consideremos,

$$U^* = (u_{ij}) = (\mu_j^{(i)} - \bar{\mu}_j), i = 1, \ldots, r,\ r = k-1 \text{ e } j = 1, 2, \ldots, p, \tag{2.55}$$

$$U^\bullet = (u_{ij}) = (\mu_j^{(i)} - \bar{\mu}_j), i = 0, \ldots, r,\ r = k-1 \text{ e } j = 1, 2, \ldots, p. \tag{2.56}$$

$$D^* = diag(p^*),\ \text{onde } p^* = (p_1, \ldots, p_r),\ r = k-1, \tag{2.57}$$

$$D^\bullet = diag(p^\bullet),\ \text{onde } p^\bullet = (p_0, \ldots, p_r),\ r = k-1, \tag{2.58}$$

$$P^* = [p_0 I_p | p_1 I_p | \ldots | p_r I_p]_{p \times [(r+1)\times p]}. \tag{2.59}$$

$$\Upsilon^* = \begin{bmatrix} \Upsilon^{(0)} \\ \Upsilon^{(1)} \\ \vdots \\ \Upsilon^{(r)} \end{bmatrix}_{[(r+1)\times p]\times p}, \tag{2.60}$$

onde $\Upsilon^{(m)} = (\sigma_{ij}^{(m)}), i, j = 1, \ldots, p; m = 0, \ldots, r$, e,

$$\Upsilon^\bullet = P^* \Upsilon^*. \tag{2.61}$$

Com isso, voltando a (2.37), temos que a matriz de variância-covariâncias populacionais para o vetor misto $\underset{\sim}{W}^* = (\underset{\sim}{Z}', \underset{\sim}{Y}')'$ é dada por:

$$\underset{\sim}{\Sigma}_{W^*} = \begin{pmatrix} \Psi & \Delta \\ \Delta' & \Gamma \end{pmatrix},$$

onde podemos reescrever as expressões (2.52) à (2.54), na forma matricial, como:

$$\Psi = D^* - p^{*\prime} p^*, \tag{2.62}$$

$$\Delta = D^* U^*, \tag{2.63}$$

$$\Gamma = \Upsilon^* + U^{*\prime} D^* U^*. \tag{2.64}$$

2.4.3 Determinação para o Caso Amostral

Determinação Algébrica

Feita a determinação das covariâncias populacionais, vamos agora, determinar os estimadores dessas covariâncias utilizando os momentos amostras.

Tomando uma amostra aleatória de tamanho n do vetor populacional misto $\underset{\sim}{W}^* = (\underset{\sim}{Z}', \underset{\sim}{Y}')'$, vamos obter os estimadores para a matriz de variâncias e covariâncias populacionais dada em (2.37). Para tanto, vamos seguir os mesmos passos da seção anterior, ou seja, vamos estimar as covariâncias a partir dos momentos de ordem 1 não centrais e dos primeiros momentos conjuntos entre as variáveis envolvidas.

Com respeito aos momentos populacionais de primeira ordem não centrados, temos de (2.41) que,

$$EZ_i = p_i, \ \forall i = 1, \ldots, r \ ,$$

assim, usando a *Definição B.8* em função do primeiro momento amostral, o estimador de p_i é dado por,

$$\hat{E}\hat{Z}_i \;=\; \hat{p}_i \;=\frac{\sum_{i=1}^{n} z_i}{n}=\frac{n_i}{n},\ \forall i=1,\ldots,r\ . \tag{2.65}$$

De (2.43), temos que,

$$EY_i = \tilde{\mu}_i = \sum_{m=0}^{r} \mu_i^{(m)} p_m,\ \forall i=1,\ldots,p\ ,$$

agora, $\mu_i^{(m)}$ é um momento populacional da variável Y_i restrito à categoria multinomial m. Por isso, o seu estimador é dado por,

$$\hat{\mu}_i^{(m)} = \frac{\sum_{j=1}^{n_m} y_j^{(m)}}{n_m} = \bar{y}_i^{(m)}, i=1,\ldots,p. \tag{2.66}$$

Como EY_i está escrita em função de $\mu_i^{(m)}$ e p_m, usando a *Definição B.7* e os resultados (2.65) e (2.66), temos que o seu estimador é,

$$\hat{E}\hat{Y}_i = \hat{\tilde{\mu}}_i = \sum_{m=0}^{r} \hat{\mu}_i^{(m)} \hat{p}_m = \frac{\sum_{m=0}^{r} \bar{y}_i^{(m)}\, n_m}{n} = \bar{y}_i, i=1,\ldots,p. \tag{2.67}$$

De (2.45) temos que,

$$\psi_{ij} = \begin{cases} -p_i p_j, & se\ i \neq j \\ p_i(1-p_i), & se\ i=j \end{cases} \forall i,j=1,\ldots,r\ .$$

Com isso, usando a *Definição B.7* e o resultado (2.65), temos que o estimador da covariância populacional entre as categorias Z_i e Z_j do vetor $\underset{\sim}{Z}$ é dado por:

$$\hat{\psi}_{ij} = \begin{cases} -\hat{p}_i\hat{p}_j \;=\; -\frac{n_i n_j}{n^2}, & se\ i \neq j \\ \hat{p}_i\,(1-\hat{p}_i) \;=\; \frac{n_i}{n}(1-\frac{n_i}{n}), & se\ i=j\ . \end{cases} \tag{2.68}$$

De (2.47), temos que,

$$\delta_{ij} = p_i(\mu_j^{(i)} - \tilde{\mu}_j),\ i=1,\ldots,r \text{ e } j=1,\ldots,p,$$

com isso, usando a *Definição B.7* e os resultados (2.65) e (2.66) e (2.67), temos que o estimador da covariância populacional entre a categoria Z_i do vetor $\underset{\sim}{Z}$ e a variável contínua Y_j é dado por:

$$\hat{\delta}_{ij} = \hat{p}_i\,(\hat{\mu}_j^{(i)} - \hat{\bar{\mu}}_j) = \frac{n_i}{n}(\bar{y}_j^{(i)} - \bar{y}_j),\ i = 1, \ldots, r \text{ e } j = 1, \ldots, p. \tag{2.69}$$

De (2.51), temos que:

$$\gamma_{ij} = \sum_{m=0}^{r} \sigma_{ij}^{(m)} p_m + \sum_{m=0}^{r} (\mu_i^{(m)} - \bar{\mu}_i)(\mu_j^{(m)} - \bar{\mu}_j) p_m, \forall i, j = 1, \ldots, p.$$

Com isso, usando a *Definição B.8* e os resultados (2.66),(2.67) e (2.65), temos que o estimador da covariância populacional entre as variáveis contínuas Y_i e Y_j é dado por:

$$\begin{aligned} \hat{\gamma}_{ij} &= \sum_{m=0}^{r} \hat{\sigma}_{ij}^{(m)} \hat{p}_m + \sum_{m=0}^{r} (\hat{\mu}_i^{(m)} - \hat{\bar{\mu}}_i)(\hat{\mu}_j^{(m)} - \hat{\bar{\mu}}_j)\,\hat{p}_m \\ &= \sum_{m=0}^{r} \hat{\sigma}_{ij}^{(m)} \frac{n_m}{n} + \sum_{m=0}^{r} (\hat{\mu}_i^{(m)} - \hat{\bar{\mu}}_i)(\hat{\mu}_j^{(m)} - \hat{\bar{\mu}}_j)\frac{n_m}{n}\ , \end{aligned} \tag{2.70}$$

onde $\hat{\sigma}_{ij}^{(m)}$ representa o estimador da covariância populacional entre Y_i e Y_j *dentro* da categoria m, sendo expresso como,

$$\hat{\sigma}_{ij}^{(m)} = \frac{1}{n} \sum_{l=0}^{n_m} (y_{li}^{(m)} - \bar{y}_i^{(m)})(y_{lj}^{(m)} - \bar{y}_j^{(m)}). \tag{2.71}$$

Podemos observar que o estimador $\hat{\gamma}_{ij}$ da covariância populacional entre as variáveis contínuas Y_i e Y_j pode ser interpretado como uma combinação de duas parcelas: A primeira corresponde ao estimador da covariância *dentro* e a segunda ao estimador da covariância *entre* categorias multinomiais.

De (2.37), temos que a matriz de variâncias e covariâncias populacionais Σ_{W^*} é dada por:

$$\underset{\sim}{\Sigma}_{W^*} = \begin{pmatrix} \Psi & \Delta \\ \Delta' & \Gamma \end{pmatrix},$$

utilizando os resultados (2.68), (2.69) e (2.70), temos que o estimador de Σ_{W^*}, que denotaremos por $\mathbf{S}_D$, é dado por,

$$\mathbf{S}_D = \begin{pmatrix} \hat{\Psi} & \hat{\Delta} \\ \hat{\Delta}' & \hat{\Gamma} \end{pmatrix}, \tag{2.72}$$

onde,

$$\hat{\Psi} = (\hat{\psi}_{ij}) = \begin{cases} -\hat{p}_i\hat{p}_j = -\frac{n_i n_j}{n^2}, & se\ i \neq j \\ \hat{p}_i\,(1-\hat{p}_i) = \frac{n_i}{n}(1-\frac{n_i}{n}), & se\ i = j\ . \end{cases} , ij = 1, \ldots, pp, \tag{2.73}$$

$$\hat{\Delta} = (\hat{\delta}_{ij}) = \hat{p}_i\,(\hat{\mu}_j^{(i)} - \hat{\bar{\mu}}_j) = \frac{n_i}{n}(\bar{y}_j^{(i)} - \bar{y}_j),\ i = 1, \ldots, r \text{ e } j = 1, \ldots, , \tag{2.74}$$

$$\begin{aligned} \hat{\Gamma} = (\hat{\gamma}_{ij}) &= \sum_{m=0}^{r} \hat{\sigma}_{ij}^{(m)} \hat{p}_m + \sum_{m=0}^{r} (\hat{\mu}_i^{(m)} - \hat{\bar{\mu}}_i)(\hat{\mu}_j^{(m)} - \hat{\bar{\mu}}_j)\,\hat{p}_m \\ &= \\ &= \sum_{m=0}^{r} \hat{\sigma}_{ij}^{(m)} \frac{n_m}{n} + \sum_{m=0}^{r} (\hat{\mu}_i^{(m)} - \hat{\bar{\mu}}_i)(\hat{\mu}_j^{(m)} - \hat{\bar{\mu}}_j)\frac{n_m}{n}\ . \end{aligned} \tag{2.75}$$

Determinação Matricial

Visando facilitar a implementação computacional, também vamos expressar os estimadores para a matriz de variâncias e covariâncias populacionais, de forma matricial. Para tanto, vamos representar a amostra aleatória de tamanho n de observações do vetor populacional misto $W_{\sim}^{*} = (Z_{\sim}', Y_{\sim}')'$ como a seguinte matriz de dados:

$$\mathbf{D} = \left(\begin{array}{cccc|ccc} z_{11} & z_{12} & \cdots & z_{1r} & y_{11} & y_{12} & \cdots & y_{1p} \\ z_{21} & z_{22} & \cdots & z_{2r} & y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & & & \vdots & \vdots & & & \\ z_{n1} & z_{n2} & \cdots & z_{nr} & y_{n1} & y_{n2} & \cdots & y_{np} \end{array} \right)_{n\times(r+p)} \tag{2.76}$$

Para obter os mesmos estimadores da determinação anterior, de forma matricial, vamos assumir que n_i observações, $i = 0, 1, \ldots, r$, estão associadas a i-ésima categoria multinomial Z_i. Com isso, reordenamos as linhas da matriz $\mathbf{D}$ de forma que as observações fiquem agrupadas em blocos, por categoria multinomial, visto que os momentos populacionais estão condicionados a estas categorias.

Fazendo esta reordenação de $\mathbf{D}$ temos,

$$\mathbf{D} = \left(\begin{array}{cccc|ccc} 1 & 0 & \cdots & 0 & y_{11}^{(1)} & y_{12}^{(1)} & \cdots & y_{1p}^{(1)} \\ \vdots & & & \vdots & \vdots & & & \\ 1 & 0 & \cdots & 0 & y_{n_1 1}^{(1)} & y_{n_1 2}^{(1)} & \cdots & y_{n_1 p}^{(1)} \\ \hline \vdots & & & \vdots & \vdots & & & \\ \hline 0 & 0 & \cdots & 1 & y_{11}^{(r)} & y_{12}^{(r)} & \cdots & y_{1pq}^{(r)} \\ \vdots & & & \vdots & \vdots & & & \\ 0 & 0 & \cdots & 1 & y_{n_r 1}^{(r)} & y_{n_r 2}^{(r)} & \cdots & y_{n_r p}^{(r)} \\ \hline 0 & 0 & \cdots & 0 & y_{11}^{(0)} & y_{12}^{(0)} & \cdots & y_{1p}^{(0)} \\ \vdots & & & \vdots & \vdots & & & \\ 0 & 0 & \cdots & 0 & y_{n_0 1}^{(0)} & y_{n_0 2}^{(0)} & \cdots & y_{n_0 p}^{(0)} \end{array} \right)_{n\times(r+p)} \tag{2.77}$$

Utilizando a notação formalizada no *Apêndice A*, podemos reescrever a matriz de dados $\mathbf{D}$, reordenada, como:

$$\mathbf{D} = \left(\begin{array}{cccc|c} 1_{n_1} & 0_{n_1} & \dots & 0_{n_1} & \mathbf{Y}^{(1)} \\ \hline & \vdots & & & \vdots \\ \hline 0_{n_r} & 0_{n_r} & \dots & 1_{n_r} & \mathbf{Y}^{(r)} \\ \hline 0_{n_0} & 0_{n_0} & \dots & 0_{n_0} & \mathbf{Y}^{(0)} \end{array} \right)_{n \times (r+p)} \tag{2.78}$$

onde,

$$\mathbf{Y}^{(i)} = \begin{pmatrix} y_{11}^{(i)} & y_{12}^{(i)} & \cdots & y_{1p}^{(i)} \\ y_{21}^{(i)} & y_{22}^{(i)} & & y_{2p}^{(i)} \\ & & & \\ y_{n_i 1}^{(i)} & y_{n_i 2}^{(i)} & & y_{n_i p}^{(i)} \end{pmatrix}_{n_i \times p} , \forall i = 0, \dots, r \ . \tag{2.79}$$

Para simplificar mais a notação, a matriz $\mathbf{D}$ ainda pode ser escrita como resultante da concatenação horizontal de sub-matrizes $\mathbf{D}^{(i)}$, $i = 0, \dots, r$, ou seja,

$$\mathbf{D} = (\mathbf{D}^{(1)\prime} | \dots | \mathbf{D}^{(r)\prime} | \mathbf{D}^{(0)\prime})_{n \times (r+p)}. \tag{2.80}$$

onde $\mathbf{D}^{(i)}$, $i = 0, \dots, r$, representa o bloco com todas as observações da amostra que estão associadas a i-ésima categoria multinomial, ou seja,

$$\mathbf{D}^{(i)} \ = \ (1_{n_i} \ 0_{n_i} \ \dots \ 0_{n_i} \ | \ \mathbf{Y}^{(i)})_{n_i \times (r+p)}, \ \forall i = 0, \dots, r \tag{2.81}$$

Utilizando a matriz $\mathbf{D}$ reordenada, o estimador da matriz de variâncias e covariâncias populacionais, dado de forma algébrica em (2.72), pode ser obtido matricialmente por,

$$\mathbf{S}_D = \frac{1}{n}(\mathbf{D}'\mathbf{H}_n\mathbf{D})_{(r+p)\times(r+p)}. \tag{2.82}$$

onde $\mathbf{H}_n$ é uma matriz de centralização (vide *Definição A.18)* particionada de forma

compatível com os blocos de D, ou seja,

$$\mathbf{H}_n = \begin{pmatrix} \mathbf{H}_{n_1} & -\frac{1}{n}1_{n_1}1'_{n_2} & \cdots & -\frac{1}{n}1_{n_1}1'_{n_r} & -\frac{1}{n}1_{n_1}1'_{n_0} \\ -\frac{1}{n}1_{n_2}1'_{n_1} & \mathbf{H}_{n_2} & \cdots & -\frac{1}{n}1_{n_2}1'_{n_r} & -\frac{1}{n}1_{n_2}1'_{n_0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{n}1_{n_r}1'_{n_1} & -\frac{1}{n}1_{n_r}1'_{n_2} & \cdots & \mathbf{H}_{n_r} & -\frac{1}{n}1_{n_r}1'_{n_0} \\ -\frac{1}{n}1_{n_0}1'_{n_1} & -\frac{1}{n}1_{n_0}1'_{n_2} & \cdots & -\frac{1}{n}1_{n_0}1'_{n_r} & \mathbf{H}_{n_0} \end{pmatrix}_{n\times n} \tag{2.83}$$

onde, como definido em (*A.14*).

$$\mathbf{H}_{n_i} = I_{n_i} - (1'_n 1_n)^{-1}\mathbf{J}_{n_i}, \textit{ tal que } \mathbf{J}_{n_i} = 1_{n_i}1'_{n_i}, i = 0, 1, \ldots, r. \tag{2.84}$$

Com respeito as propriedades verificadas, vale salientar que $\mathbf{H}_n$ é uma matriz simétrica (ver *Proposição A.2.a*) e idempotente (*Proposição A.2.b*) e, pela *Definição A.12*, é dita matriz de projeção. As sub-matrizes $\mathbf{H}_{n_i}$ são simétricas (ver *Proposição A.3*) porém, não são idempotentes (ver *contra-exemplo b1 − Apêndice A*) e, por conseguinte, também não são matrizes de projeção. (ver *Definição A.12*).

Agora, utilizando as matrizes $\mathbf{D}$ de (2.78) e $\mathbf{H}_n$ de (2.83) em (2.82), e utilizando o fato de $\mathbf{S}_D$ ser uma matriz simétrica, a recíproca matricial do estimador $\mathbf{S}_D$ dado em (2.72) é:

$$\mathbf{S}_D = \begin{pmatrix} \hat{\Psi} & \hat{\Delta} \\ \hat{\Delta}' & \hat{\Gamma} \end{pmatrix}_{(r+p)\times(r+p)},$$

onde,

$$\hat{\Psi}= \frac{1}{n}\begin{pmatrix} 1'_{n_1}\mathbf{H}_{n_1}1_{n_1} & -1'_{n_1}1_{n_1}(1'_n1_n)^{-1}1'_{n_2}1_{n_2} & \cdots & -1'_{n_1}1_{n_1}(1'_n1_n)^{-1}1'_{n_r}1_{n_r} \\ & 1'_{n_2}\mathbf{H}_{n_2}1_{n_2} & \cdots & -1'_{n_2}1_{n_2}(1'_n1_n)^{-1}1'_{n_r}1_{n_r} \\ & & \vdots & \vdots \\ & & \ddots & -1'_{n_{r-1}}1_{n_{r-1}}(1'_n1_n)^{-1}1'_{n_r}1_{n_r} \\ & & & 1'_{n_r}\mathbf{H}_{n_r}1_{n_r} \end{pmatrix}_{r\times r}, \tag{2.85}$$

$$\hat{\Delta}= \frac{1}{n}\begin{pmatrix} (1'_{n_1}(\mathbf{Y}^{(1)}-(1'_n1_n)^{-1}1_{n_1}1'_n\mathbf{Y}))_{1\times p} \\ \vdots \\ (1'_{n_r}(\mathbf{Y}^{(r)}-(1'_n1_n)^{-1}1_{n_r}1'_n\mathbf{Y}))_{1\times p} \end{pmatrix}_{r\times p}, \tag{2.86}$$

$$\hat{\Gamma}= \frac{1}{n}(\mathbf{Y}'\mathbf{H}_n\mathbf{Y})_{p\times p}. \tag{2.87}$$

Utilizando as expressões obtidas nas equações (2.73), (2.74) e (2.75) da determinação algébrica, também podemos escrever $\mathbf{S}_D$ como:

$$\mathbf{S}_D = \begin{pmatrix} \hat{\Psi} & \hat{\Delta} \\ \hat{\Delta}' & \hat{\Gamma} \end{pmatrix},$$

onde,

$$\hat{\Psi}= \begin{pmatrix} \frac{n_1}{n}(1-\frac{n_1}{n}) & -\frac{n_1n_2}{n^2} & \cdots & -\frac{n_1n_r}{n^2} \\ & \frac{n_2}{n}(1-\frac{n_2}{n}) & & -\frac{n_2n_r}{n^2} \\ & & \vdots & \\ & & \ddots & -\frac{n_{r-1}n_r}{n^2} \\ & & & \frac{n_r}{n}(1-\frac{n_r}{n}) \end{pmatrix}_{r\times r}, \tag{2.88}$$

$$\hat{\Delta}=\begin{pmatrix} \frac{n_1}{n}(\bar{y}_1^{(1)} - \bar{y}_1) & \dots & \frac{n_1}{n}(\bar{y}_p^{(1)} - \bar{y}_p) \\ \vdots & & \\ \frac{n_r}{n}(\bar{y}_1^{(r)} - \bar{y}_1) & \dots & \frac{n_r}{n}(\bar{y}_p^{(r)} - \bar{y}_p) \end{pmatrix}_{r\times p}, \tag{2.89}$$

e,

$$\hat{\Gamma}=(\hat{\gamma}_{ij})_{p\times p}\ ,\text{onde, } \forall i,j=1,2,\dots p, \hat{\gamma}_{ij}=\sum_{m=0}^{r}\hat{\sigma}_{ij}^{(m)}\frac{n_m}{n}+\sum_{m=0}^{r}(\hat{\mu}_i^{(m)} - \hat{\bar{\mu}}_i)(\hat{\mu}_j^{(m)} - \hat{\bar{\mu}}_j)\frac{n_m}{n}. \tag{2.90}$$

Capítulo 3

Distâncias para Misturas

3.1 Resumo do Capítulo

Este Capítulo está dividido em três seções: na seção 3.2 são feitas algumas considerações sobre distâncias para *misturas*. Na *seção 3.3*, que é a principal, determinamos a inversa da matriz de variâncias e covariâncias para *misturas*, introduzida no *Capítulo 2* e, apresentamos uma *Extensão da Distância de Mahalanobis* para o caso de *misturas*. Na *seção 3.4* são apresentadas e discutidas algumas das distâncias mais citadas, na literatura, para o caso de *misturas*. Vale ressaltar que os resultados apresentados na *seção 3.3* são uma contribuição desta dissertação.

3.2 Considerações Iniciais sobre Distâncias para Misturas

Segundo Kotz e Jonhson (1986), uma distância pode ser definida arbitrariamente porém, existem certas condições básicas que são naturais à definição de distância, e que permitem à mesma conservar o seu significado intuitivo. As condições citadas por Kotz e Johnson, são as que definem uma *Distância Métrica* (ver *Definição C.4*) e, por isto, o uso destes coeficientes é o mais recomendado e utilizado na literatura. Para não fugirmos a uma opinião dominante, vamos também dar preferência a utilização de distâncias deste tipo. Além desta definição, o *Apêndice C* traz uma abordagem hierarquizada dos termos

matemáticos atribuídos aos coeficientes de dissimilaridade.

3.3 A Distância de Mahalanobis para Misturas

3.3.1 Definição

A *Distância Métrica Quadrada de Mahalanobis*, definida por Mahalanobis (1936), é também conhecida como a *Métrica Euclideana Quadrada Ponderada*. Nesta *métrica*, onde a ponderação é feita pela inversa da matriz de variâncias e covariâncias, a versão para variáveis contínuas é dada da seguinte forma:

Definição 3.1: Sejam $\mathbf{x}_{(i)} = (x_{i1}, x_{i2}, \ldots, x_{iq})$ e $\mathbf{x}_{(j)} = (x_{j1}, x_{j2}, \ldots, x_{jq})$ dois vetores linhas que correspondem aos escores obtidos para os indivíduos i e j na matriz de dados $\mathbf{X}$ e $\mathbf{S}$ a matriz de variâncias e covariâncias que, necessariamente, precisa ser positiva definida. A *Distância Métrica Quadrada de Mahalanobis* (ou *Métrica Euclideana Quadrada Ponderada*, onde S^{-1} é a matriz de ponderação), denotada por $d^2_{\mathbf{S}}(\mathbf{x}_{(i)},\mathbf{x}_{(j)})$, é definida como a norma $L_{\mathbf{B}}$ (ver *Definição A.5*) ao quadrado, onde $\mathbf{B} = \mathbf{S}^{-1}$, do vetor diferença $\mathbf{x}_{(i)} - \mathbf{x}_{(j)}$, ou seja,

$$d^2_{\mathbf{S}^{-1}}(\mathbf{x}_{(i)}, \mathbf{x}_{(j)}) = [\|\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\|_{\mathbf{S}^{-1}}]^2 = (x_{(i)} - x_{(j)})\mathbf{S}^{-1}(x_{(i)} - x_{(j)})' \qquad (3.1)$$

A *Distância Métrica Quadrada de Mahalanobis*, definida acima, possui uma propriedade mais geral de invariância, que é dada na proposição abaixo:

Proposição 3.1: Seja $\mathbf{X}$ uma matriz de dados de dimensão $n \times q$. A distância de Mahalanobis, dada na *Definição 3.1*, é invariante sobre todas as transformações não singulares $\mathbf{C}$ (matriz $q \times q$) aplicadas aos vetores linhas da matriz de dados $\mathbf{X}$.

Prova:

Tomemos,

$$\mathbf{y}_{(i)} = \mathbf{C}'\mathbf{x}'_{(i)}, \ \forall i = 1, \ldots, n. \qquad (3.2)$$

onde, $\mathbf{y}'_{(i)} = \mathbf{x}_{(i)}\mathbf{C}, \forall i = 1, \ldots, n$ é um vetor linha transformado. Com isto, a nossa matriz de dados transformada é:

$$\mathbf{T} = \mathbf{XC}, \tag{3.3}$$

e a nossa matriz de variâncias e covariâncias, expressa em (2.82), fica:

$$\mathbf{S}_Y = \frac{1}{n}(\mathbf{XC})'\mathbf{H}_n\mathbf{XC} = \frac{1}{n}\mathbf{C}'\mathbf{X}'\mathbf{H}_n\mathbf{XC} = \frac{1}{n}\mathbf{C}'\mathbf{S}_X\mathbf{C} \tag{3.4}$$

Com isso, a *distância de Mahalanobis* (3.1), entre os indivíduos transformados $\mathbf{y}_{(i)}$ e $\mathbf{y}_{(j)}$ será:

$$\begin{aligned}
\partial^2_{\mathbf{S}_Y^{-1}}(\mathbf{y}_{(i)}, \mathbf{y}_{(j)}) &= (\mathbf{y}_{(i)} - \mathbf{y}_{(j)})\mathbf{S}_Y^{-1}(\mathbf{y}_{(i)} - \mathbf{y}_{(j)})' \\
&= (\mathbf{X}_{(i)}\mathbf{C} - \mathbf{X}_{(j)}\mathbf{C})(\tfrac{1}{n}\mathbf{C}'\mathbf{D}'\mathbf{H}_n\mathbf{DC})^{-1}(\mathbf{X}_{(i)}\mathbf{C} - \mathbf{X}_{(j)}\mathbf{C})' \\
&= \tfrac{1}{n}(\mathbf{X}_{(i)} - \mathbf{X}_{(j)})\mathbf{C}\mathbf{C}^{-1}\mathbf{S}_X^{-1}\mathbf{C}^{-1}\mathbf{C}'(\mathbf{X}_{(i)} - \mathbf{X}_{(j)})' \qquad (3.5) \\
&= \tfrac{1}{n}(\mathbf{X}_{(i)} - \mathbf{X}_{(j)})\mathbf{S}_X^{-1}(\mathbf{X}_{(i)} - \mathbf{X}_{(j)})' \\
&= \partial^2_{\mathbf{S}_X^{-1}}(\mathbf{x}_{(i)}, \mathbf{x}_{(j)}) \ \square
\end{aligned}$$

Considerando o caso particular em que $\mathbf{C}$ é uma matriz diagonal com elementos diagonais diferentes de zero, a matriz transformada $\mathbf{Y}$ passa a ser chamada matriz de escala, pois a j-ésima variável de $\mathbf{X}$ estará multiplicada por um escalar c_j em $\mathbf{Y}$. Com isso, a *Distância Métrica de Mahalanobis* é dita invariante com relação a escala, ou seja, ela dá a mesma distância para qualquer que seja a unidade usada para medir as variáveis. As outras métricas, em particular a *Euclideana*, não possuem esta notável propriedade.

O nosso interesse está em utilizar a *Distância de Mahalanobis* para o contexto de *misturas* onde as variáveis categóricas são tranformadas, como discutido no *Capítulo 2*, em um vetor multinomial. Para este caso, utilizando a notação definida no *Capítulo 2*,

vamos definir agora uma extensão da *Definição 3.1*.

Definição 3.2: Sejam $\mathbf{w}_{(i)} = (\mathbf{z}_{(i)}, \mathbf{y}_{(i)})$ e $\mathbf{w}_{(j)} = (\mathbf{z}_{(j)}, \mathbf{y}_{(j)})$ dois vetores linhas mistos com as mensurações obtidas para o i-ésimo e j-ésimo indivíduos pertencentes a matriz de dados $\mathbf{D}$, respectivamente, onde $\mathbf{z}_{(k)} = (z_{k1}, z_{k2}, \ldots, z_{kr})$ corresponde ao escore do bloco multinomial e $\mathbf{y}_{(k)} = (y_{k1}, y_{k2}, \ldots, y_{kp})$ corresponde ao escore do bloco contínuo, $k \in \{i, j\}$. $\mathbf{S}$ a matriz de variâncias e covariâncias que, necessariamente, precisa ser não singular. A *Extensão da Distância de Mahalanobis* para o contexto de *misturas*, denotada por $d^2_{\mathbf{S}^{-1}}(\mathbf{w}_{(i)}, \mathbf{w}_{(j)})$, é definida como,

$$d^2_{\mathbf{S}^{-1}}(\mathbf{w}_{(i)}, \mathbf{w}_{(j)}) = [(\mathbf{z}_{(i)} - \mathbf{z}_{(j)}), (\mathbf{y}_{(i)} - \mathbf{y}_{(j)})]\mathbf{S}^{-1}[(\mathbf{z}_{(i)} - \mathbf{z}_{(j)}), (\mathbf{y}_{(i)} - \mathbf{y}_{(j)})]'. \quad (3.6)$$

A exigência de $\mathbf{S}$ ser não singular é para podermos assegurar que a inversa $\mathbf{S}^{-1}$ existe. Porém, para podermos utilizar esta distância precisamos determinar a inversa $\mathbf{S}^{-1}$ da matriz de variâncias e covariâncias para *misturas* $\mathbf{S}$.

3.3.2 A Matriz de Variâncias e Covariâncias e sua Inversa

Como estamos trabalhando com dados com *misturas* de variáveis categóricas e contínuas, e estamos aplicando distâncias com respeito a vetores mistos $\mathbf{w}_{(i)} = (\mathbf{z}_{(i)}, \mathbf{y}_{(i)})$ e $\mathbf{w}_{(j)} = (\mathbf{z}_{(j)}, \mathbf{y}_{(j)})$, vamos utilizar a matriz de variâncias e covariâncias amostrais $\mathbf{S}_D$ particionada, dada em (2.72), que permite visualizar as covariâncias entre contínuas, entre categóricas e entre ambas, na forma natural de blocos, ou seja,

$$\mathbf{S}_D = \begin{pmatrix} \hat{\mathbf{\Psi}} & \hat{\mathbf{\Delta}} \\ \hat{\mathbf{\Delta}}' & \hat{\mathbf{\Gamma}} \end{pmatrix}. \quad (3.7)$$

Esta forma particionada de $\mathbf{S}_D$ permite expressar a sua inversa $\mathbf{S}_D^{-1}$ também de forma particionada, desde que as inversas necessárias existam (ver *Proposição A.1.b*), ou seja,

$$\mathbf{S}_D^{-1} = \begin{pmatrix} \boldsymbol{\Psi}^* & \boldsymbol{\Delta}^* \\ \boldsymbol{\Delta}^{*\prime} & \boldsymbol{\Gamma}^* \end{pmatrix} \tag{3.8}$$

onde,

$$\boldsymbol{\Psi}^* = (\hat{\boldsymbol{\Psi}} - \hat{\boldsymbol{\Delta}}\,\hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\Delta}}')^{-1}, \tag{3.9}$$

$$\boldsymbol{\Delta}^* = -\boldsymbol{\Psi}^*\hat{\boldsymbol{\Delta}}\,\hat{\boldsymbol{\Gamma}}^{-1} = -\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Delta}}\,\boldsymbol{\Gamma}^*. \tag{3.10}$$

$$\boldsymbol{\Gamma}^* = (\hat{\boldsymbol{\Gamma}} - \hat{\boldsymbol{\Delta}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Delta}})^{-1}. \tag{3.11}$$

Podemos observar que as expressões (3.9) e (3.11) têm as formas da matriz de resíduos das categóricas dado as contínuas e da matriz de resíduos das contínuas dado as categóricas, ou seja, são obtidas como a soma de quadrados dos resíduos de $\hat{\boldsymbol{\Psi}}$ / $\hat{\boldsymbol{\Gamma}}$ e $\hat{\boldsymbol{\Gamma}}$ / $\hat{\boldsymbol{\Psi}}$ respectivamente.

As Inversas dos Blocos Diagonais

Um ponto que vale a pena ser discutido é com relação as inversas dos blocos diagonais $\hat{\boldsymbol{\Psi}}$ e $\hat{\boldsymbol{\Gamma}}$. A existência da inversa $\hat{\boldsymbol{\Psi}}^{-1}$ para o bloco categórico $\hat{\boldsymbol{\Psi}}$ em (3.7) é importante, pois se trata de uma das inversas que são necessárias para a utilização da expressão (3.8). Para obtermos esta inversa, como discutido no *Capítulo 2*, transformamos as variáveis categóricas em um vetor multinomial com r categorias e retiramos uma categoria deste vetor para evitar a dependência linear entre as mesmas, pois o vetor multinomial completo tem posto $r-1$ (ver *Proposição B.1,a*). Esta dependência linear torna a matriz de variâncias e covariâncias $\hat{\boldsymbol{\Psi}}$ singular (ver *Proposição B.1,c*), e assim, não podemos obter a sua inversa $\hat{\boldsymbol{\Psi}}^{-1}$ de forma única. Dessa forma, o bloco $\hat{\boldsymbol{\Psi}}$ da matriz $\mathbf{S}_D$ corresponde, neste trabalho, as covariâncias entre as $r-1$ categorias multinomiais mantidas. Para obter a inversa desta matriz $\hat{\boldsymbol{\Psi}}$, vamos proceder de forma análoga a Mardia, Kent e Bibby

(1979), pág. 292, ou seja:

A matriz de variâncias e covariâncias de um vetor multinomial completo é dada por:

$$n\,\hat{\mathbf{\Psi}}^{\mathbf{c}} = diag(f^c) - \frac{f^c f^{c\prime}}{n}, \text{ onde } f^c = (n_1, n_2, \ldots, n_r)'. \tag{3.12}$$

Como $n\,\hat{\mathbf{\Psi}}^{\mathbf{c}}$ não possui inversa única devido a sua singularidade (ver *Proposição B.1.c*) retiramos, como discutido no *Capítulo 2*, a útima categoria do vetor multinomial completo. Com isso, passamos a trabalhar com uma matriz de variâncias e covariâncias $\hat{\mathbf{\Psi}}$ que, por ser não singular, possui inversa $\hat{\mathbf{\Psi}}^{-1}$ dada por:

$$(n\,\hat{\mathbf{\Psi}})^{-1} = [diag(f)]^{-1} + \frac{11'}{f_0^c}, \text{ onde } f_0 = n_0 \text{ e } f = (n_1, n_3, \ldots, n_r)', \tag{3.13}$$

o que implica que,

$$(\hat{\mathbf{\Psi}})^{-1} = n\{[diag(f)]^{-1} + \frac{11'}{f_0^c}\}, \text{ onde } f_0 = n_0 \text{ e } f = (n_1, n_3, \ldots, n_r)', \tag{3.14}$$

Podemos observar, pela equação (3.14), que apesar da última categoria multinomial ter sido retirada, a inversa da matriz de variâncias e covariâncias $\hat{\mathbf{\Psi}}$ incorpora o efeito da da mesma. Isto ocorre graças a relação existente entre as categorias do vetor multinomial completo, ou seja, $\sum_{i=1}^{r} n_i = n$.

Já com relação a inversa para o bloco contínuo $\hat{\mathbf{\Gamma}}$, temos que para $n > p$ (ver *Proposição B.3,a*) podemos assumir que $\hat{\mathbf{\Gamma}}^{-1}$ existe. Dado que o bloco $\hat{\mathbf{\Psi}}$ possui inversa e assumindo que $n > p$, para assegurar a inversa de $\hat{\mathbf{\Gamma}}$, temos que se $\mathbf{\Gamma}^*$ ou $\mathbf{\Psi}^*$ existem, ou seja, se $(\hat{\mathbf{\Gamma}} - \hat{\mathbf{\Delta}}'\,\hat{\mathbf{\Psi}}^{-1}\,\hat{\mathbf{\Delta}})$ ou $(\hat{\mathbf{\Psi}} - \hat{\mathbf{\Delta}}\,\hat{\mathbf{\Gamma}}^{-1}\,\hat{\mathbf{\Delta}}')$ possuem inversa, então $\mathbf{S}_D^{-1}$ existe (ver *Proposição A.1.a*). Além disso, segundo o *Corolário A.7.2.1* de Mardia, Kent e Bibby (1979), se $\mathbf{S}_D$ *é definida positiva* (ver *Definição A.4*) então também é *não singular* (ver *Definição A.13*), ou seja, $\mathbf{S}_D^{-1}$ existe.

3.3.3 Expressão em Blocos para a Extensão da Distância de Mahalanobis

Como expressamos a inversa da matriz de variâncias e covariâncias na forma de blocos podemos, agora, determinar uma expressão para a *Extensão da Distância de Mahalanobis* para misturas, em função desses blocos. Para tanto, aplicando a *Definição 3.2* sobre dois vetores linha mistos,

$$\mathbf{w}_{(i)} = (\mathbf{z}_{(i)}, \mathbf{y}_{(i)}) \text{ e } \mathbf{w}_{(j)} = (\mathbf{z}_{(j)}, \mathbf{y}_{(j)}), \tag{3.15}$$

temos,

$$d^2_{\mathbf{S}_D^{-1}}(\mathbf{w}_{(i)}, \mathbf{w}_{(j)}) = [(\mathbf{z}_{(i)} - \mathbf{z}_{(j)}), (\mathbf{y}_{(i)} - \mathbf{y}_{(j)})] \begin{bmatrix} \mathbf{\Psi}^* & \mathbf{\Delta}^* \\ \mathbf{\Delta}^{*\prime} & \mathbf{\Gamma}^* \end{bmatrix} \begin{pmatrix} (\mathbf{z}_{(i)} - \mathbf{z}_{(j)})' \\ (\mathbf{y}_{(i)} - \mathbf{y}_{(j)})' \end{pmatrix} \tag{3.16}$$

Para facilitar a interpretação da *Extensão da Distância de Mahalanobis* com relação às parcelas que a compõem, vamos assumir,

$$D^{cg}_{i,j} = (\mathbf{z}_{(i)} - \mathbf{z}_{(j)}) \text{ e } D^{cn}_{i,j} = (\mathbf{y}_{(i)} - \mathbf{y}_{(j)}), \tag{3.17}$$

onde $D^{cg}_{i,j}$ corresponde à diferença entre os escores observados para o $i-$ésimo e o $j-$ésimo indivíduo, somente nas variáveis categóricas. Enquanto $D^{cn}_{i,j}$ corresponde analogamente às contínuas. Com isso, substituindo (3.17) em (3.16) temos,

$$d^2_{\mathbf{S}_D^{-1}}(\mathbf{w}_{(i)}, \mathbf{w}_{(j)}) = \left(D^{cg}_{i,j},\ D^{cn}_{i,j}\right) \begin{bmatrix} \mathbf{\Psi}^* & \mathbf{\Delta}^* \\ \mathbf{\Delta}^{*\prime} & \mathbf{\Gamma}^* \end{bmatrix} \begin{pmatrix} D^{cg\prime}_{i,j} \\ D^{cn\prime}_{i,j} \end{pmatrix}, \tag{3.18}$$

e, fazendo os produtos cruzados,

$$d^2_{\mathbf{S}_D^{-1}}(\mathbf{w}_{(i)}, \mathbf{w}_{(j)}) = D^{cg}_{i,j}\mathbf{\Psi}^* D^{cg\prime}_{i,j} + D^{cn}_{i,j}\mathbf{\Gamma}^* D^{cn\prime}_{i,j} + 2D^{cg}_{i,j}\mathbf{\Delta}^* D^{cn\prime}_{i,j}. \tag{3.19}$$

Podemos observar que, ao utilizar a inversa da matriz de variâncias e covariâncias em sua forma particionada, a distância entre os indivíduos $\mathbf{w}_{(i)}$ e $\mathbf{w}_{(j)}$ é expressa como uma combinação de três parcelas distintas. Um exame destas parcelas mostra que a componente da distância, devido à parte puramente contínua, é obtida em função da inversa da matriz de resíduos da distribuição condicional das contínuas dado as categóricas. De forma análoga pode-se reconhecer, ainda que de forma menos usual e apropriada, que a componente da distância, devido à parte puramente categórica, é obtida em função da inversa da matriz dos resíduos da distribuição condicional das categóricas dado as contínuas (aproximações assintóticas da multinomial à normal tornam esta leitura menos forçada, ver *exemplo 2.9.1* de Mardia Kent e Bibby, 1979). A terceira parcela corresponde ao produto escalar entre as categóricas e as contínuas em função do bloco correspondente as inversas das covariâncias entre elas.

3.3.4 Vantagens e Desvantagens da Utilização da Distância de Mahalanobis para Misturas

Segundo Cormack (1971), existem duas razões para não se adotar uma distância, para aplicação em *Análise de Agrupamento*, baseada numa matriz de covariâncias geral: primeiro, a maioria das inter-relações existentes são, provavelmente, causadas pela existência dos grupos que estão sendo buscados e, segundo, a estrutura de inter-relação dentro dos grupos pode variar consideravelmente de grupo para grupo. Todavia, utilizando a matriz de variâncias e covariâncias $\mathbf{S}_D$ na forma particionada, como feito aqui, a adoção da *Extensão da Distância de Mahalanobis* para o contexto de *misturas* apresenta as seguintes vantagens:

a) Trata simultaneamente o vetor misto, sem atribuição arbitrária de pesos que combinem distâncias entre contínuas e entre categóricas, e sem a subjetividade, não só dos pesos como também da distância a ser adotada para cada um dos grupos.

b) A distância entre indivíduos pode ser decomposta em três parcelas que, analisadas separadamente, podem esclarecer a forma como as *misturas* afetam as distâncias através das contínuas removido o efeito das categóricas, das categóricas removido o efeito das contínuas e da presença conjunta de suas inter-relações.

c) As técnicas de agrupamento podem ser aplicadas a cada parcela separadamente, à combinação de parcelas e à distância total. Diferenças nos agrupamentos produzidos podem ser, com isso, uma auxílio à melhor caracterização dos grupos que estão sendo buscados.

3.4 Principais Distâncias para Misturas

3.4.1 Considerações Iniciais

Para facilitar a aplicação das técnicas existentes para o contexto de *misturas*, geralmente a primeira coisa a se fazer é reordenar os dados de modo a formar blocos de diferentes tipos de variáveis. A partir desta reordenação dos dados, podemos escrever dois vetores linhas, correspondentes a indivíduos distintos de uma matriz de dados $\mathbf{X}$, como:

$$\mathbf{x}_{(i)} = (\mathbf{x}_{(i)}^{(N)}, \mathbf{x}_{(i)}^{(O)}, \mathbf{x}_{(i)}^{(Q)}) \text{ e } \mathbf{x}_{(j)} = (x_{(j)}^{(N)}, x_{(j)}^{(O)}, x_{(j)}^{(Q)}), \tag{3.20}$$

onde "N" está relacionado com as *variáveis categóricas nominais*, "O" com as *categóricas ordinais* e "Q" com as *quantitativas*. Vale salientar que, como dito no *Capítulo 2*, vamos considerar apenas as *quantitativas contínuas*, ou seja, não vamos incluir as *quantitativas discretas* no nosso estudo de *misturas*. Porém, esta não é a única forma de se fazer a blocagem dos diferentes tipos de variáveis mensuradas, nos coeficientes que serão abordadas a seguir poderemos vislumbrar formas alternativas à (3.20).

3.4.2 Distância Combinada

Baseados na reordenação dos dados feita em (3.20), Green & Carmone (1970) apresentam um coeficiente combinado, de forma aditiva, por coeficientes de dissimilaridade utilizados para cada tipo de variável.

Definição 3.3: Sejam $x_{(i)}$ e $x_{(j)}$ dois vetores linhas, como em (3.20), correspondentes aos escores obtidos para o i-ésimo e o j-ésimo indivíduo da matriz de dados $\mathbf{X}$, que possuem mensurações de variáveis categóricas nominais, categóricas ordinais e contínuas. A *Função Distância Combinada*, ou simplesmente *Distância Combinada*, denotada por $d_C(x_{(i)}, x_{(j)})$, é definida como:

$$d_C(x_{(i)}, x_{(j)}) = \omega_1 d^{(N)}(x_{(i)}^{(N)}, x_{(j)}^{(N)}) + \omega_2 d^{(O)}(x_{(i)}^{(O)}, x_{(j)}^{(O)}) + \omega_3 d^{(Q)}(x_{(i)}^{(Q)}, x_{(j)}^{(Q)}), \quad (3.21)$$

onde $d^{(N)}$, $d^{(O)}$ e $d^{(Q)}$ são coeficientes de dissimilaridade para os blocos nominal, ordinal e quantitativo contínuo, respectivamente, e ω_1, ω_2 e ω_3 são seus respectivos pesos.

Segundo Bussab, Miazaki e Andrade (1990), a construção deste coeficiente combinado exige alguns cuidados especiais:

a) Os coeficientes utilizados, para cada tipo de variáveis, precisam ter intervalos de variação iguais ou próximos.

b) Os pesos precisam ser adequados e interpretáveis. Segundo Bussab, Miazaki e Andrade (1990), tem sido muito utilizada a ponderação pelo número de variáveis existentes de cada tipo.

Esta questão de ponderar pelo número de variáveis envolvidas é muito complexa, visto que geralmente existem variáveis mais importantes que outras para o estudo em questão e, isto dificulta a definição dos pesos. Um outro problema ocorre se existem valores perdidos para alguns dos objetos. Se esses valores são de mais de um tipo de variáveis,

Rubin (1967) recomenda que, no cálculo da distância, nenhuma quantidade desses objetos ou das variáveis, onde apenas um objeto tenha sido mensurado, seja considerada.

Caso as funções $d^{(N)}, d^{(O)}$ e $d^{(Q)}$ sejam métricas, pela *Proposição C.1*, para $k = 3$, temos que a *Distância Combinada* d_C também o será. Como é mais comum, na literatura, a utilização de distâncias métricas (ver *Definição C.4*), vamos apresentar agora as mais utilizadas para cada um dos blocos descritos acima.

Distâncias para Variáveis Quantitativas

Inicialmente, vamos abordar uma *distância métrica* que engloba outras bem conhecidas como casos particulares. Esta *métrica generalizada*, dada em Boyce (1969), é conhecida como *Distância Métrica de Minkowski* e derivada da norma L_w (ver *Definição A.7*).

Definição 3.4: Sejam $x_{(i)} = (x_{i1}, x_{i2}, \ldots, x_{iq})$ e $x_{(j)} = (x_{j1}, x_{j2}, \ldots, x_{jq})$ dois vetores linhas que correspondem aos escores obtidos para o i-ésimo e o j-ésimo indivíduo, na matriz de dados $\mathbf{X}$ e $w = (w_1, w_2, \ldots, w_q)$ um vetor linha de pesos. A *Distância Métrica de Minkowski* entre $x_{(i)}$ e $x_{(i)}$ será denotada por $d_M(x_{(i)}, x_{(j)})$ e definida como a norma L_w do vetor diferença $x_{(i)} - x_{(j)}$, ou seja,

$$d_M(x_{(i)}, x_{(j)}) = \|x_{(i)} - x_{(j)}\|_w = \left(\sum_{k=1}^{q} \omega_k \mid x_{ik} - x_{jk} \mid^p \right)^{1/p}, \ \forall p \geq 1, \qquad (3.22)$$

onde os ω_k's são, respectivamente, os pesos correspondentes às variáveis X_k's, $k = 1, 2, \ldots, q$.

Tomando $p = 1$, em (3.22), obtemos a *Distância Métrica City-Block*, dada em Johnson & Wall (1969), que será denotada por d_{CB} e definida como:

$$d_{CB}(x_{(i)}, x_{(j)}) = \|x_{(i)} - x_{(j)}\|_{w|_{p=1}} = \sum_{k=1}^{q} \omega_k \mid x_{ik} - x_{jk} \mid . \qquad (3.23)$$

Se além de $p = 1$, tomarmos $\omega_k = 1/q$, $\forall k = 1, 2, \ldots, q$, em (3.22), temos a *Distância Métrica City-Block Média*, dada em Cain e Harrison (1958), que será denotada por $d_{\bar{C}B}$

e definida como:

$$d_{CB}(x_{(i)}, x_{(j)}) = \|x_{(i)} - x_{(j)}\|_{w|_{p=1 \ e \ \omega_k=1/q}} = \sum_{k=1}^{q} \frac{|\, x_{ik} - x_{jk} \,|}{q} \ . \tag{3.24}$$

Tomando o limite de p quando este tende a infinito e $\omega_k = 1, \ \forall k = 1, 2, \ldots, q$, em (3.22), obtemos a *Distância Métrica de Chebychev*, dada em Spath (1980), que será denotada por d_C e definida como:

$$d_C(x_{(i)}, x_{(j)}) = \|x_{(i)} - x_{(j)}\|_{w|_{p=\infty \ e \ \omega_k=1}} = \max |\, x_{ik} - x_{jk} \,| \ . \tag{3.25}$$

Fazendo $p = 2$ e $\omega_k = 1, \ \forall k = 1, 2, \ldots, q$, em (3.22), teremos a mais conhecida de todas as distâncias existentes que é a *Distância Métrica Euclidiana Usual*, que será denotada por d_E e definida como:

$$\begin{aligned} d_E(x_{(i)}, x_{(j)}) &= \|x_{(i)} - x_{(j)}\|_{w|_{p=2 \ e \ \omega_k=1}} = \left(\sum_{k=1}^{q} (x_{ik} - x_{jk})^2 \right)^{1/2} \\ &= \sqrt{< x - y, x - y >} = \sqrt{(x-y)(x-y)'} \ . \end{aligned} \tag{3.26}$$

De acordo com Spath (1980), a *Distância Métrica Euclideana Usual* pode ser generalizada através da utilização de uma norma ponderada por uma matriz positiva definida B, denotada por L_B (ver *Definição A.5*). A vantagem desta generalização é que possibilitará a inclusão de produtos cruzados como ponderação, ou seja, poderemos levar em consideração alguma relação existente entre as variáveis como, por exemplo, utilizando a matriz de variâncias e covariâncias.

Com esta generalização, a *Distância Métrica Euclideana Usual* passa a se chamar *Distância Métrica Euclideana Ponderada por* $\mathbf{B}$, conforme definição abaixo:

Definição 3.5: Sejam $x_{(i)} = (x_{i1}, x_{i2}, \ldots, x_{iq})$ e $x_{(j)} = (x_{j1}, x_{j2}, \ldots, x_{jq})$ dois vetores linhas que correspondem aos escores obtidos para o $i-$ésimo e o $j-$ésimo objeto da matriz de dados $\mathbf{X}$ e B uma matriz positiva definida. A *Distância Métrica Euclideana Ponderada por* B, denotada por $d_{E_B}(x_{(i)}, x_{(j)})$, é definida como a norma

L_B do vetor diferença $x_{(i)} - x_{(j)}$, ou seja,

$$d_{E_B}(x_{(i)}, x_{(j)}) = \|x_{(i)} - x_{(j)}\|_B = \sqrt{(x-y)B(x-y)'} \,. \tag{3.27}$$

Para matrizes B específicas, nós podemos obter distâncias métricas bastante utilizadas. Os casos mais conhecidos são:

a) $B = I$, temos a *Distância Métrica Euclideana Usual* dada em (3.26).

b) $B = (diag(q))^{-1}$, temos a *Distância Métrica Euclideana Média*,

c) $B = (diag(s_1^2, s_2^2, \ldots, s_q^2))^{-1}$, temos a *Distância Métrica Euclideana Padronizada*,

d) $B = \mathbf{S}^{-1}$, temos a *Distância Métrica de Mahalanobis*, definida em Mahalanobis (1936)., onde $\mathbf{S}^{-1}$ é a inversa da matriz de variâncias e covariâncias.

Podemos ver que a distância definida em *b)* é apenas um reescalonamento de *a)*. Já o caso *c)* é sugerido porque a *Distância Métrica Euclideana Usual* é muito afetada pela diferença de escala entre as variáveis, por estarem sendo somadas grandezas não comparáveis. Todavia, Everitt (1974) ressalta que esta padronização pode diluir as diferenças entre grupos sobre as variáveis que são melhor discriminadoras, quando da utilização de técnicas de agrupamento. Já o caso *d)*, discutido na *seção 3.2*, possui a vantagem, sobre as outras *distâncias* apresentadas, de fazer a ponderação pelo grau de inter-relação existente entre as variáveis. Quando estas inter-relações são nulas, ela é equivalente à *Distância Métrica Euclideana Padronizada, caso c)*.

Existem muitas distâncias métricas que não são nem casos particulares da *Distância Métrica de Minkowski* e nem uma generalização da mesma. Entre estas podemos citar a *Distância Métrica de Camberra*, dada em Bray e Curtis (1957), que é definida como:

$$d_{Cam}(x_i, x_j) = \sum_{k=1}^{q} \frac{|\, x_{ik} - x_{jk} \,|}{(x_{ik} + x_{jk})} \,. \tag{3.28}$$

Além das métricas, existem várias distâncias não métricas conhecidas. Por exemplo, Sokal e Sneath (1963) propõem a seguinte função semi-métrica (ver *Definição C.3*):

$$d_{SS}(x_{(i)}, x_{(j)}) = \sqrt{\left(\frac{1}{q}\sum_{k=1}^{q}\left(\frac{x_{ik} - x_{jk}}{x_{ik} + x_{jk}}\right)^2\right)}. \tag{3.29}$$

Para uma complementação à lista aqui apresentada, Cormack (1971), Kotz e Johnson (1986) e Romesburg (1984) apresentam outras distâncias para o caso de variáveis quantitativas contínuas.

Distâncias para Variáveis Categóricas Nominais

Para este caso, como sugerido por Bussab, Miazaki e Andrade (1990), costuma-se transformar primeiro as variáveis categóricas nominais em binárias, utilizando-se a *Definição 2.3*, e então utiliza-se os coeficientes definidos para o caso binário. Entre os *coeficientes de similaridade* (ver *Definição C.6*) existentes para este caso, que possuem variação entre 0 e 1, podemos destacar o seguintes:

Coeficiente de Similaridade	Expressão	Métrica
Coincidência Simples	$\frac{a+d}{a+b+c+d}$	Sim
Jaccard	$\frac{a}{a+b+c}$	Sim
Rogers e Tanimoto	$\frac{a+d}{a+2(b+c)+d}$	Sim
Sneath e Sokal	$\frac{a+d}{a+\frac{1}{2}(b+c)+d}$	Não
Sorensen	$\frac{a}{a+\frac{1}{2}(b+c)}$	Não
Russell e Rao	$\frac{a}{a+b+c+d}$	Sim
Anderberg	$\frac{a}{a+2(b+c)}$	Sim
Kulczynski	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	Não
Ochiai	$\frac{a}{\sqrt{(a+b)(c+d)}}$	Não

Uma lista mais completa, incluindo coeficientes com outros intervalos de variação, pode ser obtida em Cormack (1971), Kotz e Johnson (1986) e Romesburg (1984). Como

todos os coeficientes de similaridade apresentados acima possuem intervalo de variação entre 0 e 1, para obter uma distância a partir dos mesmos, podemos utilizar a seguinte transformação:

$$d_{ij} = 1 - S_{ij}. \tag{3.30}$$

Distâncias para Variáveis Categóricas Ordinais

Para este caso, como também sugerido por Bussab, Miazaki e Andrade (1990), costuma-se transformar primeiro as variáveis qualitativas ordinais em binárias, utilizando-se a *Definição 2.4*, e então utiliza-se as distâncias definidas acima, para o caso binário.

3.4.3 Distância a partir do Coeficiente Geral de Similaridade de Gower

Considerando o fato de que variáveis nominais, ordinais e contínuas são de natureza distinta, Gower (1971) propôs um *Coeficiente de Similaridade Geral*, que inclui vários dos existentes como casos particulares e, que na verdade é uma forma mais elaborada do *Coeficiente Combinado*, citado na seção anterior, na versão para similaridades.

A definição do Coeficiente Geral de Similaridade de Gower

O *Coeficiente de Similaridade de Gower* exige que para cada variável X_i seja designado um coeficiente de semelhança variando entre 0 e 1. Para tanto, consideremos inicialmente a seguinte definição:

Definição 3.6: Sejam $x_{(i)}$ e $x_{(j)}$ dois vetores mistos, dispostos como em (3.20), correspondentes aos escores obtidos para o i-ésimo e o j-ésimo indivíduo da matriz de dados $\mathbf{X}$. Um *coeficiente de semelhança* para a k-ésima variável desses 2 vetores,

denotado por s_{ijk}, é definido como:

$$s_{ijk} = \begin{cases} 0, & se\ os\ escores\ i\ e\ j\ são\ diferentes\ na\ k-esima\ var. \\ \epsilon \in (0,1], & se\ os\ escores\ i\ e\ j\ tem\ uma\ parecenca\ \epsilon\ na\ k-esima\ var. \end{cases} \tag{3.31}$$

Todavia, algumas vezes uma determinada variável não é observada em ambos os indivíduos. Por isto, adicionalmente à definição anterior, Gower (1971) também utilizou a seguinte:

Definição 3.7: A possibilidade de se fazer comparações com relação a k-ésima variável é representada por uma quantidade denotada por γ_{ijk}, e definida como:

$$\gamma_{ijk} = \begin{cases} 0, & se\ a\ var.\ X_k\ não\ pode\ ser\ comparada\ entre\ o\ i-esimo\ e\ \ j-esimo\ indiv.. \\ 1, & se\ a\ var.\ X_k\ pode\ ser\ comparada\ entre\ o\ i-esimo\ e\ \ j-esimo\ indiv.. \end{cases} \tag{3.32}$$

quando $\gamma_{ijk} = 0$, convenciona-se, $s_{ijk} = 0$, pois o mesmo é desconhecido.

Finalmente, o *Coeficiente de Similaridade de Gower* entre o i-ésimo e o j-ésimo indivíduo é definido conforme segue:

Definição 3.8: O *Coeficiente de Similaridade de Gower* entre o i-ésimo e o j-ésimo indivíduo, denotado por s_{ij}, é dado pelo escore médio com relação a todas as comparações possíveis, ou seja,

$$s_{ij} = \frac{\sum_{k=1}^{q} s_{ijk}}{\sum_{k=1}^{q} \gamma_{ijk}}. \tag{3.33}$$

Quando $\gamma_{ijk} = 0, \forall k = 1, \ldots, q$, s_{ij} é indefinido, e quando todas as comparações são possíveis $\sum_{k=1}^{q} \gamma_{ijk} = q$ (número total de variáveis). Caso contrário, este somatório é igual ao número de variáveis sobre as quais a comparação é feita.

Uma forma alternativa, mas exatamente análoga a (3.33), é escrever a similaridade na forma de uma média ponderada, ou seja,

$$s_{ij} = \frac{\sum_{k=1}^{q} \gamma_{ijk} s_{ijk}}{\sum_{k=1}^{q} \gamma_{ijk}} . \tag{3.34}$$

Os escores para s_{ijk} são designados de acordo com os seguintes casos:

- **Variáveis dicotômicas:** A presença da variável é denotada por $+$ e a ausência por $-$, estando os escores dispostos da seguinte maneira:

Indivíduo\ Valor da Variável				
i-ésimo indivíduo	$+$	$+$	$-$	$-$
j-ésimo indivíduo	$+$	$-$	$+$	$-$
s_{ijk}	1	0	0	0
γ_{ijk}	1	1	1	0

- **Variáveis qualitativas:** $s_{ijk} = 1$, se o i-ésimo e o j-ésimo indivíduo são iguais na k-ésima variável e $s_{ijk} = 0$, se eles diferem.

- **Variáveis contínuas:** Para variáveis quantitativas com valores $x_1, x_2, \ldots, x_n$ da k-ésima variável, na amostra de n indivíduos, consideramos $s_{ijk} = 1 - \frac{|x_i - x_j|}{R_k}$, onde R_k é a amplitude populacional ou amostral, da k-ésima variável. Vale salientar que, quando $x_i = x_j$, temos $s_{ijk} = 1$ e quando x_i e x_j são valores extremos da amplitude temos que s_{ijk} é mínimo (será zero quando tomarmos R_k como a amplitude amostral).

Podemos observar que s_{ij} varia entre 0 e 1, onde o valor 1 indica que os 2 indivíduos não diferem em nenhuma variável enquanto que 0 indica que eles diferem ao máximo, ou seja, em todas as variáveis. Gower (1971) mostra que a matriz de similaridades $\mathbf{S}$, cujos elementos s_{ij} são os coeficientes de similaridade entre o i-ésimo e o j-ésimo indivíduo. $\forall i, j = 1, 2, \ldots, n$, é positiva semi-definida.

Relação com outros coeficientes de similaridade

Uma situação particular ocorre quando toda informação é do tipo presença (+) ou ausência (−), ou no caso de variáveis categóricas de 2 níveis. Para este caso, vamos especificar a seguinte notação:

$$\begin{cases} \text{a: \# de variáveis com escore presença em ambos os indivíduos,} \\ \text{b: \# de variáveis com presença no indiv. 1 e ausência no indiv. 2,} \\ \text{c: \# de variáveis com ausência no indiv. 1 e presença no indiv. 2,} \\ \text{a: \# de variáveis ausentes em ambos os indivíduos.} \end{cases} \tag{3.35}$$

Feita esta especificação, temos a seguinte associação:

Indivíduo 1\Indivíduo 2	+	−	Totais
+	a	b	a + b
−	c	d	c + d
Totais	a + c	b + d	a + b + c + d

Vale salientar que, para este caso de variáveis dicotômicas, o *Coeficiente de Similaridade de Gower* torna-se o *Coeficiente de Similaridade de Jaccard*, proposto em Jaccard (1901), e dado na seção anterior. Caso todas as variáveis sejam categóricas e possuam apenas 2 níveis, temos o *Coeficiente de Coincidência simples*, dado em Sokal e Michener (1958), e também definido na seção anterior. Finalmente, caso todas as variáveis sejam contínuas, teremos o seguinte coeficiente de similaridade:

$$S_{ij} = 1 - \frac{\left(\sum_{k=1}^{q} \mid x_{ik} - x_{jk} \mid / R_k \right)}{q}. \tag{3.36}$$

A definição da distância a partir do Coeficiente de Similaridade de Gower

Como o *Coeficiente de Similaridade de Gower* possui intervalo de valores entre 0 e 1, Johnson e Wishern (1982) sugerem transformá-lo em um coeficiente de dissimilaridade

usando a seguinte definição:

Definição 3.9: A *distância* entre o i-ésimo e o j-ésimo indivíduo, obtida a partir do *Coeficiente de Similaridade de Gower* s_{ij}, é denotada por d_{ij} e definida como:

$$d_{ij} = 1 - S_{ij}.$$

Do ponto de vista intuitivo, para um coeficiente que assume valores entre 0 e 1, esta transformação é adequada dado que tanto para transformar uma distância em similaridade como no caso inverso, teremos ambos os coeficientes variando neste intervalo e, enquanto um diminui em magnitude, o outro aumenta.

3.4.4 Distâncias Obtidas por Transformação de Variáveis

Outro artifício bastante usado para se obter uma quantificação da distância existente entre indivíduos, no contexto de *misturas*, é a transformação das variáveis existentes. Porém, segundo Bussab, Miazaki e Andrade (1990), é notório que este processo induz à perda de informação. Vamos abordar a seguir as duas transformações de variáveis mais sugeridas na literatura.

Distância Obtida por Transformação de Todas as Variáveis em Binárias

Este coeficiente de dissimilaridade (ver também Bussab, Miazaki e Andrade, 1990) se baseia na conversão de todas as variáveis em binárias, e no uso de uma distância definida para este caso (ver exemplos na *seção 3.3.2*). Para facilitar o trabalho, vamos adotar a blocagem de (3.20).

Para transformar o bloco nominal $x^{(N)}$ em um vetor binário, podemos utilizar a *Definição 2.3*, sugerida por Talkington (1967). Já para transformar o bloco $x^{(O)}$ em um vetor binário podemos utilizar a *Definição 2.4*, sugerida por Sokal e Sneath (1963). Para o bloco quantitativo $x^{(Q)}$, segundo Bussab, Miazaki e Andrade (1990), podemos definir zero para todos os valores abaixo da mediana e 1 para os iguais ou maiores.

Distância Obtida pela Transformação de Todas as Variáveis em Outras Assumindo Valores entre [0,1]

Nesta transformação, também sugerida por Bussab, Miazaki e Andrade (1990), a idéia é transformar todas as variáveis de modo que a amplitude de seus valores fique entre 0 e 1 e, em seguida, usar a *Distância Euclidiana Padronizada* (ver *Definição 3.5, c)* ou a usual (ver *Definição 3.5, a)*.

Para transformar o bloco nominal $x^{(N)}$ em variáveis assumindo valores entre $[0,1]$, podemos utilizar a *Definição 2.3*, ou seja, transformar em variáveis binárias. Caso as variáveis sejam binárias não será necessário fazer esta transformação. Para o bloco ordinal $x^{(O)}$ podemos utilizar a *Definição 2.4*, ou seja, também transformar em binária. Já para o bloco quantitativo $x^{(Q)}$ podemos fazer a seguinte transformação:

$$y_i = \frac{x_i^{(O)} - x_{\min}^{(O)}}{x_{\max}^{(O)} - x_{\min}^{(O)}} . \tag{3.37}$$

3.4.5 Distância de Romesburg

Romesburg (1984) propõe a mais informal de todas as distâncias apresentadas até aqui. Utilizando uma justificativa empírica, ele sugere a aplicação da *Distância Euclideana Usual* para quantificar as dissimilaridades existentes entre os indivíduos, sem levar em conta a natureza das variáveis, mas exigindo que todas possuam escores numéricos. Ele afirma que este procedimento tem a capacidade de produzir grupos semelhantes, e não muito diferentes daqueles obtidos usando distâncias mais apropriadas ao contexto de *misturas*, quando da aplicação de *Técnicas de Agrupamento*.

A grande dificuldade de se utilizar esta distância, para o caso de *misturas*, é a interpretação dos valores obtidos, pois a recodificação das variáveis nominais leva a valores distintos da distância.

Capítulo 4

Técnicas Hierárquicas de Agrupamento: Aplicação às Misturas

4.1 Resumo do Capítulo

Este capítulo tem a finalidade de apresentar uma breve discussão sobre as *Técnicas Hieráquicas de Agrupamento* e descrever alguns métodos que aceitam a *extensão da Distância de Mahalanobis para misturas* como coeficiente de dissimilaridade, não se tratando, contudo, de um texto introdutório às *Técnicas Hierárquicas de Agrupamento*. Na *seção 4.2* discutimos alguns aspectos importantes à utilização da *Análise de Agrupamento*, tais como princípios, metodologia, principais técnicas relacionadas e classificação. *Na seção 4.3* introduzimos uma abordagem mais formal às *Tecnicas Hierárquicas Aglomerativas*. Finalmente, na *seção 4.4*, a discussão fica restrita às *Técnicas Hierárquicas Aglomerativas para misturas*. Nesta última seção é apresentado um algoritmo geral para às três *Técnicas Hierárquicas Aglomerativas* mais utilizadas e ao estudo de um algoritmo geral.

4.2 Aspectos Importantes

Para melhor entendimento das técnicas de agrupamento que serão introduzidas na próxima seção, vamos abordar aqui, de forma sucinta, alguns pontos que são bastante importantes a esta compreensão. Para um estudo mais abrangente sobre análise de agrupamento, são várias as referências existentes. Dentre elas, podemos destacar Cormack (1971), Anderberg (1973), Everitt (1974), Spath (1980), Bussab, Miazaki e Andrade (1990) e Pereira (1993), entre outros.

4.2.1 Princípios Gerais da Análise de Agrupamento

Geralmente, o objetivo principal da *Análise de Agrupamento* é separar indivíduos em grupos, de modo que haja homogeneidade dentro dos grupos e heterogeneidade entre os grupos, ou seja, indivíduos de um mesmo grupo apresentem características homogêneas, e indivíduos de grupos diferentes apresentem características heterogêneas. Porém, também é possível agrupar variáveis segundo os valores obtidos pelos indivíduos, ou seja, utilizar a *Análise de Agrupamento* como uma técnica de redução de dados. Apesar disto, não existe uma definição formal, aceita por todos, do que seja um *grupo* ou um *agrupamento*, pois a mesma envolve conceitos subjetivos.

A aplicação da *Análise de Agrupamento* é bastante difundida na literatura, sendo usada para determinar indivíduos semelhantes num estágio inicial de um esquema de amostragem estratificada, para a formulação de hipóteses sobre a estrutura dos dados, e para a determinação de esquemas de classificação, entre outros. Devido a sua grande aplicação em diversas áreas, a *Análise de Agrupamento* também é conhecida por denominações diferentes, tal como *taxonomia numérica*, na biologia.

Por ser uma técnica exploratória que contém um alto grau de subjetividade, é de conhecimento geral que nenhum dos métodos existentes é o mais adequado para resolver qualquer problema de agrupamento. Por isto, muitas vezes os pesquisadores preferem definir um novo método. Por exemplo, Ferreira e Medeiros (1992) propõem um método de

agrupamento não hierárquico baseado em uma alteração do método *Isodata* proposto por Ball e Hall (1965). Um outro exemplo, também na linha não hierárquica, é o algorítmo de agrupamento proposto por Uchiyama e Arbib (1994). Ralambondrainy (1988) propõe, de forma teórica, um método de agrupamento não hierárquico para *misturas* de dados numéricos e nominais porém, não apresenta um algorítmo à sua implementação computacional.

A principal consequência da grande variedade de métodos, é que a escolha de um particular algorítmo exige o conhecimento de suas propriedades aliado aos objetivos da pesquisa que está sendo realizada. Na prática, geralmente, aplica-se vários métodos e, com base nas configurações obtidas e na opinião do pesquisador responsável, escolhe-se o agrupamento mais adequado ao problema em questão.

Com relação a implementação computacional dos algorítmos existentes, Spath (1980) traz programas, em *Fortran*, para algumas das principais técnicas de agrupamento. Devido a constante proposição de novos algorítmos de agrupamento e a falta de implentação computacional dos mesmos, nos pacotes existentes, Kaufman e Rousseuw (1990) trazem programas para alguns dos métodos não contemplados por Spath (1980).

4.2.2 Metodologia para a Obtenção de um Agrupamento

Para se ter um procedimento metodológico útil à obtenção de um agrupamento de indivíduos em um conjunto de dados, é necessário seguir um certo *ritual*. Podemos dividí-lo nas seguintes etapas:

a) Definição dos objetivos e seleção das variáveis,

b) Obtenção e preparação dos dados para análise,

c) Escolha do critério de parecença (similaridade ou dissimilaridade) a ser utilizado,

d) Escolha do algorítmo de agrupamento e,

e) Avaliação e interpretação do agrupamento obtido.

Inicialmente, antes de aplicar um algoritmo de agrupamento, é importante *definir os objetivos* dessa aplicação, pois eles terão uma influência decisiva nas etapas subsequentes do processo de agrupamento. Além disso, também é necessário *escolher as variáveis* que melhor caracterizem os indivíduos.

A *obtenção dos dados* é a fase seguinte porém, logo após a sua obtenção, é necessário *prepará-los* de forma adequada para a análise. Quando as variáveis mensuradas possuem unidades diferentes, é comum padronizá-las. Contudo, este procedimento pode diluir as diferenças existentes entre grupos com relação as variáveis que melhor os separam.

A etapa **c)** é de fundamental importância, pois o *coeficiente de parecença* (similaridade ou dissimilaridade) escolhido é que irá determinar o quão próximo ou o quão distante estão os indivíduos, e estas mensurações constituem a base para a aplicação das técnicas de agrupamento. A escolha deste coeficiente deve levar em consideração os tipos de variáveis envolvidas e, como trabalhamos com *misturas*, utilizaremos uma *extensão da Distância de Mahalanobis*, como discutido na *seção 3.5*.

A etapa **d)** trata da *escolha da técnica de agrupamento* a ser utilizada. Nesta etapa é preciso levar em consideração as características de cada método e os objetivos de se agrupar os dados, para que os grupos formados tenham significado para a pesquisa em questão. Finalmente, aceitar os grupos produzidos por uma técnica de agrupamento não é um procedimento correto pois, na maioria dos casos, os métodos sempre irão produzir grupos, independentemente da existência ou não dos mesmos. Portanto, a *avaliação e interpretação dos grupos formados* é um passo importante para a consolidação do processo de agrupamento, e o conhecimento do problema, por parte do pesquisador, é fundamental para classificar a estrutura de grupos encontrada como proveitosa ou não. Quando da utilização de métodos hierárquicos, a determinação do número de grupos também está incluída nesta etapa, visto que estes métodos apresentam uma árvore de agrupamentos (dendrograma) como configuração final. Cabe ao pesquisador determinar o número mais adequado de grupos, de acordo com o seu conhecimento do problema e da configuração apresentada pelo dendrograma.

Vale salientar que estas etapas não são interdependentes, e as vezes se torna necessário voltar às etapas anteriores para aprimorar as posteriores.

4.2.3 Principais Técnicas Relacionadas

Como mencionado anteriormente, a *Análise de Agrupamento* também pode ser utilizada para redução de dados. Para tanto, basta que seja aplicada com relação as variáveis e não com relação aos indivíduos, ou seja, deseja-se agrupar variáveis para reduzir a quantidade das mesmas. Neste contexto, esta técnica está relacionada com a *Análise de Componentes Principais*, que é outra técnica utilizada para redução de dados onde, nesta última, o objetivo é reduzir o conjunto original das variáveis em um conjunto de combinações lineares ortogonais das mesmas. Vale salientar que as *técnicas de agrupamento* usadas para agrupar variáveis são diferentes das usadas para agrupar indivíduos (ver SAS, 1989).

No contexto de *classificação*, ou seja, de encontrar o melhor procedimento para classificar n indivíduos em k populações homogêneas, com base em p variáveis observáveis, a *Análise de Agrupamento* está relacionada com a *Análise Discriminante* (ver Gnanadesikan e Kettenring, 1989). Se as categorias para classificação são conhecidas *a priori*, a *Análise Discriminante* dá a solução para o problema geral de *classificação*. Porém, se estas categorias são geradas dos dados, a *Análise de Agrupamento* é a técnica adequada. Devido a grande associação entre estas técnicas, utiliza-se o termo *classificação* tanto para um caso, quanto para o outro.

4.2.4 Classificação das Técnicas de Agrupamento

Nas últimas décadas, com o advento do computador, tem aumentado o interesse por técnicas de agrupamento. Consequentemente, houve um aumento no número de algoritmos disponíveis e melhoria em suas eficiências. Adotando a classificação usada por Cormack (1971), podemos dividir os métodos de agrupamento conforme segue:

a) **Técnicas hierárquicas** : Os grupos são constituídos de forma hierárquica, produzindo uma árvore de classificação. Esta hierarquia pode ser constituída de duas formas: Na primeira, chamada *Técnica Hierárquica Aglomerativa*, considera-se cada indivíduo como constituindo um grupo distinto e, por meio de uniões sucessivas, chega-se a um único grupo. Na segunda, chamada técnica hierárquica separativa, considera-se todos os indivíduos como formadores de um único grupo e, por meio de separações sucessivas, chega-se a grupos formados por apenas um indivíduo.

b) **Técnicas de partição** : Os grupos obtidos produzem uma partição do conjunto de indivíduos, ou seja, são mutuamente exclusivos. O uso desses métodos pressupõe o conhecimento do número k de grupos desejados. Assim, o problema passa a ser a procura de uma partição dos indivíduos em k grupos, de modo que torne ótimo o critério de adequacidade da partição.

c) **Técnicas de cobertura** : O agrupamento obtido permite uma interseção entre grupos, ou seja, um indivíduo pode estar classificado em mais de um grupo.

Devido a maior difusão, simplicidade, variedade e implementação computacional em diferentes pacotes, vamos trabalhar com as *Técnicas Hierárquicas Aglomerativas* que admitem a utilização da *extensão da Distância de Mahalanobis para misturas*, como coeficiente de dissimilaridade entre indivíduos pois, como discutido no *Capítulo 3*, este é o coeficiente de dissimilaridade que vamos adotar nesta dissertação. É importante salientar a existência de trabalhos que fazem comparações entre diversos algoritmos de agrupamento, sob diferentes situações. Um dos mais recentes foi realizado por Pereira (1993) utilizando a estatística de Rand (ver Rand, 1971) para a avaliação dos métodos de agrupamento. Ao longo das últimas décadas vários trabalhos foram publicados nesta linha, dentre eles podemos destacar: Cunningham e Olgivie (1972), Kuiper e Fisher (1975), Dubes e Jain (1976), Bayne, Beauchamp, Begowich e Kane (1980), Milligan (1980) e Dubien e Warde (1987).

4.3 Definições Básicas em Técnicas Hierárquicas Aglomerativas

Apesar do grau de subjetividade inerente não só as *Técnicas Hierárquicas Aglomerativas* como também as outras técnicas de agrupamento, vários autores têm procurado dar uma formulação matemática maior a estes métodos, através da formalização dos conceitos existentes. Entre os trabalhos publicados nesta linha, podemos destacar o de Dubien e Warde (1979). De acordo com a formalização apresentada por eles, daremos agora alguns conceitos importantes às *Técnicas Hierárquicas Aglomerativas*.

Inicialmente, de acordo com a notação estabelecida no *Apêndice A*, vamos assumir que cada um dos n objetos (indivíduos) a serem agrupados estão representados por um vetor linha p-dimensional $\mathbf{x}_{(i)}$ onde,

$$\mathbf{x}_{(i)} = (x_{i1}, x_{i2}, \ldots, x_{ip}),\ i = 1, 2, \ldots, n,$$

tal que x_{ij}, $j = 1, 2, \ldots, p$, indica o valor observado no i-ésimo objeto (indivíduo) com relação a j-ésima variável. A partir de agora, vamos nos referir a indivíduos como objetos.

Definição 4.1: O conjunto formado por todos os objetos, denotado por $\mathbf{X}$, onde $\mathbf{X} = (\mathbf{x}'_{(1)}, \mathbf{x}'_{(2)}, \ldots, \mathbf{x}'_{(n)})'$ é chamado espaço dos objetos.

Com base na teoria de conjuntos, Dubien e Warde introduzem as seguintes definições matemáticas para *grupo* e *agrupamento*.

Definição 4.2: Um *grupo* ("cluster") Y_k, é qualquer sub-conjunto não-vazio do espaço dos objetos $\mathbf{X}$. Simbolicamente, $Y_k \subseteq \mathbf{X}$, o que indica que se $\mathbf{x}_{(i)} \in Y_k$, então $\mathbf{x}_{(i)} \in \mathbf{X}$.

Definição 4.3: Um *agrupamento* ("clustering") Y, é qualquer partição do espaço dos objetos. Simbolicamente, $Y = \{Y_1, Y_2, \ldots, Y_k\}$ é uma partição de $\mathbf{X}$, se as seguintes condições forem válidas:

a) Para cada $Y_k \in Y$, $Y_k \neq \emptyset$.

b) Se $Y_k \in Y$, $Y_m \in Y$ e $Y_k \neq Y_m$, então $Y_k \cap Y_m = \emptyset$,

c) $\bigcup_{i=1}^{k} Y_k = \mathbf{X}$.

Pelas *Definições 4.2 e 4.3*, podemos observar que um *grupo* é simplesmente uma coleção de objetos, enquanto um *agrupamento* é um tipo especial de coleção de grupos. Além das definições dadas acima, Dubien e Warde também apresentam as seguintes:

Definição 4.4: O número de grupos contidos em um agrupamento Y, é chamado de *tamanho do agrupamento* Y. Para facilitar a notação, se o agrupamento Y é de tamanho k, ou seja, Y é composto de k grupos, passamos a denotá-lo por Y^k.

Definição 4.5: Uma *hierarquia*, denotada por H, sobre o espaço dos objetos $\mathbf{X}$, é uma sequência de agrupamentos ordenada de forma decrescente com relação ao tamanho. Simbolicamente, temos

$$H : Y^n, Y^{n-1}, \ldots, Y^2, Y^1, \text{ onde } Y^n \subset Y^{n-1} \subset \ldots \subset Y^1.$$

Para melhor visualização da *hierarquia* imposta sobre o espaço dos objetos, durante a aplicação de uma *Técnica Hierárquica de Agrupamento*, é utilizado um diagrama em forma de árvore denominado *dendrograma*. Vale salientar que a aplicação de um *Método de Agrupamento Hierárquico Aglomerativo* sobre o espaço dos objetos $\mathbf{X}$, resulta num tipo especial de *hierarquia* que é dada na definição abaixo:

Definição 4.6: Um *Método de Agrupamento Hierárquico Aglomerativo* é qualquer método de agrupamento que produza uma hierarquia, sobre o espaço dos objetos, sujeita às seguintes condições:

a) Y^k é o agrupamento inicial.

b) O agrupamento Y^{k-1}, $k \leq n$, é obtido pela união dos dois grupos mais próximos no agrupamento Y^k, ou seja, se Y_i e $Y_j \in Y^k$ e são considerados os grupos mais próximos, então $Y_i \cup Y_j \in Y^{k-1}$.

Ainda segundo Dubien e Warde, podemos denotar um *Método de Agrupamento Hierárquico Aglomerativo* pelo par (coeficiente de distância, algorítmo de agrupamento), pois o coeficiente de distância explica a proximidade inicial entre os grupos, e o algorítmo de agrupamento reavalia a proximidade dos grupos depois de cada união.

4.4 Técnicas Hierárquicas Aglomerativas para Misturas

4.4.1 Procedimento Geral

De acordo com Anderberg (1973), a maioria dos *Métodos Hierárquicos Aglomerativos*, incluindo-se os que serão descritos neste capítulo, seguem um procedimento geral que se baseia nos seguintes passos:

Passo 1: Supondo um conjunto de dados composto por n indivíduos, assumimos que existem $k = n$ grupos distintos $Y_1, Y_2, \ldots, Y_k$, onde cada indivíduo existente está associado a um grupo diferente.

Passo 2: Adotamos um critério de parecença entre dois grupos, que na primeira iteração corresponde à parecença entre dois indivíduos, para podermos ter uma mensuração da proximidade ou da diferença entre eles. Em seguida, calculamos este coeficiente entre todos os grupos. Vale salientar que, nesta dissertação, trabalhamos com distâncias como coeficientes de parecença, ou seja, a parecença entre dois grupos está em função das distâncias existentes entre os indivíduos de cada um deles.

Passo 3: Determinamos os grupos Y_a e Y_b onde

$$d_{Y_a Y_b} = \min_{i \neq j;\ i,j \in \{Y_1, Y_2, \ldots, Y_k\}} \{d_{ij}\}. \tag{4.1}$$

e os unimos de modo que formem o grupo $Y_l = Y_a \cup Y_b$. Para facilitar a implementação computacional, assumimos $Y_b = Y_l$ e $Y_a = \{\emptyset\}$ e fazemos $k = k - 1$, visto que um grupo deixou de existir.

Passo 4: Se obtemos $k = 1$, paramos o processo de agrupamento pois teremos chegado a um único grupo. Caso contrário, calculamos as distâncias entre o novo grupo e os demais, e voltamos ao *Passo 3*.

Vale salientar ainda que, após o *passo 2*, são calculadas apenas as distâncias entre o mais novo grupo, formado no passo 3, e os grupos restantes. Isto acontece porque entre uma iteração e outra, as distâncias entre os grupos que não sofreram alterações não mudam.

A diferença entre os métodos aglomerativos que seguem este algorítmo está no coeficiente de parecença adotado (aqui assumido como distância), *Passo 2*, e no cálculo da distância entre o grupo formado em uma determinada iteração e os outros grupos, *Passo 4*. Para os três métodos que serão descritos neste capítulo, a diferença está basicamente na distância adotada entre grupos, visto que os mesmos podem trabalhar com diferentes coeficientes de distância. Em função da proximidade entre estes algorítmos, vamos apresentar a seguir um algoritmo geral que assume, como casos particulares, os três métodos que serão descritos mais adiante.

4.4.2 O Algorítmo Geral de Lance e Williams

Definição

Em função da maioria dos *Métodos Hierárquicos Aglomerativos* diferirem, basicamente, com respeito ao cálculo das distâncias entre um grupo formado em uma determinada

iteração e os outros grupos, Lance e Williams (1967) propõe um *Algorítmo Hierárquico Aglomerativo Geral* baseado em uma forma generalizada de calcular estas distâncias. Este algorítmo assume, como casos particulares, alguns dos *Métodos Hierárquicos Aglomerativos* mais conhecidos da literatura, incluindo os três métodos que serão descritos aqui. Neste algorítmo, a distância entre dois grupos Y_a e Y_b, onde Y_b é fruto da união de dois outros grupos Y_i e Y_j, é definida como:

Definição 4.7: Sejam Y_a e Y_b dois grupos distintos na iteração associada ao agrupamento Y^{k-1}, ou seja, $Y_a \subset Y^{k-1}$ e $Y_b \subset Y^{k-1}$. Além disso, $Y_b = Y_i \cup Y_j$ é composto pelos dois grupos unidos na iteração anterior do processo hierárquico de agrupamento, ou seja, $Y_i \subset Y^k$ e $Y_j \subset Y^k$. A *distância entre os grupos* Y_a *e* Y_b, segundo Lance e Williams (1967), é dada por:

$$\partial_{Y_a Y_b} = \alpha_i \partial_{Y_a Y_i} + \alpha_j \partial_{Y_j Y_a} + \beta \partial_{Y_i Y_j} + \gamma \mid \partial_{Y_i Y_a} - \partial_{Y_j Y_k} \mid, \tag{4.2}$$

onde para α_i, β e γ particulares, obtemos alguns dos *Métodos Hierárquicos Aglomerativos* mais conhecidos da literatura.

Isto significa que na iteração associada ao agrupamento Y^{k-1}, $\forall k \in \{2, 3, \ldots, n\}$, podemos obter as distâncias entre os grupos $Y_b = Y_i \cup Y_j$ e Y_a simplesmente pela combinação das distâncias entre os grupos Y_i, Y_j e Y_a existentes na iteração anterior, associada ao agrupamento Y^k. Isto traz vantagens na hora da implementação computacional.

Sibson (1971) argumenta que o algorítmo geral de agrupamento de Lance e Williams falha por não satisfazer as sete propriedades introduzidas por Jardine e Sibson (1968), e consideradas por eles como essenciais à admissibilidade de um algorítmo de agrupamento. Porém, vários autores (ver Williams, Lance, Dale e Clifford, 1971) consideram estas condições como muito severas e sem muito sentido prático.

Propriedades Essenciais à Conduta Computacional de um Método Hierárquico

Além de propor um algorítmo geral de agrupamento, Lance e Williams mostram que a conduta computacional de um *Método Hierárquico* depende de três propriedades. Estas propriedades são as seguintes:

a) **Estratégia combinatória ou não-combinatória:** Um *Algorítmo Hierárquico* é dito possuir uma *estratégia combinatória* se, após unir dois grupos em uma determinada iteração, as distâncias entre este novo grupo e os demais puderem ser calculadas em função das distâncias obtidas na iteração anterior. Na prática, isto equivale a dizer que não é preciso voltar aos dados originais depois que o primeiro conjunto de distâncias foi calculado. Em contra-partida, ele é dito ter uma *estratégia não-combinatória* se, para obter as distâncias entre dois grupos em uma determinada iteração, for necessário voltar aos dados originais.

b) **Compatível ou incompatível:** Um método é dito compatível se as distâncias entre os grupos, calculadas nas diversas iterações do algorítmo de agrupamento, forem do mesmo tipo das calculadas no estágio inicial. Além disso, elas possuem as memas dimensões, são sujeitas as mesmas restrições e podem ser ilustradas por modelos exatamente comparáveis. Caso contrário, é dito incompatível.

c) **Distorcedor de espaço ou conservador de espaço:** Um método *distorcedor de espaço* é sub-dividido em *contrator de espaço* e *dilatador de espaço*. Ele é dito *contrator de espaço* se apresentar uma tendência maior a aproximar os objetos a grupos já existentes, ao invés de tomá-los como base para a formação de novos grupos. Em contra-partida, ele é dito *dilatador de espaço* se apresentar uma tendência a fazer com que os objetos que ainda não foram agrupados, sirvam como base para formação de novos grupos. Caso nenhum desses dois casos seja verificado, ele é dito *conservador de espaço*.

Com relação a primeira propriedade, podemos verificar que a *estratégia combinatória* traz grandes vantagens computacionais, visto que ao passar de uma iteração à outra,

no processo de agrupamento, as distâncias entre o mais novo grupo e um outro grupo qualquer pode ser obtida da combinação de três distâncias do passo anterior. Isto significa que não é necessário utilizar todas as distâncias existentes entre os indivíduos que compõe os dois grupos, ou seja, proporciona um número menor de operações computacionais.

Na propriedade b), segundo Lance e Williams (1967), a vantagem da estratégia ser compatível é que se as distâncias iniciais são interpretáveis, as distâncias entre os grupos formados, nos diversos estágios de aplicação do algoritmo, também o serão. Com relação a c), ser *contrator de espaço* significa, na prática, que um único objeto, mesmo entre dois grupos bem definidos, é o suficiente para uní-los. Esta tendência leva a formação de grupos com forma alongada, diferentes das usuais como elipsóides ou hiper-esferas. Lance e Williams consideram que um método *contrator de espaço* não é recomendável pois, os limites dos grupos existentes são encobertos. Porém, segundo Kopp (1978a), em algumas aréas estas formas podem ser bastante desejadas. Já ser *dilatador de espaço* significa que o algoritmo tende a formar um número de grupos maior, pois dificulta a união de objetos a grupos já existentes. Com isso, os grupos formados tendem a ser compactos e pequenos.

4.4.3 Uma Sub-família Bi-paramétrica de Algorítmos Hierárquicos Aglomerativos

Podemos perceber que a equação (4.2), introduzida por Lance e Williams (1967), define uma *família de Algorítmos Hierárquicos Aglomerativos de quatro parâmetros* que contempla um número infinito de algorítmos distintos (Lance e Williams apresentam 6 algorítmos conhecidos como casos particulares desta família). Com a finalidade de diminuir o número de parâmetros envolvidos e continuar a conter os principais *Métodos Hierárquicos Aglomerativos* como casos particulares, Dubien e Warde (1979) obtiveram uma *sub-família bi-paramétrica* através de restrições nos parâmetros de (4.2). Além disso, eles apresentam formalmente, para esta *sub-família*, algumas propriedades abordadas de maneira informal por Lance e Williams para a *família de Algorítmos Hierárquicos*

Aglomerativos de quatro parâmetros.

Definição

Para definir esta *sub-família bi-paramétrica*, Dubien e Warde utilizaram as seguintes restrições sobre os parâmetros α_i, α_j, β e γ de (4.2).

$$\alpha_i = \alpha_j = \alpha. \tag{4.3}$$

$$\alpha_i + \alpha_j + \beta = 1. \tag{4.4}$$

A partir destas restrições, temos que,

$$\alpha_i = \alpha_j = \frac{1}{2}(1 - \beta). \tag{4.5}$$

Com isso, alguns membros da *família de 4 paramétros* introduzida por Lance e Williams, cujos valores dos parâmetros satisfazem (4.3) e (4.4), podem ser representados por uma *sub-família bi-paramétrica* (β, γ). Dessa forma, a equação (4.2) introduzida por Lance e Williams, torna-se,

$$d_{Y_a Y_b} = \frac{1}{2}(1 - \beta)d_{Y_a Y_i} + \frac{1}{2}(1 - \beta)d_{Y_j Y_a} + \beta d_{Y_i Y_j} + \gamma \mid d_{Y_i Y_a} - d_{Y_j Y_k} \mid . \tag{4.6}$$

Sem perda de generalidade, Dubien e Warde assumiram que,

$$d_{Y_i Y_j} < d_{Y_a Y_i} < d_{Y_j Y_a}, \tag{4.7}$$

e, utilizando (4.7) e (4.6) deram a seguinte definição:

Definição 4.8: Sejam Y_a e Y_b dois grupos distintos na iteração associada ao agrupamento Y^{k-1}, ou seja, $Y_a \subset Y^{k-1}$ e $Y_b \subset Y^{k-1}$. Além disso, $Y_b = Y_i \cup Y_j$ é composto pelos dois grupos unidos na iteração anterior, ou seja, $Y_i \subset Y^k$ e $Y_j \subset Y^k$. A *distância entre os grupos* Y_a e Y_b, para os algorítmos pertencentes à *sub-família*

bi-paramétrica (β, γ), é dada por:

$$\partial_{Y_a Y_b} = \frac{1}{2}(1-\beta+2\gamma)\partial_{Y_j Y_a} + \frac{1}{2}(1-\beta-2\gamma)\partial_{Y_i Y_a} + \beta\partial_{Y_i Y_j}, \text{ onde } \partial_{Y_i Y_j} < \partial_{Y_i Y_a} < \partial_{Y_j Y_a}. \quad (4.8)$$

Dentre os *Métodos Hierárquicos Aglomerativos* pertencentes a *sub-família bi-paramétrica* (β, γ) podemos destacar o da *Ligação Simples*, o da *Ligação Completa* e o da *Ligação Média* em sua versão *equiponderada*.

Algumas Propriedades da Sub-família Bi-paramétrica (β, γ)

Para possibilitar que as propriedades apresentadas, de forma intuitiva, por Lance e Williams (1967) sejam checadas quando da aplicação de um *Algorítmo Hierárquico Aglomerativo* pertencente a *sub-família bi-paramétrica* (β, γ), Dubien e Warde apresentam as seguintes definições:

Definição 4.9: Um algorítmo pertencente a *sub-família bi-paramétrica* (β, γ) é *monótono crescente* se, e só se, $\forall \partial_{Y_a Y_b}$ definida por (4.8), $\partial_{Y_a Y_b} > \partial_{Y_i Y_j}$, $i, j \in b$.

Definição 4.10: Um algorítmo pertencente a *sub-família bi-paramétrica* (β, γ) é *conservador de espaço* se, e só se, $\forall \partial_{Y_a Y_b}$ definida por (4.8), $\partial_{Y_i Y_a} < \partial_{Y_a Y_b} < \partial_{Y_j Y_a}$, $i, j \in b$.

Definição 4.11: Um algorítmo pertencente a *sub-família bi-paramétrica* (β, γ) é *contrator de espaço* se, e só se, $\forall \partial_{Y_a Y_b}$ definida por (4.8), $\partial_{Y_a Y_b} \leq \partial_{Y_i Y_a}$, $i, j \in b$.

Definição 4.12: Um algorítmo pertencente a *sub-família bi-paramétrica* (β, γ) é *dilatador de espaço* se, e só se, $\forall \partial_{Y_a Y_b}$ definida por (4.8), $\partial_{Y_a Y_b} \geq \partial_{Y_j Y_a}$, $i, j \in b$.

Classificação para os Algorítmos Pertencentes a Sub-família Bi-paramétrica (β, γ)

Considerando que os algorítmos pertencentes a *sub-família bi-paramétrica* (β, γ) podem ser representados como um ponto no plano (β, γ), e utilizando as **Definições 4.9** a

4.12. Dubien e Warde classificam estes algorítmos em cinco classes distintas:

a) **Conservador de Espaço:** Os algorítmos classificados nesta classe pertencem a região do plano (β, γ) definida por,

$$R_1 = \{\beta = 0\} \cap \{\frac{1}{2}(\beta - 1) < \gamma < \frac{1}{2}(1 - \beta)\}. \tag{4.9}$$

b) **Contrator de Espaço:** Os algorítmos classificados nesta classe pertencem a região do plano (β, γ) definida por,

$$R_2 = [\{0 < \beta \leq 1\} \cap \{\frac{1}{2}(\beta - 1) < \gamma \leq \frac{1}{2}(\beta + 1)\}] \cup [\{0 \leq \beta < 1\} \cap \{\gamma = \frac{1}{2}(\beta - 1)\}]. \tag{4.10}$$

c) **Dilatador de Espaço:** Os algorítmos classificados nesta classe pertencem a região do plano (β, γ) definida por,

$$R_3 = [\{\beta < 0\} \cap \{\gamma \geq \frac{1}{2}(\beta - 1)\}] \cup [\{\beta = 0\} \cap \{\gamma \geq \frac{1}{2}(\beta + 1)\}]. \tag{4.11}$$

d) **Dilatador e Contrator de Espaço:** Os algorítmos classificados nesta classe pertencem a região do plano (β, γ) definida por,

$$R_4 = [\{\beta > 0\} \cap \{\gamma > \frac{1}{2}(\beta + 1)\}]. \tag{4.12}$$

e) **Contrator de Espaço ao Extremo:** Os algorítmos classificados nesta classe pertencem a região do plano (β, γ) definida por,

$$R_5 = [\bigcup_{i=1}^{4} R_i]^c. \tag{4.13}$$

4.4.4 Requisitos às Técnicas Hierárquicas Aglomerativas Abordadas

Como nosso objetivo é obter um agrupamento em indivíduos mensurados sob o contexto de *misturas*, e levando-se em consideração que nesse caso são necessários alguns cuidados especiais, como discutido nos *Capítulos 2 e 3*, as técnicas de agrupamento que serão descritas neste capítulo satisfazem as seguintes pré-suposições:

a) Ser uma *Técnica Hierárquica Aglomerativa*, pois como argumentado na *seção 4.2.4*, são as técnicas que mais se adequam ao presente trabalho.

b) Admitir a *extensão da Distância de Mahalanobis para misturas* como critério de dissimilaridade entre indivíduos, pois como discutido na *seção 3.2*, trabalhamos com este critério para o contexto de *misturas*.

c) Estar implementada computacionalmente na maioria dos softwares estatísticos. Em particular, no *SAS*, que é um dos pacotes mais reconhecidos mundialmente em se tratando não só de *Análise de Agrupamento* mas, de tratamento de dados em geral.

Mesmo pré-estabelecidos estes pontos, nosso objetivo não é o de discutir todos os *Métodos Hierárquicos Aglomerativos* que os satisfazem nem tampouco fazer comparação entre eles, visto que isto tornaria este trabalho muito extenso. Vamos apresentar aqui apenas os que achamos mais apropriados devido a simplicidade e a implementação computacional já existente na maioria dos pacotes. Outros métodos podem ser vistos em Lucas (1982) e Pereira (1993).

4.4.5 Método da Ligação Simples ("Single Linkage")

Definição

Este método, também conhecido como *Método do Vizinho mais Próximo*, foi proposto originalmente por Florek, Lukaszewisz, Perkal e Zubrzycki (1951) e depois foi revisado por

McQuity (1957), sendo um dos algorítmos mais antigos, mais simples e mais utilizados da literatura. Como já dissemos anteriormente, basicamente o que varia de um algorítmo hierárquico aglomerativo para outro é a definição de distância entre dois grupos. Com relação ao *Método da Ligação Simples*, esta definição é dada abaixo:

Definição 4.13: Sejam Y_a e Y_b dois grupos distintos na iteração associada ao agrupamento Y^{k-1}, ou seja, $Y_a \subset Y^{k-1}$ e $Y_b \subset Y^{k-1}$. Além disso, $Y_b = Y_i \cup Y_j$ é composto pelos dois grupos unidos na iteração anterior do processo hierárquico de agrupamento, ou seja, $Y_i \subset Y^k$ e $Y_j \subset Y^k$. A distância entre os grupos Y_a e Y_b, sob o *Método da Ligação Simples*, é dada por:

$$d_{Y_aY_b} = \min_{i \in Y_a;\ j \in Y_b} d_{ij}. \tag{4.14}$$

Como mostrado por Lance e Williams (1967), a distância entre o grupo Y_b, proveniente da união dos grupos Y_i e Y_j numa iteração anterior, e um outro grupo Y_a qualquer, da presente iteração, pode ser obtida como um caso particular de (4.2), ou seja, se $Y_b = Y_i \cup Y_j$, onde $Y_i, Y_j \subset Y^K$ e $Y_a, Y_b \subset Y^{K-1}$, a distância entre Y_a e Y_b é obtida tomando $\alpha_i = \alpha_j = 1/2$, $\beta = 0$ e $\gamma = -1/2$ em (4.2).

$$d_{Y_aY_b} = \frac{1}{2}d_{Y_aY_i} + \frac{1}{2}d_{Y_jY_a} - \frac{1}{2} \mid d_{Y_iY_a} - d_{Y_jY_a} \mid,\ Y_i \subset Y_b;\ Y_j \subset Y_b. \tag{4.15}$$

Como as restrições (4.3), $\alpha_i = \alpha_j$ e (4.4), $\alpha_i + \alpha_j + \beta = 1$, estabelecidas por Dubien e Warde (1979), são satisfeitas, o *Método da Ligação Simples* pertence a *sub-família bi-paramétrica* (β, γ) e, tomando $\alpha_i = \alpha_j = 1/2$, $\beta = 0$ e $\gamma = -1/2$ em (4.8) e assumindo que $d_{Y_iY_j} < d_{Y_aY_i} < d_{Y_jY_a}$, a distância entre Y_a e Y_b também é dada por:

$$d_{Y_aY_b} = \frac{1}{2}d_{Y_aY_i},\ Y_i \subset Y_b. \tag{4.16}$$

Discussão

Por ser um dos métodos mais utilizados da literatura, é natural que existam opiniões diferentes, e até mesmo conflitantes, com relação ao emprego deste *Método Hierárquico Aglomerativo*.

Lance e Williams (1967) mostram, de acordo com propriedades apresentadas por eles e descritas na *seção 4.4.2* desta dissertação, que este método é:

a) Combinatório : Dado que a distância entre um grupo Y_b, proveniente da união dos grupos Y_i e Y_j, e um grupo Y_a, é obtida diretamente da equação (4.2), tomando-se $\alpha_i = \alpha_j = 1/2$, $\beta = 0$ e $\gamma = -1/2$.

b) Compatível : Dado que todas as distâncias entre grupos, obtidas nos diversos estágios do algoritmo de agrupamento, são do mesmo tipo das distâncias obtidas na 1ª iteração (Distâncias entre os indivíduos).

c) Contrator de espaço : Dado que tem uma tendência a aproximar os objetos aos grupos já existentes, em lugar de formar novos grupos, ou seja, há uma tendência ao encadeamento dos grupos.

Segundo a classificação adotada por Dubien e Warde (1979), apresentada na *seção 4.4.3*, o *Método da Ligação Simples* também é classificado como *contrator de espaço*. Visto que o ponto $(\beta, \gamma) = (0, -0.5)$, correspondente aos valores assumidos pelos parâmetros β e γ em (4.8), está contido na região R_2 definida em (4.10).

Segundo Lance e Williams (1967), apenas às duas primeiras propriedades verificadas são desejáveis a um método de agrupamento. Porém, segundo Kopp (1978a), o fato de ser contrator de espaço, e gerar grupos com formas alongadas, é conveniente em algumas áreas de estudo. Seguindo uma abordagem mais teórica, Jardine e Sibson (1968) apresentam sete condições necessárias à admissibilidade de uma *Técnica de Agrupamento*, e mostram que o único *Método Hierárquico* que satisfaz a todas elas é o da *Ligação Simples*. Já Fisher e Van Ness (1971), apresentam nove condições necessárias à admissibilidade de

um algoritmo de agrupamento, e comparam sete *Métodos de Agrupamento* com relação a estas condições. Entre as técnicas comparadas, a da *Ligação Simples* é a que demonstra o melhor desempenho, não satisfazendo apenas a condição denominada "*convexidade admissível*". Também em uma linha mais teórica, Hartigan (1981) define grupos de "alta densidade" e mostra que, na detecção de tais grupos, o *Método da Ligação Simples* é *assintoticamente consistente*.

Do ponto de vista prático, a propriedade c) indica que este método tende a formar grupos com pouca homogeneidade interna, dado a facilidade de juntar grupos. Com isso, os grupos tendem a ser grandes, e as distâncias entre os indivíduos pertencentes a um mesmo grupo podem variar muito. Caso esta propriedade não seja desejada, para o estudo em questão, Wishart (1969) sugere uma modificação no *Método da Ligação Simples*, que permite aliviá-la. Devido a tendência de encadeamento que é inerente a esta técnica, sua aplicação geralmente impõe uma *hierarquia* sobre o *espaço dos objetos* que não proporciona muitas alternativas para a determinação dos *grupos*.

Um exemplo da aplicação deste algoritmo pode ser visto em Bussab, Miazaki e Andrade (1990).

4.4.6 Método da Ligação Completa ("Complete Linkage")

Definição

Este método, também conhecido como *Método do Vizinho mais Distante*, foi proposto originalmente por Sorensen (1948) e, pela definição adotada para distância entre dois grupos, podemos observar que ele é o oposto do *Método da Ligação Completa*.

Definição 4.14: Sejam Y_a e Y_b dois grupos distintos na iteração associada ao agrupamento Y^{k-1}, ou seja, $Y_a \subset Y^{k-1}$ e $Y_b \subset Y^{k-1}$. Além disso, $Y_b = Y_i \cup Y_j$ é composto pelos dois grupos unidos na iteração anterior do processo hierárquico de agrupamento, ou seja, $Y_i \subset Y^k$ e $Y_j \subset Y^k$. A distância entre os grupos Y_a e Y_b, sob o *Método da Ligação Completa*, é dada por:

$$d_{Y_a Y_b} = \max_{i \in Y_a;\ j \in Y_b} d_{ij}. \tag{4.17}$$

Como mostrado por Lance e Williams (1967), a distância entre o grupo Y_b, proveniente da união dos grupos Y_i e Y_i numa iteração anterior, e um outro grupo Y_a qualquer, da presente iteração, pode ser obtida como um caso particular de (4.2), ou seja, se $Y_b = Y_i \cup Y_j$, onde $Y_i, Y_j \subset Y^K$ e $Y_a, Y_b \subset Y^{K-1}$, a distância entre Y_a e Y_b é obtida tomando $\alpha_i = \alpha_j = 1/2$, $\beta = 0$ e $\gamma = 1/2$ em (4.2),

$$d_{Y_a Y_b} = \frac{1}{2} d_{Y_a Y_i} + \frac{1}{2} d_{Y_j Y_a} + \frac{1}{2} \mid d_{Y_i Y_a} - d_{Y_j Y_a} \mid,\ Y_i \subset Y_b;\ Y_j \subset Y_b. \tag{4.18}$$

Como as restrições (4.3), $\alpha_i = \alpha_j$ e (4.4), $\alpha_i + \alpha_j + \beta = 1$, estabelecidas por Dubien e Warde (1979), são satisfeitas, o *Método da Ligação Simples* pertence a sub-família bi-paramétrica (β, γ) e, tomando $\alpha_i = \alpha_j = 1/2$, $\beta = 0$ e $\gamma = 1/2$ em (4.8) e assumindo que $d_{Y_i Y_j} < d_{Y_a Y_i} < d_{Y_j Y_a}$, a distância entre Y_a e Y_b também é dada por:

$$d_{Y_a Y_b} = \frac{1}{2} d_{Y_a Y_j},\ Y_j \subset Y_b. \tag{4.19}$$

Discussão

Lance e Williams (1967) mostram, de acordo com propriedades apresentadas por eles, que este método é:

a) Combinatório : Pois a distância entre um grupo Y_b, proveniente da união dos grupos Y_i e Y_j, e um grupo Y_a, é obtida diretamente da equação (4.2), tomando-se $\alpha_i = \alpha_j = 1/2$, $\beta = 0$ e $\gamma = 1/2$.

b) Compatível : Pela mesma razão atribuída ao método da ligação simples.

c) Dilatador de espaço : Pois tem uma tendência a considerar os objetos, que ainda não tenham sido agrupados, como base para formação de novos grupos.

Segundo a classificação adotada por Dubien e Warde (1979), o *Método da Ligação Completa* também é classificado como *dilatador de espaço*. Visto que o ponto $(\beta, \gamma) = (0, 0.5)$, correspondente aos valores assumidos pelos parâmetros β e γ em (4.8), está contido na região R_3 definida em (4.11).

Segundo Jardine e Sibson (1968), este método é considerado "mal definido " e, por isto, não satisfaz uma das sete condições exigidas por eles à admissibilidade de um algoritmo de agrupamento. De acordo com Fisher e Van Ness (1971), o *Método da Ligação Completa* não satisfaz duas das condições exigidas por eles, tendo assim, o segundo melhor desempenho entre os métodos comparados. Hartigan (1981), mostra que este método é inconsistente na detecção de grupos de "alta densidade". Já Milligan (1980) verificou empiricamente que este método é muito sensível à presença de valores aberrantes e, segundo Kopp (1978b), ele tem a desvantagem de poder produzir agrupamentos diferentes quando a dissimilaridade mínima ocorre para mais de um par de grupos e, é necessário escolher um para ser unido.

Do ponto de vista prático, a propriedade c) indica que este método tende a formar mais grupos pequenos e compactos. Com isso, os grupos tendem a possuir uma maior homogeneidade interna, visto que dentro dos grupos as distâncias são pequenas. Isto é importante pois geralmente possibilita maiores alternativas para a definição dos *grupos*, ou seja, a aplicação desta técnica impõe uma *hierarquia* sobre o *espaço dos objetos* que proporciona uma boa caracterização de um número variado de *agrupamentos*.

Um exemplo da aplicação deste algoritmo também pode ser visto em Bussab, Miazaki e Andrade (1990).

4.4.7 Método da Ligação Média ("Average Linkage")

Definição

Este método, também conhecido como *Método das Médias das Distâncias*, *Método da Média das Ligações* e *Método da Média de Grupo*, foi proposto originalmente por Sokal e Michener (1958). Porém, para ficar mais de acordo com a tradução dada aos métodos

que abordamos anteriormente, vamos chamá-lo de *Método da Ligação Média*.

Definição 4.15: Sejam Y_a e Y_b dois grupos distintos na iteração associada ao agrupamento Y^{k-1}, ou seja, $Y_a \subset Y^{k-1}$ e $Y_b \subset Y^{k-1}$. Além disso, $Y_b = Y_i \cup Y_j$ é composto pelos dois grupos unidos na iteração anterior do processo hierárquico de agrupamento, ou seja, $Y_i \subset Y^k$ e $Y_j \subset Y^k$. A distância entre os grupos Y_a e Y_b, sob o *Método da Ligação Média*, é dado por:

$$d_{Y_aY_b} = \frac{1}{n_a n_b} \sum_{i \in Y_a; j \in Y_b} d_{ij}. \tag{4.20}$$

Como mostrado por Lance e Williams (1967), a distância entre o grupo Y_b, proveniente da união dos grupos Y_i e Y_j numa iteração anterior, e um outro grupo Y_a qualquer, da presente iteração, pode ser obtida como um caso particular de (4.2), ou seja, se $Y_b = Y_i \cup Y_j$, onde $Y_i, Y_j \subset Y^K$ e $Y_a, Y_b \subset Y^{K-1}$, a distância entre Y_a e Y_b é obtida tomando $\alpha_i = \frac{n_i}{n_b}$, $\alpha_j = \frac{n_j}{n_b}$, $\beta = \gamma = 0$ em (4.2),

$$\partial_{Y_aY_b} = \frac{n_i}{n_b}\partial_{Y_aY_i} + \frac{n_j}{n_b}\partial_{Y_jY_a},\ Y_i \subset Y_b;\ Y_j \subset Y_b. \tag{4.21}$$

Este método só satisfaz as restrições (4.3) e (4.4) se $\alpha_i = \frac{n_i}{n_b} = \alpha_j = \frac{n_j}{n_b}$. Caso isto ocorra, o *Método da Ligação Simples* pertence a *sub-família bi-paramétrica* (β, γ) e, tomando $\alpha_i = \alpha_j = \alpha$, $\beta = \gamma = 0$ em (4.8) e assumindo que $d_{Y_iY_j} < d_{Y_aY_i} < d_{Y_jY_a}$, a distância entre Y_a e Y_b também é dada por:

$$d_{Y_aY_b} = \frac{1}{2}d_{Y_aY_i} + \frac{1}{2}d_{Y_jY_a},\ Y_i \subset Y_b, Y_j \subset Y_b. \tag{4.22}$$

Discussão

De acordo com Lance e Williams (1967) este método é:

a) Combinatório : Pois a distância entre um grupo Y_b, proveniente da união dos grupos Y_i e Y_j, e um grupo Y_a, é obtida diretamente da equação (4.2), tomando-se

$\alpha_i = \frac{n_i}{n_k}$, $\alpha_j = \frac{n_j}{n_k}$, $\beta = \gamma = 0$.

b) Compatível : Pela mesma razão atribuída ao método da ligação simples.

c) Conservador de espaço : Pois não apresenta tendência maior nem de ser contrator de espaço e nem de ser dilatador.

Segundo a classificação adotada por Dubien e Warde (1979), o *Método da Ligação Média*, para $n_i = n_j$ (não ponderado), também é classificado como *Conservador de Espaço*. Visto que o ponto $(\beta, \gamma) = (0, 0)$, correspondente aos valores assumidos pelos parâmetros β e γ em (4.8), está contido na região R_1 definida em (4.9).

Segundo Jardine e Sibson (1968), dentre as sete condições exigidas por eles, este método não satisfaz a condição de ser uma "transformação contínua " dos dados. Porém, esta condição, exigida por Jardine e Sibson, é muito criticada por Cormack (1971) e Gower (1971). Hartigan (1981), mostra que este método também é inconsistente na detecção de grupos de "alta densidade ". Porém, Milligan e Isaac (1980), em um estudo comparativo que envolve os três métodos abordados neste trabalho, classificam o *Método da Ligação Média* como o melhor. Ainda segundo Cormack (1971), este método só deve ser utilizado com distâncias, para as quais, haja sentido se obter médias e, segundo Kopp (1978c), o mesmo tira proveito da homogeneidade do *Método da Ligação Composta* e da estabilidade do *Método da Ligação Simples*.

Do ponto de vista prático, visto que não tende nem a formar grupos pequenos e compactos (*Método da Ligação Completa*) e nem grupos grandes e heterogêneos (*Método da Ligação Simples*), este método é mais dependente da estrutura existente nos dados.

Um exemplo da aplicação deste algoritmo também pode ser visto em Bussab, Miazaki e Andrade (1990).

Capítulo 5

Aplicação e Considerações Finais

5.1 Resumo do Capítulo

Este Capítulo faz uma aplicação da teoria desenvolvida nos capítulos anteriores e apresenta algumas discussões, conclusões e perspectivas que complementam o que já foi detalhadamente apresentado e discutido ao longo desta dissertação. Na seção 5.2 é feita uma aplicação, discutidos os aspectos computacionais relacionados à mesma e, apresentada uma análise dos resultados obtidos. Já a seção 5.3, na verdade, é um complemento das conclusões e discussões iniciadas no corpo de cada capítulo

5.2 Aplicação

5.2.1 Considerações Iniciais

Ao chegar o momento de realizar uma aplicação da técnica utilizada, que neste caso é de fundamental importância visto que estamos diante de uma abordagem nova para o problema de agrupamento com *misturas*, deparamo-nos com duas alternativas passíveis de serem adotadas: Uma é a simulação de um conjunto de dados onde fossem mensuradas tanto variáveis categóricas quanto contínuas, e a outra é encontrar um conjunto de dados reais que possuísse este perfil e, no qual fizesse sentido a obtenção de agrupamentos. A primeira alternativa poderia ser adotada, dado que a maioria dos pacotes existentes possuem geradores de dados tanto de distribuições contínuas quanto de categóricas. Porém,

neste caso, nos depararíamos com a falta de sentido prático às interpretações dos resultados obtidos, além da necessidade de sermos abrangentes na escolha dos modelos a simular. A segunda, em relação ao caso de simulação, possui a vantagem de proporcionar aos resultados obtidos um maior significado prático.

Diante dessas alternativas, optamos pela segunda por considerarmos importante a interpretabilidade dos resultados obtidos. Com isso, partimos à procura de um conjunto de dados que possuísse tanto variáveis categóricas quanto contínuas e, além disso, que também fizesse sentido o agrupamento dos indivíduos.

5.2.2 Aspectos Computacionais

Considerações Iniciais

Quando se fala em agrupamento de dados onde foram mensuradas apenas variáveis contínuas, a quantidade de pacotes computacionais existentes à sua realizacão é bastante razoável porém, quando se deseja agrupar dados onde foram mensuradas tanto variáveis contínuas quanto categóricas, é necessário partir para a programação. Isto ocorre, basicamente, porque estes pacotes não possuem, implementado, nenhum coeficiente de dissimilaridade para este caso. Além disso, como estamos trabalhando com uma *Extensão da Distância de Mahalanobis para misturas*, proposta neste trabalho, a necessidade de programação torna-se óbvia. Porém, como discutido no capítulo 4, existem vários métodos hierárquicos que estão disponíveis na maioria dos pacotes existentes e que, desde que possamos definir o coeficiente de distância com o qual pretendemos trabalhar, podem perfeitamente realizar os agrupamentos para o caso de *misturas*. Com isso, o nosso trabalho passa à utilização de um pacote que possua alguns desses métodos hierárquicos implementados e que tenha, como base à aplicação dessas técnicas, a opção da entrada de uma matriz de distâncias ao invés de só admitir a matriz de dados originais. Além disso, é necessário que este pacote posssua um ambiente de programação eficiente, para que possamos definir o coeficiente com o qual pretendemos trabalhar.

A Utilização do SAS

Devido a sua grande utilização no meio científico, a existência de um procedimento específico para *Análise de Agrupamento* que possui 10 métodos hierárquicos implementados, a possibilidade de definição de qualquer coeficiente de distância por meio de programação e a opção de entrada, para o processo de agrupamento, de uma matriz triangular inferior de distâncias em lugar dos dados mensurados para cada indivíduo, a utilização do SAS tornou-se a opção mais recomendada para este caso. Além dos argumentos apresentados acima, o SAS possui um ambiente próprio para operações envolvendo matrizes, chamado *PROC IML*. Vale salientar que isto é muito importante pois, ao se programar de forma matricial, otimiza-se o tempo de CPU gasto e, ao mesmo tempo, possibilita uma programação mais clara e elegante.

Ainda com relação às facilidades apresentadas pelo SAS, vale salientar que o PROC IML permite a programação por meio de funções, o que proporciona que um número menor de vetores e matrizes fiquem armazenados na memória durante todo o tempo de execução do programa. Isto ocorre porque o SAS trabalha com os resultados obtidos nas funções, como locais, ou seja, eles ocupam a memória do micro apenas durante a execução da função exceto, obviamente, os que são retornados ao programa principal. Além disso, vale salientar que a programação por meio de funções torna a leitura do programa mais clara, visto que os cálculos auxiliares são realizados nas funções, ficando o corpo principal do programa composto apenas dos resultados mais importantes.

5.2.3 A Análise de um Conjunto de dados

Descrição do Problema

Como uma aplicação à teoria desenvolvida nesta dissertação, vamos utilizar dados dos censos demográfico e industrial dos anos de 1960, 1970 e 1980 referentes aos municípios pertencentes a região metropolitana de São Paulo. Estes dados estão relacionados a estrutura da dinâmica demográfica, econômica e social de 36 municípios, caracterizada

por 25 variáveis contínuas. Além dessas variáveis contínuas, foi incorporada uma classificação desses municípios, em categorias, obtida por um pesquisador da área. Para efeito de análise, vamos incorporar esta estrutura de classificação como uma variável categórica e obter o agrupamento dos municípios utilizando as *Técnicas Hierárquicas* discutidas no *Capítulo 4*. Vale salientar que o coeficiente de distância utilizado será a *Extensão da Distância de Mahalanobis para misturas*, descrito no *Capítulo 3*. A idéia é tentar observar a relação existente entre a classificação atribuída pelo pesquisador e o agrupamento obtido.

Com relação as 25 variáveis contínuas, provenientes do censo, suas descrições são as seguintes:

- MIGR_PEN : Porcentagem de ocupados que trabalham fora do município de residência atual em 1980;
- CRES_60_ e CRES_70_: Taxa de crescimento populacional nas décadas de 60 e 70, respectivamente;
- PEA_1SM : Porcentagem de PEA que recebia até um salário mínimo, excluindo os "sem rendimento" e os "sem declaração", no censo demográfico de 1980;
- PEA_SEC e PEA_SE70 : Porcentagem de PEA que trabalhava no setor secundário nos censos demográficos de 1980 e 1970, respectivamente;
- ÁGUA e ÁGUA70 : Porcentagem de domicílios com acesso à rede geral de água (com ou sem canalização interna) nos censos demográficos de 1980 e 1970, respectivamente;
- LUZ e LUZ70 : Porcentagem de domicílios ligados a rede de energia elétrica (somente domicílios com medidor) nos censos demográficos de 1980 e 1970, respectivamente;
- ESGOTO e ESGOTO70 : Porcentagem de domicílios que estavam ligados a rede

geral de esgoto ou que possuíam fossa séptica nos censos demográficos de 1980 e 1970, respectivamente;

- ALUGADO e ALUGA70: Porcentagem de domicílios alugados nos censos demográficos de 1980 e 1970, respectivamente;
- PEA_TE70 : Porcentagem de PEA que trabalhava no setor terciário no censo demográfico de 1970;
- EMPREG : Porcentagem de PEA (com trabalho e/ou desempregados) no município no Censo demográfico de 1980;
- VTI e VTI70 : Participação percentual no valor da transformação industrial (VTI) da região metropolitana de São Paulo nos censos industriais de 1980 e de 1970, respectivamente;
- PEPOP80 e PEPOP70 : Participação percentual na população da região metropolitana de São Paulo nos censos industriais de 1980 e 1970, respectivamente;
- PORTE e PORTE70 : Indicador de porte das indústrias (razão entre VTI e o número de estabelecimentos industriais) nos censos industriais de 1980 e 1970, respectivamente;
- CRE_DOMI : Crescimento do número de domicílios na década de 70;
- MIGR_TT : Índice de eficácia migratória (razão entre migração líquida e migração bruta) para a década de 70;
- DE_DOMI70 : Densidade domiciliar no censo demográfico de 1980;

Quanto a variável categórica incorporada, denominada CAT, os seus níveis e significados correspondentes estão dados abaixo:

categoria	significado
0	Industriais
1	Industiais/Dormitorios
2	Dormitórios I
3	Dormitórios II
4	Agrícolas

Resultados Obtidos

Em uma análise mais geral, observando os dendrogramas apresentados no *Apêndice D*, podemos fazer as seguintes considerações:

Com relação a caracterização de grupos de municípios, podemos ver que o agrupamento obtido utilizando o *Método da Ligação Completa* é o que nos dá mais alternativas à definição de grupos. Por exemplo, se considerarmos a distância máxima de 71.00 como o ponto para definição dos grupos, podemos visualizar três grandes grupos separados por dois municípios. Além disso, se formos diminuindo este valor gradativamente, a procura de um agrupamento mais adequado, poderemos visualizar a formação de um número maior de grupos, de certa forma, bem caracterizados, o que indica que os municípios que ainda não foram unidos tendem a ser tomados como base para a formação de novos grupos. Vale ressaltar que esta tendência é esperada quando da aplicação deste método, visto que ele é definido como *dilatador de espaço* (ver discussão na *seção 4.4.6*).

Com relação ao agrupamento produzido pelo *Método da Ligação Simples*, podemos ver que a tendência de encadeamento dos municípios aos grupos já existentes, é extremamente evidente pois, a partir da distância mínima de 38.7, nada menos que 26 municípios foram encadeados um a um, ao único grupo existente a este nível. Vale ressaltar que esta tendência é esperada quando da aplicação deste método, dado que ele é definido como *contrator de espaço* (ver discussão na *seção 4.4.5*).

Com relação ao *Método da Ligação Média*, se considerarmos a distância média de 60 e formos diminuindo gradativamente este valor, podemos observar que há uma formação

de um número maior de grupos, de certa forma, bem caracterizados, o que indicaria uma tendência a ser *dilatador de espaço*, como observado para no *Método da Ligação Completa*. Todavia, se formos aumentando gradativamente a distância média, a partir do valor 60, podemos observar que há um encadeamento um a um dos 10 municípios que ainda não haviam sido juntados a nenhum dos grupos existentes, o que podemos interpretar como uma tendência a ser *contrator de espaço*, como observado para o *Método da Ligação Simples*.

Diante das considerações feitas acima, em função dos dendrogramas obtidos, sugerimos, para efeito de análise, a adoção do agrupamento produzido na distância média de 71, aplicando o *Método da Ligação Completa*. Este agrupamento produz os cinco grupos descritos abaixo (entre parênteses se encontra o número referente a classificação obtida pelo pesquisador):

Grupo 1: Arujá (4), Barueri (2), Santana do Parnaíba (4), Diadema (2), Santo André (0), Embú (2), Juquitiba (4), Taboão da Serra (2) e São Caetano do Sul (0).

Grupo 2: Franco da Rocha (3).

Grupo 3: Biritiba-Mirim (4), Salesópolis (4), Francisco Morato (2), Itapecerica da Serra (2), Itaquaquecetuba (2), Mogi das Cruzes (2), Ribeirão Pires (3), Cotia (3), Jandira (4), Caieiras (3), Poá (3), Pirapora do Bom Jesus (4), Suzano (2), Guarulhos (1), Osasco (1), Mauá (2), e Guararema (4).

Grupo 4: Carapicuíba (2)

Grupo 5: Cajamar (3), Itapevi (3), Embú-Guaçú (4), Ferraz de Vasconcelos (2), Mairiporã (4), Rio grande da Serra (3), Santa Isabel (4), São Bernardo do Campo (1).

Observando os municípios dispostos nos grupos e a classificação assumida pelo pesquisador (números entre parênteses), podemos ver que não há muita concordância. Uma explicação para isso, segundo estudiosos no assunto, seria que alguns dos indicadores mensurados não eram os mais adequados e, além disso, algumas variáveis importantes não puderam

ser incluídas. Vale salientar que o nosso objetivo não é o de expressar uma decisão final sobre a validade ou não da classificação utilizada pelo pesquisador. Até porque, como já dissemos anteriormente, não existe a "melhor *Técnica de Agrupamento*", existe sim, o agrupamento mais adequado a uma determinada situacão.

5.3 Considerações Finais

5.3.1 Discussões e Conclusões

A partir dos resultados apresentados nesta dissertação, podemos dizer que a *Extensão da Distância de Mahalanobis para misturas* possui um potencial enorme para a sua utilização. Pois a expressão em blocos desta distância permite que os mesmos possam ser usados conjuntamente, ou separadamente, visto que podemos até utilizar apenas alguns deles para obtermos agrupamentos. Porém, a falta de um trabalho comparativo eficiente, com respeito aos outros coeficientes descritos no *Capítulo 3*, restringe os nossos comentários, com relação às possíveis vantagens desta extensão, ao campo da abstração.

5.3.2 Perspectivas

As perspectivas de continuação deste trabalho, na verdade, dão um sentido mais amplo ao mesmo, e há assuntos que merecem estudos complementares.

Primeiramente, pretendemos fazer um estudo mais detalhado com relação a utilização da *Extensão da Distância de Mahalanobis* aqui proposta. Bem como estudar as vantagens de utilização dos blocos separadamente. Outro estudo que pode ser feito é com relação a influência de pontos discrepantes sobre esta extensão. Uma linha de pesquisa que se abre é, sem dúvida, a de uma validação mais ampla, por meio de simulações, da utilização deste coeficiente nos *Métodos de Agrupamento* aqui utilizados. Outra linha não menos interessante que a primeira é, sem dúvida, atacar o problema de não realocação de indivíduos já agrupados, inerente aos *Métodos Hierárquicos*. Neste sentido,

nos encaminhamos na direção dos *Métodos Não Hierárquicos*, onde os pontos sementes (característicos dos *Métodos Não Hierárquicos*) podem ser obtidos utilizando-se o critério das densidades das distâncias. Além disso, podemos rediscutir a forma de realocação dos indivíduos e redefinir o que é um centróide. Ainda com relação à extensão desse estudo aos *Métodos Não Hierárquicos*, seria interessante incorporar os *parâmetros de união e separação de grupos*, definidos em Ferreira e Medeiros (1992).

Apêndice A

Álgebra Matricial: Notação, Definições e Propriedades Básicas

A.1 Considerações Iniciais

Neste Apêndice especificamos a notação utilizada nesta dissertação para o trato com estruturas matriciais e vetoriais. Isto é importante pois possibilita um maior entendimento de todas as expressões contidas neste trabalho visto que, na literatura, existem notações diferentes. Além disso, apresentamos algumas definições e resultados básicos de *Álgebra Matricial* que são usados neste trabalho. Um estudo mais detalhados deste assunto, pode ser encontrado em Graybill (1969), Mardia, Kent e Bibby (1979) e Searle (1982).

A.2 Notação Utilizada

Adotamos, com relação as estruturas matriciais e/ou vetoriais, a seguinte notação:

$\mathbf{X} = (x_{ij}) = (\mathbf{x}'_{(1)}, \ldots, \mathbf{x}'_{(n)})' = (\mathbf{x}_{1}, \ldots, \mathbf{x}_{q})$: Denota a matriz de dados com n indivíduos e q variáveis, $i = 1, \ldots, n$ e $j = 1, \ldots, q$.

x_{ij} : Denota o valor observado da j-ésima variável no i-ésimo indivíduo.

$\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})'$: Denota o vetor (coluna) formado pelas observações da j-ésima variável nos n indivíduos mensurados, $j = 1, \ldots, q$.

$\mathbf{x}_{(i)} = (x_{i1}, \ldots, x_{iq})$: Denota o vetor linha formado pelas observações do i-ésimo indivíduo nas q variáveis mensuradas, $i = 1, \ldots, n$.

$\mathbf{Y} = (\mathbf{X}^{(1)} | \mathbf{X}^{(2)})$: Denota uma matriz formada pela concatenação horizontal das submatrizes $\mathbf{X}^{(1)}$ e $\mathbf{X}^{(2)}$, que precisam possuir o mesmo número de linhas mas, que podem ter número de colunas diferentes.

$\mathbf{X}^{(i)}$: Denota a partição i, com n_i indivíduos e q variáveis observadas da matriz de dados $\mathbf{X}$, $i = 1, \ldots, k$.

$\mathbf{x}_j^{(i)}$: Denota a partição i, com n_i indivíduos, do vetor (coluna) $\mathbf{x}_j$. Possui dimensão $n \times 1$.

0_n : Denota um vetor (coluna) nulo, de n elementos.

1_n : Denota um vetor (coluna) de uns de n elementos. Também chamado vetor soma.

$\mathbf{J}_n$: Denota uma matriz de uns de dimensão $n \times n$ (ver *Definição A.17*). Também chamada matriz soma.

$\mathbf{I}_n$: Denota uma matriz identidade de dimensão $n \times n$.

$|\mathbf{X}|$: Denota o determinante da matriz $\mathbf{X}$.

$\mathbf{X}^{-1}$: Denota a inversa de $\mathbf{X}$.

$\mathbf{X}^{-}$: Denota a inversa generalizada de $\mathbf{X}$.

$\mathbf{X}'$: Denota a transposta de $\mathbf{X}$.

$\bar{\mathbf{x}}$: Denota o vetor (coluna) das médias amostrais, ou seja, vetor formado pelas médias de cada variável nos n indivíduos observados.

$\bar{\mathbf{x}}^{(i)}$: Denota o vetor (coluna) das médias amostrais nos n_i indivíduos da partição i.

$\bar{x}_j$: Denota a média da j-ésima variável nos indivíduos observados, $j = 1, \ldots, q$.

$\bar{x}_j^{(i)}$: Denota a média da j-ésima variável nos n_i indivíduos do vetor $\mathbf{x}_j^{(i)}$.

$\mathbf{H}_n$: Denota uma matriz de centralização de dimensão $n \times n$ (ver *Definição A.18*).

A.3 Definições e Propriedades de Álgebra Matricial

Inicialmente, vamos introduzir dois conceitos de *Álgebra Matricial* que têm uma aplicação importante para a formulação de muitos dos coeficientes de distância existentes.

Definição A.1: Sejam $\mathbf{x} = (x_1, x_2, \ldots, x_q)$ e $\mathbf{y} = (y_1, y_2, \ldots, y_q)$ dois vetores linha de mesma dimensão. O *produto interno* dos vetores $\mathbf{x}$ e $\mathbf{y}$, denotado por $\langle \mathbf{x}, \mathbf{y} \rangle$, é definido como:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}\mathbf{y}' = \sum_{k=1}^{q} x_k y_k \ . \tag{A.1}$$

Definição A.2: Seja $\mathbf{x} = (x_1, x_2, \ldots, x_q)$ um vetor linha qualquer. A *norma* L_2 (ou comprimento) do vetor $\mathbf{x}$, denotada por $\|\mathbf{x}\|_2$, é definida como:

$$\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\mathbf{x}\mathbf{x}'} = \sqrt{\sum_{k=1}^{q} x_k^2} \ . \tag{A.2}$$

Vamos generalizar a *norma* L_2 de duas formas. Antes, porém, consideremos as seguintes definições:

Definição A.3: Uma matriz quadrada $\mathbf{X}$ de dimensão $p \times p$ é dita *simétrica* se, e só se, $\mathbf{X} = \mathbf{X}'$, ou seja, $x_{ij} = x_{ji} \ \forall \ i, j = 1, ..., p$. Caso contrário, é dita *assimétrica*.

Definição A.4: Uma matriz quadrada $\mathbf{X}$ é dita *semi-definida positiva* se, e só se, $\mathbf{X}$ é simétrica e $\mathbf{y}\mathbf{X}\mathbf{y}' \geq 0$, para todo vetor linha $\mathbf{y} \neq 0$, tal que a igualdade ocorre para ao menos um vetor $\mathbf{y}$. Caso $\mathbf{y}\mathbf{X}\mathbf{y}' > 0, \ \forall \mathbf{y} \neq 0$, $\mathbf{X}$ é dita *definida positiva*.

A primeira generalização da *norma* L_2, chamada *norma* L_B, é dada pela utilização da *norma* L_2 *ponderada por uma matriz* $\mathbf{B}$ *positiva definida.*

Definição A.5: Seja $\mathbf{x} = (x_1, x_2, \ldots, x_q)$ um vetor linha qualquer e $\mathbf{B}$ uma matriz positiva definida. A *norma* L_B do vetor $\mathbf{x}$, denotada por $\|\mathbf{x}\|_{\mathbf{B}}$, é definida como:

$$\|\mathbf{x}\|_{\mathbf{B}} = \sqrt{\mathbf{x}\mathbf{B}\mathbf{x}'} \ . \tag{A.3}$$

O segundo tipo de generalização da *norma* L_2 é obtido utilizando o conceito de *norma* L_p.

Definição A.6: Seja $\mathbf{x} = (x_1, x_2, \ldots, x_q)$ um vetor linha qualquer. A *norma* L_p do vetor $\mathbf{x}$, denotada por $\|\mathbf{x}\|_p$, é definida como:

$$\|\mathbf{x}\|_p = \left(\sum_{k=1}^{q} \mid x_k \mid^p \right)^{1/p} , \forall p \geq 1 \ . \tag{A.4}$$

Para dar pesos às contribuições das variáveis em (*A.4*), vamos introduzir a definição de *norma* L_P *ponderada por um vetor* $\mathbf{w}$, a qual vamos chamar de norma $L_{\mathbf{w}}$.

Definição A.7: Seja $\mathbf{x} = (x_1, x_2, \ldots, x_q)'$ um vetor linha qualquer e $\mathbf{w} = (w_1, w_2, \ldots, w_q)'$ um vetor de pesos. A *norma* $L_{\mathbf{w}}$ do vetor $\mathbf{x}$, denotada por $\|\mathbf{x}\|_{\mathbf{w}}$, é definida como:

$$\|\mathbf{x}\|_{\mathbf{w}} = \left(\sum_{k=1}^{q} w_k \mid x_k \mid^p \right)^{1/p} , \forall p \geq 1 \ . \tag{A.5}$$

Com relação às estruturas matriciais, além das *Definições A.3 e A.4*, vamos destacar as seguintes:

Definição A.8: Uma matriz quadrada $\mathbf{X}$ de dimensão $q \times q$ é dita *diagonal* e denotada por $\mathbf{diag}((x_{11}, x_{22}, \ldots, x_{qq})')$ se, e só se, $x_{ij} = 0 \ \forall \ i \neq j$ e $i, j = 1, \ldots, q$.

Definição A.9: Uma matriz diagonal $\mathbf{X}$ de dimensão $q \times q$ é dita *identidade* e denotada por $\mathbf{I}_q$ se, e só se, $\mathbf{X} = \mathbf{diag}(1_q)$.

Definição A.10: Uma matriz quadrada $\mathbf{X}$ de dimensão $q \times q$ é dita *ortogonal* se, e só se, $\mathbf{X}'\mathbf{X} = \mathbf{X}\mathbf{X}' = \mathbf{I}_q$.

Definição A.11: Uma matriz quadrada $\mathbf{X}$ é dita *idempotente* se, e só se, $\mathbf{X}^2 = \mathbf{X}$.

Definição A.12: Uma matriz $\mathbf{X}$ simétrica e idempotente é dita *de projeção*.

Definição A.13: Uma matriz quadrada $\mathbf{X}$ é dita *não singular* se, e só se, $|\mathbf{X}| \neq 0$. Caso contrário é dita *singular*.

Definição A.14: A *inversa* de uma matriz quadrada $\mathbf{X}$, de dimensão $q \times q$, existe se, e só se, $\mathbf{X}$ é não singular. Caso exista ela é única, sendo denotada por $\mathbf{X}^{-1}$, e satisfaz a sequinte condição:

$$\mathbf{X}\mathbf{X}^{-1} = \mathbf{X}^{-1}\mathbf{X} = \mathbf{I}_q \ . \tag{A.6}$$

Definição A.15: Seja $\mathbf{X}$ uma matriz qualquer de dimensão $n \times q$. Apesar de geralmente não ser única, uma *inversa generalizada* de $\mathbf{X}$, denotada por $\mathbf{X}^-$, sempre existe e satisfaz a seguinte condição:

$$\mathbf{X}\mathbf{X}^-\mathbf{X} = \mathbf{X}. \tag{A.7}$$

Definição A.16: Uma matriz $\mathbf{X} = (x_{ij})$, $i = 1, ..., n$ e $j = 1, ..., q$ expressa em função de sub-matrizes $\mathbf{X_{11}}$ de dimensão $n_1 \times q_1$, $\mathbf{X_{12}}$ de dimensão $n_1 \times q_2$, $\mathbf{X_{21}}$ de dimensão $n_2 \times q_1$ e $\mathbf{X_{22}}$ de dimensão $n_2 \times q_2$ onde $n_1 + n_2 = n$ e $q_1 + q_2 = q$, é dita *particionada* e escrita como:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X_{11}} & \mathbf{X_{12}} \\ \mathbf{X_{21}} & \mathbf{X_{22}} \end{pmatrix} . \tag{A.8}$$

Com relação a uma matriz $\mathbf{X}$ particionada como na *Definição A.16* temos, segundo resultados dados em Mardia, Kent e Bibby (1979), pág. 459, que:

Proposição A.1: Dado que $\mathbf{X}$ é uma matriz particionada como na definição anterior, então:

a) O seu determinante pode ser expresso como:

$$\begin{vmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22} \end{vmatrix} = |\mathbf{X}_{11}||\mathbf{X}_{22} - \mathbf{X}_{21}\mathbf{X}_{11}^{-1}\mathbf{X}_{12}| = |\mathbf{X}_{22}||\mathbf{X}_{11} - \mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21}|.$$

b) Se todas as inversas necessárias existem, então a inversa de $\mathbf{X}$ pode ser particionada como:

$$\begin{pmatrix} (\mathbf{X}_{11} - \mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21})^{-1} & -(\mathbf{X}_{11} - \mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21})^{-1}\mathbf{X}_{12}\mathbf{X}_{22}^{-1} \\ -\mathbf{X}_{22}^{-1}\mathbf{X}_{21}(\mathbf{X}_{11} - \mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21})^{-1} & (\mathbf{X}_{22} - \mathbf{X}_{21}\mathbf{X}_{11}^{-1}\mathbf{X}_{12})^{-1} \end{pmatrix} \tag{A.9}$$

que pode ser reescrita, tomando $\mathbf{X}^{11} = (\mathbf{X}_{11} - \mathbf{X}_{12}\mathbf{X}_{22}^{-1}\mathbf{X}_{21})^{-1}$, como:

$$\begin{pmatrix} \mathbf{X}^{11} & -\mathbf{X}^{11}\mathbf{X}_{12}\mathbf{X}_{22}^{-1} \\ -\mathbf{X}_{22}^{-1}\mathbf{X}_{21}\mathbf{X}^{11} & (\mathbf{X}_{22} - \mathbf{X}_{21}\mathbf{X}_{11}^{-1}\mathbf{X}_{12})^{-1} \end{pmatrix} \tag{A.10}$$

Definição A.17: Uma matriz quadrada, de dimensão $n \times n$, é dita *soma* e denotada por $\mathbf{J}_n$ se,

$$\mathbf{J}_n = 1_n 1_n' \ . \tag{A.11}$$

Definição A.18: Uma matriz quadrada, de dimensão $n \times n$, é dita *de centralização* e denotada por $\mathbf{H}_n$ se,

$$\mathbf{H}_n = \mathbf{I}_n - (1_n' 1_n)^{-1}\mathbf{J}_n \ . \tag{A.12}$$

Com relação a matriz $\mathbf{H}_n$ vale a pena ressaltar os seguintes resultados:

Proposição A.2: Dado que $\mathbf{H}_n$ é uma *matriz de centralização*, são válidas as seguintes propriedades:

a) $\mathbf{H}_n$ *é simétrica.*

b) $\mathbf{H}_n$ é *idempotente.*

c) $\mathbf{H}_n 1_n = 0_n$ e $\mathbf{H}_n \mathbf{J}_n = \mathbf{J}_n \mathbf{H}_n = \mathbf{0}_n$.

d) $\mathbf{H}_n \mathbf{X} = \mathbf{X} - 1_n \bar{\mathbf{x}}_q'$.

e) $\mathbf{H}_n$ é *de projeção.*

Provas:

Estas demonstrações são triviais e, por isto, serão omitidas.□

A propriedade **d)** é muito importante em análise de dados, pois ela nos diz que pré-multiplicando $\mathbf{X}$ por $\mathbf{H}_n$ reescrevemos cada elemento da matriz de dados $\mathbf{X}$, como um desvio da média de sua variável correspondente, ou seja, $\mathbf{X}$ passará a ser uma matriz escrita de forma centralizada. Esta propriedade de centralizar a matriz de dados justifica o nome de *matriz de centralização* dado a $\mathbf{H}_n$.

Utilizando a *Definição A.16*, podemos particionar $\mathbf{H}_n$ da seguinte forma:

$$\mathbf{H}_n = \begin{pmatrix} \mathbf{H}_{n_1} & \boldsymbol{\Delta} \\ \boldsymbol{\Delta}' & \mathbf{H}_{n_2} \end{pmatrix}, \tag{A.13}$$

onde, para $n_1 + n_2 = n$, temos,

$$\mathbf{H}_{n_1} = \mathbf{I}_{n_1} - (1_n' 1_n)^{-1} \mathbf{J}_{n_1}, \quad \mathbf{H}_{n_2} = \mathbf{I}_{n_2} - (1_n' 1_n)^{-1} \mathbf{J}_{n_2} \ e \quad \Delta = -(1_n' 1_n)^{-1} 1_{n_1} 1_{n_2}'. \tag{A.14}$$

Podemos observar que as sub-matrizes $\mathbf{H}_{n_i}$, $i = 1, 2$ são escritas de forma similar a $\mathbf{H}_n$ (ver *Definição A.18*). Porém, com relação as propriedades verificadas para $\mathbf{H}_n$ (ver *Proposição A.2*), podemos observar que a única válida para as sub-matrizes $\mathbf{H}_{n_i}$, $i = 1, 2$ é dada abaixo:

Proposição A.3: Dado que $\mathbf{H}_n$ é uma *matriz de centralização* particionada como em (*A.13*), onde $0 < n_1 < n$ e $0 < n_2 < n$, as sub-matrizes $\mathbf{H}_{n_i}$, $i = 1, 2$ são simétricas.

Prova:

Também omitiremos por ser trivial.□

Já com relação às outras propriedades observadas para $\mathbf{H_n}$, podemos verificar que as mesmas não são válidas para as sub-matrizes $\mathbf{H}_{n_i}$, $i = 1, 2$. Para tanto, consideremos os seguintes contra-exemplos:

b1) $\mathbf{H}_{n_i}$, $i = 1, 2$ não são matrizes idempotentes.

$$\begin{aligned}
\mathbf{H}_{n_i}^2 &= [\mathbf{I}_{n_i} - (1_n' 1_n)^{-1}\mathbf{J}_{n_i}] \times [\mathbf{I}_{n_i} - (1_n' 1_n)^{-1}\mathbf{J}_{n_i}] \\
&= \mathbf{I}_{n_i} - 2(1_n' 1_n)^{-1}\mathbf{J}_{n_i} + (1_n' 1_n)^{-1}(1_n' 1_n)^{-1}\mathbf{J}_{n_i}\mathbf{J}_{n_i} \\
&= [(1_n' 1_n)^{-1}(1_n' 1_n)^{-1} - 2](1_n' 1_n)^{-1}\mathbf{J}_{n_i} \\
&\neq \mathbf{H}_{n_i}, \forall i = 1, 2 .
\end{aligned}$$

$\Rightarrow \mathbf{H}_{n_i}$, $i = 1, 2$ não são matrizes idempotentes.

c1) $\mathbf{H}_{n_i} 1_{n_i} \neq 0_{n_i}$ e $\mathbf{H}_{n_i}\mathbf{J}_{n_i} = \mathbf{J}_{n_i}\mathbf{H}_{n_i} \neq \mathbf{0}_{n_i}$

$$\begin{aligned}
\mathbf{H}_{n_i} 1_{n_i} &= [\mathbf{I}_{n_i} - (1_n' 1_n)^{-1}\mathbf{J}_{n_i}] 1_{n_i} = 1_{n_i} - (1_n' 1_n)^{-1}\mathbf{J}_{n_i} 1_{n_i} \\
&\neq 0_{n_i} \ \forall i = 1, 2 .
\end{aligned}$$

$$\begin{aligned}
\mathbf{H}_{n_i}\mathbf{J}_{n_i} &= [\mathbf{I}_{n_i} - (1_n' 1_n)^{-1}\mathbf{J}_{n_i}]\mathbf{J}_{n_i} = \mathbf{J}_{n_i}(\mathbf{I}_{n_i} - (1_n' 1_n)^{-1}\mathbf{J}_{n_i}) = \mathbf{J}_{n_i}\mathbf{H}_{n_i} \\
&= \mathbf{J}_{n_i} - (1_n' 1_n)^{-1}\mathbf{J}_{n_i}\mathbf{J}_{n_i} \neq \mathbf{0}_{n_i} .
\end{aligned}$$

d1) $\mathbf{H}_{n_i}\mathbf{X}^{(i)} = \mathbf{X}^{(i)} - 1_n\, \bar{\mathbf{x}}_q^{(i)'}$

$$\begin{aligned}
\mathbf{H}_{n_i}\mathbf{X}^{(i)} &= [\mathbf{I}_{n_i} - (1_n' 1_n)^{-1}\mathbf{J}_{n_i}]\mathbf{X}^{(i)} = \mathbf{X}^{(i)} - (1_n' 1_n)^{-1}\mathbf{J}_{n_i}\mathbf{X}^{(i)} \\
&\neq \mathbf{X}^{(i)} - 1_n\, \bar{\mathbf{x}}_q^{(i)'}
\end{aligned}$$

Apêndice B

Definições e Propriedades de Variáveis e Vetores Aleatórios

B.1 Considerações Iniciais

Neste *Apêndice* são dadas algumas *definições* e *propriedades* básicas de *Probabilidade* e *Inferência*, que são necessárias para o trato com as variáveis e vetores aleatórios abordados nesta dissertação. Um estudo mais abrangente, com relação as definições e propriedades probabilísticas, pode ser visto em Johnson e Kotz (1969), Rohatgi (1967) e James (1981). Com relação a parte inferencial, Mood, Graybill e Boes (1974) e Bickel e Doksum (1977) são recomendados.

B.2 Definições e Propriedades Probabilísticas e Inferenciais

Inicialmente, vale salientar que a notação utilizada para vetores e variáveis aleatórias é:

$\underset{\sim}{X} = (X_i)$: Denota um vetor aleatório contendo q variáveis, $i = 1, \ldots, q; \forall q \in N$.

X_i : Denota a i-ésima variável do vetor $\underset{\sim}{X}$.

Agora, consideremos a seguinte definição:

Definição B.1: Um vetor $\underset{\sim}{X} = (X_1, \ldots, X_n)$ cujas componentes são *variáveis aleatórias* definidas no mesmo espaço de probabilidade (Ω, A, P), é chamado *aleatório*, ou variável aleatória n-dimensional.

Como nesta dissertação trabalhamos com misturas de variáveis, vamos considerar a definição abaixo:

Definição B.2: Seja $\underset{\sim}{W} = (\underset{\sim}{X}', \underset{\sim}{Y}')'$ um vetor aleatório onde, $\underset{\sim}{X} = (X_1, \ldots, X_q)'$ é um vetor aleatório tal que todas as suas componentes são variáveis aleatórias de um certo tipo e $\underset{\sim}{Y} = (Y_1, \ldots, Y_p)'$ é um vetor aleatório tal que todas as suas componentes são de um tipo diferente das de $\underset{\sim}{X}$, então, $\underset{\sim}{W}$ é chamado *misto*.

Outra definição importante é:

Definição B.3: Suponha que um experimento é repetido n vezes independentemente. Cada replicação do experimento termina em um dentre k eventos mutuamente exclusivos e exaustivos $A_1, A_2, \ldots, A_k$, onde a probabilidade de ocorrência do evento A_i $(i = 1, 2, \ldots, k)$ em cada ensaio, é p_i tal que $\sum_{i=1}^{k} p_i = 1$. Se $\underset{\sim}{X} = (X_1, \ldots, X_k)$ é um vetor aleatório onde $X_i = x_i$, $i = 1, 2, \ldots, k$, indica que o evento A_i ocorreu x_i vezes em n ensaios, com $\sum_{i=1}^{k} x_i = n$ e $0 \leq x_i \leq n$, $i = 1, 2, \ldots, k$, sua distribuição conjunta é dada por:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1!, x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}, \qquad \text{(B.1)}$$

e é dito ter *distribuição multinomial* (*vetor multinomial*).

Com relação a *distribuição multinomial*, vale a pena ressaltar as seguintes propriedades:

Proposição B.1: Dado que $\underset{\sim}{X} = (X_1, \ldots, X_k)$ é um *vetor multinomial completo* e $P = (p_1, p_2, \ldots, p_k)$,

a) existe dependência nas categorias multinomiais X_i 's.

b) as distribuições marginais das variáveis X_i 's são binomiais.

c) a matriz de variâncias e covariâncias populacionais é singular.

Prova:

a)

Como $\underset{\sim}{X} = (X_1, \ldots, X_k)$ é um vetor multinomial, por definição, temos :

$$\sum_{i=1}^{k} x_i = n.$$

É fácil verificar que com k-1 categorias deste vetor podemos determinar a magnitude da restante. □

b)

Se tomarmos $s = 1$ em *b)*, podemos ver que a *distribuição marginal* de qualquer X_i, $i = 1, 2, \ldots, k$ é *binomial*, ou seja,

$$P(X_i = x_i) = \frac{n!}{x_i(n - x_i)} p_i^{x_i}(1 - p_i)^{n - x_i}, x_i = 0, 1, \ldots, n \ \square$$

c)

Temos de (2.23), que a matriz de variâncias e covariâncias populacionais de um vetor multinomial é,

$$\Sigma = diag(P) - P'P,$$

onde, $P = (p_1, p_2, \ldots, p_k)$ e $\sum_{i=1}^{k} p_i = 1$.

Para mostrarmos que Σ é singular basta que tenhamos, por exemplo,

$$\Sigma 1 = (diag(P) - P'P)1_p = 0.$$

Usando o fato de que $p_k = 1 - \sum_{i=1}^{k-1} p_i$, esta prova é direta.□

Uma outra definição importante, para o contexto de misturas, é dada abaixo.

Definição B.4: A *esperança condicional* de $\underset{\sim}{X}$ dado que $\underset{\sim}{Y}=\underset{\sim}{y}$, onde $\underset{\sim}{Y}$ é um vetor multiniomial, é a esperança da distribuição condicional de $\underset{\sim}{X}$ dado que $\underset{\sim}{Y}=\underset{\sim}{y}$, ou seja,

$$E\{E(X \setminus \underset{\sim}{Y})\} = \sum_{\underset{\sim}{y}} E(X \setminus \underset{\sim}{Y}=\underset{\sim}{y}) P(\underset{\sim}{Y}=\underset{\sim}{y}) \,. \tag{B.2}$$

Com relação a esperança condicional, é importante ressaltar a seguinte proposição:

Proposição B.2: Sejam X e Y duas variáveis aleatórias quaisquer. Então, temos que,

$$E\{E(X \setminus Y)\} = EX. \tag{B.3}$$

Prova:

Para os casos em que Y é contínua ou discreta esta prova é trivial porém, como o mesmo não acontece para o caso geral, vamos assumir que esta propriedade é válida.□

Definição B.5: A *covariância condicional* entre as variáveis aleatórias X e Y dado Z, é dada por:

$$Cov((X,Y) \setminus Z) = E(XY \setminus Z) - E(X \setminus Z)E(Y \setminus Z) \,. \tag{B.4}$$

Definição B.6: Seja $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(n)}$ uma amostra aleatória observada de uma distribuição q-variada com segundo momento finito e matriz de variâncias e covariâncias $\mathbf{\Sigma}$. Temos que a *matriz de variâncias e covariâncias amostrais* é dada por

$$\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{H}_n\mathbf{X},$$

onde $\mathbf{X}$ é a matriz de dados e $\mathbf{H}_n$ é a matriz de centralização dada na *Definição A.18*.

Com relação a matriz de Variâncias e Covariâncias amostrais, segundo resultados dados em Mardia, Kent e Bibby (1979), temos a seguinte proposição:

Proposição B.3: Seja $\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(n)}$ uma amostra aleatória observada de uma distribuição q-variada qualquer. Dado que $\mathbf{S}$ é a matriz de *variâncias e covariâncias* amostrais, temos que:

a) Se a distribuição multivariada for contínua e $n > q$, $\mathbf{S}$ é uma matriz *definida positiva*.

b) Um estimador não-viciado para a matriz de variâncias e covariâncias populacionais $\mathbf{\Sigma}$ de uma distribuição multivariada com segundo momento finito é,

$$\mathbf{S}^* = \frac{n}{n-1}\mathbf{S}\,. \tag{B.5}$$

Finalmente, vamos dar a definição de estimador pelo método dos momentos.

Definição B.7: Seja X uma variável aleatória qualquer. Suponha que $m_1(\theta), \ldots, m_r(\theta)$ são os primeiros r momentos da população da qual nós estamos tomando amostras, ou seja,

$$m_j(\theta) = E(X^j), \forall j = 1, \ldots, r\,. \tag{B.6}$$

Definimos o j-ésimo momento amostral $\hat{m_j(\theta)}$ por,

$$\hat{m_j(\theta)} = \frac{1}{n}\sum_{i=1}^{n} X_i^j\,, j \geq 1\,. \tag{B.7}$$

Considere que conseguimos expressar a função $q(\theta)$, a qual queremos estimar, como uma função ϕ dos r primeiros momentos populacionais, ou seja,

$$q(\theta) = \phi(m_1(\theta), \ldots, m_r(\theta)) \ . \tag{B.8}$$

O método dos momentos nos diz que o estimador de $q(\theta)$, tomando-se uma amostra aleatória $\mathbf{X}$, é dado por,

$$T(\mathbf{X}) = \phi\ (\hat{m_1}(\theta), \ldots, \hat{m_r}(\theta)) \ . \tag{B.9}$$

Apêndice C

Aspectos Relacionados às Distâncias Métricas

C.1 Considerações Iniciais

Este Apêndice é dividido em três sub-seções: na *seção C.2* apresentamos uma discussão hierarquizada de alguns termos matemáticos atribuídos aos coeficientes de distância, entre eles o de *distância métrica*, citado no *Capítulo 3*. Na *seção C.3* discutimos algumas propriedades relacionadas as *distâncias métricas* e, na *seção C.4*, introduzimos as definições de *função similaridade* e *função similaridade métrica*.

C.2 Abordagem Formal às Distâncias

A hierarquia que vamos discutir aqui é imposta no sentido de uma maior rigidez nas condições a serem satisfeitas pelas distâncias. Primeiramente, consideremos uma definição mais geral de *distância*, dada em Spath (1980):

Definição C.1: Seja Ω um conjunto não vazio de elementos, $\Re$ o conjunto dos números reais e d_0 um número real qualquer. Uma função que designa um número real a cada par de elementos de Ω, denotada por d e expressa como $d : \Omega \times \Omega \rightarrow \Re$, é chamada *função distância*, ou simplesmente *distância*, se, para arbitrários $x, y \in \Omega$, as seguintes condições forem válidas:

$$d(x,y) \geq d_0, \tag{C.1}$$

$$d(x.x) = d_0, \tag{C.2}$$

$$d(x.y) = d(y,x). \tag{C.3}$$

Podemos observar que as condições $(C.1)$ e $(C.2)$ indicam que o mínimo da função d (d_0) ocorre para objetos idênticos, e a $(C.3)$ indica que a função d é simétrica.

Como a maioria dos coeficientes de dissimilaridade existentes são não negativos, vamos assumir, para as definições que serão dadas a seguir que: Ω é um conjunto não vazio de elementos, $\Re^+$ é o conjunto dos números reais não negativos e d uma função que designa um número real não negativo a cada par de elementos de Ω, ou seja, $d : \Omega \times \Omega \rightarrow \Re^+$. Dessa forma, vamos redefinir o conceito de distância dado acima, utilizando uma definição também apresentada em Spath (1980).

Definição C.2: A função d é chamada *função distância*, ou simplesmente *distância*, se, para arbitrários $x, y, z \in \Omega$, as seguintes condições forem válidas:

$$d(x,y) \geq 0, \tag{C.4}$$

$$d(x,x) = 0. \tag{C.5}$$

$$d(x,y) = d(y,x). \tag{C.6}$$

A condição $(C.4)$ indica que o valor da distância entre dois indivíduos quaisquer será, necessariamente, não negativa. A condição $(C.5)$ indica que um indivíduo está a uma distância zero dele mesmo, enquanto a $(C.6)$ indica que a distância d é simétrica. Como

exemplo de uma *função distância*, podemos citar a distância definida por Johnson e Wall (1969), denominada *Distância City-Block* (ver *equação 3.23*).

Agora, vamos iniciar a abordagem hierarquizada citada acima, através da introdução do conceito de *semi-métrica* dado em Anderberg (1973).

Definição C.3: A função distância d é chamada *distância semi-métrica*, ou simplesmente *semi-métrica*, se, para arbitrários $x, y, z \in \Omega$, adicionalmente às condições $(C.4)$ e $(C.6)$, a seguinte também for válida:

$$d(x,y) = 0 \Leftrightarrow x = y. \qquad (C.7)$$

A condição $(C.7)$ indica que x está a uma distância zero dele mesmo e que quaisquer dois pontos que possuam distância zero entre si serão, necessariamente, idênticos. Podemos observar que, matematicamente, a diferença entre uma *distância semi-métrica* e uma *função distância* é que, no primeiro caso, se dois indivíduos x e $y \in \Omega$ possuem a menor distância possível entre si (zero), eles serão necessariamente idênticos. Todavia, podemos ver que isto não está explícito, formalmente, na definição de uma *função distância*. Como exemplo de uma *distância semi-métrica*, vamos citar a *semi-métrica* definida por Sokal e Sneath (1963), denominada *Distância de Sokal e Sneath* (ver *equação 3.29*).

Dando sequência à hierarquização matemática, vamos introduzir agora o conceito de *métrica* dado em Royden (1968).

Definição C.4: A função d é chamada *função distância métrica*, ou simplesmente *distância métrica*, se, para arbitrários $x, y, z \in \Omega$, adicionalmente às condições $(C.4), (C.6)$ e $(C.7)$, a seguinte condição também for válida:

$$d(x,y) \leq d(x,z) + d(y,z). \qquad (C.8)$$

Podemos ver que a diferença entre uma *distância semi-métrica* e uma *distância métrica* está na condição $(C.8)$. Esta condição é conhecida como a *desigualdade triangu-*

lar, pois requer que a distância $d(x,y)$ entre dois pontos x e $y \in \Omega$ não seja maior que a soma das distâncias $(d(x,z)+d(y,z))$ entre cada um desses pontos e um outro ponto qualquer $z \in \Omega$. Dá-se o nome de triangular pois a junção destes três pontos forma um triângulo. Como exemplo de uma *distância métrica*, vamos citar a *Distância Euclidiana* (ver *equação 3.26*).

Finalizando o enfoque matemático hierarquizado, vamos apresentar agora, a definição de *ultra-métrica* dada em Anderberg (1973).

Definição C.5: Uma *distância métrica* d é chamada *distância ultra-métrica*, ou simplesmente *ultra-métrica* se, para arbitrários $x, y, z \in \Omega$, adicionalmente às condições $(C.4), (C.6), (C.7)$ e $(C.8)$, a seguinte condição também for válida:

$$d(x,y) \leq \max\{d(x,z), d(y,z)\}. \tag{C.9}$$

Podemos observar que a condição imposta pela desigualdade da equação $(C.9)$ é consideravelmente mais forte que a imposta pela desigualdade triangular $(C.8)$, pois requer que as distâncias entre três pontos $x, y, z \in \Omega$ formem um triângulo equilátero, ou um triângulo isósceles onde a base $d(x,y)$ seja mais curta que os dois lados iguais, $d(x,z)$ e $d(y,z)$. Como exemplo de uma *distância métrica* que satisfaça a condição $(C.9)$, Johnson (1967) utilizou a *Distância Métrica Euclidiana* e mostrou que, para um conjunto de dados gerados de forma que a referida condição seja válida, esta distância se torna uma *distância ultra-métrica*.

C.3 Propriedades Relacionadas às Distâncias Métricas

Com respeito às propriedades relacionadas às *distâncias métricas*, podemos destacar as seguintes:

Proposição C.1: Sejam $x_{(i)}$ e $x_{(j)}$ 2 elementos pertencentes a um espaço mensurável Ω, e $d_k(x_{(i)}, x_{(j)}), \forall k = 1, 2, \ldots, c$ *distâncias métricas* distintas. A função

$$d(x_{(i)}, x_{(j)}) = \sum_{k=1}^{c} d_k(x_{(i)}, x_{(j)}),$$

também é uma *distância métrica*, ou seja, a soma de c *métricas* é uma *métrica*.

Prova:

É trivial, basta verificar que as 4 condições da *Definição C.4* são satisfeitas.□

Proposição C.2 : Sejam $x_{(i)}$ e $x_{(j)}$ 2 elementos pertencentes a um espaço mensurável Ω e $d_k(x_{(i)}, x_{(j)}), \forall k = 1, 2$ *distâncias métricas* quaisquer (podem ser iguais). A função

$$d(x_{(i)}, x_{(j)}) = \prod_{k=1}^{2} d_k(x_{(i)}, x_{(j)})$$

é uma *distância semi-métrica.*

Prova:

A demonstração também é simples, basta verificar que o produto de duas métricas (em particular o quadrado de uma métrica) sempre satisfaz as condições $(C.4)$, $(C.6)$ e $(C.7)$.□

Podemos ver que o produto de duas métricas (em particular o quadrado de uma métrica) não é uma métrica, visto que não necessariamente satisfaz a *desigualdade triangular* $(C.8)$. Isto pode ser observado através do seguinte contra-exemplo:

Consideremos que,

$$d_1(x_{(i)}, x_{(l)}) = d_2(x_{(i)}, x_{(l)}) = d_2(x_{(j)}, x_{(l)}) = 0.3,$$

e,

$$d_1(x_{(j)}, x_{(j)}) = 0.2, d_1(x_{(i)}, x_{(j)}) = 0.4 \text{ e } d_2(x_{(i)}, x_{(j)}) = 0.6.$$

Podemos verificar facilmente que as distâncias d_1 e d_2 são métricas. Agora, por construção, temos que:

$$\begin{cases} d(x_{(i)}, x_{(j)}) = d_1(x_{(i)}, x_{(j)}) \times d_2(x_{(i)}, x_{(j)}) = 0.4 \times 0.6 = 0.24, \\ d(x_{(i)}, x_{(l)}) = d_1(x_{(i)}, x_{(l)}) \times d_2(x_{(i)}, x_{(l)}) = 0.3 \times 0.3 = 0.09, \\ d(x_{(j)}, x_{(l)}) = d_1(x_{(j)}, x_{(l)}) \times d_2(x_{(j)}, x_{(l)}) = 0.2 \times 0.3 = 0.06. \end{cases} \quad (C.10)$$

Com isso, por $(C.10)$, podemos ver que para este caso particular o produto das métricas d_1 e d_2 não é uma métrica, pois a desigualdade triangular $(C.8)$ não é satisfeita, ou seja,

$$d(x_{(i)}, x_{(j)}) = 0.24 > 0.15 = d(x_{(i)}, x_{(l)}) + d(x_{(j)}, x_{(l)}). \square$$

Já pela proposição dada a seguir, qualquer múltiplo possível de uma métrica, também é uma métrica, ou seja,

Proposição C.3 : Sejam $x_{(i)}$ e $x_{(j)}$ 2 elementos pertencentes a um espaço mensurável Ω, $d(x_{(i)}, x_{(j)})$ uma *distância métrica* qualquer e c um escalar. A função

$$cd(x_{(i)}, x_{(j)})$$

também é uma *distância métrica*.

Prova:

Esta prova também é trivial e, por isto, a omitiremos. $\square$

C.4 Abordagem formal às similaridades

Ainda com relação a definição de *métrica*, vamos apresentar agora algumas definições formais relacionadas as *similaridades métricas*.

Definição C.6: Seja Ω um conjunto finito ou infinito de elementos, $\Re$ o conjunto dos números reais e s_0 um número real finito podendo assumir valores negativos. Uma função, que designa um número real a cada par de elementos de Ω, denotada por s e expressa como $s : \Omega \times \Omega \rightarrow \Re$ é chamada *função similaridade* se, para arbitrários $x, y \in \Omega$, as seguintes condições forem válidas:

$$s(x, y) \leq s_0, \tag{C.11}$$

$$s(x, x) = s_0, \tag{C.12}$$

$$s(x, y) = s(y, x). \tag{C.13}$$

As condições $(C.11)$ e $(C.12)$ indicam que s é máxima para objetos idênticos, ao contrário do que indicam as condições $(C.1)$ e $(C.2)$ para uma *função distância* d. A condição $(C.13)$ indica que as medidas produzidas por s são simétricas, tal como a $(C.3)$ para d.

Definição C.7: A *função similaridade* s, definida acima, é chamada *função similaridade métrica* se, adicionalmente às condições $(C.11)$, $(C.12)$ e $(C.13)$, as seguintes também forem válidas:

$$s(x, y) = s_0 \Rightarrow x = y, \tag{C.14}$$

$$(s(x, y) + s(y, z)) \times s(x, z) \geq s(x, y) \times s(y, z), \forall z \in \Omega. \tag{C.15}$$

A condição $(C.14)$ indica que a similaridade máxima pode ser possível apenas para dois objetos idênticos. Já a desigualdade $(C.15)$ é a recíproca à desigualdade triangular, dada para uma *função distância métrica*.

Intuitivamente, a relação mais óbvia entre *distâncias métricas* e *similaridades métricas* é que quanto maior a similaridade entre dois indivíduos, menor será a distância existente. Spath (1980) apresenta várias relações matemáticas existentes entre estas funções, dentre elas podemos destacar as seguintes:

- Se d é uma *distância métrica* que assume apenas valores em $\Re^+$ ou $\Re^-$, então $1/d$ e $\exp(-d)$ são *similaridades métricas.*

- Se d é uma *distância métrica* que assume apenas valores finitos em $\Re^+$ ou $\Re^-$, então $\max d - d$, $\sqrt{\max d - d}$ e $\max d - d^2$ são *similaridades métricas.*

Apêndice D

O SAS e sua Utilização para Agrupamento de Dados com Misturas

D.1 Considerações Iniciais sobre o SAS

O *SAS* (*"Statistical Analysis System"*) é um dos pacotes mais completos em se tratando de armazenamento, manipulação e análise estatística de dados sendo, dessa forma, um dos *softwares* estatísticos mais utilizados em todo o mundo. Este pacote é composto por vários módulos que foram desenvolvidos para aplicação de diferentes técnicas estatísticas nas mais diversas áreas do conhecimento, sendo revisado e atualizado constantemente por uma equipe permanente de pesquisadores altamente qualificados, ligados ao *SAS Institute Inc.* A versão utilizada nesta dissertação é a *6.08*, disponível desde 1993, e maiores detalhes sobre as informações que serão dadas e utilizadas a seguir, podem ser obtidas nos manuais do SAS (ver SAS Institute Inc., 1988 e SAS Institute Inc., 1989).

D.2 Metodologia de Trabalho no SAS

Como em qualquer pacote ligado a análise de dados, a linguagem utilizada pelo *SAS* tem seu próprio vocabulário e sintaxe. Os comandos em um programa *SAS* são dividi-

dos em dois tipos de *passos* (*"Steps"*): *Passos DATA* (*"DATA Steps"*) e *Passos PROC* (*"PROC Steps"*). Os *Passos DATA* são utilizados para transformar os conjuntos de dados que desejamos analisar em um *conjunto de dados SAS* ou, se for o caso, para alterá-los de modo que fiquem mais adequados à análise que será realizada. Já os *Passos PROC* são responsáveis pelo processamento das análises dos *conjuntos de dados SAS* e apresentação dos resultados encontrados.

Com relação a aplicação de *Técnicas de Agrupamento* de indivíduos, o SAS possui dois *Passos PROC* para este caso. Um é o *PROC CLUSTER*, que possui implementados onze métodos hierárquicos de agrupamento. O outro é o *PROC FASTCLUS* , que possui implementada uma variação do método K-means, um dos métodos não hierárquicos mais difundidos da literatura. Vale salientar que, como estamos trabalhando no contexto de *misturas* e utilizamos uma *Extensão da Distância de Mahalanobis* como coeficiente de dissimilaridade, se faz necessário a programação desta distância no *PROC IML*, dado que a mesma não está implementada no *SAS*.

D.3 Sintaxe dos Procedimentos Utilizados à Aplicação das Técnicas de Agrupamento

No contexto de *misturas*, como discutido no capítulo 4, vamos nos deter a aplicação das técnicas hierárquicas da *Ligação Simples*, *Ligação Completa* e *Ligação Média* a partir da obtenção e utilização da *Extensão da Distância de Mahalanobis para misturas*, apresentada no *Capítulo 3*, como o coeficiente de dissimilaridade entre indivíduos. Os procedimentos necessários a estas aplicações podem ser melhor entendidos se os observarmos nas seguintes etapas:

1ª Etapa (*Passo DATA ou PROC ACCESS*): Criação do *conjunto de dados SAS* a ser analisado.

2ª Etapa (*Passo DATA*): Preparação deste conjunto para o cálculo da *extensão da Distância de Mahalanobis para mistura.*

3ª Etapa (*PROC IML*): Cálculo da *Extensão da Distância de Mahalanobis para misturas.*

4ª Etapa (*Passo DATA*): Criação do *conjunto de dados SAS do tipo distância* (*"Type Distance"*) a partir da matriz de *Distâncias Estendidas de Mahalanobis* para mistura, obtida na etapa anterior.

5ª Etapa (*PROC CLUSTER e PROC TREE*): Apresentação dos grupos formados em cada iteração da *Técnica Hierárquica* utilizada (*PROC CLUSTER*) e visualização destes por meio de um *dendrograma* (*PROC TREE*).

As sintaxes dessas etapas podem ser vistas no programa apresentado na última seção deste apêndice porém, como a primeira etapa pode ser executada de quatro maneiras diferentes, dependendo da forma de entrada dos dados originais, vamos apresentar aqui a sua sintaxe para que, caso seja necessário a utilização desta etapa de uma outra forma, este problema possa ser facilmente contornado pela substituição da 1ª Etapa do programa (ver *seção D.4*) pela alternativa mais adequada entre as descritas a seguir.

D.3.1 Criação do Conjunto de Dados SAS a ser Analisado

O conjunto de dados SAS, a ser analisado, pode ser criado de quatro formas diferentes. Para melhor entendimento destas, vamos utilizar os mesmos nomes atribuídos aos conjuntos criados no programa da *seção D.4*.

a) Pela Entrada dos Dados via Passo DATA

Nesta situação os dados são digitados diretamente na tela de edição de programas (*"Program Editor"*) do *SAS*. Sua sintaxe é a seguinte:

```
data c.dados;
  input cidade $ agua luz esgoto alugado ... pepop80;
  cards;
  ARUJA                46.91 52.09 10.29 22.72 ... 0.14
  BARURI               69.24 47.85 21.92 30.92 ... 0.60
  :          :           :
  TABOAO DA SERRA   86.92 54.33 24.52 32.83 ... 0.78
  ;
run;
```

Pela Transformação de um Conjunto de Dados DBASE em SAS

Esta é a forma utilizada no programa dado na *seção D.4.*

Pela transformação de um Conjunto de Dados ASCII em SAS

Nesta situação, os dados são lidos de um arquivo externo ASCII chamado *"arq"* e transformados em um conjunto de dados SAS chamado *"dados"*. Sua sintaxe é a seguinte:

```
data c.dados;
  infile c.arq;
  input cidade $ agua luz esgoto alugado ... pepop80;
run;
```

A Partir de um Conjunto de Dados SAS já Existente

Nesta situação, os dados são lidos de um arquivo SAS já existente, por exemplo chamado *arq*, e tranformado em um outro conjunto de dados *SAS*. Isto é comum quando desejamos trabalhar apenas com um sub-conjunto de variáveis de um arquivo já existente. A sintaxe, para esta situação, é dada abaixo:

```
data c.dados;
  set c.arq;
  keep cidade $ agua luz esgoto alugado ... pepop80;
run;
```

D.4 Programa SAS

D.4.1 Considerações Iniciais

Este programa foi desenvolvido para a análise dos dados descritos no *Capítulo 5*. Porém, para facilitar a sua utilização em outras situações, procuramos torná-lo o mais geral possível.

Para os pesquisadores que trabalham com dados com *misturas* e desejam utilizar este programa, listamos abaixo alguns cuidados que devem ser tomados antes de sua execução:

- Caso se deseje salvar os programas em um diretório diferente do especificado no programa, é necessária a alteração do comando *libname*.
- Caso a entrada dos dados seja diferente da especificada no programa (em Dbase), a 1ª etapa do mesmo deve ser alterada conforme discutido na *seção D.3*.
- Caso se queira salvar os arquivos criados pelo programa com um outro nome, obviamente, os comandos relacionados com a criação de arquivos devem ser alterados de forma apropriada. Além disso, todos os comandos relacionados a leitura e manipulação destes arquivos devem ser alterados conforme as mudanças realizadas. Uma outra alternativa é alterar os nomes dos arquivos depois que os mesmos tenham sido criados pelo programa.

D.4.2 Listagem do Programa e Resultados da Aplicação

```
options ls=80 ps=500 nodate;

/* ************************************************************************** */
/* ************************  1a Etapa: PROC ACCESS  ************************* */
/* ************************************************************************** */

/* Esta etapa e responsavel pela transformacao de um conjunto de dados, origi-
nalmente digitado em DBASE, em um conjunto de dados SAS. */

libname c 'c:\users\pledson\sas';          /* Define local onde serão lidos e
                                              salvos os arquivos */

proc access accdesc=c.dados function=c;
proc access accdesc=c.dados;
run;

proc print data=c.dados;                   /* Imprime o arquivo criado onde
                                              primeiro vem indiv., var. categ.
                                              e var. contin., nesta ordem. */
run;

/* ************************************************************************** */
/* ************************  2a Etapa: Passo DATA  ************************** */
/* ************************************************************************** */

/* Esta etapa e responsavel pela separacao da variavel CIDADE num arquivo (C.CI-
DADE), criaçao de arquivo de dados sem a variável cidade (C.NEWDADOS) e
inclusao, no mesmo arquivo, da variavel ORD que indica a ordem observada dos in-
divíduos. Tambem é responsavel pela criacao de um outro arquivo somente com as
variaveis categoricas (C.CATE). */

data c.cidade;
 set c.dados;
 keep cidade;

data c.newdados;
 set c.dados;
 ord=_n_;
 drop cidade;

data c.cate;
 set c.dados;
 keep cat;

proc sort data=c.newdados;                 /* ordena arquivo c.newdados pela
 by cat;                                      variavel cat */

proc print data=c.cidade;                  /* Imprime arquivos criados */
proc print data=c.newdados;
proc print data=c.dados;
proc print data=c.cate;
run;

/* ************************************************************************** */
/* ************************  3a Etapa: PROC IML  **************************** */
/* ************************************************************************** */

/* Esta etapa e responsavel por todos os calculos e transformacoes que levam a
obtencao dos valores para a extencao da distancia de Mahalanobis entre indivídu-
os. A matriz resultante é da forma triangular inferior. */

proc iml;                          /* Da acesso ao ambiente IML */

reset deflib=c;                    /* Define local onde serão lidos e
                                      salvos os arquivos */

/* ************************  Definicao de Funcoes  ************************** */
```

```
/* A funcao fun_1, definida abaixo, recebe os argumentos n_ind (numero de indi-
viduos) como x e d (matriz de dados) como y e retorna o vetor w com o numero de
variaveis categoricas (w[1,1]) e o numero de variaveis continuas (w[2,1]). Os
argumentos recebidos nao sao alterados. */

start fun_1(x,y);
 use cate;
 read all into z;
 w=ncol(z);
 n_v=ncol(y);
 n_cn=n_v-w;
 w=w//n_cn;
 close cate;
 return(w);
finish;
/* A funcao fun_2, definida abaixo, recebe os argumentos d (matriz de dados) co-
mo y, n_g_cat (vetor a ser calculado) como z e n_c_mult (numero de variaveis ca-
tegoricas) como w. Retorna o vetor a com o numero de categorias existentes em
cada variavel categorica. Tambem calcula o vetor z e retorna-o com o numero de
categorias de cada variavel categorica e recalcula w, retornando-o com o numero
de categorias multinomiais a serem geradas. y nao e alterado. */
start fun_2(y,z,w);
 x=w;
 a=unique(y[,1])`;
 z=nrow(a);
 w=z;
 if x>=2 then do;
  do i=2 to x;
   b=unique(y[,i])`;
   c=nrow(b);
   a=a`//b`;
   z=z//c;
   w=w*c;
  end;
 end;
 w=w-1;                          /* Retira uma combinacao das cate- */
 return(a);                      /* goric. p/ evitar dependencia */
 finish;
/* A funcao fun_3, definida abaixo, recebe os argumentos n_ind (numero de indi-
duos mensurados) como x, n_c_mult (numero de categorias multinomiais a serem
criadas como y, d[,1:n_var[1,1]] ( matriz de dados categoticos) como z e z ( ma-
triz a ser calculada) como w. Retorna o vetor a com o numero de individuos por
categoria multinomial. Tambem calcula a matriz w e retorna-a com os vetores mul-
tinomiais wi's. Os outros argumentos nao sao alterados. */
start fun_3(x,y,z,w);
 i=1;
 j=0;
 do while(i<=x);
  j=j+1;
  aux=z[i,];
  aux1=0;
  do while(z[i,]=aux & j<=y);
   w[i,j]=1;
   aux1=aux1+1;
   i=i+1;
  end;
  if j>=1 & j<=y then
   if j=1 then a=aux1;
   else a=a//aux1;
  else i=x+1;
 end;
 n=nrow(a);
 aux1=x-j(1,n,1)*a;
 a=a//aux1;
 return(a);
finish;
```

```
/* A funcao fun_4, definida abaixo, recebe os argumentos n_ind (numero de indi-
viduos mensurados) como x, ncol_dtr[1,1] (numero de categorias multinomiais e-
xistentes como y e n_i_cat (vetor com o numero de individuos por categoria mul-
tinomial) como z. Retorna a matriz de projecao h. Nenhum dos argumentos recebi-
sao alterados. */

start fun_4(x,y,z);
 c=1;
 h=j(x,x,0);
 diagon=j(x,x,0);
 do i=1 to y+1;
  d=j(1,i,1)*z[1:i,1];
  a=j(z[i,1],1,1);
  hn=i(z[i,1])-(1/x)*a*a`;
  diagon[c:d,c:d]=hn;
  do j=i+1 to y+1;
   b=j(z[j,1],1,1);
   aux=-(1/x)*a*b`;
   hn=hn||aux;
  end;
  h[c:d,c:x]=hn;
  c=d+1;
 end;
 h=h`-diagon+h;
 return(h);
finish;

/* A funcao fun_5, definida abaixo, recebe os argumentos d_tr (matriz de dados
transformados reordenados) como x, s_inv (inversa da matriz de variancias e co-
variancias) como y, ncol_dtr[1,1] (vetor com o num. de categorias multinomiais)
como w e d_mah_ca, d_mah_cn e d_mah_cr (matrizes a serem calculadas), respecti-
mente como z, q e k. Calcula as matrizes triangulares inferiores z, q, e k e re-
na-as, respectivamente, como a distancia extendida de Mahalanobis entre catego-
ricas, continuas e o produto cruzado. Os outros argumentos nao sao alterados. */

start fun_5(x,y,w,z,q,k);
 a=ncol(x);
 b=nrow(x);
 do i=1 to b;
  do j=1 to i;
   d_cg=x[i,1:w]-x[j,1:w];
   d_cn=x[i,w+1:a]-x[j,w+1:a];
   delta=y[1:w,w+1:a];
   z[i,j]=d_cg*y[1:w,1:w]*d_cg`;
   q[i,j]=d_cn*y[w+1:a,w+1:a]*d_cn`;
   k[i,j]=2*d_cg*delta*d_cn`;
  end;
 end;
finish;

/* ************************** Programa Principal ************************** */

use newdados;                    /* Torna corrente arq. mun_sp */
read all into d;                 /* Passa dados para a matriz d */

ordem=d[,ncol(d)];               /* Vetor com a ordem dos indiv. */
d=d[,1:(ncol(d)-1)];             /* Matr. de dados s/ var. ordem */
n_ind=nrow(d);                   /* numero de individuos medidos */

n_var=fun_1(n_ind,d);            /* Chama funcao fun_1 */

n_n_cat=j(n_var[1,1],1,0);       /* Num. niveis das var. categ. */
n_c_mult=n_var[1,1];

niv_cat=fun_2(d,n_n_cat,n_c_mult);      /* Chama funcao fun_2 */

z=j(n_ind,n_c_mult,0);                  /* Inicializa z com zeros */
```

```
n_i_cat=fun_3(n_ind,n_c_mult,d[,1:n_var[1,1]],z);   /* Chama funcao fun_3 */

d_tr=z||d[,n_var[1,1]+1:ncol(d)];          /* Matriz de dados transform. */
ncol_dtr=n_c_mult//n_var[2,1];             /* Vetor com num. de categorias
                                              multinomiais e var. contin. */

hn=fun_4(n_ind,ncol_dtr[1,1],n_i_cat);     /* Chama funcao fun_4 */

s=(1/nrow(d_tr))*(d_tr`*hn*d_tr);          /* Calcula matriz de variancias e
                                              covariancias */

s11=s[1:ncol_dtr[1,1],1:ncol_dtr[1,1]];          /* Parte categorica de S */

s12=s[1:ncol_dtr[1,1],ncol_dtr[1,1]+1:ncol(d_tr)];   /* Parte cruzada de S */

s22=s[ncol_dtr[1,1]+1:ncol(d_tr),ncol_dtr[1,1]+1:ncol(d_tr)]; /* Parte continua de S */

s11_inv=inv(s11-s12*inv(s22)*s12`);              /* Inversa de S11 */
s22_inv=inv(s22-s12`*inv(s11)*s12);              /* Inversa de s22 */
s12_inv=-s11_inv*s12*inv(s22);                   /* Inversa de s12 */

s_inv=(s11_inv||s12_inv)//(s12_inv`||s22_inv);   /* Inversa de S */

d_tr=ordem||d_tr;

create d_trans from d_tr;              /*   Ordena  */
append from d_tr;                      /*           */
show datasets;                         /*  matriz   */
show contents;                         /*           */
close d_trans;                         /*    de     */
sort d_trans by col1;                  /*           */
use d_trans;                           /*   dados   */
read all into d_tr;                    /*           */

d_tr=d_tr[,2:ncol(d_tr)];              /* Dados transform. */

create dado_tra from d_tr;             /* Cria arquivo com dados */
                                       /* transformados de acordo com */
                                       /* originais */
append from d_tr;

d_mah_ca=j(n_ind,n_ind,0);             /* Inicializa d_mah_ca */
d_mah_cn=j(n_ind,n_ind,0);             /* Inicializa d_mah_cn */
d_mah_cr=j(n_ind,n_ind,0);             /* Inicializa d_mah_cr */

run fun_5(d_tr,s_inv,ncol_dtr[1,1],d_mah_ca,d_mah_cn,d_mah_cr); /*  Chama
                                                                   fun_5 */
create d_mahca from d_mah_ca;
append from d_mah_ca;
create d_mahcn from d_mah_cn;
append from d_mah_cn;
create d_mahcr from d_mah_cr;
append from d_mah_cr;

d_mahal=d_mah_ca+d_mah_cn+d_mah_cr;      /* Calcula a matriz com as distan-
                                            cias de Mahalan. extendidas */

create d_mahala from d_mahal;          /* Cria matriz de dados com as */
append from d_mahal;                   /*  dist. de Mahal. estendidas */

quit;                                  /* Finaliza PROC IML */
```

```
/* ********************************************************************************** */
/* ************************  4a Etapa: Passo DATA  ************************** */
/* ********************************************************************************** */

/* Esta etapa tem a funcao de criar um conjunto de dados SAS que contem as  dis-
tancias de Mahalanobis extendidas calculadas na etapa anterior e, os  nomes  das
respectivas cidades.  Também são criados conjuntos de dados para os blocos   */

data c.dist(type=distance);
  merge c.d_mahala c.cidade;
run;
data c.distca(type=distance);
  merge c.d_mahca c.cidade;
run;
data c.distcn(type=distance);
  merge c.d_mahcn c.cidade;
run;
data c.distcr(type=distance);
  merge c.d_mahcr c.cidade;
run;

/* ********************************************************************************** */
/* ********************  5a Etapa: PROC CLUSTER e PROC TREE  ****************** */
/* ********************************************************************************** */

/* Esta etapa tem a finalidade de obter os agrupamentos pelos métodos hierárqui-
cos da ligação média, ligação simples e ligação completa e, além  disso,  também
apresentar um dendograma com os mesmos. */

proc print data=c.mun_sp;              /* Imprime dados originais */

data dad_tran;
  merge c.cidade c.dado_tra;

proc print data=c.dad_tran;             /* Imprime dados transformados */

proc print data=c.dist;                /* Imprime matriz de distâncias estendidas*/
proc print data=c.distca;              /* Imprime matriz de distâncias - bloco categórico */
proc print data=c.distcn;              /* Imprime matriz de distâncias - bloco contínuo */
proc print data=c.distcr;              /* Imprime matriz de distâncias - bloco cruzado  */

options ls=130 ps=60;

proc cluster data=c.dist method=average nosquare nonorm;
  id cidade;

proc tree horizontal spaces=2;
  id cidade;

proc cluster data=c.dist method=complete nosquare nonorm;
  id cidade;

proc tree horizontal spaces=2;
  id cidade;

proc cluster data=c.dist method=single nosquare nonorm;
  id cidade;

proc tree horizontal spaces=2;
  id cidade;
run;
```

CONJUNTO DE DADOS

OBS	CIDADE	CAT	AGUA	LUZ	ESGOTO	ALUGADO	CRE_DOMI	MIGR_PEN	CRES_70
1	ARUJA	4	46.91	52.09	10.29	22.72	91.96	17.43	6.21
2	BARUERI	2	69.24	47.85	21.92	30.92	124.71	51.51	7.14
3	BIRITIBA-MIRIM	4	38.48	58.68	38.45	14.01	54.13	18.20	4.00
4	CAIEIRAS	3	76.54	64.92	64.58	29.05	87.74	35.90	4.92
5	CAJAMAR	3	53.59	42.45	41.52	29.00	126.07	17.62	7.80
6	CARAPICUIBA	2	85.19	60.96	65.20	26.31	274.52	69.56	12.97
7	COTIA	3	74.75	68.72	44.13	28.49	122.91	12.76	7.37
8	DIADEMA	2	78.54	46.78	52.40	38.42	238.88	44.23	11.23
9	EMBU	2	71.79	52.83	25.00	28.90	287.78	61.28	18.10
10	EMBU-GUACU	4	14.14	43.90	18.06	22.49	120.01	25.32	7.43
11	FERRAZ DE VASCONCELOS	2	58.13	70.36	21.54	29.59	144.62	57.13	8.08
12	FRANCISCO MORATO	2	13.14	52.03	19.29	19.81	170.77	72.41	9.77
13	FRANCO DA ROCHA	3	45.12	61.55	33.58	22.36	128.49	56.13	3.42
14	GUARAREMA	4	34.94	51.65	39.15	20.69	28.15	7.69	1.80
15	GUARULHOS	1	67.59	56.22	34.93	33.11	154.54	28.34	6.45
16	ITAPECERICA DA SERRA	2	17.28	44.09	25.79	22.65	154.57	36.47	9.10
17	ITAPEVI	3	55.31	55.78	29.49	25.72	111.58	56.90	6.84
18	ITAQUAQUECETUBA	2	65.14	59.96	3.03	25.02	168.05	45.48	9.64
19	JANDIRA	4	73.82	65.96	9.13	28.30	225.09	62.88	11.17
20	JUQUITIBA	4	18.93	34.59	30.82	9.86	72.35	3.18	5.57
21	MAIRIPORA	4	40.44	53.38	52.85	20.85	58.46	13.02	3.47
22	MAUA	2	86.75	65.62	41.05	31.67	141.57	56.60	7.30
23	MOGI DAS CRUZES	2	72.84	65.76	59.66	32.57	58.27	8.92	3.62
24	OSASCO	1	95.96	52.81	55.08	38.22	88.10	44.12	5.30
25	PIRAPORA DO BOM JESUS	4	22.95	38.05	39.85	12.94	28.36	0.00	2.62
26	POA	3	81.74	70.84	4.42	30.79	87.80	54.84	5.01
27	RIBEIRAO PIRES	3	46.30	73.36	66.02	26.48	113.34	35.62	6.89
28	RIO GRANDE DA SERRA	3	71.64	49.83	47.76	29.42	165.43	68.25	9.12
29	SALESOPOLIS	4	44.12	60.56	39.18	15.01	19.99	13.58	1.09
30	SANTA ISABEL	4	62.08	45.81	16.73	30.88	89.06	7.53	5.39
31	SANTANA DO PARNAIBA	4	28.97	48.85	62.04	19.23	84.67	26.62	6.46
32	SANTO ANDRE	0	95.90	77.32	81.15	33.68	49.20	38.39	2.82
33	SAO BERNARDO DO CAMPO	1	86.42	67.62	76.85	30.16	147.88	22.03	7.76
34	SAO CAETANO DO SUL	0	99.58	82.93	87.98	43.35	22.72	41.02	0.83
35	SUZANO	2	51.66	56.47	28.13	27.04	105.09	24.68	6.18
36	TABOAO DA SERRA	2	86.92	54.33	24.52	32.83	166.15	62.95	9.08

OBS	VTI	PORTE	MIGR_TT	PEA_SEC	PEA_ISM	PEPO_P80	EMPREG	AGUA70	ALUGA70	CRES_60_
1	0.15783	35383.15	0.43	11.14	45.28	0.14	0.14	0.00553	13.568	5.41
2	0.60737	47257.42	0.64	9.18	56.49	0.60	0.35	0.01289	25.929	8.64
3	0.00928	10224.58	0.41	14.13	32.57	0.11	0.01	0.34593	17.053	4.75
4	0.29329	57901.16	0.48	8.11	61.01	0.20	0.15	0.64890	36.409	5.19
5	0.25578	49753.67	0.53	8.47	65.82	0.17	0.13	0.22665	16.099	4.91
6	0.20518	21202.81	0.79	8.21	43.34	1.48	0.18	0.05115	24.783	14.26
7	0.99403	58960.50	0.61	9.05	46.97	0.50	0.45	0.22306	22.105	7.08
8	2.67156	44282.32	0.76	8.44	65.60	1.88	2.79	0.02304	24.003	20.44
9	0.37545	30845.95	0.86	9.33	50.77	0.76	1.18	0.15782	20.977	13.77
10	0.06016	6522.84	0.72	14.18	50.22	0.17	0.16	0.00000	11.420	8.02
11	0.20528	29514.41	0.69	9.96	55.23	0.44	0.46	0.19475	23.624	9.76
12	0.02634	8499.04	0.77	11.48	51.52	0.23	0.02	0.09164	19.751	16.01
13	0.14845	27655.66	0.61	8.46	43.15	0.40	0.02	0.45930	27.659	3.69
14	0.02317	8283.24	0.31	16.69	37.23	0.12	0.02	2.19513	17.345	5.15
15	5.49816	55685.60	0.78	8.94	54.44	4.23	4.48	0.28320	29.903	8.92
16	0.31907	24680.70	0.72	10.56	47.52	0.48	0.14	0.17256	14.860	5.96
17	0.21723	38310.97	0.55	10.23	54.92	0.42	0.07	0.01542	27.685	10.58
18	0.33702	26534.65	0.79	9.52	55.37	0.58	0.43	0.11479	22.832	9.78
19	0.26597	48862.19	0.66	9.10	58.09	0.29	0.17	0.04746	22.905	19.97
20	0.00484	2561.96	0.34	15.51	22.73	0.10	0.01	0.01152	4.488	2.21
21	0.09783	10606.77	0.45	15.16	43.24	0.22	0.02	0.24165	15.039	4.34
22	2.30701	117819.86	0.66	8.85	65.59	1.63	0.38	0.16630	23.828	13.51
23	1.46374	61464.32	0.29	13.19	44.86	1.57	0.67	0.36830	34.405	3.98
24	3.06556	79196.98	0.41	8.31	50.63	3.77	1.89	0.25532	32.634	9.45
25	0.05470	22609.00	0.41	19.23	43.02	0.04	0.03	0.23636	8.485	4.08
26	0.15376	30206.17	0.50	11.58	53.31	0.42	0.23	0.55055	27.273	6.95
27	0.41629	38506.05	0.50	10.00	59.11	0.45	0.50	0.34881	29.719	5.42
28	0.05273	26825.73	0.63	12.48	57.03	0.16	0.04	0.02650	23.659	7.90
29	0.00129	1007.52	0.11	21.35	29.69	0.08	0.01	0.27642	12.754	0.52
30	0.19430	37794.08	0.41	19.68	57.40	0.23	0.25	0.38268	20.173	3.87
31	0.09391	24842.76	0.48	10.10	51.77	0.08	0.05	0.25894	16.780	0.51
32	6.20803	83705.27	0.21	9.36	57.36	4.39	2.31	0.72792	30.548	5.60
33	12.58544	151474.36	0.52	8.57	57.98	3.38	3.21	0.77677	27.896	9.52
34	3.33124	79107.86	0.08	8.84	56.13	1.30	1.51	0.98746	40.552	2.86
35	1.41418	78595.26	0.57	13.50	52.60	0.80	0.70	0.33642	22.119	7.80
36	0.55705	42104.17	0.73	9.00	44.20	0.78	0.41	0.06113	22.349	19.12

OBS	DE_DOM70	ESGOTO70	LUZ70	PEA_SE70	PEA_TE70	PEPOP70	PORTE70	VTI70
1	4.80955	0.11055	0.47638	26.766	29.339	0.12	79.46	3417
2	5.12999	0.01723	0.70719	46.199	43.209	0.46	191.29	16069
3	4.87480	0.22882	0.46843	18.599	15.548	0.11	98.57	690
4	5.18421	0.53564	0.73451	70.422	22.089	0.19	965.96	50290
5	5.03648	0.19650	0.63181	50.153	22.433	0.13	1458.61	40841
6	5.11255	0.14870	0.77127	48.399	41.861	0.67	459.37	34912
7	5.15916	0.27144	0.64314	36.500	37.056	0.38	584.37	61944
8	5.10704	0.27427	0.82196	56.769	32.443	0.97	1073.37	212528
9	5.09632	0.17467	0.56922	46.699	36.385	0.22	315.27	26168
10	4.93282	0.01679	0.39060	33.898	33.019	0.13	48.23	5499
11	5.27815	0.21038	0.74604	50.803	37.039	0.31	146.00	8906
12	4.99600	0.03025	0.54048	45.454	43.259	0.14	12.24	355
13	7.94376	0.36915	0.84508	27.324	63.356	0.45	53.88	2910
14	4.80715	0.27729	0.41080	32.798	19.335	0.16	114.89	6664
15	5.03221	0.35547	0.83810	49.342	38.984	2.91	1096.96	789817
16	4.89159	0.16928	0.34531	36.132	30.477	0.31	58.28	13288
17	5.12245	0.15217	0.66369	44.490	40.398	0.34	1383.00	41494
18	5.20544	0.20168	0.62399	45.655	34.797	0.36	129.73	13882
19	5.15848	0.11184	0.73050	44.171	40.986	0.15	1403.57	19650
20	4.40691	0.01395	0.13584	10.496	20.575	0.09	30.45	335
21	5.03445	0.23136	0.55733	48.261	22.363	0.27	156.36	20640
22	5.10491	0.77211	0.87446	58.811	27.996	1.25	4401.59	484175
23	4.97779	0.42674	0.81754	33.543	37.578	1.70	1053.65	241288
24	4.89399	0.32226	0.93162	49.576	38.978	3.48	1972.86	581995
25	4.49576	0.00970	0.46303	35.886	36.371	0.05	94.02	3761
26	5.32188	0.33947	0.39471	43.901	45.138	0.40	666.51	21995
27	4.85266	0.47628	0.84932	47.519	36.738	0.36	543.24	54324
28	5.29779	0.10915	0.57413	46.778	33.424	0.10	447.33	4026
29	4.97501	0.17855	0.34045	19.243	21.248	0.12	58.71	822
30	4.95267	0.21270	0.50794	46.853	30.339	0.21	222.98	16501
31	4.59114	0.25043	0.45571	31.401	27.898	0.07	435.48	13500
32	4.68921	0.64082	0.97561	57.794	33.334	5.15	2291.62	1789757
33	5.06269	0.69704	0.92787	57.991	33.061	2.48	3940.35	2320870
34	4.39890	0.90955	0.99801	60.163	31.520	1.84	2237.84	1047311
35	5.19143	0.25349	0.71413	41.512	33.542	0.68	1520.91	272244
36	5.03629	0.28856	0.81070	41.721	49.569	0.50	546.57	31703

CONJUNTO DE DADOS TRANSFORMADOS

OBS	CIDADE	COL1	COL2	COL3	COL4	AGUA	LUZ	ESGOTO	ALUGADO	CRE_DOMI
1	ARUJA	0	0	0	0	46.91	52.09	10.29	22.72	91.96
2	BARUERI	0	0	1	0	69.24	47.85	21.92	30.92	124.71
3	BIRITIBA-MIRIM	0	0	0	0	38.48	58.68	38.45	14.01	54.13
4	CAIEIRAS	0	0	0	1	76.54	64.92	64.58	29.05	87.74
5	CAJAMAR	0	0	0	1	53.59	42.45	41.52	29.00	126.07
6	CARAPICUIBA	0	0	1	0	85.19	60.96	65.20	26.31	274.52
7	COTIA	0	0	0	1	74.75	68.72	44.13	28.49	122.91
8	DIADEMA	0	0	1	0	78.54	46.78	52.40	38.42	238.88
9	EMBU	0	0	1	0	71.79	52.83	25.00	28.90	287.78
10	EMBU-GUACU	0	0	0	0	14.14	43.90	18.06	22.49	120.01
11	FERRAZ DE VASCONCELOS	0	0	1	0	58.13	70.36	21.54	29.59	144.62
12	FRANCISCO MORATO	0	0	1	0	13.14	52.03	19.29	19.81	170.77
13	FRANCO DA ROCHA	0	0	0	1	45.12	61.55	33.58	22.36	128.49
14	GUARAREMA	0	0	0	0	34.94	51.65	39.15	20.69	28.15
15	GUARULHOS	0	1	0	0	67.59	56.22	34.93	33.11	154.54
16	ITAPECERICA DA SERRA	0	0	1	0	17.28	44.09	25.79	22.65	154.57
17	ITAPEVI	0	0	0	1	55.31	55.78	29.49	25.72	111.58
18	ITAQUAQUECETUBA	0	0	1	0	65.14	59.96	3.03	25.02	168.05
19	JANDIRA	0	0	0	0	73.82	65.96	9.13	28.30	225.09
20	JUQUITIBA	0	0	0	0	18.93	34.59	30.82	9.86	72.35
21	MAIRIPORA	0	0	0	0	40.44	53.38	52.85	20.85	58.46
22	MAUA	0	0	1	0	86.75	65.62	41.05	31.67	141.57
23	MOGI DAS CRUZES	0	0	1	0	72.84	65.76	59.66	32.57	58.27
24	OSASCO	0	1	0	0	95.96	52.81	55.08	38.22	88.10
25	PIRAPORA DO BOM JESUS	0	0	0	0	22.95	38.05	39.85	12.94	28.36
26	POA	0	0	0	1	81.74	70.84	4.42	30.79	87.80
27	RIBEIRAO PIRES	0	0	0	1	46.30	73.36	66.02	26.48	113.34
28	RIO GRANDE DA SERRA	0	0	0	1	71.64	49.83	47.76	29.42	165.43
29	SALESOPOLIS	0	0	0	0	44.12	60.56	39.18	15.01	19.99
30	SANTA ISABEL	0	0	0	0	62.08	45.81	16.73	30.88	89.06
31	SANTANA DO PARNAIBA	0	0	0	0	28.97	48.85	62.04	19.23	84.67
32	SANTO ANDRE	1	0	0	0	95.90	77.32	81.15	33.68	49.20
33	SAO BERNARDO DO CAMPO	0	1	0	0	86.42	67.62	76.85	30.16	147.88
34	SAO CAETANO DO SUL	1	0	0	0	99.58	82.93	87.98	43.35	22.72
35	SUZANO	0	0	1	0	51.66	56.47	28.13	27.04	105.09
36	TABOAO DA SERRA	0	0	1	0	86.92	54.33	24.52	32.83	166.15

OBS	MIGR_PEN	CRES_70_	VTI	PORTE	MIGR_TT	PEA_SEC	PEA_SM	PEPPO_70	EMPREG	AGUA_70	ALUGADO
1	17.43	6.21	0.1578	35383.15	0.43	11.14	45.28	0.14	0.14	0.00553	13.568
2	51.51	7.14	0.6074	47257.42	0.64	9.18	56.49	0.60	0.35	0.01289	25.929
3	18.20	4.00	0.0093	10224.58	0.41	14.13	32.57	0.11	0.01	0.34593	17.053
4	35.90	4.92	0.2933	57901.16	0.48	8.11	61.01	0.20	0.15	0.64890	36.409
5	17.62	7.80	0.2558	49753.67	0.53	8.47	65.82	0.17	0.13	0.22665	16.099
6	69.56	12.97	0.2052	21202.81	0.79	8.21	43.34	1.48	0.18	0.05115	24.783
7	12.76	7.37	0.9940	58960.50	0.61	9.05	46.97	0.50	0.45	0.22306	22.105
8	44.23	11.23	2.6716	44282.32	0.76	8.44	65.60	1.88	2.79	0.02304	24.003
9	61.28	18.10	0.3755	30845.95	0.86	9.33	50.77	0.76	1.18	0.15782	20.977
10	25.32	7.43	0.0602	6522.84	0.72	14.18	50.22	0.17	0.16	0.00000	11.420
11	57.13	8.08	0.2053	29514.41	0.69	9.96	55.23	0.44	0.46	0.19475	23.624
12	72.41	9.77	0.0263	8499.04	0.77	11.48	51.52	0.23	0.02	0.09164	19.751
13	56.13	3.42	0.1485	27655.66	0.61	8.46	43.15	0.40	0.02	0.45930	27.659
14	7.69	1.80	0.0232	8283.24	0.31	16.69	37.23	0.12	0.02	2.19513	17.345
15	28.34	8.45	5.4982	55685.60	0.78	8.94	54.44	4.23	4.48	0.28320	29.903
16	36.47	9.10	0.3191	24680.70	0.72	10.56	47.52	0.48	0.14	0.17256	14.860
17	56.90	6.84	0.2172	38310.97	0.55	10.23	54.92	0.42	0.07	0.01542	27.685
18	45.48	9.64	0.3370	26534.65	0.79	9.52	55.37	0.58	0.43	0.11479	22.832
19	62.88	11.17	0.2660	48862.19	0.66	9.10	58.09	0.29	0.17	0.04746	22.905
20	3.18	5.57	0.0048	2561.96	0.34	15.51	22.73	0.10	0.01	0.01152	4.488
21	13.02	3.47	0.0978	10606.77	0.45	15.16	43.24	0.22	0.02	0.24165	15.039
22	56.60	7.30	2.3070	117819.86	0.66	8.85	65.59	1.63	0.38	0.16630	23.828
23	8.92	3.62	1.4637	61464.32	0.29	13.19	44.86	1.57	0.67	0.36830	34.405
24	44.12	5.30	3.0656	79196.98	0.41	8.31	50.63	3.77	1.89	0.25532	32.634
25	0.00	2.62	0.0547	22609.00	0.41	19.23	43.02	0.04	0.03	0.23636	8.485
26	54.84	5.01	0.1538	30206.17	0.50	11.58	53.31	0.42	0.23	0.55055	27.273
27	35.62	6.89	0.4163	38506.05	0.50	10.00	59.11	0.45	0.50	0.34881	29.719
28	68.25	9.12	0.0527	26825.73	0.63	12.48	57.03	0.16	0.04	0.02650	23.659
29	13.58	1.09	0.0013	1007.52	0.11	21.35	29.69	0.08	0.01	0.27642	12.754
30	7.53	5.39	0.1943	37794.08	0.41	19.68	57.40	0.23	0.25	0.38268	20.173
31	26.62	6.46	0.0939	24842.76	0.48	10.10	51.77	0.08	0.05	0.25894	16.780
32	38.39	2.82	6.2080	83705.27	0.21	9.36	57.36	4.39	2.31	0.72792	30.548
33	22.03	7.76	12.5854	151474.36	0.52	8.57	57.98	3.38	3.21	0.77677	27.896
34	41.02	0.83	3.3312	79107.86	0.08	8.84	56.13	1.30	1.51	0.98746	40.552
35	24.68	6.18	1.4142	78595.26	0.57	13.50	52.60	0.80	0.70	0.33642	22.119
36	62.95	9.08	0.5571	42104.17	0.73	9.00	44.20	0.78	0.41	0.06113	22.349

OBS	CRES_60	DE_DOM70	ESGOTO70	LUZ70	PEA_SE70	PEA_TE70	PEPOP70	PORTE70	VTI70
1	5.41	4.80955	0.11055	0.47638	26.766	29.339	0.12	79.46	3417
2	8.64	5.12999	0.01723	0.70719	46.199	43.209	0.46	191.29	16069
3	4.75	4.87480	0.22882	0.46843	18.599	15.548	0.11	98.57	690
4	5.19	5.18421	0.53564	0.73451	70.422	22.089	0.19	965.96	50290
5	4.91	5.03648	0.19650	0.63181	50.153	22.433	0.13	1458.61	40841
6	14.26	5.11255	0.14870	0.77127	48.399	41.861	0.67	459.37	34912
7	7.08	5.15916	0.27144	0.64314	36.500	37.056	0.38	584.37	61944
8	20.44	5.10704	0.27427	0.82196	56.769	32.443	0.97	1073.37	212528
9	13.77	5.09632	0.17467	0.56922	46.699	36.385	0.22	315.27	26168
10	8.02	4.93282	0.01679	0.39060	33.898	33.019	0.13	48.23	5499
11	9.76	5.27815	0.21038	0.74604	50.803	37.039	0.31	146.00	8906
12	16.01	4.99600	0.03025	0.54048	45.454	43.259	0.14	12.24	355
13	3.69	7.94376	0.36915	0.84508	27.324	63.356	0.45	53.88	2910
14	5.15	4.80715	0.27729	0.41080	32.798	19.335	0.16	114.89	6664
15	8.92	5.03221	0.35547	0.83810	49.342	38.984	2.91	1096.96	789817
16	5.96	4.89159	0.16928	0.34531	36.132	30.477	0.31	58.28	13288
17	10.58	5.12245	0.15217	0.66369	44.490	40.398	0.34	1383.00	41494
18	9.78	5.20544	0.20168	0.62399	45.655	34.797	0.36	129.73	13882
19	19.97	5.15848	0.11184	0.73050	44.171	40.986	0.15	1403.57	19650
20	2.21	4.40691	0.01395	0.13584	10.496	20.575	0.09	30.45	335
21	4.34	5.03445	0.23136	0.55733	48.261	22.363	0.27	156.36	20640
22	13.51	5.10491	0.77211	0.87446	58.811	27.996	1.25	4401.59	484175
23	3.98	4.97779	0.42674	0.81754	33.543	37.578	1.70	1053.65	241288
24	9.45	4.89399	0.32226	0.93162	49.576	38.978	3.48	1972.86	581995
25	4.08	4.49576	0.00970	0.46303	35.886	36.371	0.05	94.02	3761
26	6.95	5.32188	0.33947	0.39471	43.901	45.138	0.40	666.51	21995
27	5.42	4.85266	0.47628	0.84932	47.519	36.738	0.36	543.24	54324
28	7.90	5.29779	0.10915	0.57413	46.778	33.424	0.10	447.33	4026
29	0.52	4.97501	0.17855	0.34045	19.243	21.248	0.12	58.71	822
30	3.87	4.95267	0.21270	0.50794	46.853	30.339	0.21	222.98	16501
31	0.51	4.59114	0.25043	0.45571	31.401	27.898	0.07	435.48	13500
32	5.60	4.68921	0.64082	0.97561	57.794	33.334	5.15	2291.62	1789757
33	9.52	5.06269	0.69704	0.92787	57.991	33.061	2.48	3940.35	2320870
34	2.86	4.39890	0.90955	0.99801	60.163	31.520	1.84	2237.84	1047311
35	7.80	5.19143	0.25349	0.71413	41.512	33.542	0.68	1520.91	272244
36	19.12	5.03629	0.28856	0.81070	41.721	49.569	0.50	546.57	31703

MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA

OBS	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9	COL10	COL11	COL12	COL1
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	38.2731	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	42.5730	57.6363	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	53.4019	57.1122	50.7011	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	48.5285	59.9170	60.9801	64.0729	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	60.4207	60.9727	62.4198	65.9601	65.8023	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	57.1468	56.7046	55.1193	65.7741	62.4298	70.3194	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	58.4538	57.0100	60.1869	60.7213	55.3999	69.0709	67.2733	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	59.2335	64.1716	63.3004	65.9457	64.2480	69.5597	64.8946	67.7693	0.0000	0.0000	0.0000	0.0000	0.0000
10	49.0469	60.3148	43.0876	65.9910	52.0209	59.1546	57.6303	58.1895	63.9225	0.0000	0.0000	0.0000	0.0000
11	38.8219	36.5393	53.5839	56.1643	49.8680	57.4099	58.3618	57.5831	52.7094	43.3212	0.0000	0.0000	0.0000
12	54.7449	50.6212	48.4834	55.0655	61.2681	63.3213	64.5502	53.8489	55.2628	44.4581	43.5204	0.0000	0.0000
13	59.6045	61.3809	62.7439	66.9623	65.2568	71.8048	69.9485	69.6579	70.2336	61.0059	58.9623	61.0335	0.0000
14	62.7834	59.0835	57.5162	67.3343	63.1406	71.2508	70.1469	69.1999	69.4543	62.0812	55.3937	61.0936	71.2988
15	55.4038	55.2968	56.4760	65.5132	70.5299	67.2694	65.3838	57.3320	64.0229	52.1345	50.1891	63.8967	66.3973
16	43.1220	45.0697	57.3674	53.6653	54.9889	57.4979	50.1409	56.4603	61.1032	43.0248	45.3616	30.9821	58.8780
17	49.0953	38.4132	47.5464	55.8064	47.1221	60.9800	60.8926	64.2595	57.7271	45.0662	54.7512	38.8045	61.9810
18	61.0291	47.9267	41.5364	58.0875	47.2073	63.9104	64.6412	66.1580	61.2986	56.1906	42.3507	50.7770	63.8920
19	50.8545	60.1867	62.7137	65.0485	69.3306	67.7216	63.6337	61.4174	67.7101	54.5140	51.6325	54.9296	67.8555
20	45.5055	58.0174	54.8955	59.0491	56.1377	65.8535	68.4807	64.7718	59.7537	54.2932	58.3130	61.3173	65.7166
21	52.8304	50.8575	54.7285	44.2707	55.0506	61.4931	56.4304	57.8958	58.5023	35.5738	28.8911	53.4645	59.5854
22	61.8072	61.8047	60.5082	66.7401	64.5349	70.7714	68.9922	67.6807	69.5111	63.0530	56.7946	63.9991	70.7775
23	64.0001	50.3883	57.5315	63.9772	66.1459	66.3963	62.4071	59.0202	65.4591	53.5613	52.7022	61.2263	65.9733
24	56.5748	59.9202	59.3979	61.5646	55.5341	68.7762	66.9654	71.6786	68.3438	60.7301	55.4007	52.8000	69.5236
25	59.2039	49.7505	62.4244	60.1983	57.5197	63.6715	57.0135	63.4899	67.1843	55.2711	57.4479	45.6357	64.9983
26	54.2885	63.5140	68.5657	56.5985	64.2434	68.3098	62.6232	65.5139	68.2643	51.5543	48.9517	58.2402	67.8806
27	47.5570	67.7022	58.4946	56.6216	55.0498	64.9477	61.4253	66.3131	66.8068	60.0523	41.5747	49.7809	65.4367
28	69.3179	47.7048	54.2241	60.4059	60.8262	67.2131	64.1847	62.5649	64.6810	51.5400	49.1335	62.4832	66.2887
29	45.7151	55.6286	39.7688	67.8820	59.1334	59.9147	59.7393	53.7998	61.2498	63.2684	39.5233	46.2226	60.8256
30	41.5595	51.9360	58.4625	56.8321	59.2742	65.5881	67.5819	62.9038	57.6050	43.8014	61.2807	58.7983	65.1669
31	48.3345	47.2964	48.9592	61.2371	60.8739	65.4637	67.9480	60.1477	57.7991	45.6098	53.0845	58.1515	64.7246
32	56.0211	61.4060	61.3180	68.0276	67.2745	71.1252	69.6567	67.3043	69.2395	59.9349	57.4933	60.1514	71.2791
33	59.5385	61.3658	62.9312	66.6773	64.9759	71.6202	69.2945	69.7928	70.5981	60.8727	57.9875	59.4276	71.9015
34	62.7179	60.7101	62.2797	65.5368	64.4797	71.7130	69.1667	69.6255	70.4647	60.2842	55.9525	61.6591	71.6635
35	50.1360	55.7105	51.3275	54.9671	57.4604	65.2418	64.4169	62.5183	61.0388	46.9169	38.3722	55.0750	64.1904
36	45.3789	53.4691	50.6916	65.9669	66.6716	63.8637	63.2256	55.4467	61.3285	52.1946	48.3184	53.2156	63.8585

MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA - CONTINUAÇÃO)

OBS	COL14	COL15	COL16	COL17	COL18	COL19	COL20	COL21	COL22	COL23	COL24	COL25	COL26
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
15	65.9947	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
16	57.1842	55.0602	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
17	62.7015	52.2685	61.3494	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
18	67.3329	53.8159	48.9572	63.2116	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
19	66.3312	68.8299	65.0767	58.2440	54.9787	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
20	64.0545	56.4439	38.5823	52.0328	56.6390	55.7661	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	59.7512	57.4983	53.1004	51.2378	56.5377	59.9070	47.9969	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
22	70.3380	65.5427	51.7322	55.1107	61.8253	64.5390	65.7652	55.9731	0.0000	0.0000	0.0000	0.0000	0.0000
23	65.7618	68.2640	54.3579	50.7511	53.6228	64.7741	52.8328	53.7975	67.0584	0.0000	0.0000	0.0000	0.0000
24	69.1642	56.9307	56.2636	63.2877	66.3851	61.7532	64.8745	57.4336	67.5684	57.5953	0.0000	0.0000	0.0000
25	63.2994	57.6755	61.2413	55.6578	52.5151	61.7338	50.1388	48.4163	64.2708	65.5693	62.2518	0.0000	0.0000
26	65.3961	62.5852	57.5871	52.4559	54.9861	66.1432	6 0.0030	64.7410	66.6415	62.3647	65.3579	62.4424	0.0000
27	64.4883	53.6607	51.7547	52.6212	60.4616	60.3659	61.2917	58.6272	63.8965	50.9683	68.2981	54.5970	65.9477
28	67.5211	65.3675	52.1219	54.7039	61.0739	61.5669	54.2365	59.3910	67.0755	69.7212	60.8213	63.4286	62.5489
29	62.8665	61.3691	52.6856	52.9747	57.6145	62.8430	49.6399	44.4249	58.7477	55.1496	56.1895	49.4478	50.0913
30	62.5449	59.8166	47.8044	58.2095	53.4065	60.4852	68.0510	52.5441	61.8027	53.6066	62.0245	50.3244	60.6191
31	65.1732	65.0672	53.3431	63.2744	60.2294	63.9739	60.3258	56.5251	60.2407	57.2035	59.5795	48.9050	56.4288
32	70.6055	68.2896	60.7300	62.4368	62.5610	69.9145	64.1923	59.5792	69.0591	65.2292	67.5061	62.5170	66.5676
33	71.2660	66.1235	60.7730	62.9519	64.2988	68.4019	64.4516	60.4829	70.2166	65.6708	69.7410	65.4330	68.0832
34	71.4054	66.1409	58.3955	60.9629	64.8332	67.1365	64.0487	61.7583	70.8870	66.8519	69.3360	65.4508	68.5108
35	65.1378	60.6058	54.7348	60.2143	64.1687	63.2975	57.4121	67.2251	59.6841	52.9489	63.1717	47.5796	62.2885
36	63.8216	68.2742	59.4830	57.8127	54.3566	69.0290	53.1747	53.2738	59.8020	60.2935	56.1360	52.4380	56.7958

(MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA - CONTINUAÇÃO)

OBS	COL27	COL28	COL29	COL30	COL31	COL32	COL33	COL34	COL35	COL36	CIDADE
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	ARUJA
2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	BARUERI
3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	BIRITIBA-MIRIM
4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	CAIEIRAS
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	CAJAMAR
6	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	CARAPICUIBA
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	COTIA
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	DIADEMA
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	EMBU
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	EMBU-GUACU
11	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	FERRAZ DE VASCONCE
12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	FRANCISCO MORATO
13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	FRANCO DA ROCHA
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	GUARAREMA
15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	GUARULHOS
16	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	ITAPECERICA DA SERR
17	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	ITAPEVI
18	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	ITAQUAQUECETUBA
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	JANDIRA
20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	JUQUITIBA
21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	MAIRIPORA
22	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	MAUA
23	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	MOGI DAS CRUZES
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	OSASCO
25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	PIRAPORA DO BOM JES
26	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	POA
27	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	RIBEIRAO PIRES
28	53.3713	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	RIO GRANDE DA SERR
29	55.2859	52.6306	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	SALESOPOLIS
30	56.4421	54.0454	48.1937	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0	SANTA ISABEL
31	53.0628	58.6647	59.2193	66.3822	0.0000	0.0000	0.0000	0.0000	0.0000	0	SANTANA DO PARNAIB
32	64.1670	64.4765	64.0818	65.5500	67.2420	0.0000	0.0000	0.0000	0.0000	0	SANTO ANDRE
33	65.8833	66.0992	61.3169	64.3875	64.5981	71.3259	0.0000	0.0000	0.0000	0	SAO BERNARDO DO CA
34	65.4554	68.1098	59.5628	62.9967	63.3155	70.4753	71.6447	0.0000	0.0000	0	SAO CAETANO DO SUL
35	63.2362	57.7559	57.4822	60.5905	65.9866	65.6709	64.8355	64.4654	0.0000	0	SUZANO
36	53.2783	56.9762	66.8362	58.9448	67.3856	67.8693	64.2238	62.2810	62.3119	0	TABOAO DA SERRA

**** **** ******** **** ****

MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA

BLOCO DAS CATEGÓRICAS

OBS	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9	COL10	COL11	COL12	COL
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	48.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	48.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	28.98	42.00	28.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	28.98	42.00	28.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	48.11	0.00	48.11	42.00	42.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	28.98	42.00	28.98	0.00	0.00	42.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	48.11	0.00	48.11	42.00	42.00	0.00	42.00	0.00	0.00	0.00	0.00	0.00	0.00
9	48.11	0.00	48.11	42.00	42.00	0.00	42.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	48.11	0.00	28.98	28.98	48.11	28.98	48.11	48.11	0.00	0.00	0.00	0.00
11	48.11	0.00	48.11	42.00	42.00	0.00	42.00	0.00	0.00	48.11	0.00	0.00	0.00
12	48.11	0.00	48.11	42.00	42.00	0.00	42.00	0.00	0.00	48.11	0.00	0.00	0.00
13	28.98	42.00	28.98	0.00	0.00	42.00	0.00	42.00	42.00	28.98	42.00	42.00	0.00
14	0.00	48.11	0.00	28.98	28.98	48.11	28.98	48.11	48.11	0.00	48.11	48.11	28.98
15	395.49	220.20	395.49	356.25	356.25	220.20	356.25	220.20	220.20	395.49	220.20	220.20	356.25
16	48.11	0.00	48.11	42.00	42.00	0.00	42.00	0.00	0.00	48.11	0.00	0.00	42.00
17	48.98	42.00	28.98	0.00	0.00	42.00	0.00	42.00	42.00	28.98	42.00	42.00	0.00
18	48.11	0.00	48.11	42.00	42.00	0.00	42.00	0.00	0.00	48.11	0.00	0.00	42.00
19	0.00	48.11	0.00	28.98	28.98	48.11	28.98	48.11	48.11	0.00	48.11	48.11	28.98
20	0.00	48.11	0.00	28.98	28.98	48.11	28.98	48.11	48.11	0.00	48.11	48.11	28.98
21	0.00	48.11	0.00	28.98	28.98	48.11	28.98	48.11	48.11	0.00	48.11	48.11	28.98
22	48.11	0.00	48.11	42.00	42.00	0.00	42.00	0.00	0.00	48.11	0.00	0.00	42.00
23	48.11	0.00	48.11	42.00	42.00	0.00	42.00	0.00	0.00	48.11	0.00	0.00	42.00
24	395.49	220.20	395.49	356.25	356.25	220.20	356.25	220.20	220.20	395.49	220.20	220.20	356.25
25	0.00	48.11	0.00	28.98	28.98	48.11	28.98	48.11	48.11	0.00	48.11	48.11	28.98
26	28.98	42.00	28.98	0.00	0.00	42.00	0.00	42.00	42.00	28.98	42.00	42.00	0.00
27	28.98	42.00	28.98	0.00	0.00	42.00	0.00	42.00	42.00	28.98	42.00	42.00	0.00
28	28.98	42.00	28.98	0.00	0.00	42.00	0.00	42.00	42.00	28.98	42.00	42.00	0.00
29	0.00	48.11	0.00	28.98	28.98	48.11	28.98	48.11	48.11	0.00	48.11	48.11	28.98
30	0.00	48.11	0.00	28.98	28.98	48.11	28.98	48.11	48.11	0.00	48.11	48.11	28.98
31	0.00	48.11	0.00	28.98	28.98	48.11	28.98	48.11	48.11	0.00	48.11	48.11	28.98
32	1578.84	1619.06	1578.84	1602.69	1602.69	1619.06	1602.69	1619.06	1619.06	1578.84	1619.06	1619.06	1602.69
33	395.49	220.20	395.49	356.25	356.25	220.20	356.25	220.20	220.20	395.49	220.20	220.20	356.25
34	1578.84	1619.06	1578.84	1602.69	1602.69	1619.06	1602.69	1619.06	1619.06	1578.84	1619.06	1619.06	1602.69
35	48.11	0.00	48.11	42.00	42.00	0.00	42.00	0.00	0.00	48.11	0.00	0.00	42.00
36	48.11	0.00	48.11	42.00	42.00	0.00	42.00	0.00	0.00	48.11	0.00	0.00	42.00

MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA
BLOCO DAS CATEGÓRICAS

- CONTINUAÇÃO -

OBS	COL14	COL15	COL16	COL17	COL18	COL19	COL20	COL21	COL22	COL23	COL24	COL25
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	395.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	48.11	220.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	28.98	356.25	42.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	48.11	220.20	0.00	42.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	0.00	395.49	48.11	28.98	48.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20	0.00	395.49	48.11	28.98	48.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	0.00	395.49	48.11	28.98	48.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	48.11	220.20	0.00	42.00	0.00	48.11	48.11	48.11	0.00	0.00	0.00	0.00
23	48.11	220.20	0.00	42.00	0.00	48.11	48.11	48.11	0.00	0.00	0.00	0.00
24	395.49	0.00	220.20	356.25	220.20	395.49	395.49	395.49	220.20	220.20	0.00	0.00
25	0.00	395.49	48.11	28.98	48.11	0.00	0.00	0.00	48.11	48.11	395.49	0.00
26	28.98	356.25	42.00	0.00	42.00	28.98	28.98	28.98	42.00	42.00	356.25	28.98
27	28.98	356.25	42.00	0.00	42.00	28.98	28.98	28.98	42.00	42.00	356.25	28.98
28	28.98	356.25	42.00	0.00	42.00	28.98	28.98	28.98	42.00	42.00	356.25	28.98
29	0.00	395.49	48.11	28.98	48.11	0.00	0.00	0.00	48.11	48.11	395.49	0.00
30	0.00	395.49	48.11	28.98	48.11	0.00	0.00	0.00	48.11	48.11	395.49	0.00
31	0.00	395.49	48.11	28.98	48.11	0.00	0.00	0.00	48.11	48.11	395.49	0.00
32	1578.84	1644.58	1619.06	1602.69	1619.06	1578.84	1578.84	1578.84	1619.06	1619.06	1644.58	1578.84
33	395.49	0.00	220.20	356.25	220.20	395.49	395.49	395.49	220.20	220.20	0.00	395.49
34	1578.84	1644.58	1619.06	1602.69	1619.06	1578.84	1578.84	1578.84	1619.06	1619.06	1644.58	1578.84
35	48.11	220.20	0.00	42.00	0.00	48.11	48.11	48.11	0.00	0.00	220.20	48.11
36	48.11	220.20	0.00	42.00	0.00	48.11	48.11	48.11	0.00	0.00	220.20	48.11

MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA
BLOCO DAS CATEGÓRICAS

- CONTINUAÇÃO -

OBS	COL27	COL28	COL29	COL30	COL31	COL32	COL33	COL34	COL35	COL36	CIDADE
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	ARUJA
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	BARUERI
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	BIRITIBA-MIRIM
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	CAIEIRAS
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	CAJAMAR
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	CARAPICUIBA
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	COTIA
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	DIADEMA
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	EMBU
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	EMBU-GUACU
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	FERRAZ DE VASCONC
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	FRANCISCO MORATO
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	FRANCO DA ROCHA
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	GUARAREMA
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	GUARULHOS
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	ITAPECERICA DA SER
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	ITAPEVI
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	ITAQUAQUECETUBA
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	JANDIRA
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	JUQUITIBA
21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	MAIRIPORA
22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	MAUA
23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	MOGI DAS CRUZES
24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	OSASCO
25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	PIRAPORA DO BOM JE
26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	POA
27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	RIBEIRAO PIRES
28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	RIO GRANDE DA SERR
29	28.98	28.98	0.00	0.00	0.00	0.00	0.00	0.00	0	0	SALESOPOLIS
30	28.98	28.98	0.00	0.00	0.00	0.00	0.00	0.00	0	0	SANTA ISABEL
31	28.98	28.98	0.00	0.00	0.00	0.00	0.00	0.00	0	0	SANTANA DO PARNAI
32	1602.69	1602.69	1578.84	1578.84	1578.84	0.00	0.00	0.00	0	0	SANTO ANDRE
33	356.25	356.25	395.49	395.49	395.49	1644.58	0.00	0.00	0	0	SAO BERNARDO DO CA
34	1602.69	1602.69	1578.84	1578.84	1578.84	0.00	1644.58	0.00	0	0	SAO CAETANO DO SUL
35	42.00	42.00	48.11	48.11	48.11	1619.06	220.20	1619.06	0	0	SUZANO
36	42.00	42.00	48.11	48.11	48.11	1619.06	220.20	1619.06	0	0	TABOAO DA SERRA

**** **** ******** **** ****

MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA

BLOCO DAS CONTÍNUAS

OBS	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9	COL10	COL11	COL12	COL1
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	83.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	42.57	130.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	78.58	97.17	82.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	46.08	108.88	64.76	64.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	83.50	60.97	113.30	99.56	108.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	71.14	88.21	75.35	65.77	62.43	95.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	75.39	57.01	104.92	87.54	91.13	69.07	85.55	0.00	0.00	0.00	0.00	0.00	0.00
9	89.69	64.17	121.55	96.50	103.71	69.56	86.90	67.77	0.00	0.00	0.00	0.00	0.00
10	49.05	115.36	43.09	85.89	44.29	92.23	66.35	85.12	104.37	0.00	0.00	0.00	0.00
11	64.69	36.54	107.25	88.83	91.44	57.41	82.48	57.58	52.71	79.18	0.00	0.00	0.00
12	63.62	50.62	85.15	81.81	96.92	63.32	82.75	53.85	55.26	63.33	43.52	0.00	0.00
13	85.13	104.82	94.50	66.96	65.26	108.79	69.95	99.86	104.17	81.25	95.01	91.16	0.00
14	62.78	114.29	57.52	98.36	66.53	104.49	89.99	96.29	110.07	62.08	91.42	80.13	102.67
15	425.15	232.71	526.35	447.98	496.59	296.39	471.83	279.07	294.61	474.18	259.64	340.17	460.01
16	68.79	45.07	110.83	99.41	109.64	57.50	87.35	56.46	61.10	78.69	45.36	30.98	108.01
17	66.34	60.86	71.02	55.81	47.12	76.97	60.89	73.47	70.67	57.03	69.81	47.94	61.98
18	58.84	47.93	67.15	74.28	72.31	63.91	72.29	66.16	61.30	64.00	42.35	50.78	83.47
19	50.85	102.43	62.71	80.26	56.91	88.00	67.66	75.55	95.36	54.51	74.70	61.00	83.41
20	45.51	114.80	54.90	100.61	70.07	100.68	98.87	93.45	101.95	54.29	95.92	81.93	107.63
21	52.83	94.56	54.73	57.55	40.70	83.23	58.53	73.49	87.62	35.57	53.41	61.00	73.21
22	102.52	61.80	129.02	111.11	117.81	70.77	104.82	67.68	69.51	113.76	56.79	64.00	118.53
23	85.62	50.39	106.95	90.02	101.10	66.40	79.91	59.02	65.46	85.18	52.70	61.23	95.41
24	303.40	143.76	406.35	320.31	357.88	204.32	349.70	199.84	205.36	359.86	171.28	235.50	339.42
25	59.20	126.92	62.42	93.94	63.63	118.87	79.57	112.55	129.76	55.27	115.43	86.63	99.09
26	73.25	117.50	93.76	56.60	64.24	115.84	62.62	106.26	112.75	65.23	95.54	98.91	67.88
27	49.93	81.75	67.10	56.62	55.05	72.55	61.43	67.13	71.36	57.14	48.23	50.52	65.44
28	77.82	65.55	68.96	60.41	60.83	78.61	64.18	67.18	73.03	54.76	59.59	67.02	66.29
29	45.72	136.32	39.77	105.09	68.71	118.64	85.77	106.38	127.35	63.27	101.04	90.74	98.38
30	41.56	102.05	58.46	83.57	58.39	93.74	83.14	84.91	93.13	43.80	92.22	72.75	92.26
31	48.33	106.89	48.96	60.42	32.43	103.09	55.95	91.63	102.80	45.61	93.50	81.57	64.25
32	1626.10	1799.71	1708.73	1616.07	1662.93	1672.17	1888.80	1676.53	1739.18	1564.77	1565.92	1788.87	1704.46
33	388.39	217.09	491.91	407.30	449.20	279.06	433.91	269.85	279.50	442.03	245.75	314.01	423.68
34	1540.42	1720.56	1617.31	1517.03	1563.59	1594.30	1791.76	1600.40	1661.95	1472.73	1485.93	1711.92	1608.29
35	56.95	55.71	85.93	80.37	91.77	65.24	81.28	62.52	61.04	63.72	38.37	55.07	92.98
36	73.33	53.47	106.43	113.95	123.56	63.86	102.66	55.45	61.33	90.14	48.32	53.22	115.23

MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA
BLOCO DAS CONTÍNUAS

- CONTINUAÇÃO -

OBS	COL14	COL15	COL16	COL17	COL18	COL19	COL20	COL21	COL22	COL23	COL24	COL25
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	459.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	93.01	297.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	85.79	368.06	89.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	75.30	339.97	48.96	61.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	66.33	490.25	87.94	65.52	49.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20	64.05	440.50	75.99	85.66	66.19	55.77	0.00	0.00	0.00	0.00	0.00	0.00
21	59.75	432.43	77.43	56.58	53.01	59.91	48.00	0.00	0.00	0.00	0.00	0.00
22	121.21	274.44	51.73	81.87	61.83	102.45	118.22	95.34	0.00	0.00	0.00	0.00
23	97.54	268.10	54.36	59.19	53.62	83.59	86.19	74.08	67.06	0.00	0.00	0.00
24	339.89	56.93	205.05	255.36	258.97	360.26	326.01	309.45	182.89	163.86	0.00	0.00
25	63.30	527.24	119.03	81.46	82.45	61.73	50.14	48.42	137.11	119.31	408.90	0.00
26	90.20	462.90	117.27	52.46	85.11	75.13	95.35	71.80	124.94	102.34	341.96	89.96
27	72.70	395.03	71.50	52.62	50.65	52.77	80.05	49.10	82.26	51.01	285.95	65.53
28	81.87	426.78	75.67	54.70	55.06	60.10	79.13	56.00	89.24	73.56	298.52	80.49
29	62.87	523.84	114.00	82.25	91.07	62.84	49.64	44.42	135.11	112.42	395.75	49.45
30	62.54	478.09	78.55	77.02	56.29	60.49	68.05	52.54	107.59	80.30	357.38	50.32
31	65.17	494.09	93.56	54.52	72.59	63.97	60.33	56.53	115.50	93.37	365.69	48.91
32	1667.38	1722.42	1769.17	1758.69	1687.46	1616.28	1532.07	1621.85	1766.62	1606.34	1598.54	1679.74
33	424.02	66.12	281.45	336.90	328.77	448.93	407.61	394.52	257.43	243.82	69.74	494.11
34	1575.80	1715.04	1688.38	1660.67	1611.28	1521.12	1439.54	1531.64	1689.99	1529.51	1595.14	1590.29
35	82.10	295.65	54.73	68.01	64.17	67.31	75.96	72.69	59.68	52.95	204.64	86.51
36	101.92	294.81	59.48	88.19	54.36	94.17	92.86	79.88	59.80	60.29	189.10	112.50

MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA
BLOCO DAS CONTÍNUAS

- CONTINUAÇÃO -

OBS	COL27	COL28	COL29	COL30	COL31	COL32	COL33	COL34	COL35	COL36	CIDADE
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	ARUJA
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	BARUERI
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	BIRITIBA-MIRIM
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	CAIEIRAS
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	CAJAMAR
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	CARAPICUIBA
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	COTIA
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	DIADEMA
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	EMBU
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	EMBU-GUACU
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	FERRAZ DE VASCONCELOS
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	FRANCISCO MORATO
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	FRANCO DA ROCHA
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	GUARAREMA
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	GUARULHOS
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	ITAPECERICA DA SERRA
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	ITAPEVI
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	ITAQUAQUECETUBA
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	JANDIRA
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	JUQUITIBA
21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	MAIRIPORA
22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	MAUA
23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	MOGI DAS CRUZES
24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	OSASCO
25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	PIRAPORA DO BOM JESUS
26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	POA
27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	RIBEIRAO PIRES
28	53.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	RIO GRANDE DA SERRA
29	69.69	73.17	0.00	0.00	0.00	0.00	0.00	0.00	0.0000	0	SALESOPOLIS
30	60.38	64.11	48.19	0.00	0.00	0.00	0.00	0.00	0.0000	0	SANTA ISABEL
31	29.44	41.17	59.22	66.38	0.00	0.00	0.00	0.00	0.0000	0	SANTANA DO PARNAIBA
32	1581.65	1600.31	1746.59	1727.93	1616.87	0.00	0.00	0.00	0.0000	0	SANTO ANDRE
33	365.42	385.68	482.90	441.77	452.73	1647.76	0.00	0.00	0.0000	0	SAO BERNARDO DO CAMPO
34	1486.39	1507.39	1649.69	1633.00	1520.56	70.48	1642.85	0.00	0.0000	0	SAO CAETANO DO SUL
35	62.64	60.96	99.94	72.47	87.34	1496.66	278.20	1417.00	0.0000	0	SUZANO
36	75.26	82.75	130.43	91.96	109.88	1604.70	269.08	1520.66	62.3119	0	TABOAO DA SERRA

**** **** ******** **** ****

MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA

BLOCO DOS PRODUTOS CRUZADOS

OBS	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9	COL10	COL11	COL12	CO
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
2	-93.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
3	0.00	-120.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
4	-54.16	-82.05	-60.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
5	-26.53	-90.96	-32.76	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
6	-71.19	0.00	-98.99	-75.60	-84.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
7	-42.98	-73.51	-49.21	0.00	0.00	-67.06	0.00	0.00	0.00	0.00	0.00	0.00	
8	-65.05	0.00	-92.84	-68.82	-77.72	0.00	-60.27	0.00	0.00	0.00	0.00	0.00	
9	-78.57	0.00	-106.36	-72.55	-81.46	0.00	-64.01	0.00	0.00	0.00	0.00	0.00	
10	0.00	-103.15	0.00	-48.88	-21.25	-81.19	-37.70	-75.04	-88.56	0.00	0.00	0.00	
11	-73.98	0.00	-101.77	-74.66	-83.57	0.00	-66.12	0.00	0.00	-83.97	0.00	0.00	
12	-56.99	0.00	-84.78	-68.74	-77.65	0.00	-60.20	0.00	0.00	-66.98	0.00	0.00	
13	-54.51	-85.44	-60.74	0.00	0.00	-78.99	0.00	-72.20	-75.94	-49.23	-78.05	-72.13	0.00
14	0.00	-103.31	0.00	-6	0.00	-32.38	-81.35	-48.82	-75.20	-88.72	0.00	-84.13	-67.14 -60.35
15	-765.23	-397.61	-865.37	-738.71	-782.31	-449.32	-762.70	-441.94	-450.79	-817.54	-429.65	-496.47	-749.86
16	-73.78	0.00	-101.57	-87.75	-96.65	0.00	-79.20	0.00	0.00	-83.77	0.00	0.00	-91.13
17	-46.22	-64.44	-52.46	0.00	0.00	-57.99	0.00	-51.21	-54.94	-40.94	-57.05	-51.13	0.00
18	-45.92	0.00	-73.72	-58.19	-67.10	0.00	-49.65	0.00	0.00	-55.92	0.00	0.00	-61.58
19	0.00	-90.35	0.00	-44.19	-16.56	-68.39	-33.01	-62.24	-75.76	0.00	-71.17	-54.18	-44.54
20	0.00	-104.90	0.00	-70.55	-42.92	-82.93	-59.37	-76.79	-90.31	0.00	-85.72	-68.72	-70.90
21	0.00	-91.81	0.00	-42.26	-14.63	-69.85	-31.08	-63.70	-77.22	0.00	-72.63	-55.64	-42.61
22	-88.82	0.00	-116.62	-86.37	-95.27	0.00	-77.82	0.00	0.00	-98.82	0.00	0.00	-89.75
23	-69.73	0.00	-97.53	-68.04	-76.95	0.00	-59.50	0.00	0.00	-79.73	0.00	0.00	-71.43
24	-642.32	-304.04	-742.45	-615.00	-658.59	-355.75	-638.98	-348.36	-357.21	-694.62	-336.07	-402.89	-626.15
25	0.00	-125.28	0.00	-62.72	-35.09	-103.31	-51.54	-97.16	-110.69	0.00	-106.10	-89.10	-63.07
26	-47.94	-95.98	-54.17	0.00	0.00	-89.53	0.00	-82.75	-86.48	-42.66	-88.59	-82.67	0.00
27	-31.35	-56.05	-37.59	0.00	0.00	-49.60	0.00	-42.81	-46.55	-26.07	-48.66	-42.74	0.00
28	-37.49	-59.85	-43.72	0.00	0.00	-53.39	0.00	-46.61	-50.35	-32.21	-52.46	-46.54	0.00
29	0.00	-128.80	0.00	-66.19	-38.56	-106.84	-55.01	-100.69	-114.21	0.00	-109.62	-92.63	-66.54
30	0.00	-98.23	0.00	-55.72	-28.09	-76.26	-44.54	-70.12	-83.64	0.00	-79.05	-62.06	-56.07
31	0.00	-107.70	0.00	-28.16	-0.53	-85.74	-16.98	-79.59	-93.11	0.00	-88.52	-71.53	-28.51
32	-3148.92	-3357.37	-3226.26	-3150.73	-3198.35	-3220.11	-3421.83	-3228.29	-3289.01	-3083.67	-3127.50	-3347.78	-3235.8
33	-724.34	-375.93	-824.47	-696.87	-740.47	-427.64	-720.86	-420.25	-429.10	-776.65	-407.96	-474.78	-708.03
34	-3056.54	-3278.91	-3133.87	-3054.18	-3101.80	-3141.65	-3325.28	-3149.84	-3210.55	-2991.29	-3049.04	-3269.32	-3139.3
35	-54.92	0.00	-82.71	-67.41	-76.31	0.00	-58.86	0.00	0.00	-64.91	0.00	0.00	-70.79
36	-76.06	0.00	-103.85	-89.98	-98.89	0.00	-81.44	0.00	0.00	-86.05	0.00	0.00	-93.37

MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA
BLOCO DOS PRODUTOS CRUZADOS

- CONTINUAÇÃO -

OBS	COL14	COL15	COL16	COL17	COL18	COL19	COL20	COL21	COL22	COL23	COL24	COL25	C
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
15	-789.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
16	-83.94	-462.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
17	-52.07	-672.04	-70.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
18	-56.08	-506.36	0.00	-40.58	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
19	0.00	-816.92	-70.98	-36.26	-43.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
20	0.00	-779.55	-85.52	-62.61	-57.66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
21	0.00	-770.43	-72.44	-34.33	-44.58	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
22	-98.98	-429.10	0.00	-68.76	0.00	-86.02	-100.56	-87.48	0.00	0.00	0.00	0.00	0.0
23	-79.89	-420.04	0.00	-50.44	0.00	-66.93	-81.47	-68.39	0.00	0.00	0.00	0.00	0.0
24	-666.22	0.00	-368.99	-548.33	-412.78	-694.00	-656.63	-647.51	-335.52	-326.46	0.00	0.00	0.0
25	0.00	-865.06	-105.90	-54.79	-78.04	0.00	0.00	0.00	-120.94	-101.85	-742.14	0.00	0.
26	-53.79	-756.57	-101.68	0.00	-72.12	-37.97	-64.33	-36.04	-100.30	-81.97	-632.85	-56.50	0.
27	-37.20	-697.62	-61.74	0.00	-32.19	-21.39	-47.74	-19.45	-60.36	-42.04	-573.91	-39.91	0.
28	-43.33	-717.67	-65.54	0.00	-35.99	-27.52	-53.87	-25.59	-64.16	-45.84	-593.95	-46.05	0.
29	0.00	-857.97	-109.42	-58.26	-81.57	0.00	0.00	0.00	-124.47	-105.38	-735.05	0.00	-59.
30	0.00	-813.77	-78.85	-47.79	-51.00	0.00	0.00	0.00	-93.90	-74.80	-690.85	0.00	-49.
31	0.00	-824.52	-88.33	-20.23	-60.47	0.00	0.00	0.00	-103.37	-84.28	-701.60	0.00	-21.
32	-3175.62	-3298.72	-3327.51	-3298.95	-3243.96	-3125.20	-3046.72	-3141.11	-3316.62	-3160.17	-3175.62	-3196.06	-3104.8
33	-748.24	0.00	-440.88	-630.20	-484.67	-776.03	-738.65	-729.53	-407.41	-398.35	0.00	-824.17	-714.
34	-3083.23	-3293.49	-3249.05	-3202.39	-3165.51	-3032.82	-2954.33	-3048.73	-3238.17	-3081.72	-3170.39	-3103.68	-3008.2
35	-65.08	-455.25	0.00	-49.80	0.00	-52.12	-66.66	-53.58	0.00	0.00	-361.67	-87.04	-81.
36	-86.21	-446.74	0.00	-72.37	0.00	-73.25	-87.79	-74.71	0.00	0.00	-353.16	-108.17	-103

MATRIZ DE DISTÂNCIAS DE MAHALANOBIS ESTENDIDA
BLOCO DOS PRODUTOS CRUZADOS

- CONTINUAÇÃO -

OBS	COL27	COL28	COL29	COL30	COL31	COL32	COL33	COL34	COL35	COL36	CIDADE
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	ARUJA
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	BARUERI
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	BIRITIBA-MIRIM
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	CAIEIRAS
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	CAJAMAR
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	CARAPICUIBA
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	COTIA
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	DIADEMA
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	EMBU
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	EMBU-GUACU
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	FERRAZ DE VASCONCELOS
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	FRANCISCO MORATO
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	FRANCO DA ROCHA
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	GUARAREMA
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	GUARULHOS
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	ITAPECERICA DA SERRA
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	ITAPEVI
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	ITAQUAQUECETUBA
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	JANDIRA
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	JUQUITIBA
21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	MAIRIPORA
22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	MAUA
23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	MOGI DAS CRUZES
24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	OSASCO
25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	PIRAPORA DO BOM JESUS
26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	POA
27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	RIBEIRAO PIRES
28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	RIO GRANDE DA SERRA
29	-43.38	-49.52	0.00	0.00	0.00	0.00	0.00	0.00	0	0	SALESOPOLIS
30	-32.92	-39.05	0.00	0.00	0.00	0.00	0.00	0.00	0	0	SANTA ISABEL
31	-5.36	-11.49	0.00	0.00	0.00	0.00	0.00	0.00	0	0	SANTANA DO PARNAIBA
32	-3120.18	-3138.52	-3261.35	-3241.23	-3128.47	0.00	0.00	0.00	0	0	SANTO ANDRE
33	-655.78	-675.83	-817.08	-772.87	-783.62	-3221.02	0.00	0.00	0	0	SAO BERNARDO DO CAM
34	-3023.63	-3041.97	-3168.97	-3148.84	-3036.08	0.00	-3215.79	0.00	0	0	SAO CAETANO DO SUL
35	-41.40	-45.20	-90.57	-59.99	-69.47	-3050.06	-433.56	-2971.60	0	0	SUZANO
36	-63.98	-67.78	-111.70	-81.13	-90.60	-3155.90	-425.05	-3077.44	0	0	TABOAO DA SERRA

ANÁLISE DE AGRUPAMENTO PELO MÉTODO HIERÁRQUICO DE LIGAÇÃO MÉDIA

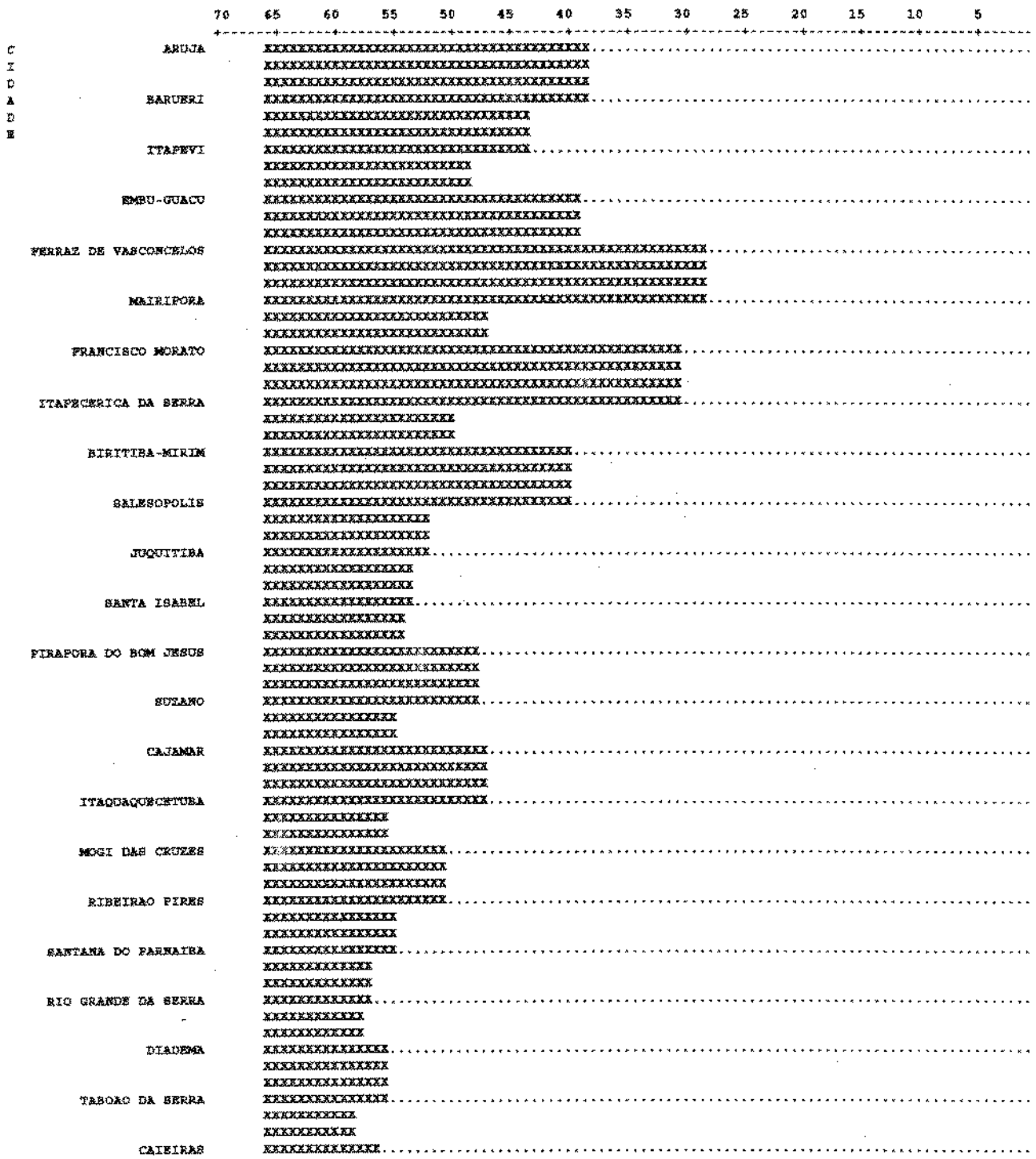
Average Linkage Cluster Analysis

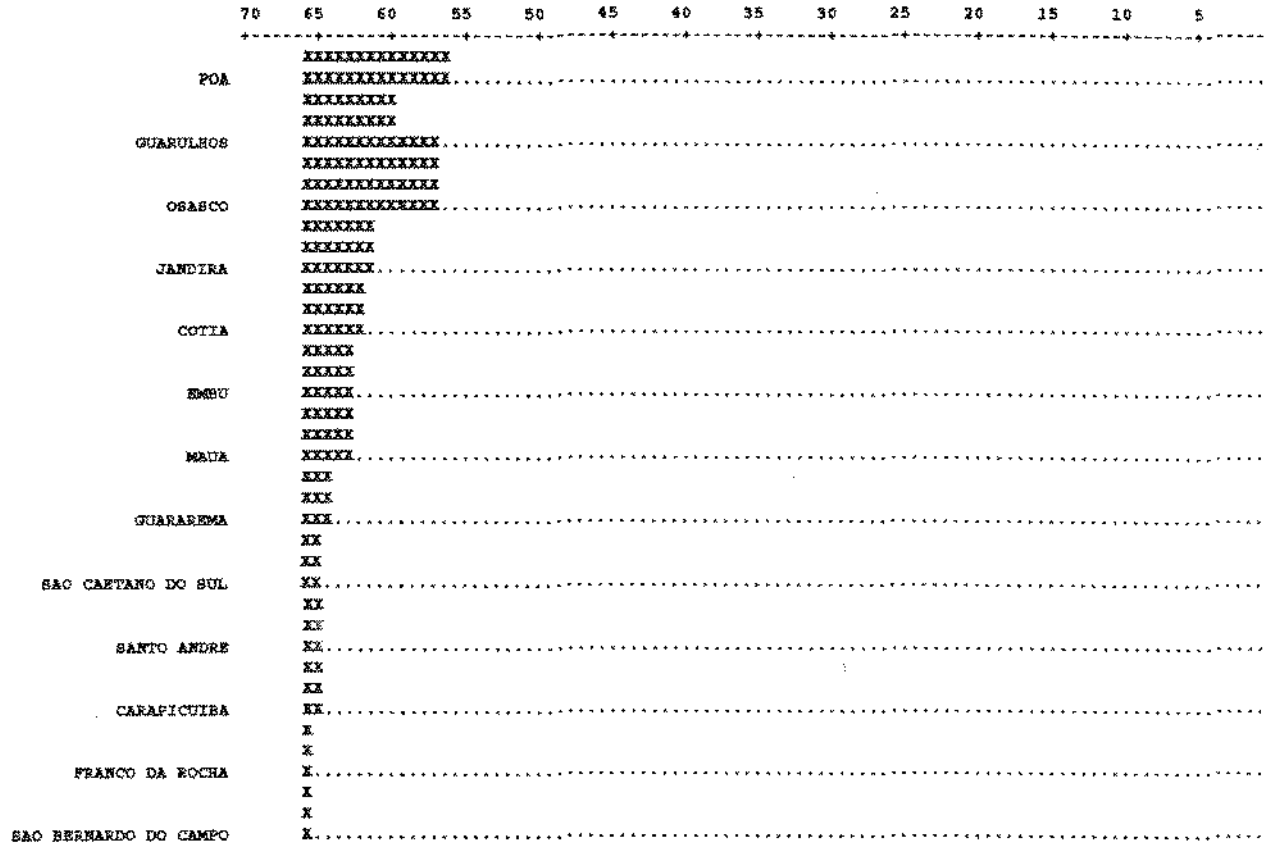
Number of Clusters	Clusters Joined		Frequency of New Cluster	Average Distance	Tie
35	FERRAZ DE VASCONCELOS	MAIRIPORA	2	28.89112	
34	FRANCISCO MORATO	ITAPECERICA DA SERRA	2	30.98212	
33	ARUJA	BARUERI	2	38.27310	
32	EMBU-GUACU	CL35	3	39.44747	
31	BIRITIBA-MIRIM	SALESOPOLIS	2	39.76883	
30	CL33	ITAPEVI	3	43.75427	
29	CL32	CL34	5	47.15499	
28	CAJAMAR	ITAQUAQUECETUBA	2	47.20726	
27	PIRAPORA DO BOM JESUS	SUZANO	2	47.57962	
26	CL30	CL29	8	48.87852	
25	CL26	CL31	10	50.34060	
24	MOGI DAS CRUZES	RIBEIRAO PIRES	2	50.96834	
23	CL25	JUQUITIBA	11	52.05937	
22	CL23	SANTA ISABEL	12	53.69464	
21	CL22	CL27	14	54.17320	
20	CL21	CL28	16	54.71894	
19	CL24	SANTANA DO PARNAIBA	3	55.13315	
18	DIADEMA	TABOAO DA SERRA	2	55.44667	
17	CL20	CL19	19	56.02546	
16	CAIEIRAS	POA	2	56.59850	
15	GUARULHOS	OSASCO	2	56.93074	
14	CL17	RIO GRANDE DA SERRA	20	57.17761	
13	CL14	CL18	22	58.18701	
12	CL13	CL16	24	58.78112	
11	CL12	CL15	26	59.95449	
10	CL11	JANDIRA	27	61.13059	
9	CL10	COTIA	28	61.93053	
8	CL9	EMBU	29	62.54772	
7	CL8	MAUA	30	62.85180	
6	CL7	GUARAREMA	31	64.20303	
5	CL6	SAO CAETANO DO SUL	32	64.67037	
4	CL5	SANTO ANDRE	33	64.89212	
3	CL4	CARAPICUIBA	34	65.25407	
2	CL3	FRANCO DA ROCHA	35	65.49374	
1	CL2	SAO BERNARDO DO CAMPO	36	65.69123	

DENDROGRAMA

Average Linkage Cluster Analysis

Average Distance Between Clusters



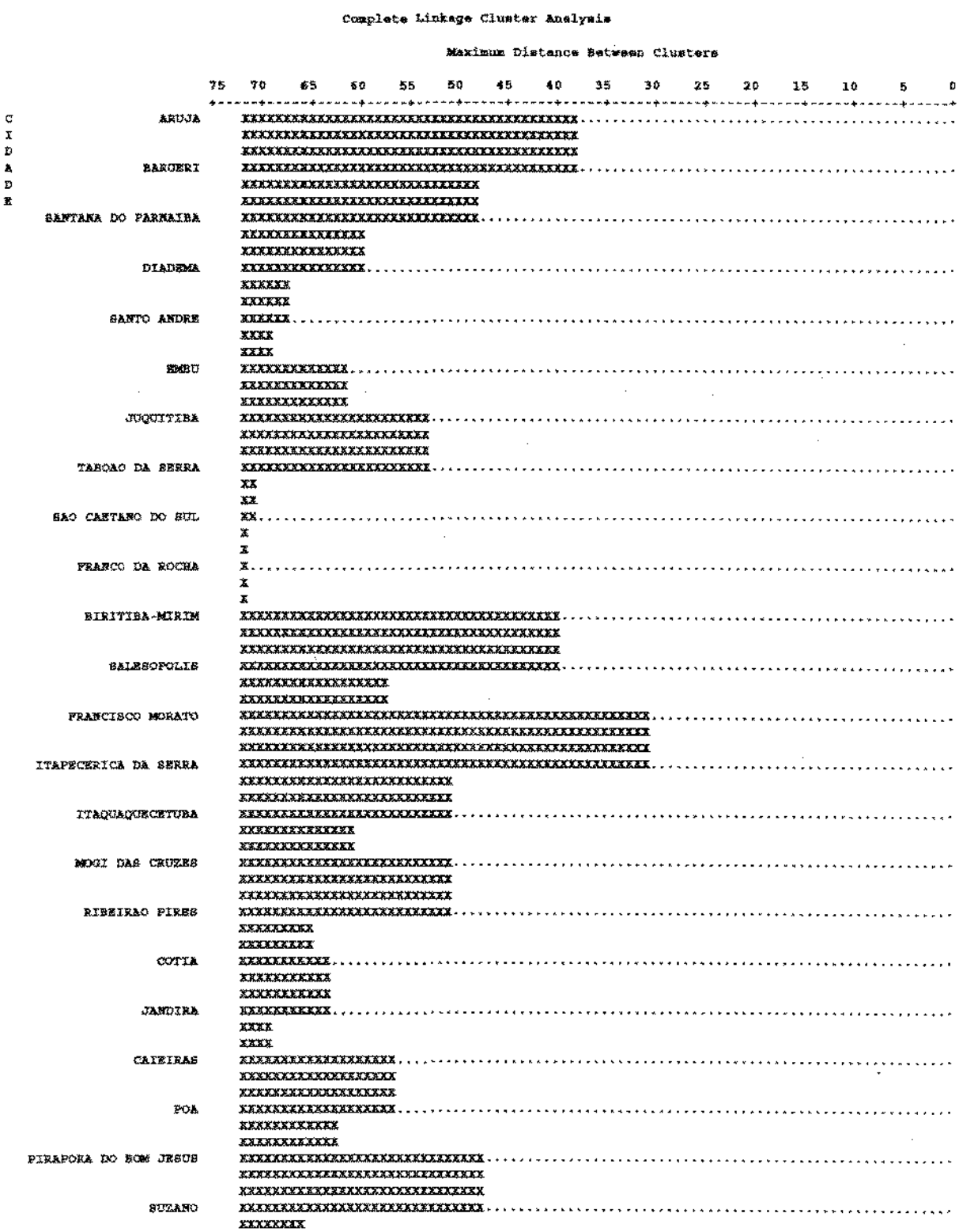

70 65 60 55 50 45 40 35 30 25 20 15 10 5
POA
GUARULHOS
OSASCO
JANDIRA
COTIA
EMBU
MAUA
GUARAREMA
SAO CAETANO DO SUL
SANTO ANDRE
CARAPICUIBA
FRANCO DA ROCHA
SAO BERNARDO DO CAMPO

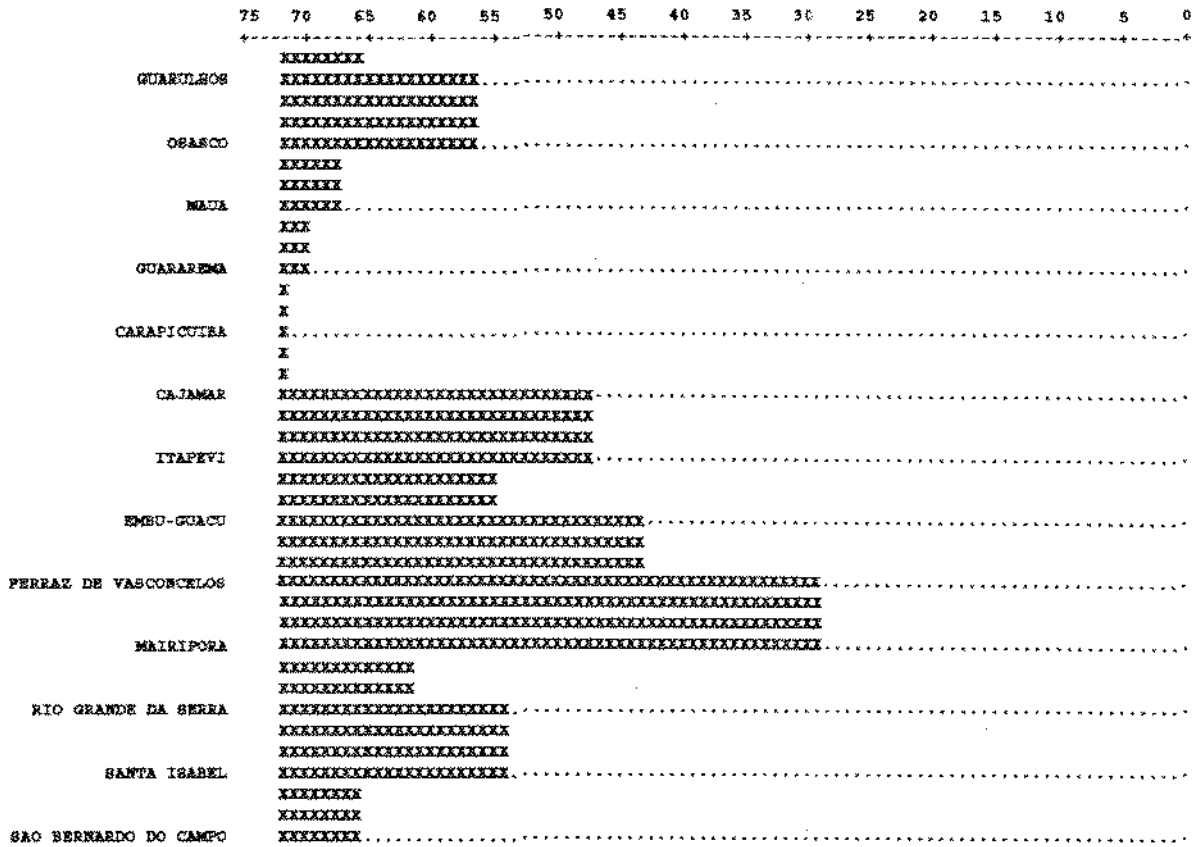
ANÁLISE DE AGRUPAMENTO PELO MÉTODO HIERÁRQUICO DE LIGAÇÃO COMPLETA

Complete Linkage Cluster Analysis

Number of Clusters	Clusters Joined		Frequency of New Cluster	Maximum Distance	Tie
35	FERRAZ DE VASCONCELOS	MAIRIPORA	2	28.89112	
34	FRANCISCO MORATO	ITAPECERICA DA SERRA	2	30.98212	
33	ARUJA	BARUERI	2	38.27310	
32	BIRITIBA-MIRIM	SALESOPOLIS	2	39.76883	
31	EMBU-GUACU	CL35	3	43.32118	
30	CAJAMAR	ITAPEVI	2	47.12215	
29	PIRAPORA DO BOM JESUS	SUZANO	2	47.57962	
28	CL33	SANTANA DO PARNAIBA	3	48.33453	
27	CL34	ITAQUAQUECETUBA	3	50.77703	
26	MOGI DAS CRUZES	RIBEIRAO PIRES	2	50.96834	
25	JUQUITIBA	TABOAO DA SERRA	2	53.17469	
24	RIO GRANDE DA SERRA	SANTA ISABEL	2	54.04542	
23	CL30	CL31	5	55.05058	
22	CAIEIRAS	POA	2	56.59850	
21	GUARULHOS	OSASCO	2	56.93074	
20	CL32	CL27	5	57.61452	
19	CL28	DIADEMA	4	60.14773	
18	CL20	CL26	7	61.22631	
17	CL23	CL24	7	61.28066	
16	EMBU	CL25	3	61.32852	
15	CL22	CL29	4	62.44240	
14	COTIA	JANDIRA	2	63.63370	
13	CL18	CL14	9	65.07667	
12	CL15	CL21	6	65.51320	
11	CL17	SAO BERNARDO DO CAMPO	8	66.09923	
10	CL19	SANTO ANDRE	5	67.30434	
9	CL12	MAUA	7	67.56840	
8	CL13	CL9	16	68.99219	
7	CL10	CL16	8	69.23948	
6	CL8	GUARAREMA	17	70.33797	
5	CL7	SAO CAETANO DO SUL	9	70.47535	
4	CL6	CARAPICUIBA	18	71.25084	
3	CL4	CL11	26	71.62023	
2	CL5	FRANCO DA ROCHA	10	71.66348	
1	CL2	CL3	36	71.90148	

DENDROGRAMA



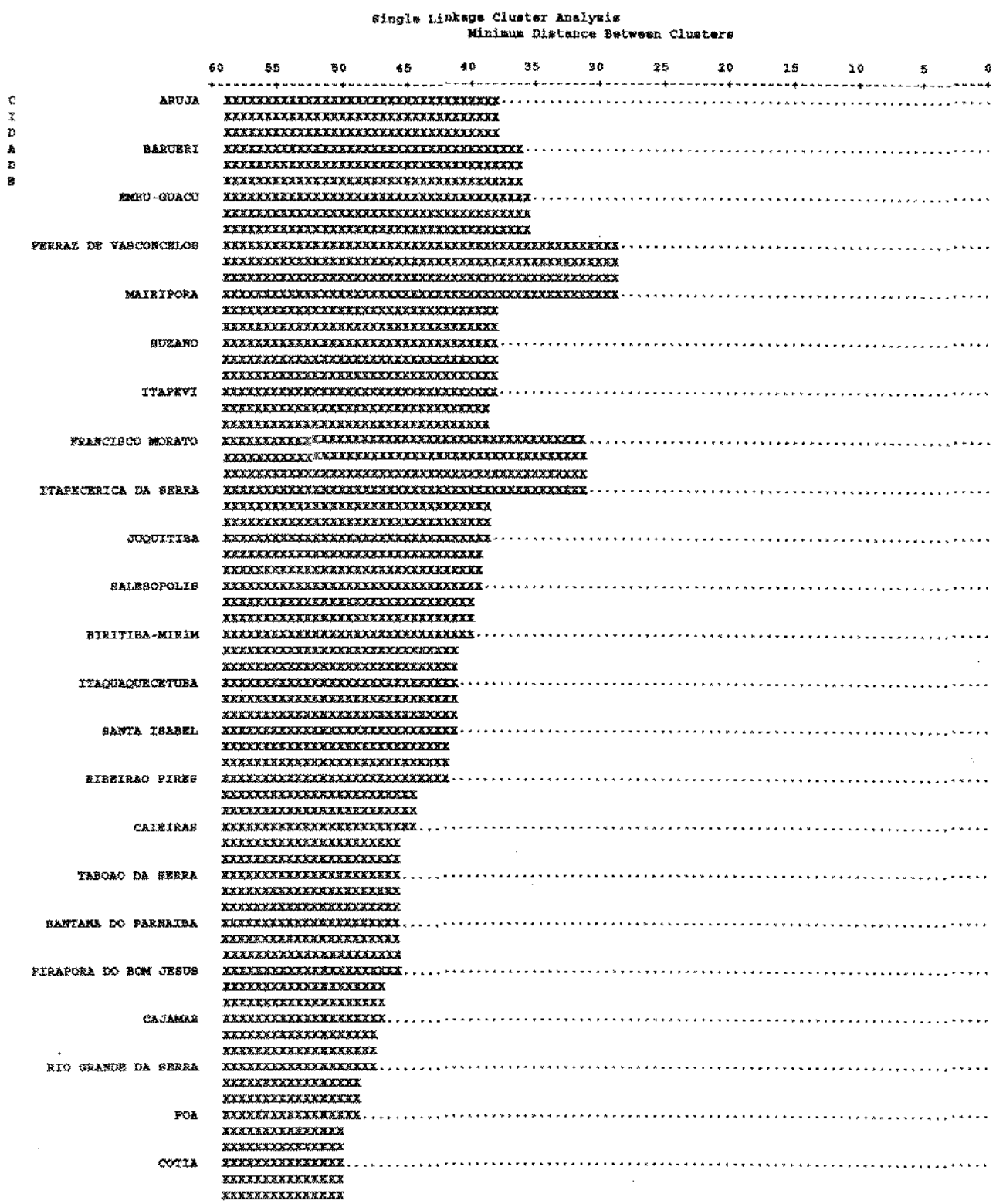

75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0
GUARULHOS
OSASCO
MAUA
GUARAREMA
CARAPICUIBA
CAJAMAR
ITAPEVI
EMBU-GUACU
FERRAZ DE VASCONCELOS
MAIRIPORA
RIO GRANDE DA SERRA
SANTA ISABEL
SAO BERNARDO DO CAMPO

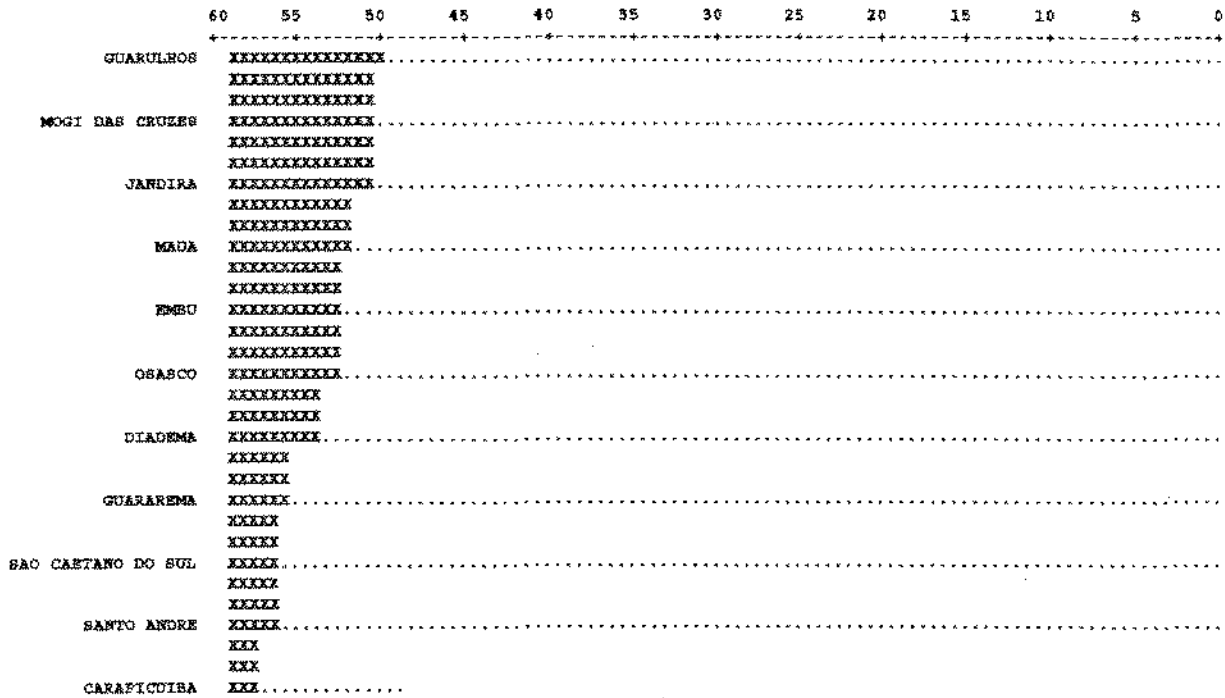
ANÁLISE DE AGRUPAMENTO PELO MÉTODO HIERÁRQUICO DE LIGAÇÃO SIMPLES

Single Linkage Cluster Analysis

Number of Clusters	Clusters Joined		Frequency of New Cluster	Minimum Distance	Tie
35	FERRAZ DE VASCONCELOS	MAIRIPORA	2	28.89112	
34	FRANCISCO MORATO	ITAPECERICA DA SERRA	2	30.98212	
33	EMBU-GUACU	CL35	3	35.57376	
32	BARUERI	CL33	4	36.53933	
31	ARUJA	CL32	5	38.27310	
30	CL31	SUZANO	6	38.37220	
29	CL30	ITAPEVI	7	38.41323	
28	CL34	JUQUITIBA	3	38.58226	
27	CL29	CL28	10	38.80452	
26	CL27	SALESOPOLIS	11	39.52331	
25	CL26	BIRITIBA-MIRIM	12	39.76883	
24	CL25	ITAQUAQUECETUBA	13	41.53642	
23	CL24	SANTA ISABEL	14	41.55951	
22	CL23	RIBEIRAO PIRES	15	41.57470	
21	CL22	CAIEIRAS	16	44.27068	
20	CL21	TABORO DA SERRA	17	45.37892	
19	CL20	SANTANA DO PARNAIBA	18	45.60983	
18	CL19	PIRAPORA DO BOM JESUS	19	45.63566	
17	CL18	CAJAMAR	20	47.12215	
16	CL17	RIO GRANDE DA SERRA	21	47.70477	
15	CL16	POA	22	48.95171	
14	CL15	COTIA	23	50.14090	
13	CL14	GUARULHOS	24	50.18908	
12	CL13	MOGI DAS CRUZES	25	50.38828	
11	CL12	JANDIRA	26	50.85449	
10	CL11	MAUA	27	51.73222	
9	CL10	EMBU	28	52.70939	
8	CL9	OSASCO	29	52.79996	
7	CL8	DIADEMA	30	53.79977	
6	CL7	GUARAREMA	31	55.39370	
5	CL6	SAO CAETANO DO SUL	32	55.95253	
4	CL5	SANTO ANDRE	33	56.02113	
3	CL4	CARAPICUIBA	34	57.40986	
2	CL3	SÃO BERNARDO DO CAMPO	35	57.98750	
1	CL2	FRANCO DA ROCHA	36	58.87796	

DENDROGRAMA




60 55 50 45 40 35 30 25 20 15 10 5 0
GUARULHOS
MOGI DAS CRUZES
JANDIRA
MAUA
EMBU
OSASCO
DIADEMA
GUARAREMA
SAO CAETANO DO SUL
SANTO ANDRE
CARAPICUIBA

Bibliografia

[1] Anderberg, M. R. (1973). *Cluster analysis for applications*. Academic Press, New York.

[2] Balakrishnan, V. e Sangghvi, L. D. (1968). Distance between populations on the basis of attribute data. *Biometrics*, **24**, 859-865.

[3] Ball, G. H. e Hall, D. J. (1965). ISODATA, a novel method of data analysis and pattern classification. *Technical Report*. Stanford Research Institute, Menlo Park.

[4] Bayne, C. K.; Beauchamp, J. J.; Begowich, C. L. e Kane, V. E. (1980). Monte carlo comparasions of selected clusterig procedures. *Patter recognition*, **12**, 51-62

[5] Bickel, P. E. e Doksum, K. A. (1977). *Mathematical statistics*. Rolden Day, San Francisco.

[6] Bolshev, L. N. (1969). Cluster analysis. *Bull. I.S I.*, **43**, Book 1, 411-425.

[7] Boyce, A. J. (1969). Mapping diversity: A comparative study of some numerical methods. In numerical taxonomy (A. J. Cole, ed), pp. 1-30. Academic Press, New York.

[8] Bray, J. R. e Curtis, J. T. (1957). An ordination of the upland forest communities of S. Wisconsin. *Ecol. Monogr.*, **27**, 325-349.

[9] Bussab, W. O., Miazaki, E. S. e Andrade, D. F. (1990). Introdução à análise de agrupamentos. in: 9º *Simpósio Brasileiro de Probabilidade e Estatística*, IME, USP. São Paulo.

[10] Cain. A. J. e Harrison. G. A. (1958). An analysis of the taxonomist's judgment of affinity. *Proc. Zool. Soc. Lond.*, **131**, 85-98.

[11] Cochran. W. J. e Hopkins, C. (1961). Some classification problems with multivariate qualitative data. *Biometrics*. **17**, 10-32.

[12] Cormack, R. M. (1971). A review of classification. *J. R. Statist. Soc., Series A*, **134**, 321-367.

[13] Cox. D. R. (1972). *The analysis of multivariate binary data.* Applied Statistics. **21**, 113-120.

[14] Cunningham, K. M. e Olgivie, J. C. (1972). Evaluation of hierarchical grouping technics: a preliminary study. *The computer journal*, **15**, 209-213.

[15] Dubes. R. e Jain, A. K. (1976). *Clusterig technics: the user's dilemma.* Pattern recognition, 8, 247-260.

[16] Dubien, J. L. e Warde, W. D. (1979). A mathematical comparision of the members of infinity family of aglomerativa clustering algorithms. The canadian journal of statistics, **7**, 29-38.

[17] Dubien, J. L. e Warde, W. D. (1987). A comparision of aglomerative clustering algorithms with respect to noise. *Comm. stat-theory and methods*, **16**, 1433-1460.

[18] Everitt, B. (1974). Cluster analysis. Heineman Educational Books, London.

[19] Ferreira, A. M. e Medeiros. P. G. (1992). O algoritmo "centra " de análise de agrupamento não hierárquico. In: *Atas dos resumos* - 10º *Simpósio Brasileiro de Probabilidade e Estatística*, UFRJ. Rio de Janeiro.

[20] Fisher. L. e Van Ness, J. W. (1971). Admissible clustering procedures. *Biometrika*. **58**, 91-104.

[21] Florek, K.; Lukaszewisz, J.; Perkal, J e Zubrzycki, S. (1951). Sur la liaison et la division des oints d'un ensemble fini. In: *Colloquium mathematicae*, **2**, 282-285.

[22] Gnanadesikan, R. e Kettenring, J. R. (1989). Discriminant analysis and clustering. *Statistical Science*, **4**. 34-69.

[23] Gower, J. C. (1970). Classification e Geology. *Rev. I.S.I.*. **38**, 35-41.

[24] Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857-872.

[25] Graybill F.A. (1969). *Matrices with application in statistics*. Second edition. Wadsworth, Belmont.

[26] Green, P. E. e Carmone, F. J. (1970). *Multidimensional scaling and related techniques in marketing analysis*. Allyn & Bacon, Boston.

[27] Hartigan, J. A. (1981). Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.*, **76**, 388-394.

[28] Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques regions voisines. *Bull. Soc. Vaud. Sci. Nat.*, **37**. 241-272.

[29] James, B. R. (1981). Probabilidade: um curso em nível intermediário. IMPA, Rio de Janeiro.

[30] Jardine, N. e Sibson, R. (1968). The construction of hierarchic and non-hierarchic classification. *Computer journal*, **11**, 177-184.

[31] Johnson, N. L. e Kotz, S. (1969). Discrete Distributions. Houghton Mifflin, New York.

[32] Jonhson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, **32**, 241-254.

[33] Johnson, R. L. e Wall, D. W. (1969). Cluster analysis of semantic diferential data. Educ. *Psychol. Measur.*, **29**, 769-780.

[34] Johnson, R. A. e Wishern, D. W. (1982). Applied multivariate statistical analysys. Prentice-Hall, Englewood Clifs.

[35] Kaufman, L. e Rousseeuw, P. J. (1990). Finding groups in data : an introduction to cluster analysis. John Wiley & Sons, Inc., New York.

[36] Koop, B. (1978a). Hierarchical classification I: single linkage method. *Biometrical Journal*, **20**. 485-501.

[37] Koop, B. (1978b). Hierarchical classification II: complete linkage method. *Biometrical Journal*, **20**. 597-602.

[38] Koop, B. (1978c). Hierarchical classification III: average linkage, median, centroid, ward, flexibe - strategy. *Biometrical Journal*, **20**, 703-711.

[39] Kotz, S. e Johnson, N. L. (1985). Encyclopedia of statistical sciences. vol 7. John Wiley & Sons, New York.

[40] Kotz, S. e Johnson, N. L. (1986). Encyclopedia of statistical sciences. vol 5. John Wiley & Sons, New York.

[41] Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *J. Amer. Statist. Assoc.*, **70**, 782-790.

[42] Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, **36,** 493-499.

[43] Krzanowski, W. J. (1983). Distance between populations using mixed continuous and categorical variables. *Biometrika*, **70**, 235-243.

[44] Krzanowski, W. J. (1988). *Principles of multivariate analysis: a user's perspectives.* Oxford university press, Oxford.

[45] Krzanowski, W. J. (1993). The location model for Mixtures of categorical and continuous variables. *Journal of classification*, **36,** 493-499.

[46] Kuiper, F. K. e Fisher, L. (1975). A monte carlo comparasion of clustering procedures. *Biometrics*, **31**, 777-783.

[47] Kurczynski, T. W. (1970). Generalized distance and discrete variables. *Biometrics*. **26**, 525-534.

[48] Lance, G. N. e Williams, W. T. (1967) . A general theory of classificatory sorting strategies I. Hierarchical systems. *Computer journal*, **9**, 373-380.

[49] León, M. A. V. (1993). Discriminaçâo com mistura de variáveis contínuas e categóricas. *Dissertaçâo de mestrado*, IMECC, UNICAMP, Campinas.

[50] Lucas, L. C. S. (1982). Análise de agrupamento. *R. Bras. Estat.*, **43** 589-723.

[51] Mahalanobis, P. C. (1936). On the generalizad distance to statistics. *Proc. Nat. Inst. Sci. India*, **2**, 49-55.

[52] Mardia, K. V., Kent, J. T. e Bibby, J. M. (1979). *Multivariate analysis*. Academic Press, London.

[53] McQuity, L. L. (1957). Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. *Educational and psychological measurement*. **17**, 207-229.

[54] Milligan, G. W. (1980). An examination of the effect of six types of error pertubation on fiftten clustering algorithms. *Psychometrika*, **45**, 325-342.

[55] Milligan, G. W. e Isaac, P. D. (1980). The validation of four ultrametric clustering algorithms. *Pattern Recognition*, **12**, 41-50.

[56] Mood, D. A. M., Graybill, F. A. e Boes, D. C. (1974). Introduction to the theory of statistics. Third edition, McGraw-Hill, Singapure.

[57] Olkin, I. e Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Statistic.*, **32**, 348-365.

[58] Pereira, J. R. G. (1993). Um estudo sobre alguns métodos hierárquicos para análise de agruparnentos. *Dissertação de mestrado*, IMECC, UNICAMP, Campinas.

[59] Ralambondrainy, H. (1988). A clustering method for nominal data and mixture of numerical and nominal data. *Classification and related methods of data analysis. H.H.* Bock (editor). Elsevier Sciense Publishers B. V., North-Holland.

[60] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Stat. Assoc.*, **66**. 846-850.

[61] Rohatgi, V. K. (1967). *An introduction to probability theory and Mathematics*. John e Willey. New York.

[62] Romesburg, H. C. (1984). *Cluster analysis for researchers*. Lifetime Learnig Publications, California.

[63] Royden, H. (1968). *Real analysis*. 2nd edition. Macmillian, New York.

[64] Rubin, J. (1967). Optimal classification into groups: an approach for solving the taxonomy problem. *J. Theor. Biol.*, **15**, 103-144.

[65] SAS Institute Inc. (1988). SAS/IML User's guide. Release 6.03. Cary, NC.

[66] SAS Institute Inc. (1989). SAS/STAT User's guide. Version 6, Fouth Edition. Volume 1 e 2. Cary NC.

[67] Searle, S. R. (1982). Matriz algebra useful for statistics. John e Wiley, New York.

[68] Shen, H. C., Bie, C. Y. C. e Chiu, D. K. Y. (1993). A texture-based distance measure for classification. *Pattern recognition*, **26**, 1429-1437.

[69] Sibson, R. (1971). Some observations on a paper by Lance e Williams. *The computer Journal*, **14**, 156-157.

[70] Sokal, R. R. e Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas sciense bulletin*, **38**. 1409-1438.

[71] Sokal, R. R. e Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. Freeman and Co. San Francisco.

[72] Sorensen, T. (1948). A method of establishing groups of equal amplitude in plan sociology based on similarity of species content and its application to analysis of the vegetation on danish common. *Biologiske skrifter*, **5**, 1-34.

[73] Spath, H. (1980). Cluster Analysis Algorithms. Ellis Horwood, Chichester.

[74] Talkington, L. (1967). A method of scaling for a mixed set of discrete and continuous variables. *Syst. Zoology*, **16**, 149-152.

[75] Tate, R. F. (1954). Correlation between a discrete and a continuous variable. *Ann. Math. Stat.*, **25**, 603-607.

[76] Uchiyama, T. e Arbib, M. A. (1994). An algorithm for competitive learning in clustering problems. *Pattern Recognition*, **27**, 1415-1421.

[77] Vlachonikolis, I. G. e Marriott, F. H. C. (1982). Discrimination with mixed binary and continuous data. *Appl. Statist.*, **31**, 23-31.

[78] 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**, 162-165.

[79] Wishart, D. (1969). An algorithm for hierarchical classification. *Biometrics*, **25**, 165-170.