Universidade da Madeira
Departamento de Matemática e Engenharias

Análise de clusters aplicada ao Sucesso/Insucesso em Matemática

Dissertação submetida com vista à obtenção do grau de Mestre em Matemática para o Ensino

Mestranda: Guida Maria da Conceição Caldeira Quintal

Orientadora: Professora Doutora Rita Vasconcelos

Funchal, Dezembro de 2006

Este trabalho de dissertação foi co-financiado por:

CITMA
CENTRO DE CIÊNCIA E TECNOLOGIA DA MADEIRA

PROGRAMA OPERACIONAL
PLURIFUNDOS DA REGIÃO
AUTÓNOMA DA MADEIRA

UNIÃO EUROPEIA
FUNDO SOCIAL

Agradecimentos

Quero formular o meu agradecimento a todos os que de algum modo contribuíram para a realização deste trabalho, nomeadamente:

- à Professora Doutora Rita Vasconcelos pela orientação, confiança, críticas e sugestões dadas ao longo do trabalho;
- à comissão de mestrado, mais concretamente à sua presidente Professora Doutora Margarida Faria, à Professora Doutora Teresa Gouveia e à Professora Doutora Sandra Mendonça, por estarem sempre disponíveis a ouvir e a ajudar a resolver dificuldades que surgiram;
- ao Departamento de Matemática e Engenharias que tornou possível a realização do mestrado, nomeadamente ao seu Presidente e aos membros da Secretaria deste Departamento;
- à Guida Rodrigues, pela sua disponibilidade quer no esclarecimento de dúvidas quer em discussões produtivas acerca do trabalho;
- ao centro de Ciência e Tecnologia da Madeira (CITMA) pela bolsa de estudo concedida;
- à Secretaria Regional de Educação;
- às Direcções Executivas das Escolas onde foram realizados os inquéritos, por terem autorizado a realização dos mesmos;
- aos alunos que responderam aos inquéritos;
- aos meus filhos Sofia, Margarida e Pedro pelo amor e carinho;
- ao meu marido Lúcio pela paciência, apoio e confiança;
- ao meu irmão Francisco pela amizade e força anímica que transmite;
- à minha família;
- às minhas amigas, em particular à Teresa Roque;
- à Ivanilda e à Sofia pela colaboração na análise exploratória dos dados;
- aos colegas de mestrado.

Resumo

De acordo com [Mirkin B., 1996], classificação é um agrupamento existente ou ideal daqueles que se parecem (ou são semelhantes) e separação dos que são dissemelhantes. Sendo o objectivo/razão da classificação: (1) formar e adquirir conhecimento, (2) analizar a estrutura do fenómeno e (3) relacionar entre si diferentes aspectos do fenómeno em questão.

No estudo do sucesso/insucesso da Matemática está de algum modo subjacente nos nossos objectivos "classificar" os alunos de acordo com os factores que se pretende que sejam determinantes nos resultados a Matemática.

Por outro lado, voltamos a recorrer à classificação quando pretendemos estabelecer os tipos de factores determinantes nos resultados da Matemática.

Os objectivos da Análise de Clusters são: (1) analisar a estrutura dos dados; (2) verificar/relacionar os aspectos dos dados entre si; (3) ajudar na concepção da classificação.

Pensámos que esta técnica da análise exploratória de dados poderia representar uma ferramenta muito potente para o estudo do sucesso/insucesso da Matemática no Ensino Básico.

O trabalho desenvolvido nesta dissertação prova que a Análise de Clusters responde adequadamente às questões que se podem formular quando se tenta enquadrar socialmente e pedagogicamente o sucesso/insucesso da Matemática.

Palavras chave

Análise de clusters; clusters; medidas de semelhança/dissemelhança; métodos hierárquicos; métodos não hierárquicos; dendograma.

Abstract

According to [Mirkin B., 1996], classification is a concrete or ideal grouping of those which look alike (or which are similar) and a separation of the dissimilar ones. Being the aim/reason of the classification: (1) to form and to acquire knowledge, (2) analyze the structure of the phenomenon and (3) to relate different aspects of the phenomenon being analyzed, among themselves.

In the study of the success/failure in Mathematics it is somehow underlying in our objectives to "classify" the students according to the elements/factors intended to be decisive in the results in Mathematics.

On the other hand, we come back again to classification when we intend to establish the types of decisive factors in the results in Mathematics.

The objectives of Clusters Analysis are: (1) to analyze the structure of the data; (2) verify/relate the aspects of the data among themselves; (3) to help in the conception/ generation of the classification.

We thought that this technique of exploratory data analysis could represent a very powerful tool for the study of the success/failure of Mathematics in first grade school[1].

The work developed in this thesis proves that Clusters Analysis appropriately answers to the questions that can be formulated when one tries to frame socially and pedagogically the success/failure of Mathematics.

Keywords

Cluster analysis; clusters; measures of similarity/dissimilarity; hierarchical methods; non hierarchical methods; dendogram.

[1] The first grade school (or primary school) consists of 9 years in Portugal

Índice geral

Índice das tabelas

Índice das figuras

Índice dos Anexos

Nota Introdutória

Esta tese consta de duas partes.

Na **primeira parte** será apresentada a técnica estatística multivariada denominada Análise de Clusters. Esta apresentação, sem ser exaustiva, pretende abranger os métodos mais importantes para a construção de clusters, de forma a tornar clara a opção que fizemos para a abordagem de uma base de dados obtidos para estudar o sucesso/insucesso da Matemática ao nível do 9º ano de escolaridade.

A Análise de Clusters é a arte de encontrar grupos nos dados, [Kaufman L. e Rousseeeuw P., 1990].

A Análise de Clusters tem como objectivo identificar subgrupos homogéneos (clusters) na população de objectos[2], de variáveis[3] ou de ambos, de tal forma que a variabilidade nos elementos no mesmo grupo seja mínima e a variabilidade entre os grupos seja máxima. Não há uma indicação prévia dos membros dos grupos. As dificuldades normalmente encontradas na aplicação desta análise a uma base de dados dispostos numa matriz $n \times p$ (em que as n linhas correspondem à informação sobre os n objectos relativamente às p variáveis observadas (colunas)), consistem em escolher o tipo de medida de proximidade a usar, o método a aplicar para a obtenção dos clusters, a determinação do número de clusters e a interpretação das características dos mesmos. Não existem regras estipuladas, mas sim alguns guias que ajudam a escolher o método. O algoritmo a escolher depende do tipo de variáveis, do objectivo a atingir e da dimensão da amostra.

A Análise de Clusters é uma técnica da Estatística Descritiva (não inferencial); é usada como uma ferramenta exploradora e descritiva. Em oposição aos testes estatísticos que são usados para confirmar hipóteses, esta técnica é

[2] podem ser pessoas, flores, palavras, países, plantas, rochas, mercadorias, etc.
[3] pode ser altura, peso, sexo, habilitações, etc.

usada para tentar perceber o que os dados nos dizem; o que interessa é descobrir grupos e interpretar as características dos seus elementos.
Assim no capítulo 1 será apresentada a análise de clusters como técnica da análise exploratória de dados. No capítulo 2, será tratada a representação gráfica de dados multivariados. No capítulo 3, serão abordadas as medidas de proximidade e nos capítulos 4 e 5 serão estudados os diversos métodos para a formação dos clusters. Segue-se o capítulo 6 da descrição e apresentação dos resultados de uma Análise de Clusters. Neste capítulo serão, também, apresentadas as considerações finais.

Na **segunda parte** será feita a aplicação de grande parte do que foi apresentado na primeira parte, ao sucesso/insucesso em Matemática através da informação contida numa amostra dos alunos do 9º ano que frequentaram algumas escolas do Funchal no ano lectivo 2005/2006. Para este efeito realizaram-se inquéritos a alunos de 9º ano com o objectivo de recolher informação julgada pretinente para o desempenho na disciplina de Matemática. Posteriormente foram recolhidas as classificações em Matemática do 3º Período e do exame final, assim como se os alunos foram aprovados ou não aprovados.

Depois de se ter feito uma primeira análise descritiva dos dados, procederemos à análise de clusters e à interpretação dos resultados obtidos.

Os resultados obtidos serão analisados de forma crítica e esperamos que o estudo sirva para alertar/reflectir/discutir sobre alguns factores que levam ao sucesso/insucesso em Matemática, e que leve a algumas alterações tendo em vista o aumento do sucesso em Matemática.

O *software* utilizado foi o SPSS, versão 13.0.

PARTE I

ANÁLISE DE CLUSTERS

Capítulo 1
Conceitos preleminares

1.1 Introdução

Agrupar dados semelhantes com vista à classificação é uma das características primitivas e básicas do ser humano. No nosso dia-a-dia lidamos com agrupamentos em muitos aspectos da nossa vida.

Colocar objectos/indivíduos em grupos é uma habilidade necessária nas mais variadas situações. Esta habilidade é, por exemplo, essencial aos biólogos devido à enorme diversidade de organismos, moléculas, doenças, etc.. A correcta classificação de organismos é tão importante em Biologia, que surgiu um novo ramo da Biologia designado por Taxonomia. Os resultados obtidos na análise de clusters podem contribuir para a definição de uma classificação (como a taxonomia relativa a animais, insectos, plantas) ou sugerir modelos estatísticos para descrever a população.

A Análise de Clusters é um procedimento da Estatística Multivariada que tenta agrupar um conjunto de dados em subgrupos homogéneos, chamados clusters; os dados podem ser objectos ou variáveis. Trata-se, pois de uma técnica matemática que foi concebida com a finalidade de revelar estruturas de classificação nos dados recolhidos em fenómenos do mundo real.

Uma questão que surge frequentemente aos investigadores é como organizar dados observados, em estruturas com significado. A análise de clusters é usada com esse objectivo por investigadores de várias áreas: para descobrir uma estrutura nos dados sem uma explicação/interpretação prévia.

Este método da análise exploratória procede ao agrupamento dos objectos em função da informação existente, de tal modo que objectos pertencentes a um mesmo cluster sejam o mais semelhantes possível e os objectos pertencentes a clusters diferentes sejam o mais dissemelhantes possível. Os membros de cada grupo possuem certas características em comum e espera-se que o resultado da classificação forneça pistas para a interpretação dos grupos. Podem ser reveladas associações nos dados, não evidentes previamente, mas que são importantes e úteis uma vez encontradas. A ideia chave é que os clusters tenham significado e sejam interpretáveis.

A ideia é sobretudo gerar hipóteses, mais do que testá-las.

Como veremos ao longo deste trabalho, a matemática subjacente aos métodos usados na análise de clusters é relativamente simples. Porém, para a concretização de alguns algoritmos para a formação dos clusters é necessário recorrer a *software* adequado devido aos cálculos laboriosos envolvidos.

Como veremos, a classificação obtida depende da medida de semelhança/dissemelhança e do método usado para formar clusters. Consequentemente, não há apenas uma classificação correcta.

A especificidade de cada situação prática faz com que para além dos critérios disponíveis para a formação de clusters, haja necessidade de considerar aspectos importantes da análise como sejam a selecção de variáveis, a escolha da medida de proximidade, o contexto do estudo e o objectivo do trabalho.

Na análise de clusters, os dados são colocados numa matriz multivariada $n \times p$, $X = [X_{ij}]$, $i = 1,...,n$; $j = 1,...,p$ contendo os valores de cada variável observada em cada objecto, sendo X_{ij} o valor da variável j para o objecto i. As variáveis podem ser contínuas ou categorizadas.

Aos valores desta matriz multivariada são aplicadas medidas de proximidade (semelhança[4] / dissemelhança[5]), convertendo a matriz inicial X numa matriz $n \times n$ de semelhanças/dissemelhanças entre objectos. Note-se que se pode actuar de forma idêntica para as variáveis.

À matriz $n \times n$ de semelhanças/dissemelhanças entre objectos ou entre variáveis são aplicados métodos para a constituição de clusters. Estes métodos podem ser hierárquicos ou não hierárquicos, como veremos neste estudo.

1.2 A importância da Análise de Clusters na classificação

A classificação é uma actividade conceptual básica dos seres humanos; as crianças aprendem a distinguir desde muito cedo, por exemplo, a mãe/de outras mulheres, a distinguir pai/de outros homens, a distinguir entre quem lhes dá atenção/estranhos, quente/frio, limpo/sujo, animal/planta, casa/carro.

De acordo com [Mirkin B., 1996], classificação é um agrupamento existente ou ideal daqueles que se parecem (ou são semelhantes) e a separação daqueles que são dissemelhantes; sendo o objectivo/razão da classificação: (1) formar e adquirir conhecimento, (2) analizar a estrutura do fenómeno e (3) relacionar entre si diferentes aspectos do fenómeno em questão.

A classificação tem tido, também, um papel fundamental em muitos ramos da Ciência. Ela é necessária ao desenvolvimento do homem e da ciência pois ajuda-nos a reconhecer e discutir diferentes tipos de acontecimentos, objectos, pessoas que encontramos na nossa vida.

[4] Mede o quão próximos estão dois objectos.
[5] Mede a distância a que dois objectos estão um do outro.

Por exemplo: na Biologia, a classificação é conhecida por taxonomia, é a base da Teoria de Darwin; na Química, a classificação representa um método conveniente para organizar um grande conjunto de dados de forma eficiente num pequeno número de grupos. A classificação de elementos na tabela periódica, por Mendeleiev, em 1860, teve um profundo impacto na compreensão da estrutura do átomo; na Medicina, na Psicologia, na Psiquiatria a classificação é necessária no diagnóstico e tratamento de doenças;

Na classificação é atribuído um nome ao grupo. Esse nome colecciona indivíduos/casos em grupos, de tal modo que quem tiver esse nome terá uma ou várias características comuns que são essenciais na descrição desse grupo[6].

A classificação pode envolver pessoas, animais, elementos químicos, modelos de comportamento, estrelas, etc., como entidades que podem ser agrupadas.

A Análise de Clusters inclui uma série de procedimentos estatísticos sofisticados que podem ser usados para classificar objectos sem preconceitos, isto é, observando apenas as semelhanças ou dissemelhanças entre eles, sem definir previamente critérios de inclusão em qualquer agrupamento.

A Análise de Clusters, pode ser usada não só para identificar uma estrutura presente nos dados, mas também para impôr uma estrutura num conjunto de dados mais ou menos homogéneos que têm de ser separados.

Assim, para além da estruturação dos dados em grupos e a consequente redução da dimensão do espaço associado às nossas variáveis, dado um objecto qualquer, a comparação das suas propriedades com as propriedades dos elementos dos subgrupos permite identificar o subgrupo onde incluí-lo, uma vez que elementos pertencentes ao mesmo subgrupo têm propriedades semelhantes.

[6]Por exemplo, o grupo dos cães, o grupo dos gatos.

1.3 Exemplos da aplicação da análise de clusters

Para termos uma percepção mais sólida da relevância da Análise de Clusters, serão dados, de seguida, alguns exemplos de aplicação. Alguns deles estão relacionados com os exemplos que apresentamos no parágrafo 1.2 sobre classificação.

- Na Arqueologia, a identificação de grupos de artefactos semelhantes usados por povos já desaparecidos, ajuda a compreender muitos aspectos das civilizações antigas [Hodson F.R., 1971].
- Nas Ciências Sociais, os métodos de análise de clusters foram utilizados pelos antropólogos para definirem áreas culturais homogéneas.
- Na Sismologia, a análise de clusters tem sido usada na predição de abalos sísmicos [Wardlaw et al., 1991].
- Na classificação de documentos, a procura de informação em grandes bases de dados, nomedamente na Web, fica facilitada se os documentos estiverem agrupados em clusters [Willet, 1990].
- No Data mining, a análise de clusters constitui um dos primeiros passos deste processo. Data mining é o processo de identificar grupos de registos e extrair conhecimento de grandes bases de dados [Han and Kamber, 2001].
- Na Biologia e na Química a análise de clusters pode contribuir para uma definição de classificação tal como a taxonomia relativa a minerais, insectos, plantas etc..
- Na Medicina, na Psicologia, na Psiquiatria, a classificação obtida de uma análise de clusters permite identificar as causas das doenças, os sintomas, e consequentemente criar/ melhorar os seus tratamentos.
- Na Análise de Mercados, os segmentos de consumidores ou produtos são em geral clusters, sendo necessário conhecê-los para perceber a estrutura de mercado [De Sarbo et al., 1993] e [Arabie e Hubert, 1996].

- Em Marketing, a análise de clusters tem sido aplicada para proceder à segmentação de mercados a partir das características geográficas, demográficas e psicográficas dos consumidores, para identificar mercados potenciais para determinados produtos, determinar mercados identicos em países diferentes ou encontrar grupos de consumidores que possam servir de referência na previsão de vendas.
- Na Análise da Política e da Economia, etc.;
- Em geral, quando precisamos de classificar uma "montanha" de informação, a análise de clusters é muito útil.

1.4 Etapas da análise de clusters

De um modo geral, verificamos que a Análise de Clusters compreende os seguintes procedimentos: selecção de objectos; selecção de variáveis; transformação de variáveis; selecção da medida de semelhança/dissemelhança; escolha do método de formação de clusters a aplicar; e discussão e apresentação dos resultados. Contudo, podemos omitir, por exemplo, a transformação de variáveis numa primeira análise. Podemos repetir a análise e trabalhar com variáveis estandardizadas.

➢ Selecção de objectos

A selecção de objectos depende dos objectivos da análise. Se forem utilizados dados de análises anteriores pode ser necessário analisá-los e retirar os objectos que não tenham relevância para o estudo. No entanto, devemos ter o cuidado de não deixar objectos importantes para o estudo, de fora do conjunto a analisar. Outras vezes o conjunto de objectos é uma amostra da população, que é desejável que seja representativa para que os grupos

resultantes possam ser considerados representativos dos grupos existentes na população.

A Análise de Clusters é uma análise exploratória de dados, que pretende criar grupos e descobrir relações entre os elementos desses grupos, que provavelmente não seria possível serem detectadas sem esta análise. Não é uma estatística inferencial e como tal o seu objectivo não é inferir as conclusões do estudo à população; para tal deverão ser aplicadas técnicas de inferência estatísticas adequadas.

➢ Selecção de variáveis

As variáveis caracterizam os objectos. A selecção das variáveis é um dos aspectos que mais influencia os resultados da análise de clusters. Não se trata de um problema apenas do Matemático, mas também do responsável pelo estudo que proporciona os dados. O Matemático deverá ter em atenção, essencialmente, o tipo de variáveis utilizadas e a escala. O responsável pelo estudo deverá atender aos conhecimentos que possui acerca do tema em estudo, para seleccionar as variáveis mais importantes para obter um resultado digno de confiança.

Acerca do **número de variáveis**, é de referir as opiniões contrárias de que aumentando o número de variáveis obtém-se uma melhor identificação dos clusters [Everitt B., 1987] e a que sustenta que se obtém uma fraca identificação dos clusters, nas mesmas condições [Price,1993].

Estas opiniões podem parecer contrárias, mas não o são. Pensamos que se podem dar as duas situações. O que deve acontecer é que com muitas variáveis a caracterização dos grupos poderá ser muito boa para aquele conjunto específico de dados, mas se usarmos outros dados da mesma população e as mesmas variáveis o resultado da análise de clusters já pode ser diferente. E, neste sentido pode dizer-se que a identificação dos grupos é fraca. Por outro lado, com poucas variáveis podemos estar a omitir informação importante e obter poucos clusters que serão interpretados de uma forma muito geral e portanto pouco

exclarecedora. Deve haver um número de variáveis que corresponde à situação mais "equilibrada" da análise de clusters, no sentido em que será robusta para outras bases de dados da mesma população.

A **atribuição de peso às variáveis** influencia a semelhança/dissemelhança entre os objectos e consequentemente influencia a formação de clusters.

De acordo com [Romesburg C., 1990], há quatro formas de atribuir pesos às variáveis. Na primeira forma, o investigador pode deixar livremente algumas variáveis de fora da matriz de dados original, isto é, atribuir-lhes peso igual a zero.

A segunda, consiste em fazer uma análise de correlação que encontrará variáveis altamente correlacionadas. Se assim for, nas variáveis quantitativas mais correlacionadas, usa-se uma análise de componentes principais para obter um novo conjunto de variáveis não correlacionadas que são as componentes principais (em menor número do que as variáveis originais, em princípio). Embora estas componentes sejam abstractas, é uma base para descrever os objectos.

A terceira forma é escolher uma função de estandardização (descrita na etapa seguinte sobre a transformação de variáveis). Esta irá influenciar a contribuição das variáveis no estudo que está a ser feito.

Na quarta forma, é possível atribuir pesos para fazer com que as variáveis contribuam de uma forma que se baseia na semelhança entre objectos. O peso de uma variável pode ser aumentado pela sua repetição na matriz. Suponhamos que a matriz dos dados tem apenas duas variáveis e suponhamos que queremos que a primeira contribua 60 por cento e a segunda 40 por cento. Podemos acrescentar duas colunas com os respectivos valores da variável 1 e uma coluna com a variável 2. A nova matriz de dados contém 3 colunas com a variável 1 e duas colunas com os valores da variável 2.

Se aplicarmos a distância Euclideana à matriz original dos dados, obtemos a distância Euclideana ponderada, $d_{ij} = \left[\frac{\sum_{k=1}^{p} w_k \left(x_{ki} - x_{kj}\right)^2}{\sum_{k=1}^{p} w_k} \right]^{1/2}$, em que w_k são inteiros, 1,2,3,.... No exemplo dado acima, p=2 e $w_1 = 3$ e $w_2 = 2$, não sendo, pois, preciso que a matriz de dados seja escrita com 5 variáveis.

Atendendo a que o total de variabilidade compromete a variação dentro do grupo e entre os grupos, uma determinação comum dos pesos da matriz X, consiste em definir o peso, w_k, da k-ésima variável por forma a ser inversamente proporcional à variância desta variável, $k = 1,2,3,..., p,$ sendo p o número de variáveis a considerar na Análise de Clusters. Deste modo, a importância da variável decresce com o aumento da sua variabilidade (instabilidade).

Outras medidas de variabilidade podem ser usadas para definir pesos (como o desvio padrão, a amplitude da amostra).

Milligan e Cooper (1988) estudaram o peso para variáveis contínuas e concluíram que o peso baseado simplesmente no intervalo de variação da variável era o mais efectivo.

Há autores tais como Peter Bryant, que defendem a não atribuição de pesos às variáveis, mas defendem a sua estandardização (das variáveis quantitativas) por serem a maioria das vezes medidas em unidades de medida diferentes.

Esta abordagem faz sentido, essencialmente, quando a recolha de dados foi feita com a intenção de realizar uma análise de clusters e, portanto, houve uma selecção cautelosa das variáveis a incluir na análise. No entanto, por vezes, não foi este o objectivo da recolha de dados. É sim, uma proposta do Matemático para responder às questões que lhe propõem. Nesta situação podem existir na base de dados variáveis menos importantes do que outras.

A atribuição de pesos às variáveis depende do objectivo e do contexto da investigação, reflectindo a importância que o investigador atribui às diferentes variáveis na tarefa da classificação. A atribuição de pesos pode ser o resultado de

um estudo feito pelo investigador ou baseada nalguns aspectos dos dados da matriz X. Como vimos acima, o investigador pode determinar os pesos directamente especificando-os, ou indirectamente recorrendo a outras análises estatísticas, como por exemplo recorrendo à Análise em Componentes Principais.

➢ Transformação de variáveis

Como acabámos de ver, um processo utilizado para anular a influência das diferentes unidades de medida e das diferentes variâncias das variáveis, sobre os resultados da análise de clusters, é recorrer à estandardização das variáveis.

Com a estandardização, todas as variáveis terão o mesmo peso no que diz respeito às unidades de medida e à variância. No entanto, nalgumas situações poderão existir variáveis com uma importância superior a qual deverá ser mantida.

Estandardizar é um meio de mudar os dados originais. Há outras duas formas de mudar os dados: transformar os dados usando uma função de transformação como por exemplo $Z_{ij} = \log(X_{ij})$ ou $Z_{ij} = \sqrt{X_{ij}}$ e identificando e depois removendo *outliers*. A função de estandardização usa parâmetros como $\overline{X_j}$ ou S_j, a média da amostra e o desvio padrão da amostra para a variável X_j, respectivamente. A uma matriz de dados pode ser aplicada a estandardização das variáveis e a transformação destas ou apenas uma delas. A identificação e remoção de outliers pode ser feita observando os dados ou aplicando métodos estatísticos.

Para estandardizar uma matriz de dados, devemos escolher primeiro a equação, chamada função de estandardização e aplicá-la à matriz de dados.
A escolha da função de estandardização depende do contexto do problema de investigação e do objectivo da investigação. Numa matriz de dados, com n objectos, $i = 1,2,...,n$ e p variáveis, $j = 1,2,...,p$, a função de estandardização mais

usada é: $Z_{ij} = \frac{X_{ij} - \overline{X_j}}{S_j}$ para qualquer objecto i e qualquer variável j, para cada valor da matriz dos dados, X_{ij} , sendo $S_j = \left(\frac{\sum_{i=1}^{n} \left(X_{ij} - \overline{X_j} \right)^2}{n-1} \right)^{\frac{1}{2}}$ o desvio padrão (com $S_j \neq 0$) e $\overline{X}_j = \frac{\sum_{i=1}^{n} X_{ij}}{n}$ a média dos valores da variável em estudo que constituem a nossa amostra. A estandardização converte as variáveis originais em variáveis sem unidades de medida (Z_{ij} não tem unidade de medida porque o numerador e o denominador estão na mesma unidade de medida).

Outra função de estandardização dos dados iniciais da matriz pode ser feita utilizando proporções, por exemplo[7] $Z_{ij} = \frac{X_{ij}}{R\max_j}$ ou $Z_{ij} = \frac{X_{ij} - R\min_j}{R\min_j - R\min_j}$ ou $Z_{ij} = \frac{X_{ij}}{\sum_{j=1}^{t} X_{ij}}$ (sendo R a amplitude do intervalo de valores que a variável toma e $\min_j$ o valor mínimo da variável j e $\max_j$ o valor máximo da variável j).

De seguida é apresentado um exemplo em que é aplicada a análise de clusters a uma matriz de dados iniciais e, seguidamente, a estes dados estandardizados. Podemos verificar através deste exemplo que obtemos clusters diferentes quando procedemos à estandardização dos dados.

[7] poderá encontrar outras funções de estandardização em [Romesburg C. , 1990], na página 83.

Exemplo 1.1 Ilustra a estandardização pela equação $Z_{ij} = \dfrac{X_{ij} - \overline{X_j}}{S_j}$,

$i = 1,2,...,n$ e $j = 1,2,...,p$,

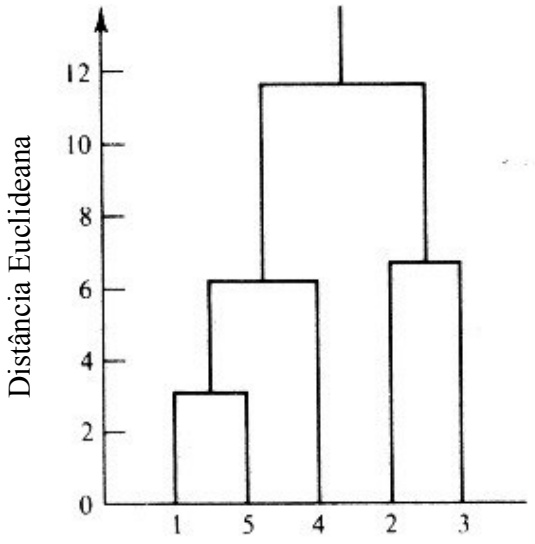
a) Matriz inicial dos dados

Objectos	Variável 1	Variável 2
1	20	19
2	24	6
3	21	12
4	19	24
5	23	18

b)Matriz de dados estandardizados

Objectos	Variável 1	Variável 2
1	-0,68	0.46
2	1.25	-1.41
3	-0.19	-0.55
4	-1.16	1.18
5	0.77	0.32

c) Árvore originada na Matriz inicial dos dados



d) Árvore originada na Matriz de dados estandardizados

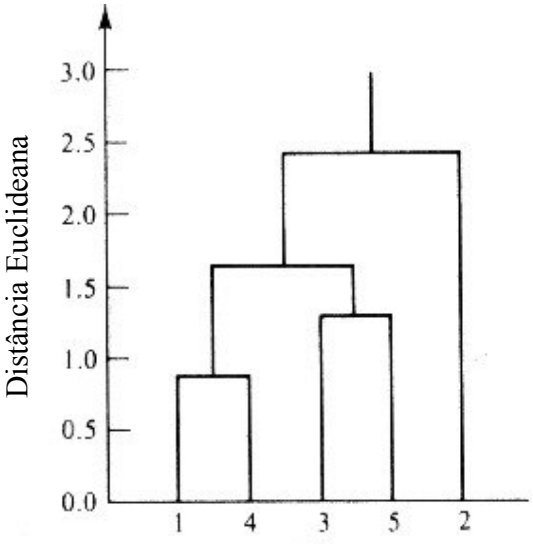


Figura 1 - Representação de diagramas de árvore a partir dos dados originais e a partir destes dados estandardizados, respectivamente.

Uma vez que que a análise de clusters é um método de descrição, a transformação dos dados e a identificação de *outliers* é menos importante do que em métodos inferenciais.

Um problema que surge frequentemente na selecção de variáveis é a **ausência de alguns valores nas variáveis seleccionadas** pelo facto de não se ter conseguido esta informação quando a amostra foi recolhida ou pode ter sido perdida em fase posterior à recolha. Na matriz dos dados pode ser colocado nos valores em falta um símbolo, por exemplo "N". Se esta matriz de dados for estandardizada, colocamos novamente "N" nos valores que faltam.

Se aplicarmos a distância euclidiana à matriz de dados verificamos que o valor de $d_{ij} = \left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right)^{\frac{1}{2}}$ diminui porque menos termos são somados havendo portanto um aumento artificial da semelhança. Mas se aplicarmos $e_{ij} = \left(\frac{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2}{n^{o} \ valores \ que \ são \ comparados} \right)^{\frac{1}{2}}$ dá uma média dos valores que efectivamente são comparados, não se verificando portanto um aumento artificial da semelhança.

Uma generalização deste raciocínio consiste em usar o coeficiente de semelhança geral de Gower para construir a matriz de semelhanças/dissemelhanças com objectos que tenham pelo menos um valor da variável.

O coeficiente de semelhança geral de Gower é dado por: $s_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} s_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$,

em que s_{ijk} é a semelhança entre os objectos i e j na variável k e w_{ijk} é um ou zero consoante a comparação entre os objectos é considerada válida ou não, respectivamente. w_{ijk} é zero se faltar o valor na variável k para um dos objectos ou

para ambos os objectos i e j. Também pode ser zero nas variáveis binárias em que seja apropriado excluir valores negativos. Se as variáveis qualitativas têm mais do que dois níveis, s_{ijk} deverá ser igual a 1 se os dois objectos têm o mesmo valor e deverá ser igual a zero se os dois objectos não têm o mesmo valor. Para as variáveis contínuas Gower sugeriu $s_{ijk} = 1 - \left| x_{ik} - x_{jk} \right| / R_k$, em que R_k é a amplitude do intervalo de variação das observações da variável k.

Embora seja preferível que não faltem dados na matriz de dados, a falta de alguns valores, de preferência poucos, não é crítica.

- As etapas seguintes:
  - Selecção de uma medida de semelhança/dissemelhança.
  - Escolha do método a aplicar aos dados.
  - Discussão e apresentação de resultados serão apresentados nos capítulos 3, 4, 5 e 6 respectivamente.

1.5 Propriedades das medidas de semelhança e de dissemelhança

Consoante o estudo que está a ser feito e o contexto do mesmo, o orientador ou o responsável do estudo escolherá uma medida de semelhança ou uma medida dissemelhança.

Podem ser usadas tanto semelhanças como dissemelhanças no processo de construção de clusters, no entanto a maior parte do *software* disponível para a Análise de Clusters usa dissemelhanças.

A análise teórica das relações de semelhança/ dissemelhança tem sido dominada por modelos geométricos. Nestes modelos os objectos são representados como pontos no espaço de forma que as dissemelhanças

observadas entre os objectos correspondam a distâncias métricas entre os respectivos pontos.

A semelhança mede o grau de proximidade entre os objectos. Por outro lado a dissemelhança reflecte o grau de diferença ou afastamento entre dois objectos. Dois objectos pertencem ao mesmo cluster se são semelhantes e pertencem a clusters diferentes se são dissemelhantes.

Os números d_{ij} (valor de uma medida de dissemelhança entre o objecto i e objecto j) ou s_{ij} (valor de uma medida de semelhança entre o objecto i e objecto j) são colocados numa matriz $n \times n$, conhecida por matriz de proximidades.

1.5.1 Propriedades das medidas de dissemelhança

As dissemelhanças podem ser obtidas de diversas maneiras. Podem ser obtidas de forma objectiva aplicando uma medida de dissemelhança (o que será descrito no terceiro capítulo). Também podem ser obtidas de uma escala subjectiva à qual se atribui uma classificação sobre quanto é que certos objectos diferem[8].

Dada uma colecção de objectos, define-se dissemelhança entre dois objectos da colecção, i e j, como a função dos objectos cujos valores d_{ij} satisfazem as propriedades métricas:

1. $d_{ij} \geq 0$, $\forall i, j \in I$

[8]Por exemplo, se pedirmos a catorze estudantes universitários que indiquem o grau de dissemelhança entre 0 (idênticas) e 10 (muito diferentes) para 11 disciplinas; seguidamente será feita a média das respostas que serão posteriormente colocadas numa matriz 11×11.

2. d_{ii}=0, $\forall i \in I$ (Identidade)

3. $d_{ij} = d_{ji} \quad \forall i, j \in I$ (Simetria)

Poderá acontecer d_{ij}=0 sendo $i \neq j$, por exemplo se dois objectos diferentes tiverem as mesmas medidas nas variáveis em estudo.
Por vezes não se verifica a propriedade 3[9], a simetria. Esta pode ser restabelecida tomando: $d'_{ij} = \frac{d_{ij} + d_{ji}}{2}$.

As dissemelhanças que normalmente são usadas na Análise de Clusters, satisfazem as primeiras três propriedades, embora nenhuma destas propriedades seja realmente essencial havendo métodos de *clustering* que não exigem nenhuma delas, ver [Kaufman L. e Rousseeuw P., 1990].

Se além das propriedades anteriores se verificar a propriedade triangular seguidamente apresentada, então a dissemelhança satisfaz as propriedades de uma **semi-distância:**

4. $d_{ij} \leq d_{ik} + d_{kj}, \forall i, j, k \in I$ (Desigualdade triangular)
Mas, em geral, esta propriedade não é verificada.

Se a semi-distância satisfazer a propriedade seguinte:
5. $d_{ij} = 0 \quad \Leftrightarrow i = j$
então a dissemelhança é **uma distância.**

Se a distância satisfaz a propriedade seguinte, então a dissemelhança é uma **ultra métrica.**

[9] Por exemplo em estudos de atitudes.

6. $d_{ij} \leq \max(d_{ik}, d_{jk}), \forall i, j, k$

No entanto, é suficiente, na prática, que a dissemelhança satisfaça as propriedades 1, 2 e 3.

1.5.2 Propriedades das medidas de semelhança

As semelhanças podem ser obtidas de diversas maneiras. Podem ser o resultado de opiniões subjectivas ou o resultado da aplicação de fórmulas aos objectos que foram avaliados.

Dada uma colecção de objectos, define-se semelhança entre dois objectos da colecção, i e j, como a função dos objectos cujos valores, s_{ij} satisfazem as propriedades métricas:

1. $0 \leq s_{ij} \leq 1, \ \forall i, j$

Os valores entre 0 e 1 indicam vários graus de semelhança. Se s_{ij}=0 significa que i e j não são semelhantes e s_{ij}=1 reflecte a máxima semelhança.

Quando a semelhança depende de grandezas do tipo da correlação, pode acontecer que $-1 \leq s_{ij} \leq 1$.

2. $s_{ij} = s_{ji} \ \ \forall i, j \in I$ (Simetria)

Esta propriedade nem sempre é verificada, como foi dito nas dissemelhanças.

3. $s_{ii} = 1, \ \ \forall i \in I$ (Identidade)

Suponhamos que os dados estão dispostos numa matriz de semelhanças, mas o algoritmo que temos para aplicar exige que a matriz de dados seja uma matriz de dissemelhanças, então é necessário transformar as semelhanças em dissemelhanças. É possível estabelecer uma relação entre as semelhanças e dissemelhanças dos mesmos objectos. A dissemelhança d_{ij} pode obter-se da semelhança s_{ij}, usando uma função decrescente de s_{ij} e $s_{ij} \geq 0$; por exemplo $d_{ij} = 1 - s_{ij}$, ou $d_{ij} = \sqrt{1 - s_{ij}}$. Uma consequência desta última transformação é a matriz de dissemelhanças ser mais homogénea e perder, assim, "clareza" na formação de clusters.

Capítulo 2

Representação gráfica de dados multivariados

2.1 Introdução

A representação gráfica de dados multivariados é importante em todas as fases da Análise de Clusters. Mas, quando esta é feita na fase inicial da investigação, pode revelar-nos a estrutura dos dados, permite que os dados falem por eles próprios. É útil na detecção de padrões.

Antes de realizar uma Análise de Clusters, a representação gráfica ajuda a escolher a medida de semelhanança/dissemelhança que melhor reflecte o "comportamento" dos dados ou o método mais indicado para a construção dos clusters, evidenciando a escolha mais adequada do número de clusters.

As técnicas gráficas para visualização e identificação de clusters, que serão abordadas, são numa ou duas dimensões. Para os métodos hierárquicos há várias versões de dendogramas, que serão abordados no capítulo 4, em 4.4.1, mas para os métodos não hierárquicos pouco se tem feito a nível da sua representação gráfica.

Existe variado *software* que faz representações gráficas de dados multivariados de forma interactiva, tal como o SPSS e o ClustanGraphics8.

2.2 Representação gráfica prévia à Análise de Clusters

2.2.1 Uma ou duas variáveis

De acordo com [Everitt, 2001], uma distribuição unimodal corresponde a uma população homogénea com um cluster. Pelo contrário, a existência de várias modas, indica uma população heterogénea, população com vários clusters, na qual cada moda corresponde a um cluster.

Quando os objectos em estudo são medidos numa variável, normalmente constrói-se o histograma. Para dados univariados, existem outras possibilidades, destacando-se a construção do gráfico de barras, gráficos circulares, gráficos de caule e folhas.

Para dados bivariados (dados relativos a duas variáveis), é possível construir um histograma bidimensional ou um gráfico de barras bidimensional ou um diagrama de dispersão, dependendo da natureza dos dados.

2.2.2 Três ou mais variáveis

Quando existem três ou mais variáveis, poderíamos pensar em construir tantos gráficos de barras quantas as variáveis[10]. No entanto, cada histograma poderia sugerir uma classificação diferente. Quando a observação dos objectos é feita em três ou mais variáveis, o histograma e o diagrama de dispersão não podem ser usados directamente uma vez que estes não podem envolver todas as variáveis em simultâneo.

Se o número de variáveis observadas for p>2, uma possibilidade é projectar os dados a duas dimensões, preservando a estrutura dos dados tanto quanto possível. Há várias análises estatísticas que nos permitem atingir estes resultados, mas uma das mais usadas é a análise em componentes principais.

Caras de Chernof, estrelas e curvas de Andrews são três maneiras de representar dados multivariados.

Nas caras de Chernof a associação entre as variáveis e as caras é subjectiva, e maneiras diferentes de associar as variáveis conduzem a gráficos de diferentes aspectos, o que poderá levar à formação de diferentes clusters (composição e número de clusters) e portanto levar a interpretações diferentes.

Nas estrelas, polígonos ou raios de sol, cada objecto é associado a um círculo de raio constante e o valor das variáveis é indicado ao longo dos raios do círculo. Ao ligar as extremidades dos raios obtém-se um polígono ou estrela.

Nas curvas de Andrews cada objecto do estudo é associado a um objecto familiar do nosso dia a dia ou a entidades matemáticas. Andrews associou ao objecto $x_r = \left(x_{r_1}, x_{r_{21}}, ..., x_{r_p}\right)$ a função harmónica:

$$f_r(t) = \frac{x_{r_1}}{\sqrt{2}} + x_{r_2} sent + x_{r_3} \cos t + x_{r_4} sen(2t) + x_{r_5} \cos(2t) + \ldots \text{, onde } t \in]-\pi, \pi[.$$

[10] Este procedimento é aplicado na Parte II-2, na Análise Descritiva feita à base de dados obtida a partir do inquérito feito aos alunos de 9º ano.

Depois de feita a representação gráfica, verifica-se que a distância euclideana entre dois objectos i e j é proporcional à distância euclideana entre as respectivas funções $f_i(t)$ e $f_j(t)$.

Uma grande desvantagem das curvas de Andrews e das caras de Chernoff é a interpretação muito difícil quando há muitos objectos e a mudança da representação obtida quando muda a ordem das variáveis. Novamente a estandardização é útil quando as variáveis são medidas em unidades diferentes.

Além dos três métodos atrás referidos na representação gráfica de três ou mais variáveis, existem outros tais como caixas de bigodes, bolhas, diagramas de contornos.

2.3 Representação gráfica indirecta

Como já referimos no parágrafo anterior, muitos dos métodos de Análise Multivariada levam à redução do número de dimensões do espaço inicial de trabalho. Este resultado é muito útil porque a redução do número de variáveis facilita a interpretação da estrutura dos dados e as análises subsequentes. Uma outra vantagem que resulta desta redução da dimensão traduz-se na possibilidade de os objectos serem representados graficamente em espaços de pequena dimensão, eventualmente em espaços de duas dimensões, sendo, então, mais fácilmente visualizados.

Será apresentada em seguida, uma breve abordagem a três métodos de análise multivariada que podem levar à redução do número de variáveis. Não é nosso objectivo entrar detalhadamente nos três métodos, mas apenas apresentar a informação necessária à percepção do papel que podem desempenhar na realização de Análise de Clusters. Informação mais completa poderá ser encontrada em [Johnson e Wichern, 2002], [Jobson, 1992], [Everitt e Dunn, 2001].

2.3.1 Componentes principais

É uma técnica da Estatística Multivariada, onde se pretende transformar um conjunto de variáveis relacionadas $X_1, X_2, ..., X_p$ num outro, que desejamos que tenha um menor número de variáveis. Estas novas variáveis não estão relacionadas. As novas variáveis $Y_1, Y_2, ..., Y_p$ chamam-se componentes principais. Cada componente principal é combinação linear das variáveis originais. Uma medida da quantidade de informação transmitida por cada componente principal é a sua variância. Por esta razão as componentes principais aparecem ordenadas segundo a magnitude de variância ($\text{var}\, y_1 \geq \text{var}\, y_2 \geq ... \geq \text{var}\, y_p$). Assim, a componente principal mais informativa é a primeira, e a menos informativa é a última. Tendo em conta isto, o investigador pode optar por analisar apenas as primeiras componentes principais. Será analisado uma menor quantidade de informação mas, em contrapartida, ganha em termos de simplificação e compreensão imediata.

As componentes principais podem também ser usadas para detectar outliers multivariados, pois os valores muito grandes ou muito pequenos das componentes principais retidas são candidatos a outliers. Podem, ainda ser usadas para testar a Normalidade. Se as componentes principais não forem Normalmente distribuídas, então as variáveis originais também não são.

Este método parte de p variáveis iniciais $X_1, X_2, ..., X_p$ observadas em n objectos e encontra p combinações lineares $Y_1, Y_2, ..., Y_p$ com $Y_i = a_{i1}X_1 + ... + a_{ip}X_p$, $i = 1,2,...., p$ não correlacionadas entre si.

As características principais deste método são as seguintes:

- As novas variáveis, as componentes principais $Y_1, Y_2, ..., Y_p$ são definidas como:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p$$
$$Y_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$

.

.

.

$$Y_p = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p$$

- Os coeficientes são obtidos a partir dos vectores próprios da matriz de covariâncias S, mas se as variáveis tiverem diferentes escalas; são encontrados a partir da matriz de correlações R. Na prática, é mais frequente obtermos os valores próprios da matriz das correlações.
- As variâncias de $Y_1, Y_2, ..., Y_p$ são dadas pelos vectores próprios de S ou R; Os primeiros componentes contribuem para explicar uma grande proporção da variância total.

A projecção dos dados, de dimensão p, num espaço de dimensão menor, q, $q < p$, obtida a partir duma Análise de Componentes Principais, fornece-nos uma visão mais informativa de possíveis agrupamentos naturais de dados.

2.3.2 Multidimensional scaling (MDS)

A técnica MDS (*multidimensional scaling*) também é útil na apresentação dos clusters a duas dimensões afim de que haja uma melhor visualização da sua estrutura.

É um conjunto de técnicas que usam proximidades entre objectos para produzir uma representação espacial dos objectos.
A matriz de proximidades é, geralmente, uma matriz de dissemelhanças.
A representação espacial que se obtém consiste numa configuração geométrica de pontos num mapa, cada ponto correspondendo a um dos objectos.
Quanto maior for a semelhança entre os objectos, mais próximos eles se encontrarão no mapa.
As medidas de proximidade usadas para relacionar os objectos, em muitas aplicações são baseadas não em medições directas, mas em avaliações de semelhança originadas em apreciações feitas por pessoas.

Distingue-se, pois, duas formas de MDS:

- a MDS métrica que é baseada em valores das medidas de proximidade;
- a MDS não-métrica, que é baseada em proximidades subjectivas obtidas a partir de apreciações humanas;

Na MDS métrica a representação espacial tenta preservar as distâncias entre objectos, enquanto que na MDS não-métrica a representação espacial apenas preserva a ordem da grandeza das dissemelhanças.

A MDS é uma técnica da análise exploratória de dados. Procura determinar e compreender as dimensões, subjacentes ao nosso conjunto de dados, que contribuem para as diferenças observadas (perceptíveis) entre os objectos.

A MDS métrica começa com uma matriz de proximidades $n \times n$, **D**, de dissemelhanças δ_{rs}, r,s=1,2,…,n, que associa a todos os pares possíveis dos n objectos, uma medida de dissemelhança. Os elementos da diagonal principal de **D** são, portanto, zero. O objectivo da MDS métrica é definir um conjunto de q dimensões (q≤p) subjacentes (q variáveis), $X_1, X_2, ..., X_q$, tais que:

- As coordenadas dos n objectos nestas q dimensões obtidas originam uma matriz de distâncias euclideanas;

- Os elementos da matriz de distâncias euclideanas são equivalentes, ou muito próximos, aos elementos δ_{rs} de D.

Em geral, quanto maior for o número de dimensões que usamos para reproduzir a matriz de distâncias, melhor será o ajustamento da matriz obtida à matriz observada. De facto, se usarmos tantas dimensões quantas as variáveis na nossa base de dados, então podemos reproduzir perfeitamente a matriz de distâncias observadas. Mas o nosso objectivo é diminuir a complexidade do fenómeno observado, isto é, explicar a matriz de distâncias em termos de um menor número de dimensões subjacentes. Temos também de ter em conta que a configuração final deve ser clara para ser interpretada sem ambiguidades. Como podemos constatar, ao contrário de outras técnicas de análise multivariada, a matriz **X**, n x p , de observações, é obtida a partir de uma matriz **D** de dissemelhanças dada. As dissemelhanças observadas δ_{rs} são usadas para construir um conjunto de distâncias derivadas, d_{rs} , que estão relacionadas de uma forma muito próxima com as dissemelhanças observadas δ_{rs} através de uma função monótona crescente f , isto é,

$$d_{rs} \approx f(\delta_{rs}),$$

em que f é uma função tal que: $\delta_{rs} < \delta_{uv} \Leftrightarrow f(\delta_{rs}) < f(\delta_{uv})$.

2.3.3 Análise factorial

A análise factorial pode ser usada para representar graficamente as variáveis, com o objectivo de encontrar agrupamentos de variáveis.
Foi inventada há cerca de cem anos pelo Psicólogo Charles Spearman, que considerou a hipótese de que a grande variedade de testes de capacidade mental, medidas de aptidão matemática, de vocabulário, outras aptidões verbais, aptidões

artísticas, capacidade de raciocínio lógico, etc.; poderia ser explicado por um "factor" subjacente a que ele chamou δ. Provou-se que isto não chega, são necessários três factores importantes de capacidade mental: capacidade verbal, matemática e lógica. E os psicólogos estão de acordo em que muitos outros factores poderiam também ser identificados.

Nesta análise o objectivo é semelhante ao da análise em componentes principais – descrever a variabilidade entre muitas variáveis em termos de um pequeno número de variáveis aleatórias, chamadas factores, subjacentes, mas não observáveis.

O modelo da análise factorial, pode ser expresso algebricamente por um conjunto de p equações lineares:

$$X_1 - \mu_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + ... + \lambda_{1k}F_k + E_1$$
$$X_2 - \mu_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + ... + \lambda_{2k}F_k + E_2$$
.
.
.
$$X_p - \mu_p = \lambda_{p1}F_p + \lambda_{p2}F_2 + ... + \lambda_{pk}F_k + E_p$$

Sendo: $X_1, X_2, ..., X_p$ um conjunto de p variáveis observadas; $F_1, F_2, ..., F_k$ um conjunto de k variáveis ($k < p$) não observadas, chamadas factores comuns; $E_1, ..., E_p$ um conjunto de p factores específicos não observados. $\lambda_{1k}, \lambda_{2k}, ..., \lambda_{pk}$ são os pesos do factor k. Os factores $F_1, F_2, ..., F_k$ são comuns a todas as variáveis $X_1, X_2, ..., X_p$, enquanto que os erros ou os factores específicos são únicos para cada variável X_i. $\mu_1, \mu_2, ..., \mu_p$ são os valores médios das variáveis $X_1, X_2, ..., X_p$.

Baseia-se na matriz das correlações ou na matriz das covariâncias. Se as variáveis X estiverem estandardizadas, $\mu_1 = \mu_2 = ... = \mu_p = 0$ e trabalhamos com a matriz das correlações. Todas as correlações ou covariâncias são explicadas pelos factores comuns. A parte da variância que não é explicada pelos factores comuns é incluída no erro residual ou factor específico. Assume-se que os factores específicos são não correlacionados. O modelo de análise factorial

assume, então, que a matriz das covariâncias ou matriz das correlações pode ser dividida em duas partes. A 1ª parte é gerada pelos factores comuns e a 2ª parte é gerada pelos erros ou factores específicos.

Este modelo pode ser usado para explorar a estrutura dos dados e investigar a relação entre as variáveis observadas e os factores.

A Análise factorial procura descobrir se as variáveis observadas podem ser em grande parte ou na totalidade aplicadas à custa de um muito menor número de variáveis a que chamamos factores.

Além dos três métodos que acabámos de referir, a representação gráfica indirecta de objectos e variáveis na análise multivariada pode utilizar, a análise de correspondências e o método que assenta na representação gráfica biplot. Estes dois métodos não serão abordados neste trabalho.

Capítulo 3
Medidas de proximidade

3.1 Introdução

Em relação à escolha da medida de proximidade, medida de dissemelhança/semelhança a aplicar aos dados, não se conhece uma fórmula para tal; por isso, e apesar de tanta oferta, os investigadores continuam a propor novos coeficientes.

Tal como Gower e Legendre (1986) referiram "um coeficiente tem de ser considerado no contexto do estudo estatístico, incluindo a natureza dos dados e do tipo de análise pretendido". Sugeriram alguns critérios para ajudar a fazer a escolha. Em primeiro lugar a matriz dos dados deve influenciar profundamente a escolha da medida de proximidade. Em segundo lugar, a medida deverá depender da escala dos dados. Em terceiro lugar, o método a aplicar para a construção de clusters deve ter alguma relação com a medida escolhida.

Gower and Legendre, apresentaram uma análise detalhada acerca da escolha da medida de semelhança ou de dissemelhança e uma tabela de medidas de proximidade que poderá ajudar nestas situações. Contudo, concluíram que não é possível dar uma resposta definitiva acerca de qual a melhor medida a ser usada.

As medidas de proximidade podem ser entre objectos ou entre variáveis, consoante se pretende obter clusters de objectos ou clusters de variáveis, respectivamente.

3.2 Medidas de proximidade entre objectos

As medidas de proximidade entre objectos são medidas quantitativas, referidas geralmente como semelhanças ou dissemelhanças.

São obtidas a partir de uma matriz multivariada X de dimensão $n \times p$ resultante da observação de p variáveis em n objectos, e são escolhidas de acordo com o tipo de variáveis.

Dois objectos estão próximos quando a sua dissemelhança ou distância é pequena ou a sua semelhança é grande.

Há muitas medidas de proximidade, mas nos parágrafos seguintes serão descritas apenas algumas delas.

A maior parte dos métodos para a construção de clusters usam algoritmos que operam sobre dissemelhanças. Portanto, o utilizador deverá transformar a matriz de semelhanças numa matriz de dissemelhanças quando for mais conveniente começar por construir uma matriz de semelhança, com vista à utilização daqueles algoritmos.

Como já vimos anteriormente, podemos converter uma semelhança numa dissemelhança, tomando por exemplo $d_{ij} = 1 - s_{ij}$. Também podemos converter uma dissemelhança numa semelhança, fazendo por exemplo $s_{ij} = 1 - d_{ij}$.

Por vezes parece natural a utilização da semelhança em vez da dissemelhança, ou vice-versa. Geometricamente a distância é mais perceptível.

Muitas das medidas de dissemelhança são inspiradas em modelos geométricos, o que leva a que a dissemelhança seja visualizada como a distância entre pontos no espaço, como vimos no capítulo anterior.

Como já vimos na introdução a este capítulo, as medidas de proximidade dependem em primeiro lugar, da natureza das características que são observadas nos objectos. Por isso, serão apresentadas várias medidas de proximidade entre objectos para:

- variáveis qualitativas ,
- variáveis quantitativas e
- variáveis de diferentes tipos .

3.2.1 Variáveis qualitativas

Num conjunto de dados com variáveis qualitativas, são usadas normalmente medidas de semelhança. Estas medidas de semelhança geralmente têm valores pertencentes ao intervalo $[0,1]$, embora por vezes sejam expressas em percentagem, estando portanto os seus valores no intervalo $[0,100]$.

Dois objectos i e j têm um coeficiente de semelhança igual a um, $s_{ij} = 1$, se têm valores idênticos para todas as variáveis.

Dois objectos i e j têm um coeficiente de semelhança igual a zero, $s_{ij} = 0$, se diferem no máximo para todas as variáveis.

3.2.1.1 Medidas de semelhança para variáveis nominais com dois níveis (binárias)

Existem vários coeficientes/medidas de semelhança utilizados em dados binários[11] (serão apresentados na tabela 2). Na matriz de dados, estas variáveis são geralmente codificadas de um ou zero. Quando a f-ésima variável é binária, os objectos terão valor $x_{if} = 0$ (se atributo a que se refere a variável não está presente) ou $x_{if} = 1$ (se atributo a que se refere a variável está presente). Poderá ser indicado um terceiro código no caso de desconhecermos a resposta.

Num determinado problema podem ser aplicados vários coeficientes. A escolha depende do tipo de problema em estudo, dos objectivos a atingir e da experiência do investigador, ou melhor, do bom senso do investigador. O investigador terá de ter conhecimento profundo do assunto em estudo e usá-lo de forma criteriosa e eficiente para considerar um coeficiente que integre as suas hipóteses de trabalho e vá de encontro aos objectivos que pretende.

Note-se que a escolha de diferentes coeficientes de semelhança pode originar resultados muito diferentes.

Na escolha do coeficiente de semelhança deverá ter-se em conta a utilidade da informação que o respectivo valor fornece ao nosso estudo. Por exemplo, quando a falta de uma característica nos dois objectos em estudo é considerada informativa para o nosso estudo, é normalmente usado o coeficiente de concordância simples [Everitt B. e al, 2001].

Se as variáveis binárias têm o mesmo peso, consideramos uma tabela dicotómica, tabela de dupla entrada ou tabela de contingência para cada uma delas (ver tabela 1, apresentada em baixo).

Todas as medidas são definidas a partir das entradas cruzadas para dois objectos i e j.

[11] Exemplos de dados binários: masculino/feminino, sim/não, fumador/não fumador.

Tabela 1 - Tabela de contingência ou tabela de associação

	Objecto j		
	1	0	
Objecto i 1	a	b	a + b
Objecto i 0	c	d	c + d
	a + c	b + d	a+b+c+d=p

- as variáveis binárias têm valor 1 se a característica está presente e valor 0 se a característica não está presente;
- os valores a, b, c e d são contagens;
- os valores a, b, c e d correspondem a :

  - **a** número de variáveis, de entre as p observadas, que tomam o valor 1 para os dois objectos i e j, isto é, $x_{if} = x_{jf} = 1$; ou, dito de outra forma, o número de atributos, de entre os p observados, que estão presentes nos dois objectos, i e j.

  - **b** número de variáveis, de entre as p observadas, que tomam o valor 1 no objecto i e o valor 0 no objecto j, isto é, o número de atributos que estão presentes apenas no objecto i; $x_{if} = 1$ e $x_{jf} = 0.$

  - **c** número de variáveis, de entre as p observadas, que tomam o valor 0 no objecto i e o valor 1 no objecto j, isto é, o número de atributos que estão presentes apenas no objecto j ; $x_{if} = 0$ e $x_{jf} = 1.$

  - **d** número de variáveis, de entre as p observadas, que tomam o valor 0 para os dois objectos i e j, isto é, se a característica não está presente nos dois objectos i e j, $x_{if} = x_{jf} = 0$, ou dito de outra forma, o número de atributos, de entre os p observados, que não estão presentes em nenhum dos objectos, i e j.

Tendo em conta a informação descrita anteriormente na tabela 1, a tabela 2 apresenta uma lista de coeficientes de semelhança. Será usado o símbolo C_{ij} para representar qualquer coeficiente de semelhança entre os objectos i e j. Cada coeficiente será uma função diferente dos valores de a, b, c, e d da tabela anterior. Uma lista mais extensa poderá ser encontrada em Gower e Legendre (1986).

Tabela 2 - Equações e intervalo de variação dos coeficientes de semelhança mais usados para variáveis binárias.

Coeficiente	**Intervalo de variação**
1. Jaccard $C_{ij} = \frac{a}{a+b+c}$	$[0,1]$
2. Concordância Simples $C_{ij} = \frac{a+d}{a+b+c+d}$	$[0,1]$
3. Yule $C_{ij} = \frac{ad-bc}{ad+bc}$	$[-1,1]$
4. Hamann $C_{ij} = \frac{(a+d)-(b+c)}{(a+d)+(b+c)}$	$[-1,1]$
5. Sorenson $C_{ij} = \frac{2a}{2a+b+c}$	$[0,1]$
6. Rogers e Tanimoto $C_{ij} = \frac{a+d}{a+2(b+c)+d}$	$[0,1]$
7. Sokal e Sneath $C_{ij} = \frac{2(a+d)}{2(a+d)+b+c}$	$[0,1]$
8. Russel e Rao	

$C_{ij} = \frac{a}{a+b+c+d}$	$[0,1]$
9. Baroni-Urbani e Buser $C_{ij} = \frac{a+(ad)^{1/2}}{a+b+c+(ad)^{1/2}}$	$[0,1]$
10. Distância Binária de Sokal $C_{ij} = \left(\frac{b+c}{a+b+c+d}\right)^{1/2}$	$[0,1]$
11. Ochiai $C_{ij} = \frac{a}{[(a+b)(a+c)]^{1/2}}$	$[0,1]$
12. Phi $C_{ij} = \frac{ad-bc}{[(a+b)(a+c)(b+d)(c+d)]^{1/2}}$	$[-1,1]$
13.Gower e Legendre $C_{ij} = \frac{a+d}{a+1/2(b+c)+d}$	$[0,1]$
14. Gower e Legendre $C_{ij} = \frac{a}{a+1/2(b+c)}$	$[0,1]$

- Note-se que, como já foi dito no início de 3.2, a partir da semelhança s_{ij}, cujo intervalo de variação é [0,1], podemos obter a correspondente dissemelhança, d_{ij}, fazendo d_{ij}=1-s_{ij}. Por exemplo, o coeficiente de semelhança de Jacard, como foi visto na tabela anterior, é dado por $s_{ij} = \frac{a}{a+b+c}$, então o respectivo coeficiente de dissemelhança é $d_{ij} = 1 - \frac{a}{a+b+c} = \frac{b+c}{a+b+c}$.

➢ Algumas considerações sobre as medidas apresentadas na tabela 2:

1. O coeficiente de Jaccard, definido por $C_{ij} = \frac{a}{a+b+c}$, $0 \le C_{ij} \le 1$.

C_{ij}=1 dá-nos a semelhança máxima quando os dois objectos têm valores idênticos, isto é, no caso em que b=c=0.

C_{ij}=0 dá-nos a dissemelhança máxima quando a=0, isto é, quando nenhum dos atributos está nos dois objectos simultaneamente.

Este coeficiente é muito usado na Taxonomia Numérica e na Ecologia.

O coeficiente de Jacard, definido como:

$$JAC_{ij} = \frac{\text{nº de variáveis que tomam valor 1 em ambos os objectos}}{\text{nº de variáveis que estão presentes em pelo menos um dos objectos}}$$

dá mais importância à situação de os atributos estarem presentes em ambos objectos e, não dá importância à situação em que os atributos não estão presentes em nenhum dos dois objectos. Se ambos os objectos têm muitos atributos em falta, pode não ser desejável dizer que são semelhantes. Por exemplo na taxonomia numérica o coeficiente de Jacard é usualmente o preferido. Uma vez que um peixe e um pássaro têm poucos atributos em comum, ninguém quereria dizer que as duas espécies são semelhantes.

2. O coeficiente Concordância Simples, definido por $C_{ij} = \frac{a+d}{a+b+c+d}$,

$0 \le C_{ij} \le 1$.

C_{ij}=1 dá-nos a semelhança máxima e ocorre quando b=c=0.

É usado por exemplo em estudos sobre medicamentos em Farmácia.

3. Coeficiente de Yule $C_{ij} = \frac{ad - bc}{ad + bc}$, $-1 \leq C_{ij} \leq 1$

C_{ij} =1 dá-nos a perfeita semelhança e ocorre quando b=0 ou c=0

C_{ij} =-1 dá-nos a máxima dissemelhança que ocorre quando a=0 ou d=0

C_{ij} =0 valor intermédio entre os valores extremos, quando ad=bc.

Este coeficiente tem sido usado na Investigação em Psicologia.

4. Coeficiente de Haman $C_{ij} = \frac{(a+d)-(b+c)}{(a+d)+(b+c)}$; $-1 \leq C_{ij} \leq 1$

C_{ij} =1 dá-nos a perfeita semelhança que ocorre quando b=0 ou c=0

C_{ij} =-1 dá-nos a máxima dissemelhança e ocorre quando a=d=0

C_{ij} =0 valor intermédio entre os valores extremos, quando a+d=b+c

5. Coeficiente de Sorenson $C_{ij} = \frac{2a}{2a+b+c}$; $0 \leq C_{ij} \leq 1$

C_{ij} =1 dá-nos a perfeita semelhança que ocorre quando b=c=0

C_{ij} =0 dá-nos a máxima dissemelhança; ocorre quando a=0, isto é quando os objectos não têm atributos comuns.

É usado por exemplo em Botânica, em estudos de análise de similaridade entre comunidades florestais.

6. Coeficiente de Rogers e Tanimoto $C_{ij} = \frac{a+d}{a+2(b+c)+d}$; $0 \leq C_{ij} \leq 1$

C_{ij} =1 dá-nos a perfeita semelhança; ocorre quando b=c=0

C_{ij} =0 dá-nos a máxima dissemelhança; ocorre quando a=d=0

É usado por exemplo em Botânica, em Agro-pecuária.

7. Coeficiente de Sokal e Sneath $C_{ij} = \frac{2(a+d)}{2(a+d)+b+c}$; $0 \le C_{ij} \le 1$

C_{ij} =1 dá-nos a perfeita semelhança; ocorre quando b=c=0

C_{ij} =0 dá-nos a máxima dissemelhança; ocorre quando a=d=0

É usado por exemplo em estudos sobre Zoologia, Genética.

8. Coeficiente de Russell e Rao $C_{ij} = \frac{a}{a+b+c+d}$; $0 \le C_{ij} \le 1$

C_{ij} =1 dá-nos a perfeita semelhança; ocorre quando b=c=d=0

C_{ij} =0 dá-nos a máxima dissemelhança; ocorre quando a=0

9. Coeficiente de Baroni-Urbani e Rao $C_{ij} = \frac{a}{a+b+c+d}$; $0 \le C_{ij} \le 1$

C_{ij} =1 dá-nos a perfeita semelhança; ocorre quando b=c=d=0

C_{ij} =0 dá-nos a máxima dissemelhança; ocorre quando a=0

10. Coeficiente de distância Binária de Sokal $C_{ij} = \frac{2(a+d)}{2(a+d)+b+c}$; $0 \le C_{ij} \le 1$

C_{ij} =1 dá-nos a perfeita semelhança; ocorre quando b=c=0

C_{ij} =0 dá-nos a máxima dissemelhança; ocorre quando a=d=0

11. Coeficiente de Ochiai $C_{ij} = \frac{a}{[(a+b)(a+c)]^{1/2}}$; $0 \le C_{ij} \le 1$

C_{ij} =1 dá-nos a perfeita semelhança; ocorre quando b=c=0

C_{ij} =0 dá-nos a máxima dissemelhança; ocorre quando a=0

12. Coeficiente de Phi $C_{ij} = \frac{ad - bc}{[(a+b)(a+c)(b+d)(c+d)]^{1/2}}$; $-1 \leq C_{ij} \leq 1$

C_{ij} =1 dá-nos a perfeita semelhança; ocorre quando b=c=0

C_{ij} =-1 dá-nos a máxima dissemelhança; ocorre quando a=d=0

É usado por exemplo em estudos sobre Psicologia e Psiquiatria.

- Muitos dos coeficientes da tabela 2 estão correlacionados.
- Quando d=0, ou seja quando não há pares (0,0) e, portanto, pelo menos um dos atributos está sempre presente, os coeficientes de Jacard e de Concordância Simples têm os mesmos valores, conduzindo portanto aos mesmos resultados na Análise de Clusters.
- Gower e Legendre (1986), defenderam que o coeficiente de Concordância Simples (ou Matching Coefficient), o Coeficiente de Rogers e Tanimoto, o coeficiente de Gower e Legendre: (a+d)/(a+1/2(b+c)+d) estão relacionados monotonicamente assim como os coeficientes de Jacard, Sokal e Sneath e o coeficiente de Gower e Legendre : a/(a+1/2(b+c)).
- O coeficiente de Sneath e Sokal está relacionado de uma forma monótona com o Coeficiente de Matching, sendo sempre maior ou igual a este.
- Tanto os coeficientes de Jaccard, de Sorenson e de Baroni-Urbani e Buser estão relacionados de forma monótona entre si. Os coeficientes de Simple Macthing, de Sokal e Sneath, de Tanimoto e de Hamann também estão relacionados entre si.
- Embora a escolha do coeficiente de semelhança deva ser feita de forma lógica, há uma tendência entre os investigadores de se usar certos coeficientes em certas áreas, mesmo que seja com menos lógica. Isto explica porque razão os geólogos usam de preferência o coeficiente do cosseno e os ecologistas o coeficiente de Bray-Curtis (coeficiente de

semelhança para variáveis quantitativas analisado de seguida em 3.2.). Alguns investigadores optam por usar um coefiente de semelhança que outros na sua área já usaram ou porque há opção desse coeficiente no *software* que estão a usar.

3.2.1.2 Medidas de semelhança para variáveis nominais com mais de dois níveis

Quando as variáveis têm mais do que dois níveis, normalmente a estratégia é decompor cada variável em variáveis binárias, tantas quantos os níveis dessa variável, e construir a partir do vector de variáveis binárias resultante, um coeficiente de semelhança (atrás referido).

Outra maneira de construir um coeficiente de semelhança para variáveis categorizadas com mais de dois níveis é fazer $s_{ij} = \frac{1}{p}\sum_{k=1}^{p} s_{ijk}$, em que s_{ijk} toma valor zero ou 1 para cada variável k quando os dois objectos i e j assumem o mesmo nível, sendo p o número total de variáveis. A desvantagem deste coeficiente é tratar de forma igual todas as variáveis, quer elas tenham muitos ou poucos níveis.

Com o fim de corrigir o desequílibrio causado pelo diferente número de níveis de cada variável, faz-se intervir no cálculo do coeficiente o número de níveis de cada variável. Suponhamos que há p variáveis, $y_1,...,y_p$ com $l_1,...,l_p$ níveis respectivamente, então o coeficiente de semelhança s_{ij} será

$$s_{ij} = \frac{\sum_{k=1}^{p} l_k I(y_k(i), y_k(j))}{\sum_{K=1}^{p} l_k},$$

sendo I a função indicatriz dos níveis dos dois objectos, i e j , na variável k, isto é,

$$I(y_k(i), y_k(j)) = \begin{cases} 1 \quad se \quad y_k(i) = y_k(j) \\ 0 \quad se \quad y_k(i) \neq y_k(j) \end{cases}$$

onde $y_k(i) \;\; e \;\; y_k(j)$ são os níveis do objecto i e do objecto j na variável k, respectivamente.

Outra proposta, é considerar:

$$s_{ij} = \frac{\sum_{k=1}^{p} \ln l_k I(y_k(i), y_k(j))}{\sum_{k=1}^{p} \ln l_k}$$

Usa-se o logaritmo para atenuar os resultados.

3.2.1.3 Medidas de semelhança para variáveis ordinais

Considera-se que se um objecto tem um certo nível, tem naturalmente os níveis inferiores (ex: níveis de escolaridade). Assim, as variáveis ordinais terão de ser tratadas como variáveis normais e cada uma terá de ser decomposta em

tantas variáveis binárias quantos os níveis que tem. Seguidamente constrói-se uma tabela de duas entradas para cada par de objectos onde serão apresentados os números de níveis comuns a ambos os objectos, o número de níveis que só estão num dos objectos e o número de níveis que não estão em nenhum dos objectos. Agora, então, podemos calcular os valores dos coeficientes de semelhança atrás referidos para variáveis binárias.

Consideremos o seguinte exemplo:
Suponhamos que duas pessoas, A e B, têm o 3º ciclo e licenciatura, respectivamente. A variável ordinal tem nove níveis:

1-Nenhum nível de ensino
2-Básico-1º ciclo/4ªclasse
3-Básico - 2º ciclo
4-Básico - 3º ciclo
5-Secundário
6-Superior, Curso médio ou apenas alguns anos da niversidade (bacharelato)
7-Superior - Licenciatura
8-Superior – mestrado
9-Superior – doutoramento

As nove variáveis binárias permitem definir os vectores associados às duas pessoas:

A	1	1	1	1	0	0	0	0	0
B	1	1	1	1	1	1	1	0	0

De onde se obtém:

		Pessoa B	
		1	0
Pessoa A	1	4	0
	0	3	2

Portanto o coeficiente de Jaccard seria $s_{AB} = \frac{4}{4+0+3} = 0{,}57$, enquanto que o coeficiente de concordância simples $s_{AB} = \frac{4+2}{4+0+3+2} = 0{,}66$. O coeficiente de Jaccard revelou-se mais pequeno, porque não dá importância ao que não há em comum.

3.2.2 Variáveis quantitativas

Quando as variáveis são contínuas, as medidas de proximidade entre os objectos são normalmente medidas de dissemelhança ou medidas de distância, δ_{ij} ou d_{ij}. As medidas de distância gozam das propriedades anteriormente referidas em 1.5.

É possível a conversão de dados contínuos em categorizados e usar medidas para variáveis nominais. No entanto há que ter em conta a perda de informação na conversão.

Na tabela 3, apresenta-se uma lista de coeficientes/medidas de dissemelhança para dados quantitativos.

Tabela 3 - Lista de alguns coeficientes de dissemelhança para dados quantitativos

Medida	**Fórmula**
Distância Euclideana	$d_{ij} = \left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right)^{\frac{1}{2}}$

Distância City Block	$d_{ij} = \sum_{k=1}^{p} \left\| x_{ik} - x_{jk} \right\|$
Distância de Minkowski	$d_{ij} = \left[\sum_{k=1}^{p} \left\| x_{ik} - x_{jk} \right\|^{r} \right]^{\frac{1}{r}}, \forall r \in IR : r \geq 1$
Distância de Camberra	$d_{ij} = \sum_{k=1}^{p} \frac{\left\| x_{ik} - x_{jk} \right\|}{\left\| x_{ik} \right\| + \left\| x_{jk} \right\|}$
Coeficiente de correlação de Pearson	$r_{ij} = \frac{\sum_{k=1}^{p} \left(x_{ik} - \bar{x}_{i.} \right)\left(x_{jk} - \bar{x}_{j.} \right)}{\left[\sum_{k=1}^{p} \left(x_{ik} - \bar{x}_{i.} \right)^2 \sum_{k=1}^{p} \left(x_{jk} - \bar{x}_{j.} \right)^2 \right]^{\frac{1}{2}}}$, onde $\bar{x}_{s.} = \frac{\sum_{k=1}^{p} x_{sk}}{p}$
Coeficiente de separação angular	$c_{ij} = \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\left(\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{jk}^2 \right)^{\frac{1}{2}}}$
Coeficiente de Bray-Curtis	$d_{ij} = \frac{\sum_{k=1}^{p} \left\| x_{ik} - x_{jk} \right\|}{\sum_{k=1}^{p} \left(x_{ik} + x_{jk} \right)}$

Analisa-se de seguida alguns destes coeficientes que são medidas de dissemelhança para dados contínuos.

3.2.2.1 Distância Euclideana

É a medida de proximidade mais usada para este tipo de variáveis. A distância euclideana entre dois objectos i e j, define-se por:

$$d_{ij} = \left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right)^{\frac{1}{2}} = \sqrt{\left(x_{i1} - x_{j1} \right)^2 + \left(x_{i2} - x_{j2} \right)^2 + ... \left(x_{ip} - x_{jp} \right)^2} \ ,$$

para uma matriz de dados X, onde x_{ik} e x_{jk} correspondem aos valores observados da k-ésima variável nos dois objectos i e j, respectivamente; p é o número de variáveis que foram observadas.

Geometricamente, pode ser interpretada como a distância entre dois pontos num espaço de dimensão p: $\left(x_{i1}, x_{i2}, ..., x_{ip} \right)$ e $\left(x_{j1}, x_{j2}, ..., x_{jp} \right)$.

A distância euclideana satisfaz as propriedades usuais da função semi-distância:

- $d_{ij} \geq 0, \forall i, j$, a distância são números não negativos
- $d_{ii} = 0, \forall i$, a distância de um objecto a si próprio é zero
- $d_{ij} = d_{ji}, \forall i, j$, (simetria)
- $d_{ij} \leq d_{ih} + d_{hj}, \forall i, j, h$,

Note-se que nem sempre $d_{ij} = 0$ implica que $i = j$. Pode acontecer, por exemplo, dois objectos diferentes terem os mesmos valores para as variáveis em estudo.

A distância euclideana depende da escala das variáveis e pode ser "destorcida" por outliers. Por razões de interpretação, há quem prefira o quadrado da distância Euclideana, d_{ij}^2.

Consideremos o caso especial para $p = 2$, a distância entre dois objectos i e j, representados por: $\left(x_{i1}, x_{i2} \right)$ e $\left(x_{j1}, x_{j2} \right)$ é dada, como é ilustrado na figura seguinte, pelo comprimento da hipotenusa do triângulo.

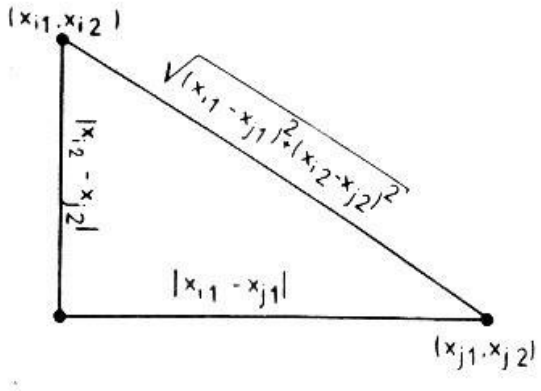


Figura 2 - Ilustração da distância Euclideana

A distância euclideana pode ser influenciada pela falta de valores na matriz inicial de dados, como foi visto no Capítulo 1. Por isso, quando é necessário fazer face a este problema, usamos a **média da distância euclideana**, definida por:

$$e_{ij} = \left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 / n \right)^{\frac{1}{2}}$$

como medida de dissemelhança.

Note-se que **a forma vectorial da distância euclideana** é definida por:

$$d_{ij} = \left[\left(x_i - x_j \right)^T \left(x_i - x_j \right) \right]^{\frac{1}{2}}$$

A distância euclideana não deverá ser usada quando as variáveis são medidas em unidades diferentes, quando são correlacionadas ou quando têm variâncias muito diferentes porque nestas condições as variáveis intervêm com pesos muito diferentes na determinação dos valores da dissemelhança.
Além disso, esta dissemelhança é sensível a mudanças de escala. Mudando a escala, não só muda os valores da distância como também as ordens destes valores e consequentemente os resultados da análise de clusters.

Para ultrapassar os inconvenientes da distância euclideana atrás mencionados, usam-se as seguintes medidas de dissemelhança dela derivadas:

- **Distância Euclideana ponderada ou estandardizada**

Sendo w_j o peso associado à variável j, a distância euclideana ponderada define-se por:

$$d_{rs} = \left[\sum_{j=1}^{p} w_j \left(x_{rj} - x_{sj}\right)^2\right]^{\frac{1}{2}} = \left[w_1\left(x_{i1} - x_{j1}\right)^2 + w_2\left(x_{i2} - x_{j2}\right)^2 + ... + w_p\left(x_{ip} - x_{jp}\right)^2\right]^{\frac{1}{2}} =$$

$$= \left[\sum_{j=1}^{p} \frac{1}{s_j{}^2}\left(x_{rj} - x_{sj}\right)^2\right]^{\frac{1}{2}} = \left[\left(x_r - x_s\right)^T D^{-1}\left(x_r - x_s\right)\right]^{\frac{1}{2}} = \left[\left(x_r - x_s\right)^T \Delta^{-1}\left(x_r - x_s\right)\right]^{\frac{1}{2}}$$

em que cada variável tem um peso de acordo com a sua importância. Δ^{-1} é a matriz dos pesos; $s_j{}^2$, $j = 1,2,...,p$, representa a variância da variável x_j através dos n objectos e D é a matriz diagonal das variâncias $D = diag(s_1{}^2, s_2{}^2, ..., s_p{}^2)$.
Os pesos associados às variáveis são, aqui, dados pelo inverso da respectiva variância.

Este processo consegue eliminar a dependência dos resultados da Análise de Clusters das unidades de medição.

- **Distância Estatística ou de Mahalanobis**

Define-se por:

$$d_{rs}{}^2 = \left(x_r - x_s\right)^T S^{-1}\left(x_r - x_s\right)$$

em que $S^{-1} = \left[\text{diag}\left(s_1^2, s_2^2, ..., s_p^2\right)\right]^{-1}$, em que S é a estimativa da matriz de covariâncias das p variáveis em estudo.

A distância de Mahalanobis além de reduzir a dependência das unidades de medição, reduz também a correlação entre variáveis.

3.2.2.2 Distância City Block ou Manhattan

Esta medida mede a distância numa configuração rectilínea, numa configuração de cidade (os quarteirões). Dados dois objectos i e j, a distância de City Block define-se por:

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| = \left| x_{i1} - x_{j1} \right| + \left| x_{i2} - x_{j2} \right| + ... + \left| x_{ip} - x_{jp} \right|$$

Esta distância é menos afectada por outliers do que a distância euclideana e de mais fácil interpretação.
Satisfaz todas as propriedades de uma semi-distância já referidas anteriormente.

3.2.2.3 Dissemelhanças usando distâncias de Minkowski

As dissemelhanças usando métricas[12] de Minkowski são uma generalização da distância Euclideana e da distância de Manhattan. A família de métricas de Minkowski é dada pela fórmula geral

$$d_{ij} = \left[\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{r} \right]^{\frac{1}{r}} = \left[\left| x_{i1} - x_{j1} \right|^{r} + \left| x_{i2} - x_{j2} \right|^{r} + ... + \left| x_{ip} - x_{jp} \right|^{r} \right]^{\frac{1}{r}}, \forall r \in IR : r \geq 1 .$$

Variando r, obtém-se uma infinidade de dissemelhanças. Para $r = 1$, obtém-se a distância City Block, conhecida por métrica do quarteirão. Por exemplo para $r = 2$ obtém-se a distância Euclideana.

[12] Funções dos objectos cujos valores satisfazem as propriedades métricas apresentadas em 1.5, da página 28.

Quando r tende para infinito, tem-se a **métrica de Chebychev, métrica** L_∞ **ou métrica do supremo.** É definida por:

$$\lim_{r\to\infty}\left[\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^{r}\right]^{\frac{1}{r}}=\sup_{k=1,\ldots,p}\left|x_{ik}-x_{jk}\right|$$

A fim de reduzir o efeito resultante das diferenças nas escalas de medição e o efeito de correlação entre características dos objectos aplica-se a **métrica de Minkowski ponderada,** que é definida por**:**

$$d_{rs}=\left[\sum_{j=1}^{p}w_k\left|x_{rj}-x_{sj}\right|^{r}\right]^{\frac{1}{r}}$$

3.2.2.4 Distância de Canberra

Define-se por:

$$d_{ij}=\sum_{k=1}^{p}\frac{\left|x_{ik}-x_{jk}\right|}{\left|x_{ik}\right|+\left|x_{jk}\right|}$$

Este coeficiente de dissemelhança varia no intervalo $[0,1]$ e $d_{ij}=0$ indica a máxima semelhança que ocorre apenas quando os objectos i e j são idênticos.

3.2.2.5 Coeficiente de correlação de Pearson

Dados dois objectos i e j , este coeficiente define-se por

$$r_{ij} = \frac{\sum_{k=1}^{p}\left(x_{ik} - \bar{x}_{i.}\right)\left(x_{jk} - \bar{x}_{j.}\right)}{\left[\sum_{k=1}^{p}\left(x_{ik} - \bar{x}_{i.}\right)^2 \sum_{k=1}^{p}\left(x_{jk} - \bar{x}_{j.}\right)^2\right]^{\frac{1}{2}}}, \text{ onde } \bar{\chi}_s = \frac{\sum_{k=1}^{p} x_{sk}}{p}$$

Este coeficiente de semelhança varia no intervalo $[-1,1]$, em que $r_{ij} = 1$ indica a semelhança máxima, mas não necessariamente identidade entre as características dos objectos i e j. Se $r_{ij} = -1$ indica o máximo de dissemelhança.

3.2.2.6 Coeficiente de separação angular (ou cosseno)

Este coeficiente de semelhança entre os objectos i e j, é definido por:

$$c_{ij} = \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\left(\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{jk}^2\right)^{\frac{1}{2}}}$$

Este coeficiente de semelhança define-se no intervalo $[-1,1]$ em que

c_{ij} =1 indica que a semelhança é máxima (dissemelhança mínima)

c_{ij} =-1 indica que a semelhança é mínima (dissemelhança máxima),

c_{ij} o cosseno do ângulo formado pelas duas semi-rectas que unem a origem aos respectivos objectos, representados como pontos no espaço.

Tanto o coeficiente de correlação de Pearson como o coeficiente de separação angular (ou cosseno) podem ser usados para quantificar semelhanças entre os objectos observados i e j , num espaço de dimensão p .

3.2.2.7 Coeficiente de Bray-Curtis

Define-se por:

$$d_{ij} = \frac{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|}{\sum_{k=1}^{p} \left(x_{ik} + x_{jk} \right)}$$

Este coeficiente de dissemelhança varia no intervalo $[0,1]$, em que $d_{ij}=0$ indica a máxima semelhança, que ocorre apenas quando os objectos i e j são idênticos.

3.2.3 Outras dissemelhanças

Para além dos coeficientes de dissemelhança apresentados na tabela 3, podem ser utilizados os seguintes coeficientes:

- **Coeficiente de Sokal e Sneath**

  Define-se por:

$$d_{ij} = \left[\frac{1}{p} \sum_{k=1}^{p} \left(\frac{x_{ik} - x_{jk}}{x_{ik} + x_{jk}} \right)^2 \right]^{\frac{1}{2}}$$

- **Coeficiente de Soergel**

  Define-se por:

$$d_{ij} = \frac{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|}{\sum_{k=1}^{p} \max(x_{ik}, x_{jk})}$$

- **Métrica de Gower**

  Define-se por:

$d_{ij} = \sum_{k=1}^{p} \frac{\left| x_{ik} - x_{jk} \right|}{r_k}$, em que r_k é a amplitude do intervalo de variação da k-ésima variável.

3.2.4 Variáveis de diferentes tipos

As estratégias para aplicar a uma matriz de dados que contém tanto variáveis quantitativas como qualitativas, são as seguintes:

Ignorar a diferença nas variáveis qualitativas, tratá-las como se fossem quantitativas e usar um coeficiente de semelhança (estratégia de Romesburg, como veremos na página seguinte).

Colocar os diferentes tipos de variáveis em diferentes matrizes, aplicar a Análise de Clusters a cada uma delas e tentar chegar a um consenso pelo estudo das duas árvores.

Categorizar as variáveis quantitativas, tornando-as qualitativas e, depois, usar um coeficiente de semelhança.

- Usar o coeficiente geral de Gower, podendo ter variáveis quantitativas e qualitativas.

- Usar uma matriz de semelhança combinada: atribuir um coeficiente de semelhança em separado consoante o tipo de variáveis e então combinar os dois coeficientes de semelhança usando pesos apropriados. Aplicar, então a Análise de Clusters.

Seguidamente apresentamos estas estratégias de forma mais detalhada.

3.2.4.1 Estratégia de Romesburg

A Estratégia de Romesburg, em 1984, era esquecer a natureza das variáveis de tipo misto e considerá-las todas de tipo quantitativo, depois usar um coeficiente para variáveis quantitativas, por exemplo distância euclideana.
Um grande inconveniente deste método é a interpretação dos coeficientes de semelhança depender da codificação previamente atribuída às variáveis qualitativas.

3.2.4.2 Realizar análises separadas

A construção de uma medida de semelhança para cada grupo de variáveis, variáveis categorizadas e variáveis contínuas, efectuando análises de clusters em separado, é uma abordagem ao caso de termos em estudo variáveis de diferentes tipos.

Se os resultados revelarem concordância significa que a estratégia atrás referida pode ser adoptada, mas se se verificar o contrário devemos processar os dados conjuntamente para realizarmos uma única Análise de Clusters.

3.2.4.3 Reduzir todas as variáveis quantitativas a variáveis categorizadas

Vejamos o caso da redução das variáveis quantitativas a variáveis binárias.
Temos que dividir o domínio de cada variável em dois blocos e aplicar a seguinte regra

se $x_{ij} < c_j, então \quad y_{ij} = 0$

se $x_{ij} \geq c_j, então \quad y_{ij} = 1$

em que x_{ij} é o valor original da variável j no objecto i, c_j é o valor crítico que divide o domínio da variável j em dois e y_{ij} é o valor que a variável binária criada assume no objecto i.
A desvantagem deste procedimento é a perda de informação resultante da redução de dados completos a dados binários.
Para o caso de estabelecermos mais categorias, o raciocínio é análogo.

3.2.4.4 Construir um coeficiente de semelhança combinado

Calcular coeficientes de semelhança para cada grupo de variáveis do mesmo tipo e usá-los de forma combinada para construir um único coeficiente de semelhança, usando pesos apropriados.
O coeficiente de semelhança combinado para os objectos i e j, é:

$$s_{ij} = w_1 s_{ij}^q + w_2 s_{ij}^n + w_3 s_{ij}^0$$

onde s_{ij}^q, s_{ij}^n e s_{ij}^0 são os coeficientes de semelhança calculados para as variáveis quantitativas , nominais e ordinais, e w_k, $k = 1,2,3$ são os pesos associados.

Bassab et al, (1990), trabalharam este procedimento para construir uma matriz de semelhanças combinada.

Gower,1971 apresentou o coeficiente de semelhança combinado, por $s_{ij} = \frac{\sum_{k=1}^{p} w_{ij} s_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$, em que s_{ijk} é a semelhança entre os objectos i e j com base na variável k.

Normalmente w_{ijk}, na variável k, toma valor 1 se a comparação entre os valores i e j é válida e toma valor 0 se a comparação entre os valores i e j não é válida, isto é se o valor da variável k é omisso em pelo menos um dos objectos i e j.

Quando as variáveis são binárias ou do tipo nominal com mais de dois níveis, os coeficientes s_{ijk} tomam o valor 1 se os dois objectos têm o mesmo valor na variável k e tomam valor zero, caso contrário.

Para variáveis contínuas, Gower (1971), sugere o uso do coeficiente de semelhança, com base na métrica de City-Block, para a variável k.

$s_{ij} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$, em que R_k é a amplitude do intervalo de variação dos valores da variável k.

3.3 Medidas de proximidade entre variáveis

Para agruparmos variáveis basta fazermos a transposta da matriz de dados $X_{n\times p}$ e efectuar a análise de clusters sobre as linhas de $X_{p\times n}$. As variáveis tomam

o lugar dos objectos e podemos aplicar as medidas de proximidade usadas na análise de objectos.
No entanto, geralmente as medidas de proximidade mais adequadas, para variáveis, são medidas de correlação e associação.

3.3.1 Variáveis quantitativas

Quando as variáveis são quantitativas podemos aplicar um dos seguintes coeficientes para agrupar as variáveis:

Coeficiente de separação angular ou cosseno

Define-se por

$s_{ij} = \dfrac{\sum_{k=1}^{n} x_{ki} x_{kj}}{\left(\sum_{k=1}^{n} x_{ki}^2 \sum_{k=1}^{n} x_{kj}^2 \right)^{\frac{1}{2}}} = \cos\alpha$ em que α representa o ângulo entre os vectores representativos das variáveis i e j, $(x_{1i},...,x_{ni})^T$ e $(x_{1j},...,x_{nj})^T$.

Coeficiente de correlação de Pearson

Define - se por :

$$r_{ij} = \frac{\sum_{k=1}^{n} \left(x_{ki} - \bar{x}_{.i} \right) \left(x_{kj} - \bar{x}_{.j} \right)}{\left(\sum_{k=1}^{n} \left(x_{ki} - \bar{x}_{.i} \right)^2 \sum_{k=1}^{n} \left(x_{kj} - \bar{x}_{.j} \right)^2 \right)^{\frac{1}{2}}}$$, em que $\overline{x}_{s.} = \dfrac{\sum_{k=1}^{p} x_{sk}}{p}$

Varia entre [-1,1] e não depende das unidades de medida. Este coeficiente é aplicado quando há uma relação linear entre as variáveis i e j.

3.3.2 Variáveis qualitativas

Quando as variáveis são qualitativas podemos aplicar um dos seguintes coeficientes:

3.3.2.1 Variáveis nominais com dois níveis

Consideremos novamente a tabela 1 dicotómica, anteriormente apresentada, em que i e j representam a i-ésima e a j-ésima variáveis. Os coeficientes de separação angular e de correlação de Pearson anteriormente apresentados tomam a forma:

$$s_{ij} = \frac{a}{((a+b)(a+c))^{\frac{1}{2}}} = \cos\alpha$$

$$r_{ij} = \frac{ad - bc}{((a+b)(c+d)(a+c)(b+d))^{\frac{1}{2}}}$$

3.3.2.2 Variáveis nominais com mais de dois níveis

Consideremos a seguinte tabela dicotómica, em que g e h representam variáveis com r e s categorias, respectivamente (s e r têm de ser maiores do que 2).

A classificação dos n objectos é normalmente representada numa tabela de contingência, em que n_{ij}, $n_{i.}$, $n_{j.}$ representam as frequências absolutas; f_{ij}, $f_{i.}$, $f_{.j}$ são frequências relativas;

n_{ij} é o número de objectos que surgem no nível i da variável g e no nível j da variável h;

$n_{i.}$ é o número de objectos que surgem no nível i da variável g (para todos os níveis da variável h);

$n_{.j}$ é o número de objectos que surgem no nível j da variável h (para todos os níveis da variável g);

Tabela 4 - Tabela de contingência

		Variável h				
		1	2	…j...	s	
	1	.	.	…….	.	.
	2	.	.	…….	.	.
	.	.	.	…….	.	.
Variável g	i	.	.	$n_{ij}(f_{ij})$	.	$n_{i.}(f_{i.})$
	.	.	.	…….	.	.
	r	.	.	…….	.	.
				$n_{.j}(f_{ij.})$		$n(1)$

Seguidamente são apresentadas várias medidas para variáveis nominais e variáveis ordinais apresentadas por Agresti (1981). As mais usadas são a do Qui-Quadrado de Pearson e as derivadas desta.

A medida do **Qui – Quadrado** é definida por:

$$X^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

A partir desta medida, podem obter-se outras medidas, como por exemplo:

- **Coeficiente de contingência quadrático**

$$\Phi^2 = \frac{X^2}{n}$$

- **Coeficiente de contingência de Pearson**

$$P = \left[\frac{\Phi^2}{1+\Phi^2} \right]^{\frac{1}{2}}$$

- **Coeficiente de Tschuprow**

$$T = \left[\frac{\Phi^2}{(r-1)(s-1)} \right]^{\frac{1}{2}}$$

- **Coeficiente de Cramer**

$$C = \left[\frac{\Phi^2}{\min(r-1, s-1)} \right]^{\frac{1}{2}}$$

3.3.3 Variáveis ordinais

Existem várias medidas de associação entre variáveis ordinais. Uma das mais usadas é a do **coeficiente de correlação ordinal de Spearman**, definida por :

$$r_{ij} = 1 - \frac{6\sum_{k=1}^{n} d_k{}^2}{n(n^2-1)}$$

sendo d_k a diferença entre as ordens dos valores que o objecto k assume nas duas variáveis i e j; r_{ij} é o coeficiente de correlação de Pearson entre as ordens dos valores assumidos por cada uma das variáveis i e j.

O coeficiente de correlação de Spearman toma valores no intervalo [-1,1] e não depende das unidades de medida. Este coeficiente é aplicado quando há uma relação monótona entre as variáveis i e j.

Outro coeficiente de correlação que se poderá usar quando as variáveis estão numa escala ordinal, é o coeficiente de correlação de Kendall tau.

O coeficiente de Kendall tau, $\tau_{jk} = \frac{S_{jk}}{0{,}5p(p-1)}$, $-1 \le \tau_{jk} \le 1$, em que S_{jk} é o coeficiente de semelhança entre os objectos j e k; p é o número de variáveis.

Os coeficientes de correlação podem ser convertidos em dissemelhanças, $d_{ij} = \frac{1-r_{ij}}{2}$ ou nalguns casos $d_{ij} = 1 - |r_{ij}|$.

Capítulo 4

Métodos hierárquicos

4.1 Introdução

Os métodos de construção dos clusters podem ser hierárquicos ou não hierárquicos.

A selecção dos métodos depende do objectivo da investigação e das propriedades dos vários métodos. O que recomendam muitos autores é utilizar vários métodos, comparar os resultados, verificar se são consistentes e seleccionar o que tiver interpretação mais fácil. Os vários métodos e algoritmos poderão revelar vários aspectos a estudar, diferentes aspectos da estrutura dos dados.

Nos métodos hierárquicos, os grupos formam uma hierarquia porque dados dois grupos quaisquer, os grupos ou são disjuntos ou um deles está contido no outro.

Nestes métodos, não é requerido um conhecimento prévio do número de clusters, no entanto, sempre que um objecto é atribuído a um cluster, não pode sair mais desse cluster.

Os métodos hierárquicos podem ser divididos em dois tipos de métodos: os métodos aglomerativos ou ascendentes e os métodos divisivos ou descendentes. Ambos os tipos de métodos procuram o conjunto óptimo de clusters, operando em cada etapa na subdivisão ou fusão progessiva dos dados com base na matriz de proximidades (semelhanças ou dissemelhanças entre casos).

Nestes métodos as divisões ou fusões, uma vez feitas, são irreversíveis, isto é, os algoritmos aglomerativos que juntam dois objectos não podem depois separá-los e os algoritmos divisivos que separam dois objectos não os podem depois unir. Kaufman e Rousseuw, (1990), defenderam que os métodos hierárquicos têm o defeito de nunca poderem reparar o que foi feito em conjuntos prévios.

Nos algoritmos aglomerativos parte-se de n grupos com um objecto cada. Estes vão sendo agrupados sucessivamente até se encontrar um grupo que inclua a totalidade dos objectos. Estes métodos procedem à **fusão** sucessiva de grupos.

Nos algoritmos divisivos parte-se de um grupo que inclui todos os objectos em estudo e por um processo de divisões sucessivas obtém-se n grupos de um elemento cada. Nestes métodos procede-se à **divisão** sucessiva de grupos.

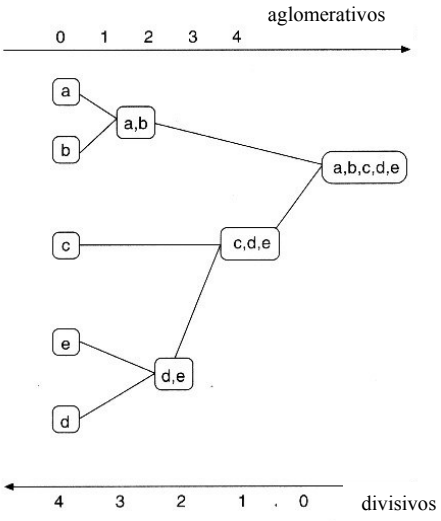


Figura 3 - exemplo da estrutura de uma árvore hierárquica (retirado de [Kaufman e Rousseeuw, 1990]).

Dois objectos estão próximos quando a sua dissemelhança ou distância é pequena ou a sua semelhança é grande.

Os métodos de análise de clusters mais utilizados são os hierárquicos aglomerativos. São normalmente usados em amostras pequenas (com n<250)

A representação hierárquica de um conjunto de dados é usualmente feita através de um dendograma, também conhecido por diagrama de árvore, árvore hierárquica ou fenograma.

Este tipo de representação hierárquica foi desenvolvida primeiramente em Biologia, aparecendo em vários formatos gráficos consoante o *software* que os produz. É um gráfico bidimensional que ilustra as fusões ou divisões em cada nível da análise de clusters.

Quando o dendograma configura uma árvore invertida com raiz para cima e ramos para baixo, os nós internos representam clusters e a altura dos troncos indicam a distância a que os clusters se ligam. As alturas pequenas significam que a aglutinação é feita entre clusters razoavelmente homogéneos.

Quando o dendograma está numa posição horizontal, a árvore fica com os ramos à esquerda e a raíz a direita, lembrando a classificação em famílias, espécies e subespécies usadas em zoologia e botânica.

Os dendogramas mostram como os sucessivos grupos se vão formando ao longo do processo hierárquico. A ordem dos objectos é arbitrária, o que faz com que o esquema do dendograma não seja rígido mas sim um esquema móvel em torno dos eixos de ligação.

4.2 Métodos aglomerativos

São provavelmente os métodos hierárquicos mais usados. Pressupõem uma série de partições sucessivas: começam com n clusters (cada cluster é constituído por um objecto) e terminam com um cluster (com n objectos).

De acordo com [Branco J., 2004] e com [Jonhson e Wichern, 2002], os procedimentos de um método aglomerativo podem representar-se essencialmente em quatro passos:

1º Passo – Considerar os n objectos iniciais como n *clusters* (grupos) singulares e calcular matriz de proximidade, por exemplo a matriz de dissemelhança, entre o objecto i e objecto j, $i = 1,...,n$ e $j = 1,...,n$; $D_{nxn} = [d_{ij}]$

2º Passo – Procurar na matriz D os pares de objectos i e j (ou de variáveis) mais semelhantes (com menor dissemelhança d_{ij}). Se houver vários valores iguais optamos pelo objecto que tiver menor valor alfanumérico. Se tivermos em conta que cada objecto, inicialmente, constitui um grupo, poderemos descrever este procedimento da seguinte maneira: identificar os dois grupos mais semelhantes, por exemplo A e B; a sua dissemelhança representa-se por d_{AB};

3º Passo – Unir os pontos A e B à distância crítica - distância entre os dois grupos, d_{AB}. Actualizar a matriz D, eliminando as linhas e as colunas correspondentes aos grupos A e B e introduzindo uma nova linha e coluna com as dissemelhanças calculadas entre o novo grupo (AB) e cada um dos restantes grupos. Com esta operação a ordem da matriz baixa uma unidade;

4º Passo – Repetir os passos 2 e 3 num total de n-1 vezes até obter um único grupo que, desta maneira, incluirá todos os objectos.

A questão que se levanta é como definir a distância entre dois grupos. Existem várias propostas e cada uma delas proporciona um método hierárquico aglomerativo diferente.

Seguidamente serão apresentados os **métodos hierárquicos aglomerativos mais utilizados,** de acordo com [Anderberg, 1973]:

- Ligação simples ou critério do vizinho mais próximo ou Menor distância.
- Ligação completa ou critério do vizinho mais afastado ou Maior distância.
- Ligação média ou UPGMA (unweighted pair-group method using the average approach) ou distância média entre clusters.
- Método do Centroide ou UPGMC (*unweighted pair-group method using the centroid approach*).
- Ligação mediana ou WPGMC (*weighted pair-group method using the centrid approach*) ou Distância Mediana.
- Critério de Ward.
- Distância média dentro dos clusters.

4.2.1 Ligação simples ou critério do vizinho mais próximo

Este método usa como ponto de partida uma matriz de dissemelhanças. A distância entre os grupos A e B é definida pela distância mais próxima entre dois grupos. Pode ser definida como $d_{AB} = \min \left\{ d_{ij} : i \in A, j \in B \right\}$.

Inicialmente, considera-se uma matriz de proximidades D_1, considerando cada objecto como um grupo singular; procuram-se os dois objectos mais próximos, isto é, os que têm menor valor da matriz, por exemplo i e j.

Seguidamente, dados dois grupos (i, j) e (k), a distância entre dois grupos é a menor das distâncias entre os seus elementos, ou seja :

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

Este método é portanto, um sistema contractor do espaço.

É muito utilizado nas Ciências Sociais e muito popular na área da Taxonomia numérica [Sokal and Sneath, 1963].

A dissemelhança entre dois grupos é determinada pelos objectos mais próximos (os vizinhos mais próximos). Cada vez que um objecto é adicionado a

um grupo, as distâncias do novo grupo aos restantes ou tornam-se menores ou ficam inalterados.

Se houver duas dissemelhanças que sejam iguais e, além disso, que sejam menores que todas as outras, $d_{AB} = d_{CB}$, o resultado final não se altera qualquer que seja a opção de escolha por d_{AB} ou d_{CB}, isto é, quer se considere A e B como os vizinhos mais próximos ou quer se considere C e B como os vizinhos mais próximos.

Será dado um exemplo que servirá para ilustrar o procedimento geral dos métodos hierárquicos. Também poderão ser aplicados aos métodos divisivos, começando por um cluster com todos os objectos e depois separá-los em dois clusters cuja distância do vizinho mais próximo seja máxima, mas este método será apresentado no ponto 4.3 deste capítulo.

Exemplo 4.1 - Para ilustrar o funcionamento do algoritmo aglomerativo aplicado ao método de ligação simples, construíu-se a seguinte matriz de dissemelhanças, relativa a 5 objectos.

$$D_1 = [d_{ij}] = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 6 & 5 & 0 & & \\ 10 & 9 & 4 & 0 & \\ 9 & 8 & 5 & 3 & 0 \end{bmatrix}$$

Passo 1: Considera-se que cada objecto é um grupo singular e procuram-se os dois objectos mais próximos, identificados pelo elemento de menor valor da matriz D, isto é $d_{ij} = \min \{d_{ij} : i, j = 1,...,5\}$, obtém-se $d_{12} = 2$. Os dois objectos 1 e 2, fundem-se, ao nível crítico $d_{12} = 2$, para formar um novo grupo (12). Passa-se então à construção da nova matriz de dissemelhanças, D_2. Primeiro obtêm-se as dissemelhanças $d_{(12)i}$ entre o grupo (12) e os grupos singulares ou objectos,

i= 3,4,5

$$d_{(12)3} = \min\{d_{13}, d_{23}\} = d_{23} = 5$$
$$d_{(12)4} = \min\{d_{14}, d_{24}\} = d_{24} = 9$$
$$d_{(12)5} = \min\{d_{15}, d_{25}\} = d_{25} = 8$$

A nova matriz, D_2, obtém-se a partir de D_1, eliminando as linhas e as colunas correspondentes aos objectos 1 e 2 e acrescentando a nova linha e a nova coluna correspondente ao grupo (12).

$$D_2 = \begin{matrix} (12) \\ 3 \\ 4 \\ 5 \end{matrix} \begin{bmatrix} 0 & & & \\ 5 & 0 & & \\ 9 & 4 & 0 & \\ 8 & 5 & 3 & 0 \end{bmatrix}$$

Passo 2: O menor elemento da matriz D_2, para os objectos 4 e 5, $d_{45} = 3$, formará um novo conjunto de distâncias.

$d_{(12)(45)} = \min\{d_{(12)4}, d_{(12)5}\} = \min\{9,8\} = 8$, representa a distância entre os grupos (12) e o recém-formado (4,5).

$$d_{(45)3} = \min\{d_{34}, d_{35}\} = d_{34} = 4$$

A nova matriz de dissemelhanças, D_3, será então:

$$D_3 = \begin{matrix} (12) \\ 3 \\ (45) \end{matrix} \begin{bmatrix} 0 & & \\ 5 & 0 & \\ 8 & 4 & 0 \end{bmatrix}$$

Passo 3: O menor elemento da matriz D_3, é 4, correspondente à distância entre o objecto 3 e o cluster (45), isto é, o cluster que contém os objectos 4 e 5. Forma-se

então o cluster [3(45)] ou (345). Então a distância entre os grupos [3(45)] (ou (345)) e o grupo (12) é 5, que é $\min\{d_{3(12)}, d_{(45)(12)}\} = \min\{5,8\}$.

Seguidamente representaremos o dendograma correspondente a este processo e as partições produzidas em cada fase. A altura, neste diagrama, representa a distância a que cada fusão é feita.

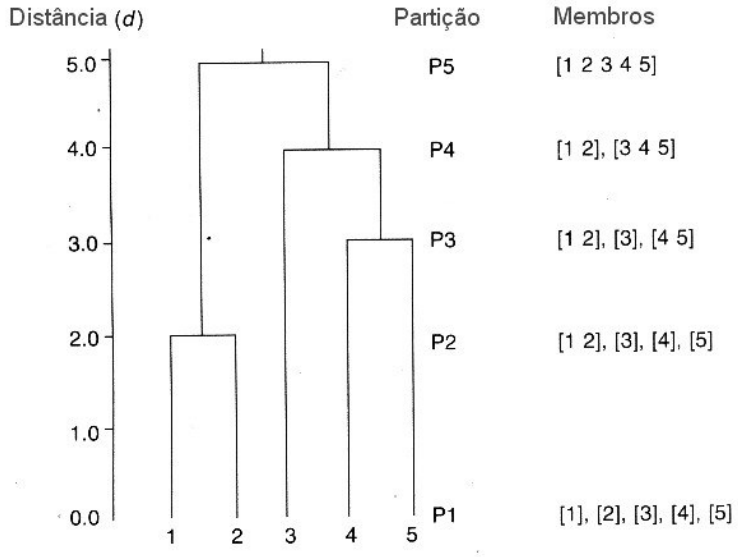


Figura 4 - Dendograma referente ao exemplo dado anteriormente usando o Método de Ligação Simples.

4.2.2 Ligação completa ou critério do vizinho mais afastado

Este método também usa como ponto de partida uma matriz de dissemelhanças. A distância entre dois grupos é definida como sendo a distância entre os seus elementos mais afastados ou menos semelhantes.

Dados dois grupos (ij) e (k), a distância entre eles é a maior distância entre os seus elementos $d_{AB} = \max \left\{ d_{ij} : i \in A, j \in B \right\}$

Nesta estratégia cada elemento de cada grupo é mais semelhante aos restantes elementos do grupo do que a qualquer elemento dos restantes grupos. Este método tem tendência para encontrar clusters compactos compostos de objectos muito semelhantes entre si.

Quando um objecto é acrescentado a um grupo, a distância do novo grupo aos restantes aumenta ou então fica inalterada. O método de ligação completa tende a formar grupos pequenos que depois serão aglutinados para formar grupos maiores.

4.2.3 Ligação média entre clusters

A ligação média entre clusters também é conhecida por UPGMA (unweighted pair-group method using the average approach) ou distância média entre clusters, usa como ponto de partida uma matriz de dissemelhanças. A distância entre dois grupos A e B, é defenida como sendo a média das distâncias entre todos os pares de objectos constituídos por um objecto de cada grupo, isto é,

$d_{AB} = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d_{ij}}{n_A n_B}$, sendo n_A e n_B iguais ao número de objectos do grupo A e do grupo B, respectivamente.

Esta estratégia é intermédia relativamente às duas descritas anteriormente. Enquanto que no vizinho mais próximo e no vizinho mais afastado a inclusão de um novo indivíduo num grupo depende de um único valor de dissemelhança, o menor ou o maior, respectivamente, no critério da distância média entre grupos, incluímos a média das dissemelhanças.

A grande vantagem deste critério é tornear as consequências da existência de valores extremos e considerar toda a informação dos grupos. Só que, dependendo do tipo de clusters que se espera obter, esta propriedade também pode ser vista como uma desvantagem.

De seguida é apresentada uma imagem que retrata os métodos de Ligação média, Ligação simples e Ligação completa.

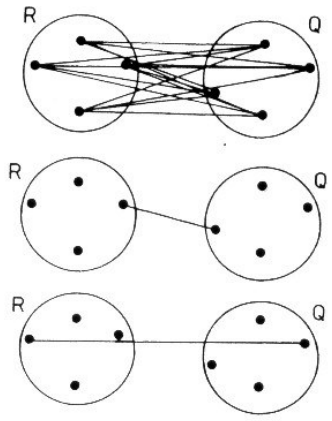


Figura 5 – Representação dos métodos: Ligação média, Ligação simples, Ligação completa, respectivamente (retirado de [Kaufman e Rousseeuw, 1990])

4.2.4 Método do Centróide

O método do centróide, também é conhecido por UPGMC (unweighted pair-group method using the centroid approach). Neste critério, a distância entre dois grupos A e B é definida como a distância euclideana entre os seus centróides (pontos definidos pelas médias das p variáveis para os indivíduos de cada grupo). Aliás, a representação das populações pelas suas médias ou centróides, é um procedimento muito comum em Estatística.

Este critério parte de uma matriz de proximidades, sendo a distância euclideana a mais usada neste critério.

Dados dois grupos A e B, a distância entre eles é igual à distância entre os seus centróides, isto é:

$d_{AB} = d\left(\bar{x}_A, \bar{x}_B\right)$ em que $\bar{x}_A$ e $\bar{x}_B$ são os centróides dos grupos A e B, respectivamente.

$$\bar{x}_A = \frac{\sum_{i \in A} x_i}{n_A} \text{ e } \bar{x}_B = \frac{\sum_{i \in B} x_i}{n_B}$$

x_i é o vector das p observações do objecto i

$\bar{x}_A$ é um vector formado pelas médias aritmétricas das p variáveis, calculadas para os n_A objectos pertencentes ao grupo A. O mesmo se pode dizer relativamente a $\bar{x}_B$ e aos n_B objectos pertencentes ao grupo B.

4.2.5 Distância Mediana

O método da distância mediana, também é conhecido por WPGMC (weighted pair-group method using the centroid approach). O critério da mediana é semelhante ao do centróide com a diferença de que ao aglutinar os dois grupos A e B, os seus centróides, $\overline{x_A}$ e $\overline{x_B}$, recebem pesos iguais antes de produzirem o centróide do novo cluster resultante da aglutinação. O novo centróide $\overline{x}$ fica a meio dos centróides dos grupos aglutinados,

$$\overline{x} = (\overline{x_A} + \overline{x_B})/2$$

com o objectivo de evitar que o grupo com maior número de objectos absorva o grupo com menor número de objectos.

A mediana referida não corresponde à mediana estatística mas sim à mediana de um triângulo, isto é um segmento de recta que une um vértice de um triângulo ao ponto médio do lado oposto.

4.2.6 Critério de Ward

Uma medida de proximidade alternativa, baseada na distância euclideana entre centróides, usa o facto de que existe um total de $n_A \times n_B$ distâncias entre os grupos A e B. Uma medida das distâncias totais entre os dois grupos é dada por $n_A n_B d_{AB}^2$, em que $d_{AB}^2 = \sum_{j=1}^{p} \left(\bar{x}_{jA} - \bar{x}_{jB} \right)^2$ é o quadrado da distância euclideana entre os dois centróides. Uma vez que existem $n_A + n_B$ objectos, uma média daqueles valores é dada por $\frac{n_A n_B d_{AB}^2}{n_A + n_B}$. Esta medida da distância média é equivalente à alteração na soma dos quadrados dos erros dentro do grupo, ou soma de quadrados dos erros incremental, que resulta da combinação dos grupos A e B.

Para o grupo A a soma de quadrados dos erros dentro do grupo é dada por:

$$ssw_A = \sum_{i=1}^{n_A} \sum_{j=1}^{p} \left(x_{ijA} - \bar{x}_{jA} \right)^2$$

De forma semelhante, para o grupo B, a soma de quadrados dos erros dentro do grupo é dada por:

$$ssw_B = \sum_{i=1}^{n_B} \sum_{j=1}^{p} \left(x_{ijB} - \bar{x}_{jB} \right)^2$$

Se os grupos A e B forem combinados para formar um novo grupo C, obteremos um novo centróide $\left(\bar{x}_{1C}, \bar{x}_{2C}, ..., \bar{x}_{pC}\right)$ e a soma dos quadrados dos erros dentro do grupo C é dada por:

$$ssw_C = \sum_{i=1}^{n_A+n_B} \sum_{j=1}^{p} \left(x_{ijC} - \bar{x}_{jC}\right)^2$$

O aumento na soma total de quadrados dos erros dentro do grupo resultante da união dos grupos A e B é dada por:

$ssw_C - \left(ssw_A + ssw_B\right)$.

Este aumento na soma total de quadrados dos erros dentro do grupo é equivalente à distância total média $\frac{n_A n_B d_{AB}^2}{n_A + n_B}$.

Este incremento da soma de quadrados dos erros, é geralmente usado como uma medida de proximidade entre grupos. O objectivo é minimizar este incremento quando dois grupos são unidos.

Na tabela seguinte será apresentada uma comparação dos seis métodos típicos.

Tabela 5 - Métodos hierárquicos aglomerativos mais utilizadas (baseada em [Everitt B. e al, 2001]).

Método	Nome alternativo*	Utilizado normalmente com:	Distância entre clusters definida como:	Notas
Ligação simples Sneath (1957)	Vizinho mais próximo	Similaridade ou distância	Distância mínima entre pares de objectos, um num cluster, outro no outro cluster	Tende a produzir clusters não balanceados e irregulares (em corrente), especialmente em conjuntos grandes de dados. Não tem em conta a estrutura do cluster.

Ligação completa Scorensen (1948)	Vizinho mais afastado	Similaridade ou distância	Distância máxima entre pares de objectos, um num cluster, outro no outro cluster	Tende a encontrar clusters compactos com diâmetros iguais (distância máxima entre objectos). Não tem em conta a estrutura do cluster.
Ligação média (grupo) Sokal e Michener (1958)	UPGMA	Similaridade ou distância	Média das distâncias entre pares de objectos, um num cluster, outro no outro cluster	Tende a juntar clusters com variâncias pequenas. Intermédio entre ligações simples e complexas. Tem em conta a estrutura do cluster. Relativamente robusto.
Ligação centróide Sokal e Michener (1958)	UPGMC	Distância (requer dados brutos)	Quadrado da distância euclideana entre vectores médios (centróides)	Assume que os pontos podem ser representados no espaço euclideano (para interpretação geométrica). O cluster mais numeroso dos dois grupos unidos num cluster domina o cluster fundido.
Ligação mediana Gower (1967b)	WPGMC	Distância (requer dados brutos)	Quadrado da distância euclideana entre centróides ponderados.	Assume que os pontos podem ser representados no espaço euclideano para interpretação geométrica. O novo grupo fica intermédio entre grupos fundidos.
Método de Ward Ward (1963)	Soma mínima de quadrados	Distância (requer dados brutos)	Aumento na soma dos quadrados dos erros dentro dos clusters, após	Assume que os pontos podem ser representados no espaço

			fusão, somada sobre todas as variáveis	Euclideano para interpretação geométrica. Tende a encontrar clusters esféricos com o mesmo tamanho; sensível a outliers.

*U, unweighted (não ponderado / pesado); W, weigthed (ponderado / pesado); PG, grupo par; A, average (média); C, centróide

4.2.7 Conclusão

O método de ligação completa (ou critério do vizinho mais afastado), é menos sensível a erros de observação do que o Método de ligação simples (ou critério do vizinho mais próximo), embora ambos os métodos sejam muito susceptíveis a observações extremas. No método de ligação completa, pequenas alterações na localização de certos pontos ou erros podem ter um impacto substancial na solução hierárquica.

A ligação simples depende das distâncias mais pequenas, por isso é preciso fazer as medições com menor erro para que este método tenha sucesso [Hartigan, 1975]. Um único outlier que se situe entre dois clusters pode resultar numa eventual fusão dos dois grupos. É fácil de aplicar a um grande conjunto de dados, é fácil de programar embora seja menos satisfatório que os restantes métodos, atendendo ao que já foi dito.

O Método da ligação média, o método do centróide e o método de Ward são usualmente preferíveis ao método de ligação simples e ao método de ligação completa devido à sua insensibilidade a valores extremos ou outliers. Como vimos, isto às vezes pode tornar-se uma desvantagem, dependendo do tipo de clusters que esperamos obter.

Gordon (1998), concluíu que não há nenhum método que possa ser recomendado em detrimento dos outros.

O método ou os métodos a usar deverão depender do tipo de dados, do contexto do estudo que está a ser feito, do conhecimento e da experiência do investigador.

Numa análise prática se aplicarmos os seis métodos anteriormente abordados: centróide, mediana, Ward, ligação simples, ligação completa e ligação média e os representarmos em dendogramas para compararmos os resultados, isto é , compararmos o número de clusters, a sua composição e o nível de fusão dos clusters, podemos verificar se os dados possuem ou não uma estrutura de grupos unanime.

4.2.8 Fórmula de Recorrência de Lance Williams

Lance e Williams, (1967), introduziram uma regra geral para definir dissemelhança, $d_{C(AB)}$, entre o grupo C e o grupo (AB) (formado pela fusão dos grupos A e B), como

$$d_{C(AB)} = \alpha_A d_{CA} + \alpha_B d_{CB} + \beta d_{AB} + \gamma \left| d_{CA} - d_{CB} \right|$$

em que α_A , α_B , β e γ são parâmetros que ou são constantes ou dependem do número de objectos em cada grupo, $n_A, n_B \;\; e \;\; n_C$. d_{AB} é a dissemelhança entre os grupos A e B.

Lance e Williams analisaram resultados e concluíram que uma escolha adequada para um grande número de situações seria

$$\alpha_A + \alpha_B + \beta = 1, \;\; \alpha_A = \alpha_B, \;\; \beta < 0, \;\; \gamma = 0 .$$

É, portanto, uma estratégia flexível.

Por variação do β podem ser obtidos vários esquemas de clusters com várias características:

- Quando $\beta = 1$, obtém-se o método de ligação simples.
- Quando β **decresce para zero** ou se torna negativo, os grupos tornam-se mais homogéneos.

Lance e Williams sugeriram valores negativos para β, tal como -0,25. Contudo, por volta de 1985, Scheibler e Schnieder sugeriram -0,50.

As medidas de distância inter-grupos usadas por algumas técnicas hierárquicas usuais de construção de clustes, são obtidas pela escolha adequada dos parâmetros $\alpha_i, \alpha_j, \beta \ e \ \gamma$, apresentados na tabela seguinte com os valores dos parâmetros para os diferentes métodos aglomerativos.

Tabela 6 - Parâmetros de Lance-Williams para vários métodos hierárquicos aglomerativos

	Parâmetros de Lance-Williams			
Método	α_A	α_β	β	γ
Ligação simples	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Ligação completa	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Ligação média	$\frac{n_A}{n_A + n_B}$	$\frac{n_B}{n_A + n_B}$	0	0
Centróide	$\frac{n_A}{n_A + n_B}$	$\frac{n_B}{n_A + n_B}$	$-\frac{n_A n_B}{(n_A + n_B)^2}$	0
Mediana	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{n_C + n_A}{n_C + n_A + n_B}$	$\frac{n_C + n_B}{n_C + n_A + n_B}$	$-\frac{n_C}{n_C + n_A + n_B}$	0

Lance-Williams	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	<1	0

Por exemplo, o critério de ligação simples corresponde aos parâmetros $\alpha_i = \alpha_j = \frac{1}{2}; \beta = 0 \;\; e \;\; \gamma = -\frac{1}{2}$; uma vez que estas condições equivalem a $d_{C(AB)} = \min\{d_{ij} : i \in C, j \in (AB)\}$.

4.3 Métodos divisivos

Os métodos divisivos operam no sentido contrário aos métodos aglomerativos: começam com um único cluster, contendo todos os objectos. Esse grupo é dividido em dois subgrupos distintos com base em algum critério de dissemelhança.

Continua-se com divisões sucessivas, repetindo o processo descrito anteriormente até obtermos n clusters (n é o número total de objectos).

É computacionalmente exigido que em cada passo sejam calculadas $2^k - 1$ dissemelhanças correspondentes à divisão dos k objectos em dois grupos distintos. No primeiro passo são $2^n - 1$ dissemelhanças, uma vez que inicialmente existe um único cluster com os n objectos.

A nível gráfico, move-se da raiz do dendograma para os ramos, ao contrário do que acontece nos métodos aglomerativos, que se movem dos ramos para a raiz.

Os métodos divisivos têm a vantagem de ao fim dos primeiros passos do processo fornecerem grandes grupos, informando, assim, sobre a estrutura principal dos dados, o que geralmente interessa ao investigador. No entanto, são menos utilizados que os métodos aglomerativos.

4.3.1 Métodos divisivos Monotéticos (com uma variável)

Nos métodos divisivos monotéticos todas as variáveis são binárias. A escolha da variável na qual a sub-divisão é feita, depende do critério de optimização que reflecte quer a homogeneidade do cluster quer a associação com outras variáveis. Isto faz com que diminua o número de quebras que deverão ser feitas.
Lance e Williams, definiram um exemplo de um critério de homogeneidade por:

$$C = pn\log n - \sum_{k=1}^{p}\left[f_k \log f_k - (n - f_k)\log(n - f_k)\right], \text{ em que:}$$

- C representa o caos,
- n é o numero de objectos,
- p é o número de variáveis e
- f_k o número de objectos que têm o k–ésimo atributo.

Se um grupo X for dividido em dois sub grupos A e B, a redução em C é $C_X - C_A - C_B$, em que $C_X, C_A \text{ e } C_B$ são o valor de C calculado para os objectos pertencentes a cada um dos grupos $A, B \text{ e } X$, respectivamente.

O conjunto ideal de clusters terá membros (objectos ou variáveis, conforme a análise que está a ser feita) com características idênticas e C igual a zero. Contudo os clusters são divididos em cada fase de acordo com a presença do atributo que leva a uma maior redução em C.

Em vez da homogeneidade do cluster, o atributo pode ser escolhido com as suas associações a outros dos atributos que restam em cada conjunto; é a chamada análise de associação, especialmente utilizada em Ecologia [Williams and Lambert, 1959]. Por exemplo para um par de variáveis $v_i \;\; e \;\; v_j$, com valores 0 e 1, as frequências podem ser a, b, c, d, de acordo com a tabela seguinte.

Tabela 7 - Tabela de contingência

v_j \ v_i	1	0
1	a	b
0	c	d

As medidas de associação mais comuns são as seguintes:

1. $|ad-bc|$
2. $(ad-bc)^2$
3. $(ad-bc)^2 n/[(a+b)(a+c)(b+d)(c+d)]$
4. $\sqrt{(ad-bc)^2 n/[(a+b)(a+c)(b+d)(c+d)]}$
5. $(ad-bc)^2/[(a+b)(a+c)(b+d)(c+d)]$

As duas primeiras medidas têm a vantagem de não haver problema a nível computacional se algum total dar zero. As últimas três estão relacionadas com o teste do Qui-Quadrado, a sua raiz quadrada e o coeficiente de correlação de Pearson, respectivamente.

Os métodos divisivos monotéticos, são métodos de classificação fácil de novos membros, e da inclusão de casos com falta de valores.

Se faltar um valor da variável v_1, é substituído pelo valor da variável v_2 para a mesma observação (associação positiva entre v_1 e v_2) ou $1-v_2$ (associação negativa).

Este método é normalmente utilizado em Medicina, em diagnósticos, [Payne e Preece, 1980] e nos estudos de Mortuário em Arqueologia [O'Shea, 1985], nos quais se argumenta que o estrato social pode ser deduzido a partir do conjunto de objectos que existem na sepultura.

4.3.2 Métodos divisivos Politéticos

Os métodos divisivos politéticos estão relacionados com os métodos aglomerativos, desde que usem todas as variáveis em simultâneo e que se possa trabalhar com uma matriz de proximidades.
MacNaughton-Smith et al, (1964), consideram todas as possibilidades de quebra, um problema dos métodos divisivos politéticos. Estes métodos detectam o objecto que dentro do grupo está mais longe dos restantes e utilizam isto para separar o grupo.

4.4 Aplicação dos métodos hierárquicos

Para que haja uma melhor aplicação dos métodos hierárquicos, tanto aglomerativos como divisivos, o utilizador necessita, muitas das vezes, de ter em conta os seguintes aspectos:

1 – Representação gráfica dos métodos de clusters
2 – Comparação de dendogramas
3 – Propriedades matemáticas dos métodos
4 – Escolha da partição
5 – Algoritmos computacionais

4.4.1 Dendogramas

O dendograma, é uma representação matemática pictural do procedimento de construção de clusters completo. O dendograma inclui uma árvore e o valor da medida de proximidade em cada passo do processo hierárquico.

Seguidamente é apresentada a forma básica de um dendograma com a respectiva terminologia.

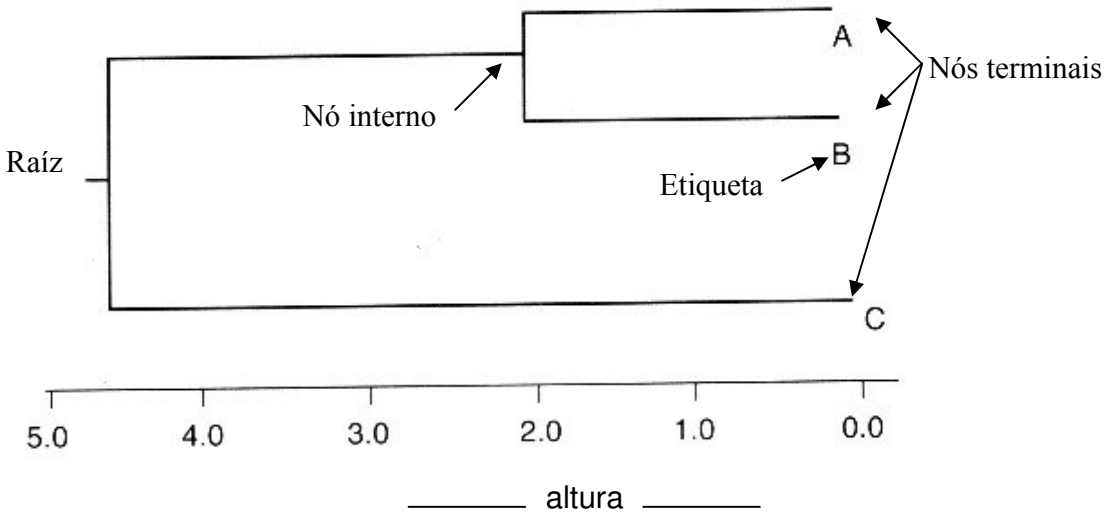


Figura 6 – Forma de um dendograma e alguma terminologia associada.

Os nós do dendograma representam os clusters e o tamanho do tronco (altura) representa a distância a que os clusters se juntam. Dendogramas que não têm informação numérica junto aos troncos são designados por “diagramas sem peso”. A ligação entre nós e troncos dá-nos a disposição da árvore, a sua topologia.

Os nós terminais são usualmente etiquetados enquanto que os nós internos normalmente não são etiquetados.

Os membros típicos ou representativos dos clusters podem ser associados aos nós internos, chamados centrótipos e são definidos como os objectos que têm a semelhança média máxima dentro do cluster (ou mínima dissemelhança).

De acordo com Everitt (2001), se aplicarmos os mesmos procedimentos de construção de clusters ao mesmo conjunto de dados podemos obter 2^{n-1} dendogramas com aparências diferentes dependendo da ordem dos nodos.

A maioria do *software* existente no mercado disponibiliza algoritmos que desenham automaticamente os dendogramas.

Na árvore aditiva (abaixo representada) os comprimentos dos caminhos entre os nodos representam as proximidades entre objectos, e onde é válida a desigualdade triangular. A desigualdade triangular é uma condição necessária e suficiente para um conjunto de proximidades serem representadas na forma de árvore aditiva. A desigualdade triangular é representada da seguinte forma $d_{xy} + d_{uv} \leq d_{xu} + d_{yv} + d_{yu}$, $\forall \quad x, y, u, v$.

A árvore aditiva representada em baixo, retrata a associação genética entre trinta populações humanas.

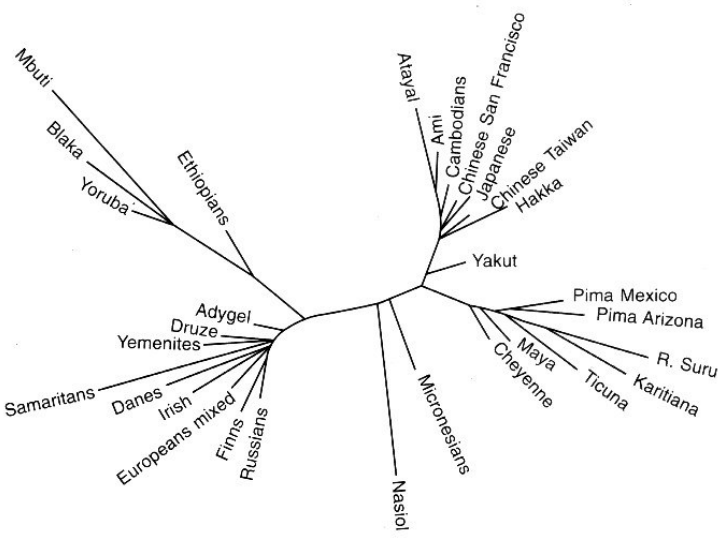


Figura 7 - Árvore aditiva representando as distâncias genéticas entre 30 humanos (retirado de [Everitt B. E al., 2001])

4.4.2 Comparação de dendogramas

Os métodos hierárquicos da Análise de Clusters impõem uma estrutura hierárquica aos dados. Então é necessário verificar se este tipo de estrutura é aceitável ou se introduz uma distorção inaceitável das relações nas proximidades originais entre os objectos.

Uma forma de validar o agrupamento consiste em comparar a matriz de proximidades original com a matriz de proximidades derivada (ou seja, do agrupamento). O Método mais usado consiste em calcular uma correlação de Pearson entre os valores da matriz de proximidades original e os valores da matriz de proximidades derivada. O coeficiente de correlação assim definido chama-se coeficiente de correlação cofenético, e a matriz obtida chama-se a matriz cofenética. O valor deste coeficiente deve ser muito próximo de 1 para uma solução de alta qualidade. Esta medida também pode ser usada para comparar soluções de agrupamento alternativas obtidas através de algoritmos diferentes. Podemos, pois, comparar os dois dendogramas.

O coeficiente de correlação cofenético, é dado por

$$r_{X,Y} = \sum XY - \frac{1}{n}\sum X \sum Y, \qquad -1 \le r_{X,Y} \le 1$$

X é a matriz seguinte:

$$X = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 5 & 5 & 0 & & \\ 5 & 5 & 4 & 0 & \\ 5 & 5 & 4 & 3 & 0 \end{bmatrix}$$

e Y a matriz de proximidade, corresponde à matriz D_1 do exemplo 4.1.

Colocando as matrizes X e Y na forma vectorial, obtemos:

X: (2, 5, 5, 5, 5, 5, 5, 4, 4, 3)

Y: (2, 6, 10, 9, 5, 9, 8, 4, 5, 3)

Fazendo os cálculos, o coeficiente de correlação cofenética entre a matriz X e Y é 0,82.

Outra medida utilizada para o mesmo fim é o coeficente de Goodman and Kruskal's, γ, definido como $\frac{S_+ - S_-}{S_+ + S_-}$, em que S_+ e S_- são o número de concordâncias e discordâncias, respectivamente.

Um método alternativo para comparar os dois dendogramas e, portanto, dois conjuntos de proximidades, no caso de ser usada a distância euclideana, é calcular a medida de Stress $\sum_{i<j}^{n}\sum_{j=1}^{n}\left(p_{ij} - \hat{p}_{ij}\right)^2 / \sum_{i<j}^{n}\sum_{j=1}^{n}\hat{p}_{ij}^{\,2}$, em que p_{ij} é o valor original da medida de proximidade entre os objectos i e j e $\hat{p}_{ij}$ é o valor correspondente na proximidade derivada.

4.4.3 Propriedades Matemáticas dos Métodos

Para os métodos hierárquicos são usadas medidas de *clustering*, para as quais podem ser definidas várias propriedades matemáticas, como vimos em 1.5.1. Uma destas é a propriedade ultra métrica, primeiramente introduzida por [Hartigan, 1967], e que tem sido desde então relacionada com vários aspectos da análise de cluster, em particular, a habilidade de representar a hierarquia por um dendograma.

A propriedade ultra métrica consiste no seguinte:

$d_{ij} \leq \max(d_{ik}, d_{jk})$, $\forall i, j, k$ em que d_{ij} é a distância entre os clusters i e j.

A propriedade não necessita ser válida para os elementos da matriz de proximidades.

Mas os valores da medida de proximidade num processo hierárquico não aumentando de forma monótona se o método de agrupamento hierárquico satisfaz a propriedade ultramétrica

Na consequência da falta da propriedade ultra métrica é que podem ocorrer inversões no dendograma.

[Morgan and Ray, 1995] descreveram alguns estudos de inversão para alguns métodos. Inversões não são sempre um problema, principalmente se houver interesse numa particular partição mais do que na estrutura hierárquica completa e podem ser úteis em áreas em que não é clara a estrutura, [Gower, 1990]. Contudo para [Murtagh, 1985], a inversão pode tornar a interpretação da hierarquia difícil.

4.4.4 Escolher o número de clusters

Muitas vezes os investigadores não estão interessados na hierarquia completa, mas apenas numa ou em algumas partições. É necessário decidir acerca do número de clusters.

Nos métodos aglomerativos ou divisivos, é comum surgir as questões: **"**Onde cortar a árvore?" "Qual é o melhor corte?"

A decisão é subjectiva e o investigador depara-se com o dilema:

- Se árvore for cortada numa fase em que nos dê poucos clusters, corremos o risco de estarmos a ser muito generalistas e estarmos a perder informação importante;
- Se árvore for cortada numa fase em que nos dê muitos clusters, corremos o risco de estarmos a pormenorizar demasiado e consequentemente deparamo-nos com uma interpretação muito complicada.

Tem havido muita investigação à volta deste assunto, mas ainda não se conseguiu uma resposta concisa a esta questão.

Dependendo da aplicação que vai ter, a solução hierárquica pode requerer estudo adicional antes de se fazer a escolha definitiva do agrupamento a apresentar. Na secção 4.4.2 já vimos a utilidade do coeficiente de correlação cofenético para validar uma solução ou até comparar duas soluções para o agrupamento. Apresentámos, também, o γ de Goodman e Kruskal e ainda a medida de Stress.

A cada passo do processo hierárquico, uma medida de proximidade derivada indica o valor da medida de proximidade do grupo correspondente aos dois clusters unidos naquele passo. Se o método utilizado na construção dos clusters satisfaz a propriedade ultramétrica, então as medidas derivadas aumentam de forma monótona ao longo do processo. O *software clustangraphycs8* usa estes valores das medidas de proximidade derivadas para obter duas estatísticas de teste para o número óptimo de clusters.

Mojena (1977), sugeriu os seguintes procedimentos para ajudar a escolher o número de grupos.

1º Método: é baseado no tamanho relativo do nível de fusão do dendograma (valor da proximidade derivada) e é conhecido por "upper tail rule". O objectivo é seleccionar o número de grupos correspondentes ao primeiro "andaime" do dendograma satisfazendo $\alpha_{j+1} > \overline{\alpha} + ks_{\alpha}$, em que $\alpha_0, \alpha_1, \alpha_2, ..., \alpha_{n-1}$ são os níveis de fusão correspondentes aos estados com n, n-1,…,1 clusters, respectivamente.

Os termos $\overline{\alpha},\ ks_{\alpha}$, são a média e o desvio padrão enviesado dos níveis de fusão j prévios e k é uma constante que varia no intervalo 2.75 a 3.5. Contudo Milligan e Cooper sugeriram fazer $k = 1.25$.

Alternativamente podemos usar a distribuição t-student para realizar um teste estatístico. Porém esta assume que as medidas de proximidade derivadas seguem uma distribuição Normal.

Uma aproximação visual é identificar quebras no gráfico de valores $(\alpha_{j+1} - \overline{\alpha}) / s_\alpha$ contra o número de clusters j.

2º Método: é baseado numa aproximação média de movimento.

Aqui a regra é usar a partição correspondente à primeira fase j, na sequencia parcial de clusters de $j = r$ a $j = n - 2$ clusters, satisfazendo $\alpha_{j+1} > \overline{\alpha} + L_j + b_j + ks_j$

em que $\overline{\alpha}$, s_j são, respectivamente, a média e o desvio padrão dos valores de fusão.

L_j e b_j são correcções da média para além da tendência dos valores de fusão.

b_j é o declive da relação linear entre os s_j (sob a hipótese de na realidade, não existirem clusters, espera-se que os s_j sejam aproximadamente lineares). $L_j = (r-1)b_j / 2$.

De acordo com Wishart, (1987), esta regra tem a vantagem de o nível de fusão considerado não entrar numa única estatística

No capítulo 5, em 5.2.4, voltaremos a abordar a questão da escolha do número de *clusters*.

4.4.5 Algoritmos hierárquicos

Nos métodos hierárquicos, podem ser usados diferentes algoritmos computacionais para conseguir o mesmo resultado.

Vários investigadores têm proposto algoritmos de optimização, que poderão ser encontrados em [Day, 1996].

A fórmula recursiva de Lance and Williams pode ser usada para programar métodos aglomerativos hierárquicos.

De acordo com Everitt, (2001), os métodos divisivos têm uma programação mais difícil, por isso são muito menos utilizados.

Os cálculos da Análise de Clusters podem ser realizados utilizando os pacotes de software *SPSS, ClustanGraphics8, NTSYSpc* e outros.

Capítulo 5
Métodos não hierárquicos

5.1 Introdução

Os médodos não hierárquicos da Análise de Clusters assentam em diferentes princípios e os seus resultados não constituem hierarquias.
Relativamente aos métodos hierárquicos, têm a vantagem de poderem ser aplicados a matrizes de dados muito grandes, uma vez que não é preciso calcular e armazenar uma nova matriz de dissemelhança em cada passo do algoritmo. Outra vantagem dos métodos não hierárquicos, é serem capazes de reagrupar os objectos em clusters diferentes daqueles em que foram colocados inicialmente.
Mas apresentam a desvantagem de, à priori, termos que definir o número de grupos, o que é muitas vezes difícil pois desconhecemos a estrutura dos dados. Uma opção viável, é aplicar um método hierárquico aos nossos dados para determinarmos o número de *clusters*, número esse que corresponderá ao número

de partições. Outra opção pode ser uma escolha aleatória. Outra opção poderá basear-se na experiência/conhecimento do investigador.

Os métodos não hierárquicos podem ser métodos de partição, métodos baseados em modelos, métodos difusos, métodos de sobreposição. Diferem essencialmente na forma como se desenrola a primeira agregação dos objectos em clusters, e no modo como as distâncias entre os centróides dos clusters e os objectos são medidas.

5.2 Métodos de Partição, método das k-means

Estes métodos aplicam-se a objectos (e não a variáveis). Operam sobre uma matriz de dados e exigem que o número de grupos seja fixado à partida. São métodos que usualmente não usam a matriz de proximidades inicial, mas sim a matriz dos dados inicial. Isto é um contraste com os métodos hierárquicos.

A partir de um conjunto de dados é construída uma partição, isto é, uma colecção de grupos disjuntos de objectos cuja reunião constitui o conjunto de objectos inicial.

Os grupos devem satisfazer, em geral, os critérios de homogeneidade, coesão interna, isolamento dos grupos e heterogeneidade entre grupos, que servem de guia à formação de clusters.

Os métodos de partição são por vezes usados como um complemento de um método hierárquico: para escolher o melhor nível para "cortar a estrutura", e uma resposta para a questão "quantos clusters apresentam?"

Um método de partição geralmente usado é o Método das "K-*means*", que mede a proximidade entre grupos usando a distância euclideana entre os centróides dos grupos.

Os métodos de partição usam procedimentos que em geral seguem os passos seguintes [Johson e Wichern, 2002]:

1. Seleccionar uma partição[13] inicial dos n objectos em k *clusters*.
2. Calcular os centróides para cada um dos k *clusters* (no SPSS as primeiras k observações são usadas como centróides dos k *clusters* no primeiro passo; ou o analista pode definir o cálculo da distância euclideana dos centróides a cada objecto na base dados).
3. Agrupar os objectos aos *clusters* de cujos centróides se encontram mais próximos, depois voltar ao passo (2) até não ocorrer variação significativa na distância mínima de cada objecto da base de dados a cada um dos centróides dos k *clusters* (ou até que o número máximo de iterações ou o critério de convergência, definido pelo analista, seja alcançado).

Deverá ser feito um esforço para minimizar a variância à volta do centróide do *cluster*. Não devemos esquecer realçar uma desvantagem/problema do método k-*means* que é a procura de *clusters* esféricos com o mesmo tamanho.
Algumas das vantagens do método k-*means* consistem em poder ser aplicado em conjuntos de dados muito grandes e tipicamente convergir rapidamente.

No que se refere à **escolha da partição inicial,** os k *clusters* (grupos) da partição inicial, como já foi dito anteriormente, podem ser o resultado da aplicação prévia de outro método de análise ou de uma pré-selecção de k objectos que são então usados para colocar os restantes objectos, ou do conhecimento prévio do problema em estudo ou, ainda, podem ser escolhidos ao acaso, de forma aleatória.

Seria proibitivo analisar todas as possíveis partições se o número de objectos n é grande, conforme vimos anteriormente. Na prática, no entanto, pode-se

[13] $P(n,k) = \dfrac{\left[k^n - \sum_{i=1}^{k-1} \dfrac{k!}{(k-i)!} P(n,i)\right]}{k!}$, é o número de partições possíveis de n objectos em k grupos.

analisar algumas soluções iniciais bem escolhidas. É uma forma de assegurarar a validade da solução final.

Relativamente à **deslocação de objectos para grupos,** os procedimentos mais comuns consistem em deslocar um objecto de cada vez ou grupos de objectos simultaneamente.

- Será apresentado um exemplo deste método, método das K-means, na Parte II - 4.2 deste trabalho.

5.2.1 Critérios de formação de clusters para dados contínuos

Os critérios de formação de clusters que constituem uma partição que mais se destacam, são os critérios de formação de clusters usados na análise de uma matriz de dados contínuos, $X_{n\times p}$, que usam a decomposição da matriz de dispersão T, dada por:

$T = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}\right)\left(x_{ij} - \bar{x}\right)^T$, em que x_{ij} é o vector de dimensão p das observações do objecto i no grupo j e $\bar{x}$ é o vector de dimensão p das médias de cada variável.

$\bar{x} = \dfrac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{\sum_{j=1}^{k} n_j}$ que é o vector das médias das p variáveis nos n objectos

Esta matriz da variabilidade total pode ser decomposta em:

- matriz da dispersão dentro do grupo, W, definida por :

$W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j \right) \left(x_{ij} - \bar{x}_j \right)^T$, em que $\bar{x}_j$ é o vector de dimensão p das médias das variáveis dentro do grupo j.

- matriz da dispersão entre grupos, B, definida por :

$B = \sum_{j=1}^{k} \sum_{i=1}^{n_j} n_j \left(\bar{x}_j - \bar{x} \right) \left(\bar{x}_j - \bar{x} \right)^T$, com $\sum_{j=1}^{k} n_j = n$

Então

$$T = B + W ,$$

onde T, W e B são as matrizes associadas à variabilidade total dos dados, à variabilidade dentro dos grupos e à variabilidade entre os grupos, respectivamente.

Para dados univariados, $p = 1$, a equação $T = B + W$ representa a decomposição da soma total dos quadrados da variável em soma dos quadrados dentro dos grupos e a soma dos quadrados entre grupos, que é fundamental na análise de variância.

Como T é fixo, porque não depende do agrupamento que se realize, a melhor partição é aquela em que W é mínimo ou B máximo, isto é quanto maior a homogeneidade interna dos grupos maior é a separação entre os grupos.

➢ *Minimização do traço de W*

No caso da análise multivariada, $p>1$, generaliza-se o caso sugerido na análise univariada, embora o critério $T=B+W$, não seja tão claro como para $p=1$.

Para determinar as três somas de quadrados acima referidas relativamente às p variáveis, necessitamos da soma dos elementos da diagonal principal destas matrizes. As três somas de quadrados são dadas por: trT, trW e trB.

A extensão óbvia no caso da análise multivariada, é a minimização da soma dos quadrados dentro dos grupos, que é equivalente a minimizar o traço da matriz W, trW, ou a maximizar o traço de B.

Minimizar o traço da matriz W é equivalente a minimizar a soma dos quadrados das distâncias euclideanas entre os objectos e as médias dos respectivos grupos,

$$E=\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\bar{x}_j\right)\left(x_{ij}-\bar{x}_j\right)^T=\sum_{j=1}^{k}\sum_{i=1}^{n_j}d_{ij,j}^2=\sum_{l=1}^{p}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ijl}-\bar{x}_{jl}\right)^2,$$ em que

$d_{ij,j}$ é a distância euclideana do objecto i do grupo j à média do grupo j. O critério pode também ser derivado do príncipio fundamental da matriz de distâncias:

$$E=\sum_{j=1}^{k}\frac{1}{2n_j}\sum_{i=1}^{n_j}\sum_{v=1,v\neq i}^{n_j}d_{ij,vj}^2$$

em que $d_{ij,vj}$ é a distância euclideana entre o objecto i e o objecto v no grupo j. Assim, a minimização do traço de W é equivalente à minimização do critério de perda de homogeneidade para distâncias euclideanas usada por Ward no processo hierárquico para a formação de clusters.

➢ *Minimização do determinante de W*

Na análise de variância múltipla, um teste para verificar se os vectores de médias são identicos para os grupos considerados, é baseado na razão entre os determinantes da matriz da variabilidade total e da matriz da variabilidade dentro dos grupos $|T|/|W|$.

Grandes valores de $\det(T)/\det(W)$ significa que os vectores de médias não são idênticos em todos os grupos.

Uma vez que para todas as partições de n objectos em k grupos, **T** permanece igual, a maximização de $\det(T)/\det(W)$, equivale a minimizar $\det(W)$. Este critério foi estudado por Marriot, (1971).

➢ *Maximização do traço de BW^{1}*

Uma função usada, também, na análise de variância múltipla, é o $tr(BW^{-1})$ sendo B a matriz que representa a variabilidade entre os grupos e W, a matriz que representa a variabilidade de dentro dos grupos. Grandes valores de $tr(BW^{-1})$ significa que os vectores de médias não são idênticos em todos os grupos. Baseando-nos neste critério, a melhor partição será a que corresponde ao máximo de $tr(BW^{-1})$. Quanto maior é o $tr(BW^{-1})$ e quanto menor $|W|$, maior é a diferença entre as médias dos grupos.

5.2.2 Propriedades do critério de *clustering*

Um dos critérios mais usados é o da minimização do traço da matriz W por ser simples e fácil de tratar computacionalmente. Consiste em minimizar a soma dos quadrados das distâncias euclideanas entre os objectos e os centróides dos respectivos grupos.

Apesar de ser o critério mais usado, este critério apresenta alguns inconvenientes, tais como:

- dependência da escala, ou seja, obtêm-se soluções diferentes com os mesmos dados estandardizados ou não estandardizados. É um grande inconveniente uma vez que o recurso à estandardização é muito frequente em análise de clusters;
- imposição de uma estrutura esférica, aos clusters observados, mesmo quando a estrutura "natural " dos dados tem outra forma, porque nos cáculos só tem em conta os elementos da diagonal principal de W e B.

Friedman e Rubin, (1967), procuraram um critério alternativo à minimização do trW de tal forma que o resultado fosse independente da escala. Tal critério baseou-se na maximização de $\det(T)/\det(W)$ ou na maximização de $tr(BW^{-1})$ com a utilização dos valores próprios $\lambda_1, \lambda_2, ..., \lambda_p$, da matriz BW^{-1}**:**

$$\text{traço}(\text{B W}^{-1}) = \sum_{l=1}^{p} \lambda_l$$

$$\frac{\det(T)}{\det(W)} = \prod_{l=1}^{p} (1 + \lambda_l)$$

Uma vez que os valores próprios da matriz BW^{-1} são os mesmos independentemente do facto desta matriz ser obtida da matriz original X ou a partir da matriz estandardizada, não são afectados pela escala.

O critério de minimização do $\det(W)$, tem sido o mais usado e não se restringe a clusters esféricos, ao contrário do critério do trW, que só identifica clusters esféricos. Estes critérios produzem grupos com o mesmo número de objectos, e o critério do determinante embora permita a formação de clusters elípticos, assume

que todos os clusters têm a mesma forma; o que poderá, como é evidente causar alguns problemas. Por isso serão necessários outros critérios, abordados no ponto seguinte.

5.2.3 Critérios alternativos para clusters de diferentes formas e tamanhos

Para ultrapassar o problema da forma dos clusters gerados pelos métodos referidos em 5.2.2, Scott e Symons, (1971), sugeriram um método de clustering baseado na minimização de

$$\prod_{j=1}^{k}\left[\det\left(W_j\right)\right]^{n_j},$$

em que W_j é a matriz de variabilidade dentro do j-ésimo grupo.

$$W_j=\sum_{i=1}^{n_j}\left(x_{ij}-\overline{x_j}\right)\left(x_{ij}-\overline{x_j}\right)^T$$

sendo n_j o número de objectos no j-ésimo grupo.

Este método é restringido a soluções em que cada cluster contém no mínimo $p+1$ objectos (p é o número de variáveis na base de dados), o que é necessário para evitar a matriz de dispersão singular, cujo determinante será zero.

Maronna e Jacovkis, (1974), descreveu um critério alternativo:

- A minimização de

$$\sum_{i=1}^{g}(n_i - 1)[\det(W_i)]^{1/p}$$

Symons, (1981), sugeriu dois outros critérios, que consistem na minimização de:

$\prod_{j=1}^{k}\left[\det\left(W / n_j^{\,2}\right)\right]^{n_j}$ ou de

$\prod_{j=1}^{k}\left[\det\left(W_j / n_j^{\,2}\right)\right]^{n_j}$

Todos os critérios analisados são mais apropriados quando todas as variáveis são medidas em escalas contínuas.

Quando as variáveis não são contínuas, a matriz de dissemelhanças pode ser obtida utilizando uma medida de dissemelhança analisada no capítulo 3 e aplicado um critério de formação de clusters operando na base da dissemelhança. Alternativamente, a matriz de dissemelhanças pode ser transformada numa matriz de distâncias euclideanas.

5.2.4 Escolha do número de clusters

Um dos problemas da Análise de Clusters, como aliás já falamos, é determinar o número de clusters. Enquanto nos métodos hierárquicos esta determinação é feita no fim da análise, depois de se ter aplicado um método hierárquico a uma matriz de proximidade (semelhança ou dissemelhança) e de ser feita a sua reprentação gráfica num diagrama de árvore, o dendograma, surge a questão "onde cortar a árvore?"

Nos métodos não hierárquicos de partição, esta determinação surge no início da análise "Quantos grupos deverá ter a partição?"

Seja qual for o critério usado no método hierárquico (ligação média, ligação completa, Wards's, ligação simples, etc). Se a matriz de proximidades original é o

quadrado da distância euclideana, então as matrizes das somas de quadrados T, W e B podem ser usadas para construir uma variedade de medidas para indicar o nº de clusters.

Poderá usar-se um critério de optimização, por exemplo trW**,** contra o número de clusters. Examina-se o gráfico e a zona cotovelo sugere-nos o número de clusters.

Poderá também construir-se um critério baseado na razão entre a soma dos quadrados entre os k clusters e a soma total de quadrados.

$$R_k^2 = \frac{trB_k}{trT} = 1 - \frac{trW_k}{trT},$$

Para $k = n$ clusters, vem $trW = 0$ e $R_n^2 = 1$.

Durante a diminuição do número de clusters de n até 1, os clusters vão ficando mais separados, e um decréscimo grande de R_k^2 é indicador de clusters bem distintos, informando-nos, portanto, do número de clusters.

Tem sido feita muita investigação à volta deste assunto, isto é, da determinação do número de clusters, tendo sido criadas várias fórmulas com a intenção de ajudar a resolver este problema. Milligan e Cooper (1985) apresentam uma discussão detalhada do assunto.

5.3 Outros Métodos

5.3.1 Pesquisa de densidades

Em aplicações em que se espera observar clusters naturais, usam-se métodos que procuram regiões de alta densidade de pontos (objectos). O agrupamento natural sugere que deverão existir muitos pontos no espaço que

estão muito próximos de outros pontos e que estes clusters estão separados por áreas com muitos poucos pontos. Neste tipo de problemas, em geral usa-se o método da ligação simples para a obtenção dos clusters. Inicialmente escolhe-se um raio r e o número de pontos k. À volta de cada ponto ou objecto, determina-se uma esfera de raio r e calcula-se o número de pontos contidos na esfera. Todos os pontos com, pelo menos, outros k pontos contidos na esfera, são chamados pontos de densidade. Os clusters iniciais são definidos pelos pontos de densidade de forma a que se um ponto de densidade pertence a mais do que um cluster então estes clusters são unidos. Também se unem os clusters em que a distância entre eles é inferior à média das $2k$ distâncias mais pequenas entre os n pontos originais. Qualquer ponto que esteja separado de todos os pontos de densidade de uma distância superior a r, forma o seu próprio cluster. Depois de se ter determinado a primeira solução (solução inicial), pode-se aumentar o valor de r e repetir o processo.

5.3.2 Métodos difusos

Este método geralmente começa a partir de uma matriz de proximidade. Inicialmente escolhe-se um valor de proximidade p e um inteiro k. Unem-se todos os pontos que estão a um nível de proximidade maior ou igual a p. Obtém-se um cluster determinando o maior subconjunto possível de pontos que estão unidos a todos os outros pontos no conjunto. Se um cluster tem pontos que estão unidos a pelo menos k pontos noutro cluster, então unem-se estes dois clusters. Neste método de agrupamento uma parte de um cluster pode sobrepor-se a uma parte de outro cluster.

Este método exige algoritmos complexos e o *output* é volumoso, o que torna difícil a sua interpretação.

5.3.3 Métodos baseados em modelos

Supõe-se uma estrutura com k grupos, havendo um modelo subjacente responsável pela criação de cada um dos clusters. As hipóteses colocadas com maior frequência são:

1ª O vector de observações x tem função densidade de probabilidade $f_j(x;\theta_j)$ se provém do grupo j, j = 1,...,k, onde θ_j representa um vector de parâmetros desconhecidos.

2ª O vector de observações x tem função de densidade de probabilidade

$$f(x;p,\theta)=\sum_{j=1}^{k} p_j f_j\left(x;\theta_j\right)$$

correspondente a uma mistura finita de densidades $f_j(x;\theta_j)$ de cada um dos clusters. O peso p_j é a probabilidade associada a cada componente da mistura, sendo $\sum_{j=1}^{k} p_j = 1$.

O problema consiste em estimar parâmetros em cada um dos modelos a partir das observações feitas.

A situação mais comum é assumir que as densidades envolvidas são de distribuições Normais multivariadas.

Capítulo 6

Considerações finais

6.1 Introdução

Na Análise de Clusters as dificuldades mais comuns são[14]: a escolha da medida de proximidade, a escolha da medida de proximidade, a escolha da medida de proximidade (três vezes!), a escolha do método de formação dos clusters e a determinação do número de clusters e sua interpretação. Além disto, há que ter em conta que o resultado de uma Análise de Clusters não depende só dos factores que acabámos de referir. Também depende das variáveis seleccionadas e da estrutura de agrupamento subjacente aos dados, se existir.

A formação de clusters é um processo subjectivo, o que torna a formação de clusters difícil. O mesmo conjunto de dados poderá ser dividido de forma

[14] Peter Bryant, Professor da Universidade de Colorado, na Summer School em Análise de Clustes, Oeiras, 2006.

diferente consoante o método, dependendo do investigador e das técnicas usadas.

6.2 Validação dos resultados

Uma vez que a aplicação de diferentes medidas de semelhança e diferentes métodos de construção dos clusters podem conduzir a diferentes resultados, é importante avaliar esses resultados para verificar se constituem um resumo útil dos dados ou se pelo contrário revelam uma estrutura imposta a esses dados.

Como sabemos, em Estatística, para validar resultados efectuamos testes sobre os modelos usados ou aplicamos os mesmos procedimentos a outros dados. Mas isto não tem sido feito em Análise de Clusters devido à natureza dos dados, ao facto de estes serem descritos por variáveis de diferentes tipos impedindo que seja verificada a hipótese de Normalidade que abre caminho à construção de testes estatísticos. Além disso, muitas vezes o conjunto de objectos a analisar é a população total, sendo portanto inapropriado dividir a população que está a ser usada em duas partes: uma parte para a análise exploratória (relembramos que a Análise de Clusters é exploratória) e a outra parte para ser usada na confirmação dos resultados obtidos.

Jain e Dubes, (1988), apresentaram os quatro tipos de critérios seguintes, para validação dos resultados da Análise de Clusters:

i) Critérios externos que medem o desempenho dos resultados obtidos comparando a estrutura dos dados com informação exterior não utilizada na Análise. Avaliar os algoritmos para a construção dos clusters usando amostras de clusters conhecidos, é um exemplo de uma avaliação externa do método utilizado.

ii) Critérios internos que comparam a estrutura obtida com a matriz de proximidades para os dados iniciais. Como vimos no capítulo 4, em 4.4.4, no caso dos métodos hierárquicos é habitual utilizar o coefiente de correlação cofenética. Os critérios internos visam a qualidade do ajustamento da solução obtida relativamente à matriz de proximidades original.

iii) Critérios relativos – comparam diferentes estruturas construídas a partir dos mesmos objectos, mas utilizando métodos diferentes.
Dispondo de duas partições, constrói-se uma tabela de contingência com os objectos pertencentes às duas classificações, e avalia-se o grau de associação entre estas duas partições.

iv) Critério da replicabilidade – compara os resultados obtidos com uma metade da amostra com os resultados obtidos com a outra metade.

6.3 Apresentação de resultados

Os resultados deverão ser apresentados num relatório no qual deverá constar: representação gráfica; uma descrição do problema em estudo; indicação do método de selecção dos objectos e das variáveis; a(s) medida(s) de proximidade e o método usado para a construção dos clusters; o *software* utilizado; descrição do critério usado na determinação do número de clusters.

A utilização da Análise de Clusters requere uma participação activa do analista para interpretar e dar significado aos resultados. Esta fase do processo é subjectiva, intuitiva e heurística [Anderberg B., 1973].

O resultado da análise não é apenas um mero conjunto de clusters. Estes resultados precisam de ser interpretados e podem ser considerados relevantes ou irrelevantes. Os que forem considerados relevantes poderão

aumentar o conhecimento e a organização de factos conhecidos permitindo uma descrição mais pormenorizada do assunto que está a ser estudado.

6.4 Sugestões para a utilização da Análise de Cluster

Não há estratégia óptima para aplicação dos critérios de formação de clusters, mas podemos apresentar algumas sugestões as quais poderão ajudar em muitas situações. De acordo com Miligan, (1996), deverá ter-se em conta:

- Os objectos - pois estes deverão ser representativos da estrutura do cluster;
- As variáveis a serem usadas – só deverão ser incluídas se houver uma boa razão para se pensar que estas definirão os clusters, se forem irrelevantes deverão ser excluídas do estudo.
- A estandardização das variáveis não é sempre necessária. A redução das variáveis (ou seja, aos valores observados para a variável subtrai-se a média e divide-se pelo desvio-padrão da amostra), é uma alternativa ao método mais usado, a divisão pelas variâncias. Muitas vezes a estandardização dos dados reduz as diferenças entre os objectos, mascarando os agrupamentos naturais que possam existir entre os dados originais. Se uma variável estiver medida numa escala numericamente mais elevada, a sua dispersão deverá ser também maior e, consequentemente, esta variável terá maior peso na solução final da análise de clusters. Com a estandardização, todas as variáveis estandardizadas passam a ter o mesmo peso. Ora, em algumas situações poderão existir variáveis com uma importância intrínseca superior às outras variáveis, e essa importância deve ser mantida e não anulada. Outra opção é um método de cluster que seja invariante às escalas.

- A medida de proximidade – não há um guia único para a sua escolha, mas o tipo de dados, o contexto da análise, assim como o *software* disponível irá sugerir a escolha;
- O método de formação de clusters – designado para descobrir os clusters, pouco sensível aos erros (e disponível no *software* utilizado) será condicionado pelos mesmos factores atrás referidos;
- O número de clusters é uma das decisões mais difíceis de tomar, em que diferentes regras sugerem diferentes números. A interpretação dos resultados é importante na decisão do número de clusters. É esta a meta final da análise.

PARTE II

Uma aplicação da Análise de Clusters

1. Introdução à Parte II

Na parte II do trabalho será usado o *software* SPSS para a formação de clusters usando a base de dados obtida do inquérito feito a alunos de 9º ano de escolaridade de oito escolas do Funchal: a duas escolas privadas e seis escolas públicas escolhidas após uma rápida consulta aos resultados obtidos nos exames nacionais de Matemática de 9ºano nas diversas escolas do Funchal no ano lectivo transacto, 2004/2005 e tendo em conta a sua localização geográfica. Este inquérito pode ser consultado no anexo I deste trabalho.

Nas escolas foram seleccionadas duas turmas, uma com melhores e outra com piores resultados em Matemática no final do 2º Período do ano lectivo 2005/2006.

Na Região Autónoma da Madeira[15], no ano lectivo 2005/2006, estavam inscritos 11 376 alunos, dos quais 5 370 eram alunos inscritos em escolas no Concelho do Funchal.

No final do 3º Período desloquei-me às escolas onde tinham sido feitos inquéritos, para recolher a informação das pautas relativa às classificações de frequência, de exame e final daqueles alunos. Alguns alunos não fizeram correctamente a sua identificação o que impossibilitou a aquisição de informação relativa às classificações.

2. Sucesso/Insucesso em Matemática

Há muitas definições/interpretações sobre sucesso/insucesso em Matemática; é sem dúvida algo de subjectivo. Consideramos que um aluno teve sucesso em Matemática num determinado nível de ensino quando obteve classificação positiva (nota igual ou superior a 3, pois no 3º ciclo as classificações são dadas entre 1 e 5).

Tem-se falado muito sobre o insucesso dos alunos Portugueses em Matemática, baseado essencialmente nos resuldos dos alunos nos exames de Matemática do 12º ano e do 9º ano. O exame nacional de Matemática do 9º ano realizou-se pela primerta vez a 22 de Junho de 2005, embora se realizem Provas de Afericção desde 2001.

Apresentaram-se ao exame cerca de 90 000 jovens, dos quais 99% foram à 1ª chamada[16]. 71% dos alunos tiveram nível negativo.

Os resultados obtidos são preocupantes e por isso foram alvo de reflexão pelos docentes de Matemática com o objectivo de compreender algumas causas e essencialmente implementar estratégias para a promoção da melhoria da

[15] Ver site www.madeira-edu.pt

[16] para mais informação, consultar www.gave.pt

aprendizagem[17]. Nesta reflexão os docentes apresentaram como principais causas as dificuldades a nível dos conhecimentos específicos da Matemática, a ausência de hábitos de trabalho, a falta de interesse e motivação e a existência de percursos escolares anteriores deficientes; as condições de funcionamento do Sistema Educativo, a insuficiente carga horária de Matemática para cumprir o programa e o elevado número de alunos por turma; o fraco envolvimento das famílias, as condições sócio-económicas das mesmas e a imagem negativa da Matemática junto das famílias.

As medidas que deverão ser intoduzidas visam essencialmente incidir na resolução de problemas, usar as tecnologias de informação, de materiais manipuláveis; a promoção de clubes de Matemática, salas de estudo e a utilização da aula extra curricular, estudo acompanhado, para o desenvolvimento de actividades relacionadas com a Matemática; a criação de horários de apoio aos alunos e atribuição de espaços e de tempo para os docentes trabalharem em equipa; maior responsabilização dos Encarregados de Educação no acompanhamento dos seus educandos bem como na importância que atribuem à Matemática. Foi sugerido ao poder central que desdobrasse as turmas, que as turmas fossem menos heterogéneas no que concerne às capacidades e interesses dos alunos, reforço da carga horária de Matemática, aquisição de materiais específicos de Matemática e a introdução de critérios mais rigorosos na transição.

O desempenho dos examinandos foi satisfatório nos aspectos da competência matemática (restringem-se a conceitos, procedimentos e raciocínos desde que estes sejam simples); o desempenho foi fraco na resolução de problemas e muito fraco no raciocínio dedutivo.

Não é de modo algum correcto estabelecer rankings das várias escolas, pois para que a análise estatística seja correcta, além dos resultados deve se ter em

[17] para mais informação poderá consultar o livro publicado pelo GAVE (gabinete de avaliação educacional)*: Reflexão dos Docentes do 3º Ciclo sobre os Resultados do Exame do 9º ano 2005 1ª chamada,* Editorial do Ministério da Educação, Lisboa.

conta aspectos sócio-culturais, económicos e académicos das famílias, as condições dos alunos de cada escola; as características das escolas em termos organizativos, humanos e naturais; a localização geográfica da escola no panorama nacional e a estabilidade do corpo docente.

Os resultados obtidos nos exames nacionais são semelhantes aos do estudo internacional PISA 2000 e 2003. Neste estudo constatou-se um afastamento dos alunos portugueses, de 15 anos, para pior, em relação aos desempenhos dos restantes colegas do espaço OCDE.

Em termos de formação académica, Portugal é um dos países da União Europeia com menor percentagem de licenciados e de pessoas com a escolaridade obrigatória; com maior taxa de insucesso escolar e de maior abandono escolar precoce, enfim um dos países com mais dificuldades ao nível da formação académica dos habitantes.

Outro aspecto que nos diferencia de outros países é a formação em Matemática dos professores do 1º ciclo, que nem sempre mereceu particular atenção por parte da Comunidade Científica, ao contrário do que acontece noutros países, como por exemplo Inglaterra, em que a formação matemática destes é indispensável.

Considero que os professores do 1º ciclo deveriam ter mais formação em Matemática e no caso de isto já não ser possível, por exemplo para professores que já tenham terminado a sua formação académica, a Matemática de 1º ciclo deveria ser dada por um professor que tenha feito pelo menos uma licenciatuara em Matemática.

Além disso, os alunos têm de estudar mais e melhor, para que haja melhor aprendizagem; para que os novos conhecimentos se interliguem com os que já estão acumulados na mente. Aliás, cerca de 80% dos alunos que responderam ao inquérito, portanto a esmagadora maioria dos alunos entrevistados, reconhecem que a falta de estudo assim como a falta de atenção/concentracção são as principais causas do insucesso escolar. E neste aspecto, como é óbvio, é fundamental o incentivo por parte dos Encarregados de Educação.

De acordo com a Psicóloga e investigadora, Margarida Pocinho da Universidade da Madeira «a automatização adquirida através da prática, interfere

com o desempenho e dá lugar a uma maior capacidade de processamento da informação, disponibilizando e libertando recursos cognitivos que serão rentabilizados em processos de pensamento mais complexos, como melhor atenção ao que interessa e maior capacidade para resolver problemas». Pensamos que esta capacidade para resolver problemas deve ser adquirida ao longo do ensino básico através da identificação de questões da vida real com os raciocínios matemáticos já desenvolvidos. A resolução de problemas deve então, tornar-se numa parte intrínseca da nossa capacidade mental que será utilizada intuitivamente. Este é na nossa opinião, o percurso para desdramatizar a resolução de problemas.

É necessário repetir as matérias, pois se assim não for não há assimilação. A situação é semelhante à que acontece com um jogador de futebol, ou outro atleta. Para ser bem sucedido tem que treinar, o mesmo se verifica a nível cognitivo. A nossa memória tem que ser trabalhada. Quanto mais limitados cognitivamente são os jovens, mais têm que repetir, não de forma mecânica, mas sim compreensiva.
Hoje, ao contrário do que se defendeu nas últimas décadas «assiste-se a uma valorização da memória através das teorias do processamento da informação e da psicologia cognitiva. A compreensão não basta para assimilar determinadas matérias mais complexas e que exigem mais trabalho, como as Línguas, a Matemática, a Física, a História ou mesmo a Música».
Nos anos 60/70 considerava-se que tudo traumatizava o jovem e a criança, nomeadamente fazê-los estudar em excesso. Hoje verifica-se que é ao contrário. O insucesso é que traumatiza a criança, os adolescentes, os próprios pais e a sociedade em geral.

Assim como a nossa felicidade está em nós e não nas coisas que possuímos, também a riqueza de um país está na sua população. Por isso é urgente que todos nós, assim como os nossos governos, invistam mais e melhor na Educação.

3. Análise Descritiva dos resultados obtidos no inquérito

De seguida mostramos uma primeira abordagem descritiva aos resultados obtidos para cada uma das perguntas do inquérito:

1.2 Sexo: Masculino □ Feminino □

Tabela PII.3.2 – sexo

		Frequency	Percent	Valid Percent
Valid	masculino	146	46,6	46,6
	feminino	167	53,4	53,4
	Total	313	100,0	100,0

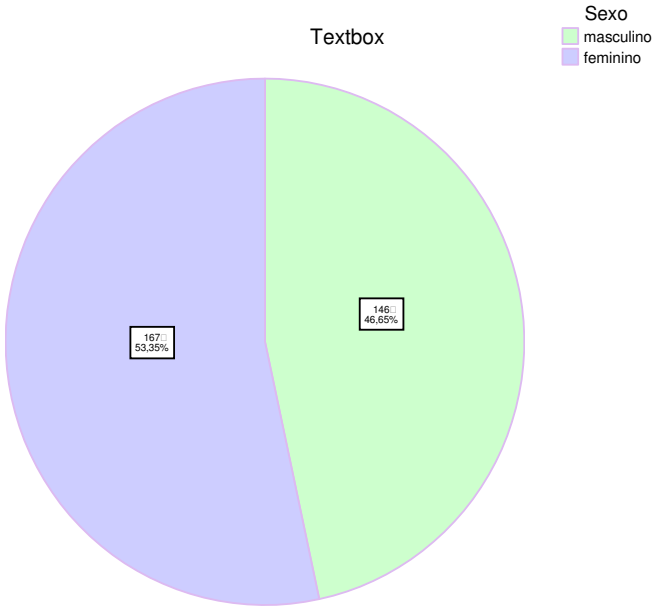


Figura PII.3.1 – Percentagemde rapazes e raparigas que responderam ao inquérito

53% dos alunos que responderam ao inquérito são rapazes e 47% são raparigas

1.3 Idade: Observamos que os alunos a quem foi pedido que respondessem ao inquérito, se encontram numa faixa etária que vai desde os 14 aos 18 anos. A tabela seguinte dá-nos a frequência e a percentagem das idades relativas aos alunos do 9º ano da nossa amostra.

Tabela PII.3.3 – Idade

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	14	106	33,9	33,9	33,9
	15	112	35,8	35,8	69,6
	16	50	16,0	16,0	85,6
	17	33	10,5	10,5	96,2
	18	12	3,8	3,8	100,0
	Total	313	100,0	100,0	

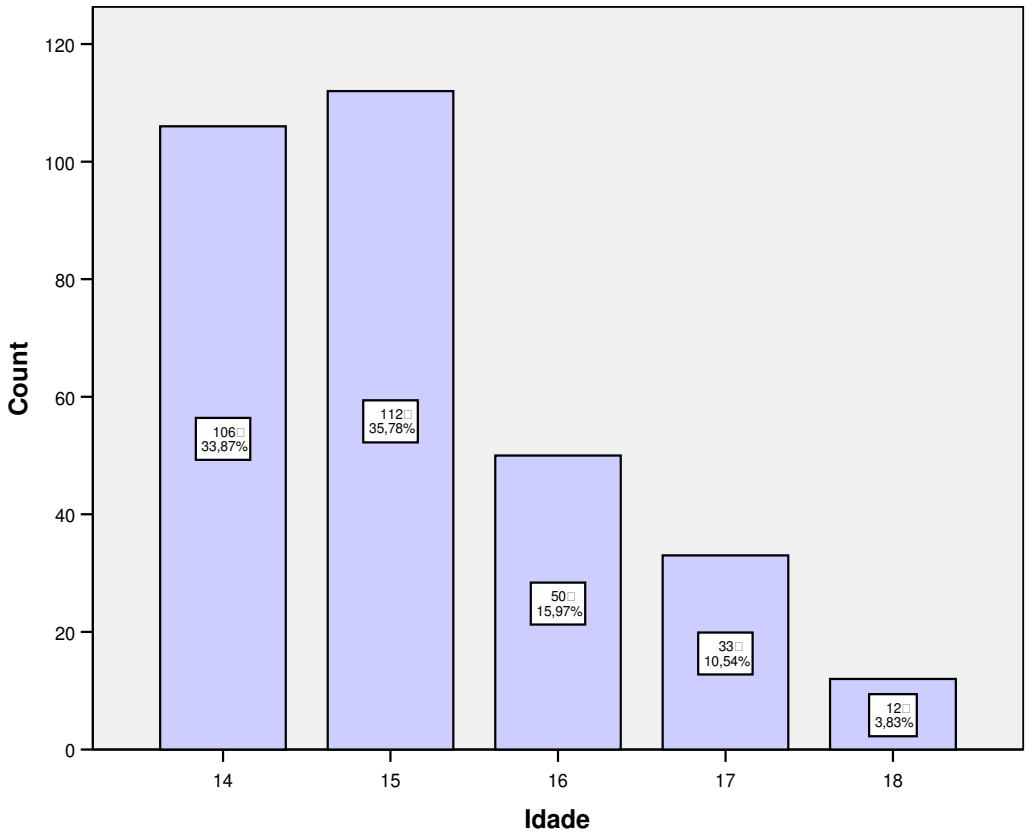


Figura PII.3.2 - Distribuição dos alunos pela sua idade.

1.4. Já reprovou?

Tabela PII.3.5 - Já reprovou?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	não	194	62,0	62,0	62,0
	sim	119	38,0	38,0	100,0
	Total	313	100,0	100,0	

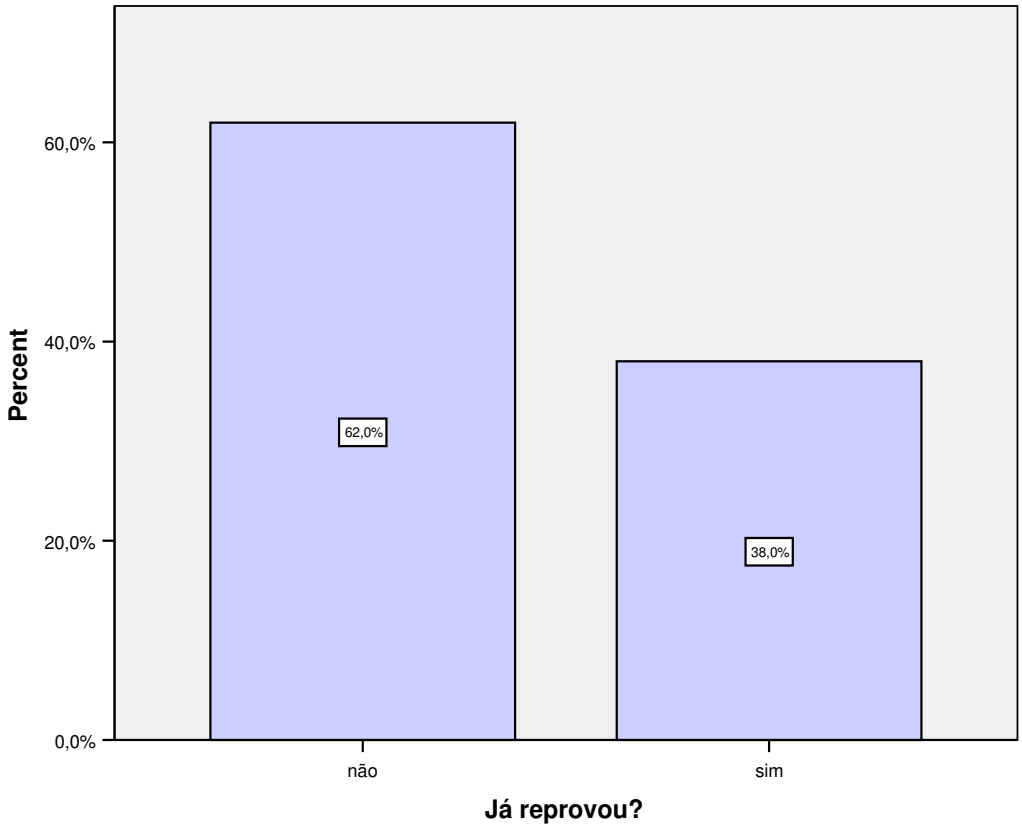


Figura PII.3.3 – Percentagem dos alunos que reprovaram

Através do gráfico observamos que 62% nunca reprovaram, enquanto que 38% dos alunos já reprovaram.

Se sim, quantos anos reprovou?

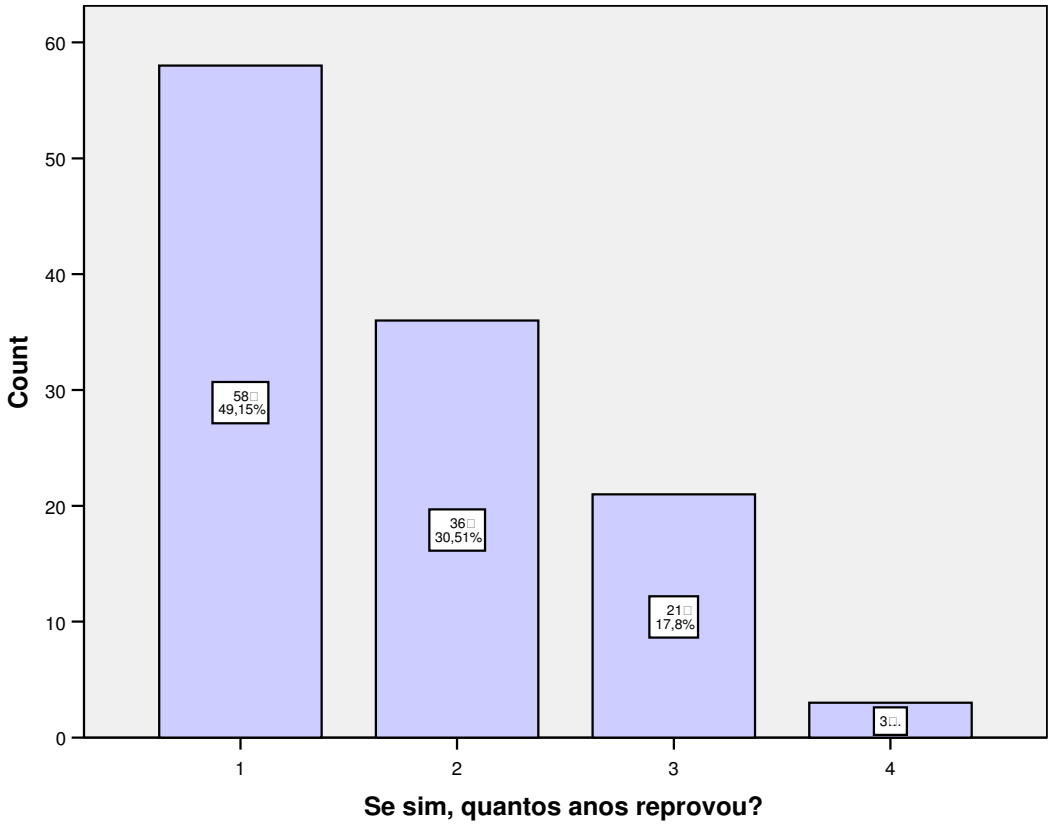


Figura PII.3.4 - Percentagem dos alunos que já perderam um ano, dois anos, três.

Como podemos observar, 49,15% dos estudantes que já reprovaram, reprovaram apenas um ano. 30,51% reprovaram 2 anos e 17,8% reprovaram 3 anos.

Nos anos em que reprovou, teve nota negativa em Matemática?

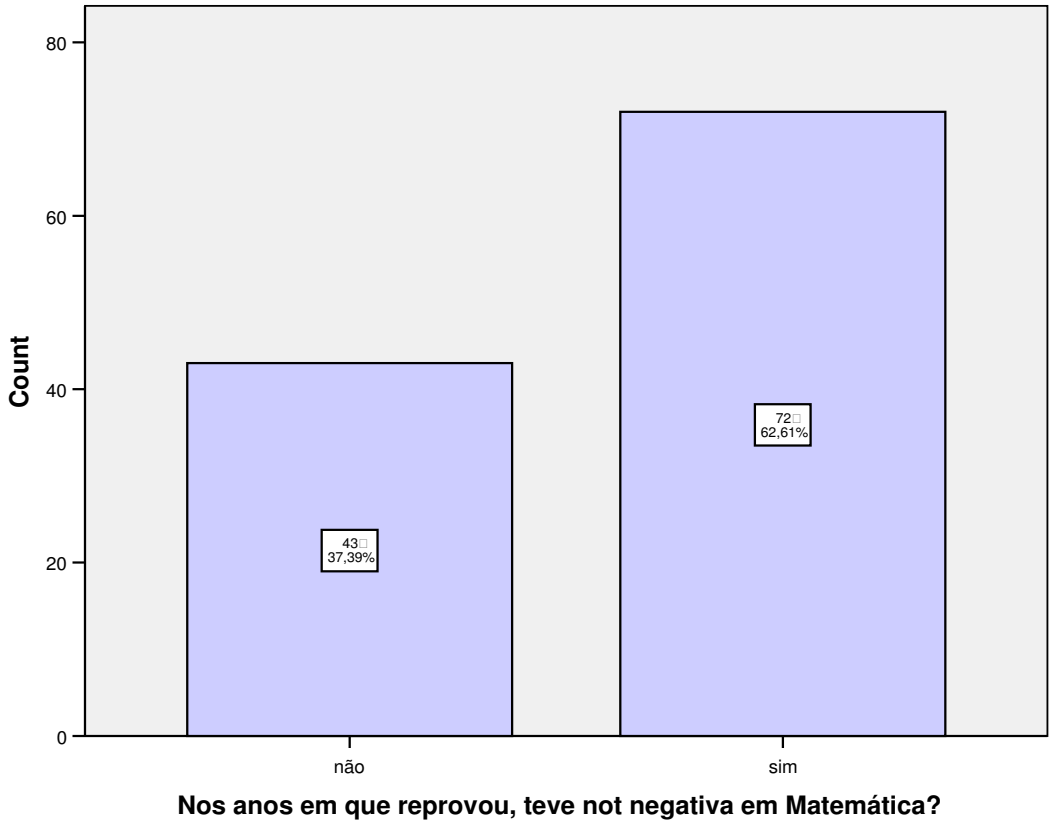


Figura PII.3.5 – Percentagem de alunos que tiveram nota negativa a Matemática nos anos em que reprovaram.

62,61% dos alunos que reprovaram tiveram nota negativa a Matemática nos anos em que reprovaram.

1.5 Com quem reside?

Mãe e pai ☐ Sem a mãe ☐ Sem o pai ☐

Avós ☐ Outros ☐

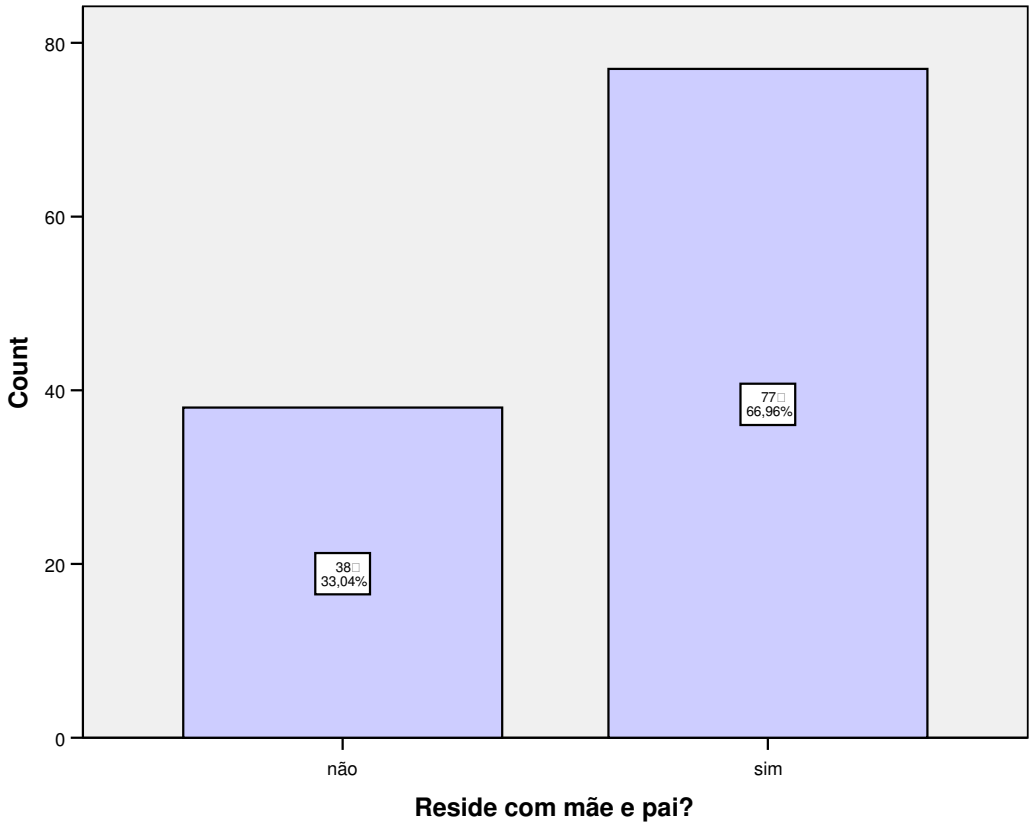


Figura PII.3.6 – percentagem de alunos que residem com mãe e pai

66,96% dos alunos inquiridos vivem com a mãe e o pai. 21,74% residem sem o pai e 5,22% residem sem a mãe.
8,7% dos alunos que responderam ao inquérito residem com os avós. 37,39% dos jovens têm outras pessoas (distintas dos pais e dos avós) a viver com eles.

1.6. Tem bom relacionamento com quem reside?

Sim ☐ Não ☐ Às vezes ☐

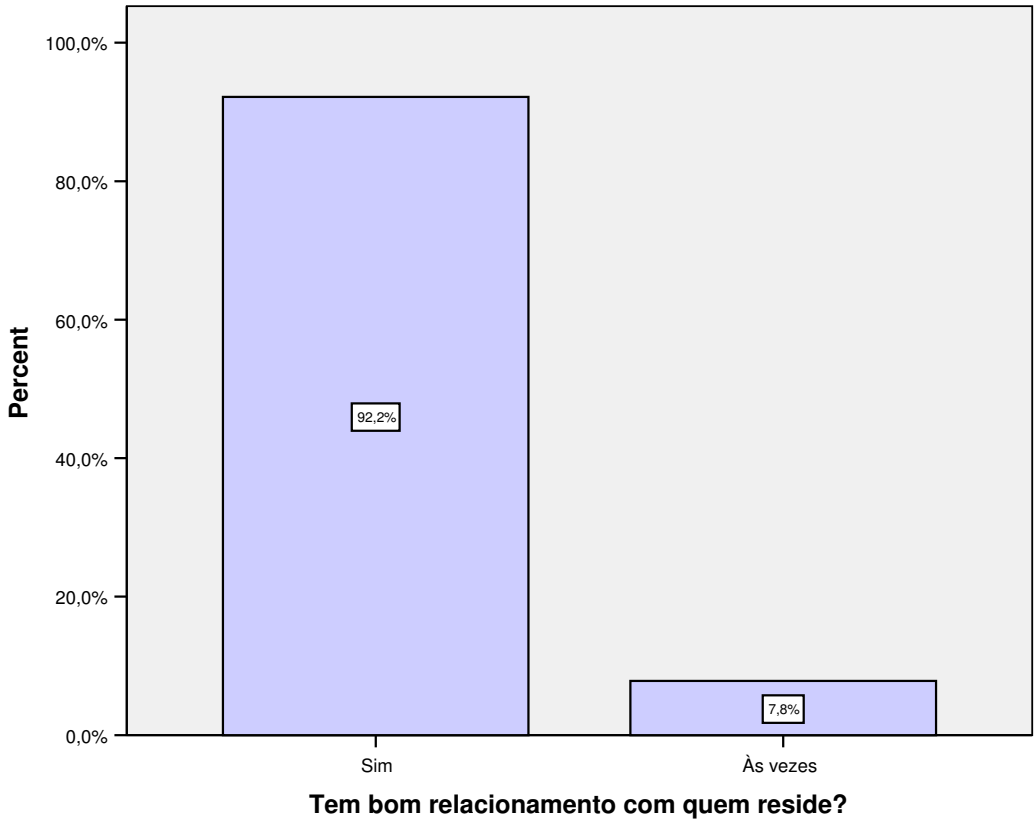


Figura PII.3.7 – Percentagem dos alunos de acordo com o relacionamento com quem vivem

À pergunta se tem bom relacionamento com quem reside, observamos que a maioria dos alunos respondeu que sim e uma pequena percentagem às vezes, não sendo observada a resposta não.

1.7. Como classificaria o seu ambiente familiar:

Muito Bom ☐ **Bom** ☐ **Razoável** ☐

Mau☐ **Muito mau** ☐ **Outro**☐

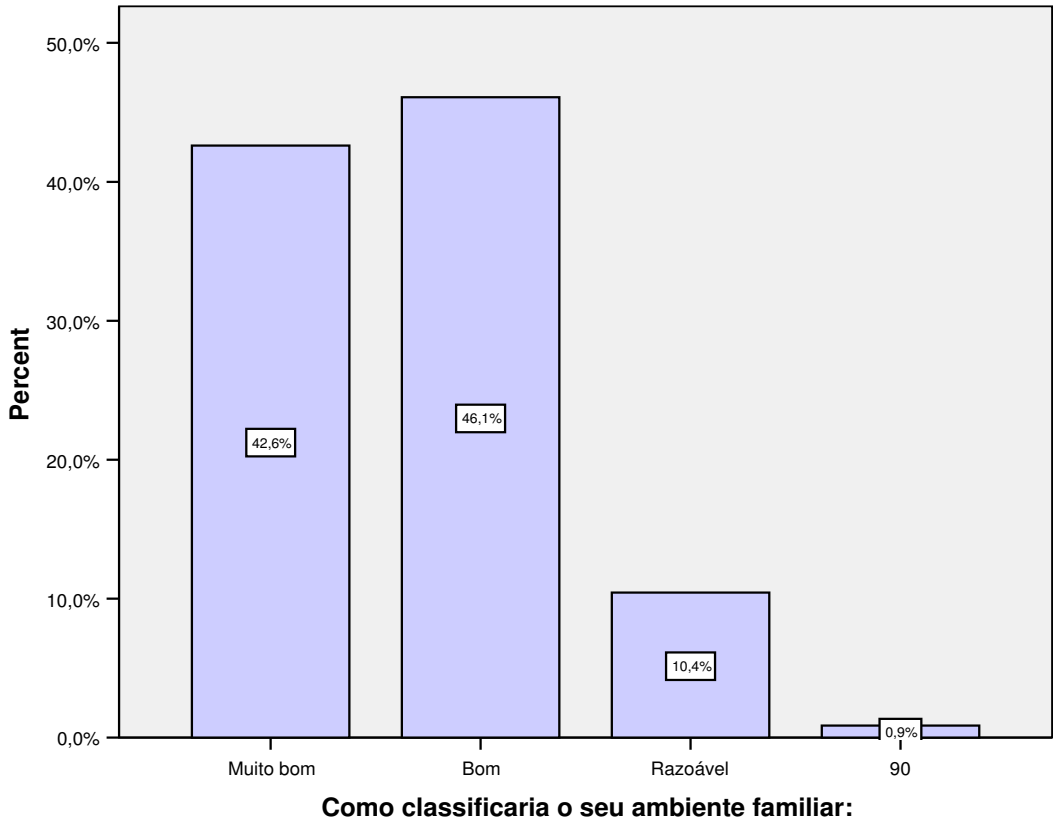


Figura PII.3.8 – Classificação do ambiente familiar, feita pelos alunos

42,8% dos alunos classificaram o seu ambiente familiar de muito bom, 46,1% de bom, 10,4% responderam ser razoável e 0,9% optaram por não responder à pergunta.

1.8. Habilitações dos pais (mesmo se já falecidos) ou pessoa com quem reside

Para as habilitações dos pais obtivemos os seguintes gráficos que nos indicam que a maioria dos pais destes alunos só tem o ensino básico.

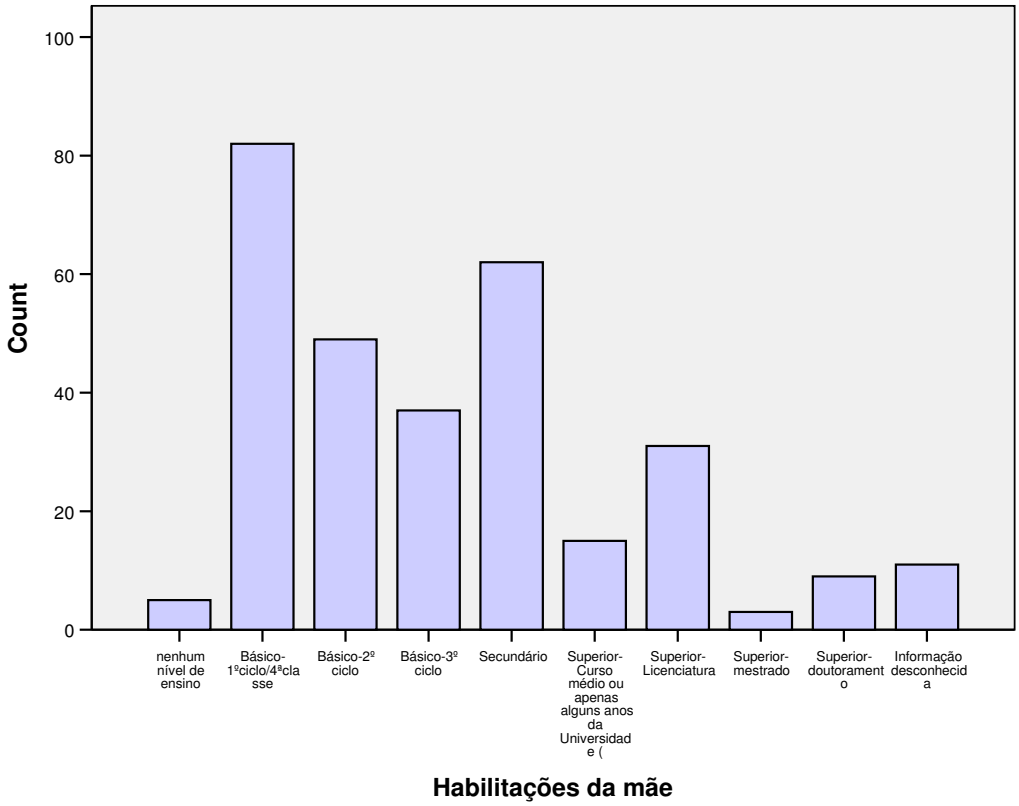


Figura PII.3.9 – habilitações da mãe

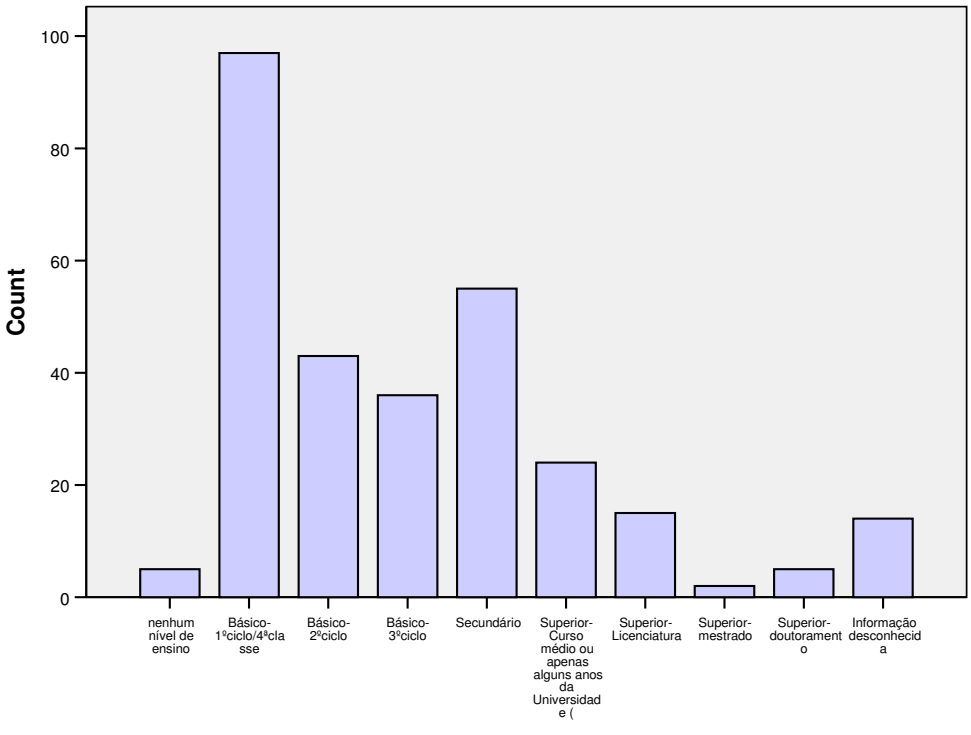


Figura PII.3.10- habilitações do pai

1.9. Tempo semanal de estudo em Matemática (T.P.C. e estudo em matéria dada)

0h ☐ 1h ☐ 2h ☐ mais do que 2h ☐

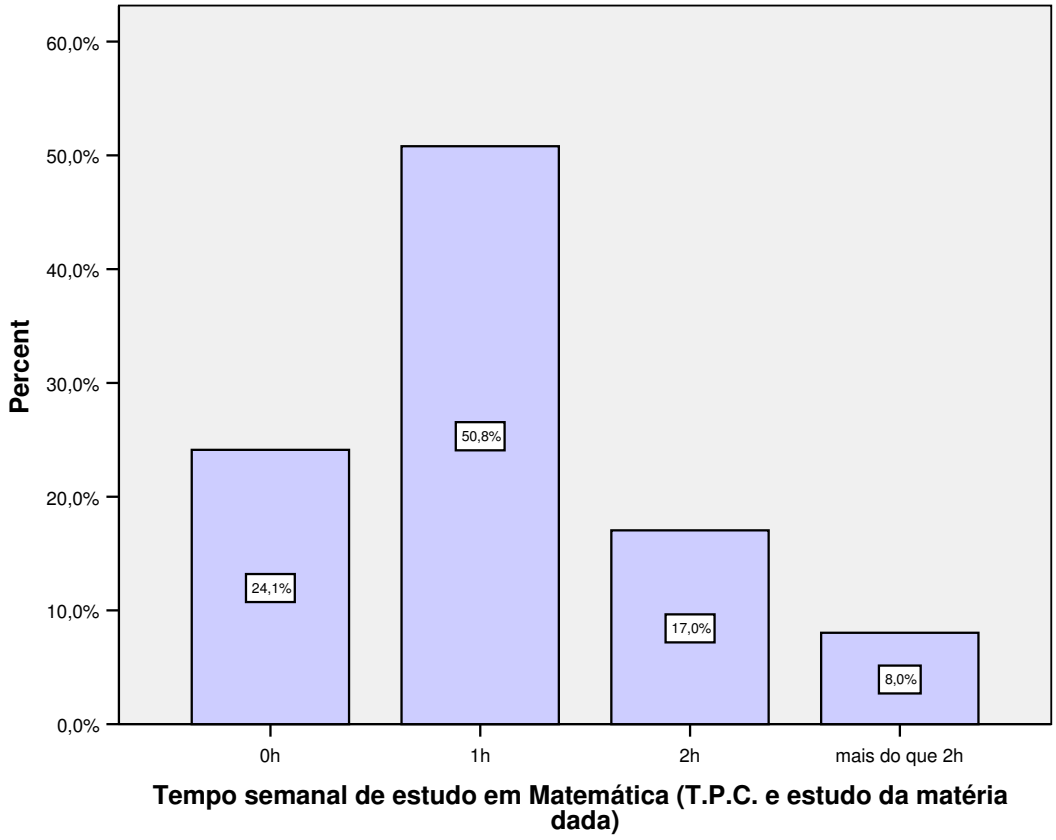


Figura PII.3.11 – tempo semanal de estudo

50,8% dos alunos responderam que estudam 1h por semana. Mas 24,1% responderam que não estudam. Apenas 17% dos alunos estudam 2 horas por semana e 8% estudam mais do que 2 horas por semana.

1.10. Quando tem dúvidas ou dificuldades em Matemática, a quem recorre para esclarecê-las?

Às pessoas com quem reside ☐ **colegas ☐**

Ao professor de Matemática ☐ **outros ☐**

Nos próximos 4 gráficos, podemos ver as percentagens com que foi dada cada uma destas respostas.

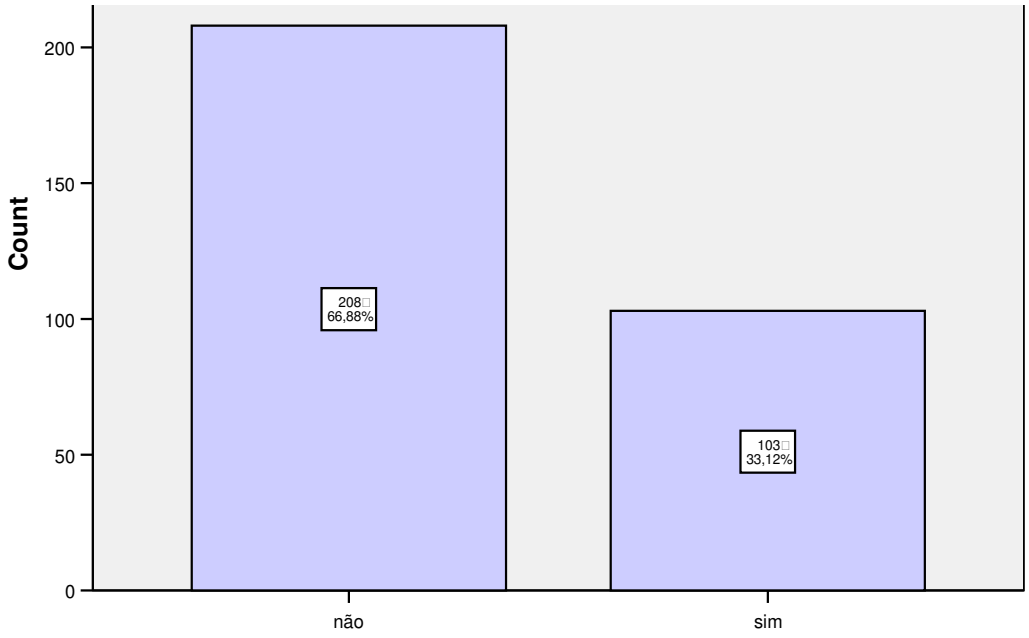


Figura PII.3.12

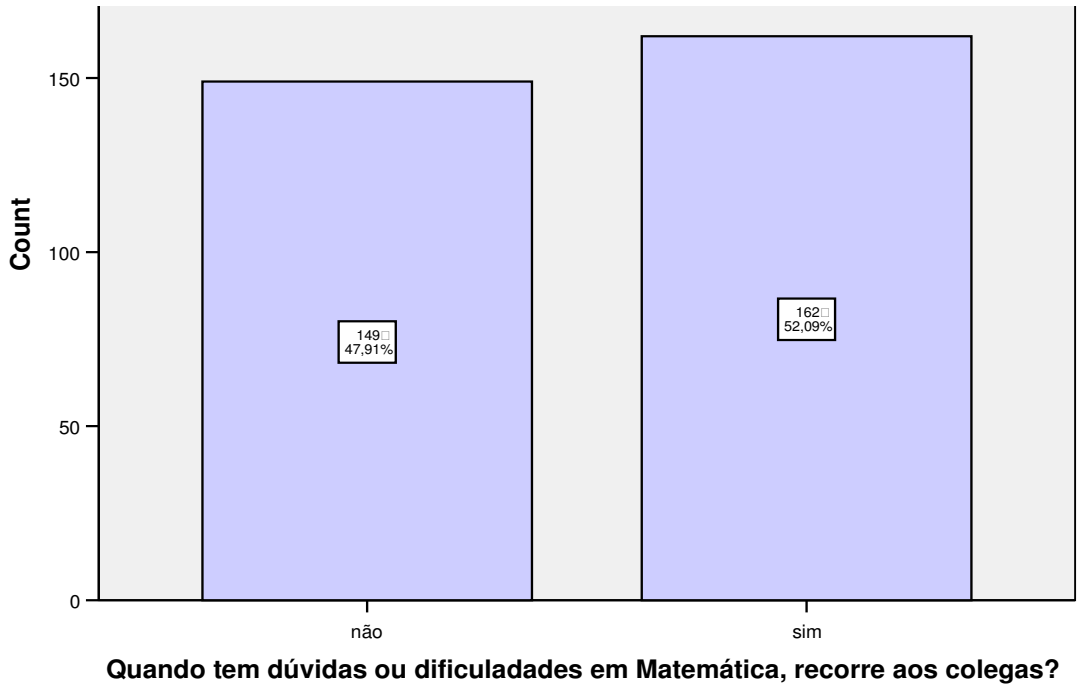


Figura PII.3.13

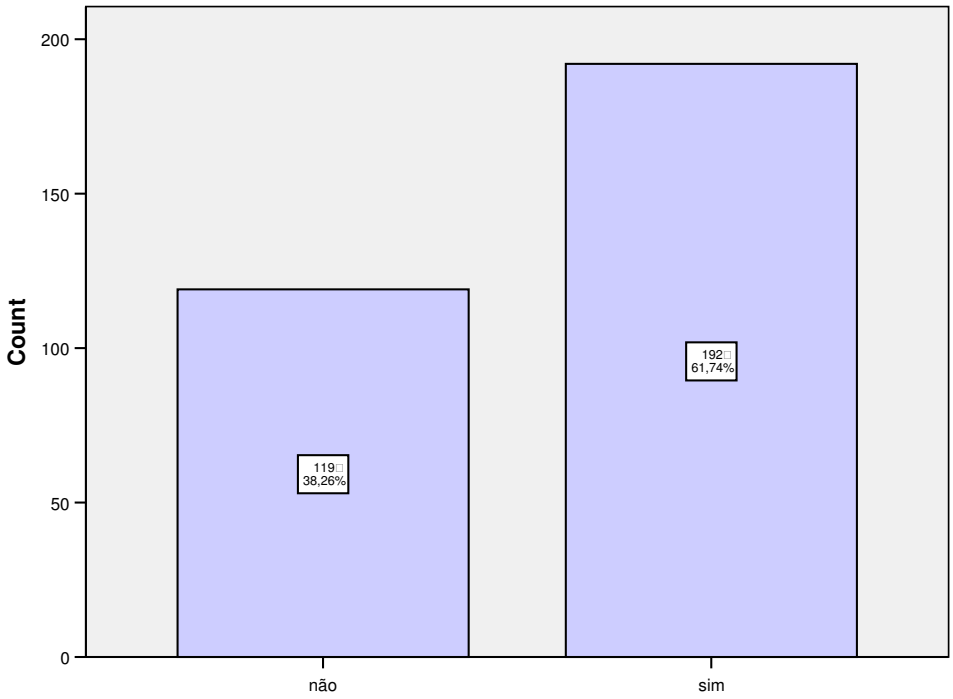


Figura PII.3.14

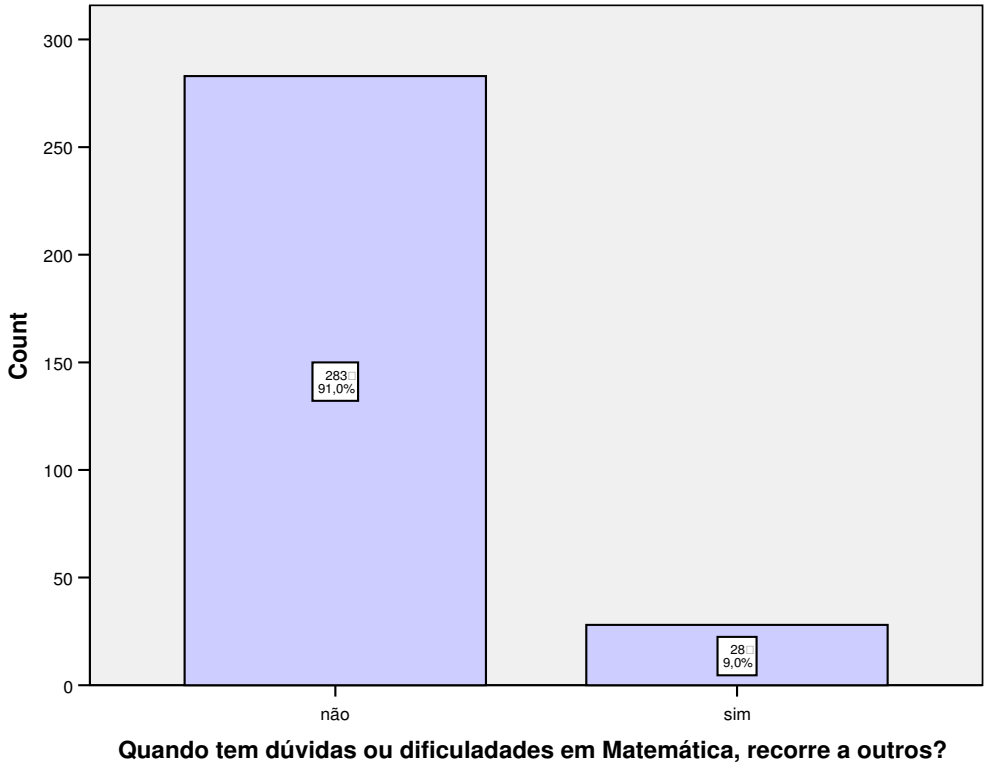


Figura PII.3.15

1.11. Quantas horas de sono dorme em média por noite?

Menos de 7 horas ☐ **entre 7h e 8h ☐** **entre 8h e 9h ☐**

Entre 9h e 10h ☐ **mais do que 10h ☐**

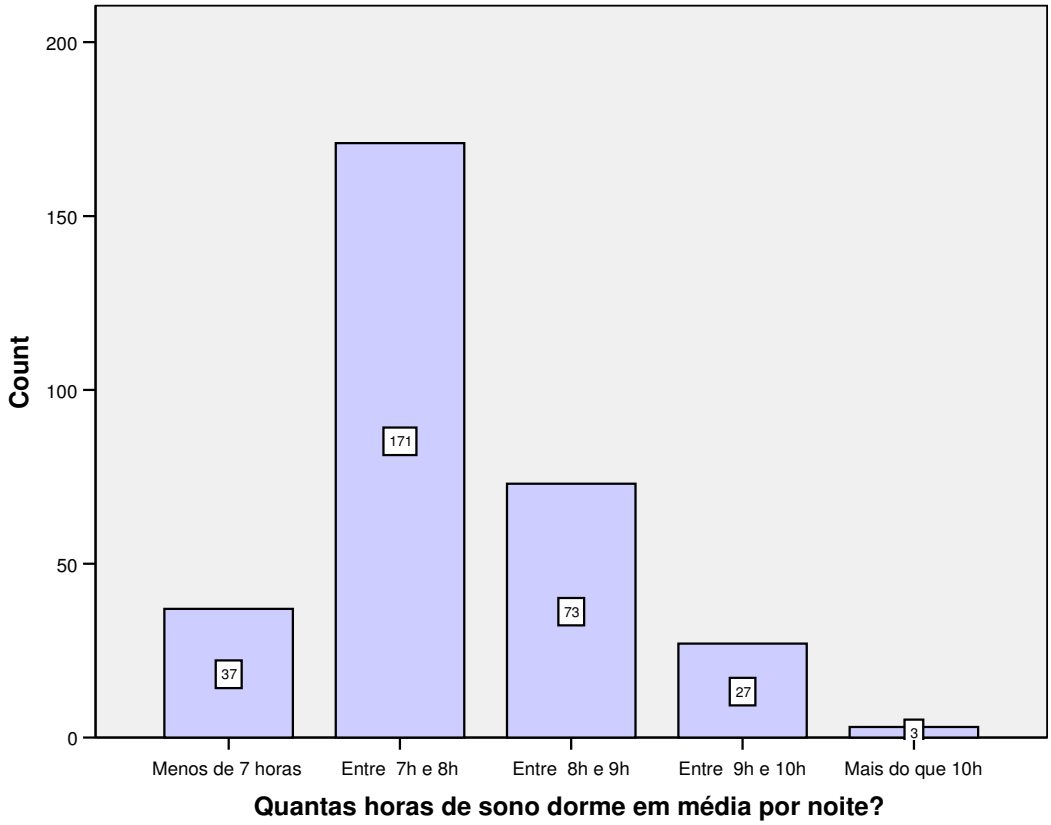


Figura PII.3.16 – horas de sono por noite

Cerca de 10% dos alunos dormem menos de 7 horas, cerca de 50% dormem entre 7 e 8 h.

2.1. Na sua escola existe sala de estudo/oficina de aprendizagem ou aulas de apoio em Matemática?

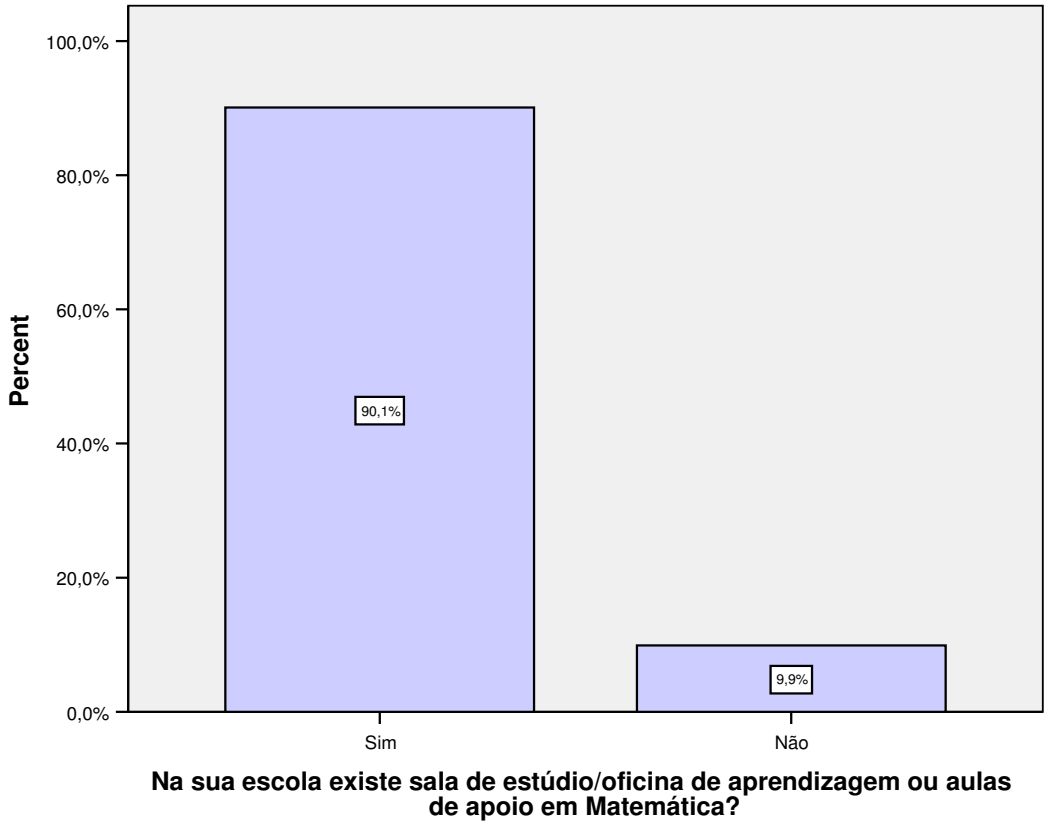


Figura PII.3.17 – Existência ou não de sala de estudo/aulas de apoio

A grande maioria das escolas, 90%, têm sala de estudo ou aulas de apoio em Matemática.

2.2 Gostaria que na sua escola existissem aulas de Apoio ministradas pelo seu professor de Matemática?

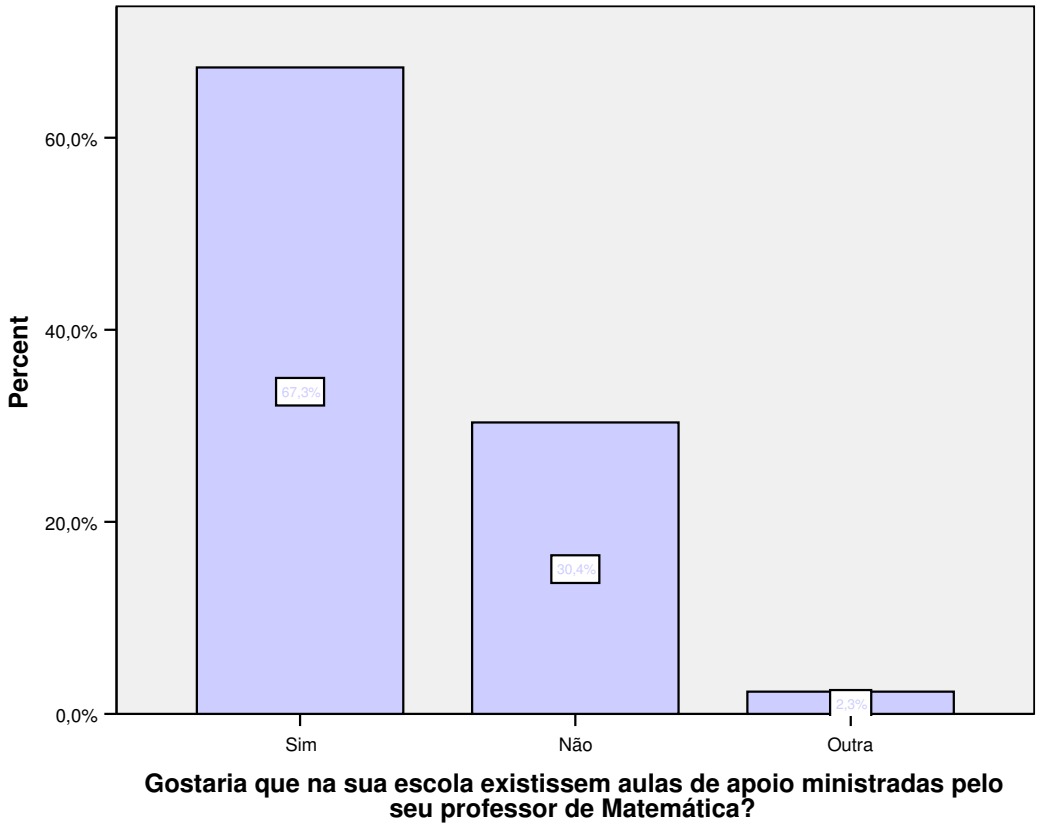
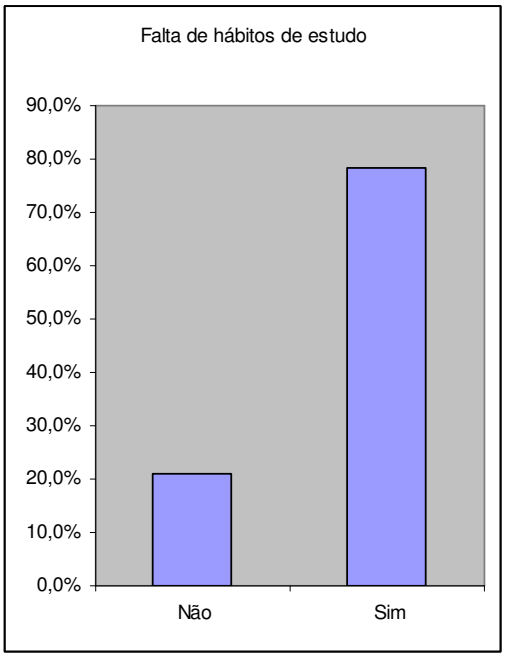


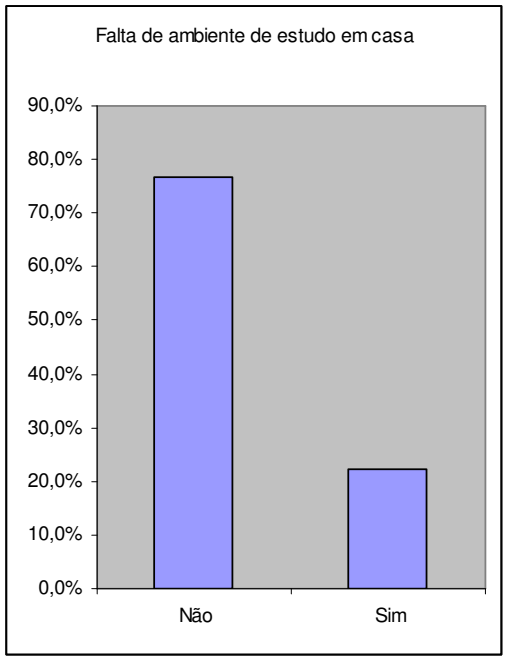
Figura PII.3.18 –

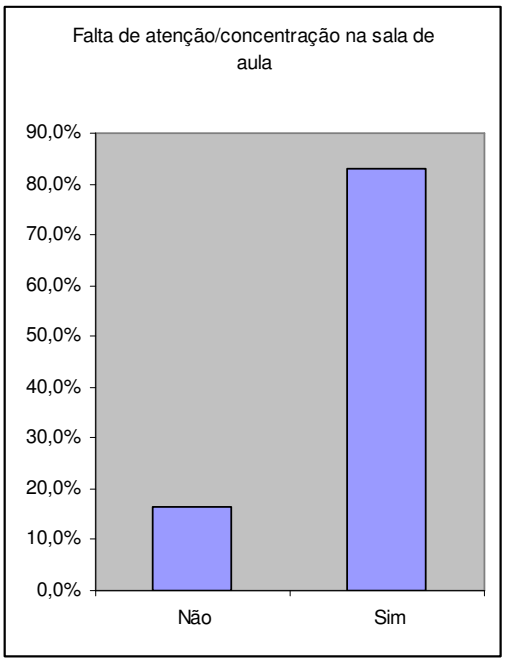
67,3% dos alunos gostariam que na sua escola existissem aulas de apoio ministradas pelo seu professor de Matemática, 30,4% não gostariam e 2,3% responderam que gostariam que fossem tomadas outras medidas.

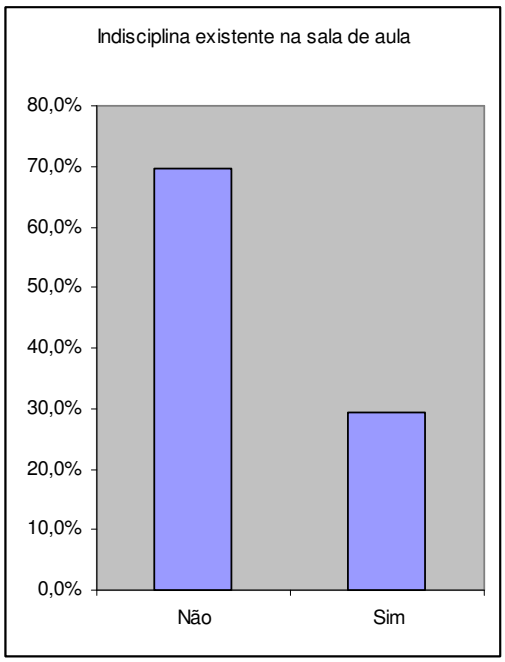
Na pergunta **2.3** deste inquérito tínhamos 10 questões e era pedido ao aluno para indicar, na sua opinião, quais as que contribuíam para o insucesso em Matemática. Para cada uma das alíneas obtiveram-se os seguintes gráficos.

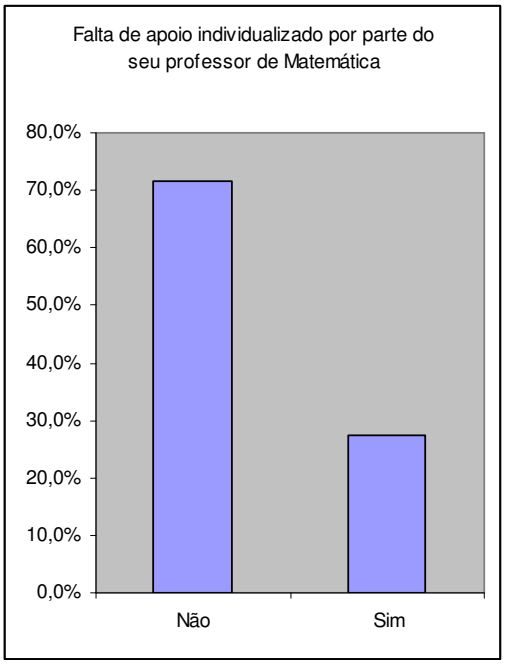
Figura PII.3.19

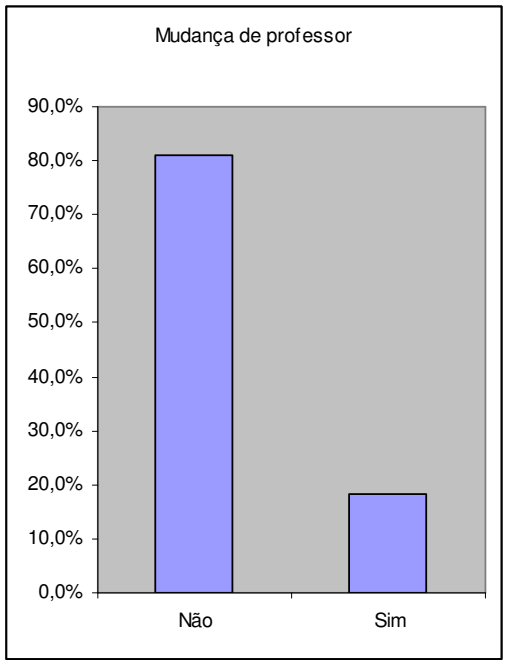

Falta de hábitos de estudo
90,0%
80,0%
70,0%
60,0%
50,0%
40,0%
30,0%
20,0%
10,0%
0,0%
Não
Sim

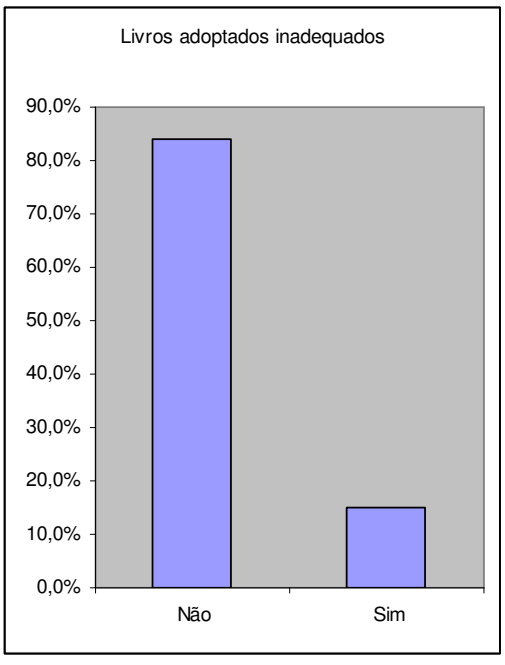

Falta de ambiente de estudo em casa
90,0%
80,0%
70,0%
60,0%
50,0%
40,0%
30,0%
20,0%
10,0%
0,0%
Não
Sim

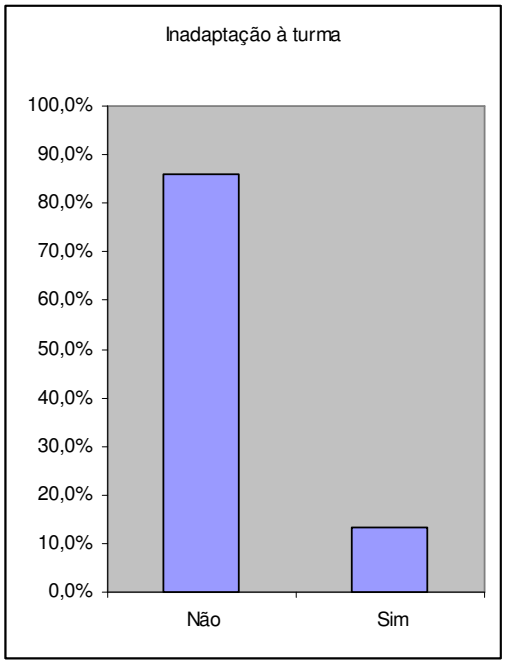

Falta de atenção/concentração na sala de aula
90,0%
80,0%
70,0%
60,0%
50,0%
40,0%
30,0%
20,0%
10,0%
0,0%
Não
Sim

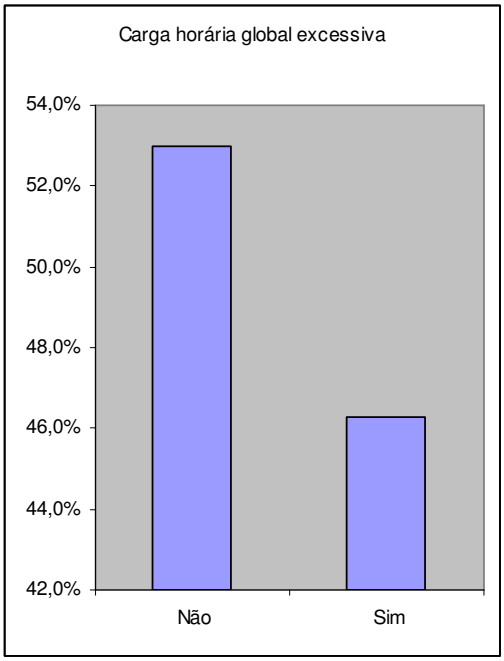

Indisciplina existente na sala de aula
80,0%
70,0%
60,0%
50,0%
40,0%
30,0%
20,0%
10,0%
0,0%
Não
Sim


Falta de apoio individualizado por parte do seu professor de Matemática
80,0%
70,0%
60,0%
50,0%
40,0%
30,0%
20,0%
10,0%
0,0%
Não
Sim


Mudança de professor
90,0%
80,0%
70,0%
60,0%
50,0%
40,0%
30,0%
20,0%
10,0%
0,0%
Não
Sim


Livros adoptados inadequados
90,0%
80,0%
70,0%
60,0%
50,0%
40,0%
30,0%
20,0%
10,0%
0,0%
Não
Sim


Inadaptação à turma
100,0%
90,0%
80,0%
70,0%
60,0%
50,0%
40,0%
30,0%
20,0%
10,0%
0,0%
Não
Sim



Na pergunta **2.4** foi perguntado aos alunos que, entre as 5 hipóteses dadas, enumerassem de 1 até 3 as que consideravam mais importantes para aumentar o sucesso dos alunos em Matemática e diminuir o abandono escolar precoce. A 1ª mais importante, teve a seguinte distribuição: opção 1 corresponde à "Possibilidade de os alunos que têm nota negativa no final do Período terem apoio com o seu professor de Matemática pelo menos duas vezes por semana (em horário extra-aulas); a opção 2 corresponde a - "Turmas serem mais homogéneas (alunos com interesses de aprendizagem idêntico"; a opção 3 corresponde a – "Turmas terem menos alunos"; a opção 4 corresponde a "Serem criadas turmas com currículos alternativos para alunos que estejam muito desinteressados/desmotivados na escola"; a opção 5 corresponde a "Os alunos trabalharem em grupos, pelo menos nalgumas aulas ou no 2º tempo de cada bloco de 90minutos".

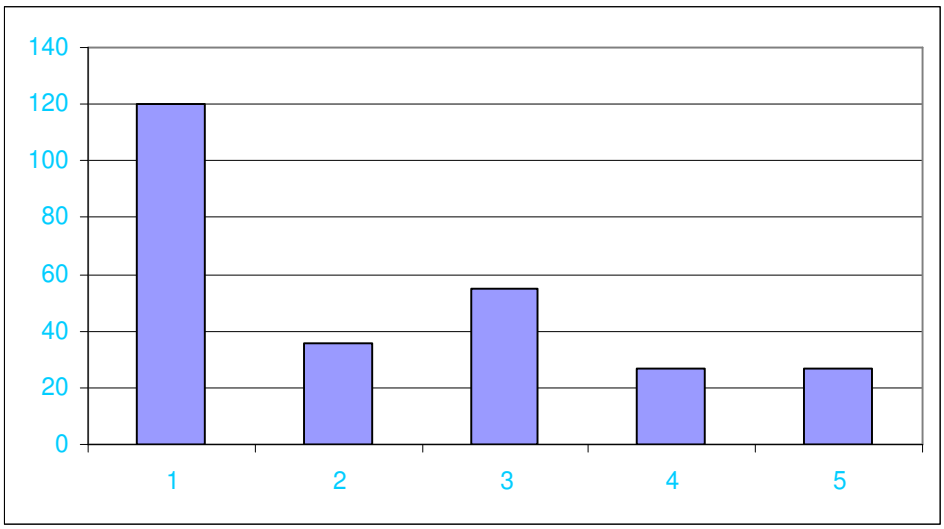


Figura PII.3.20

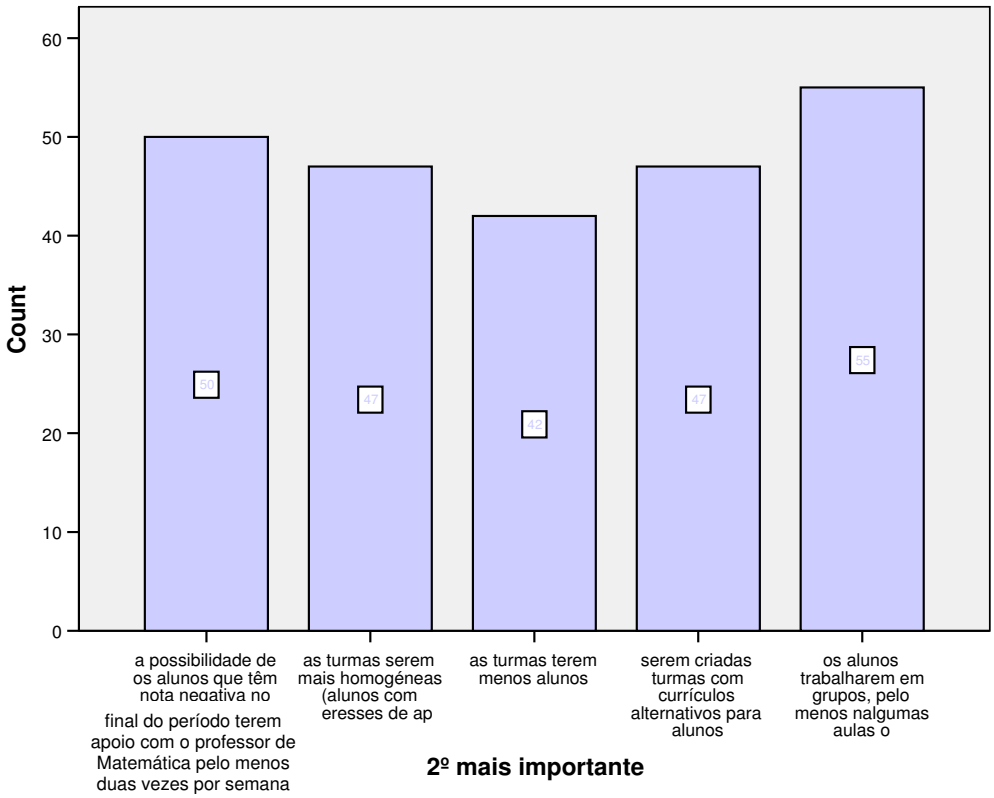


Figura PII.3.21

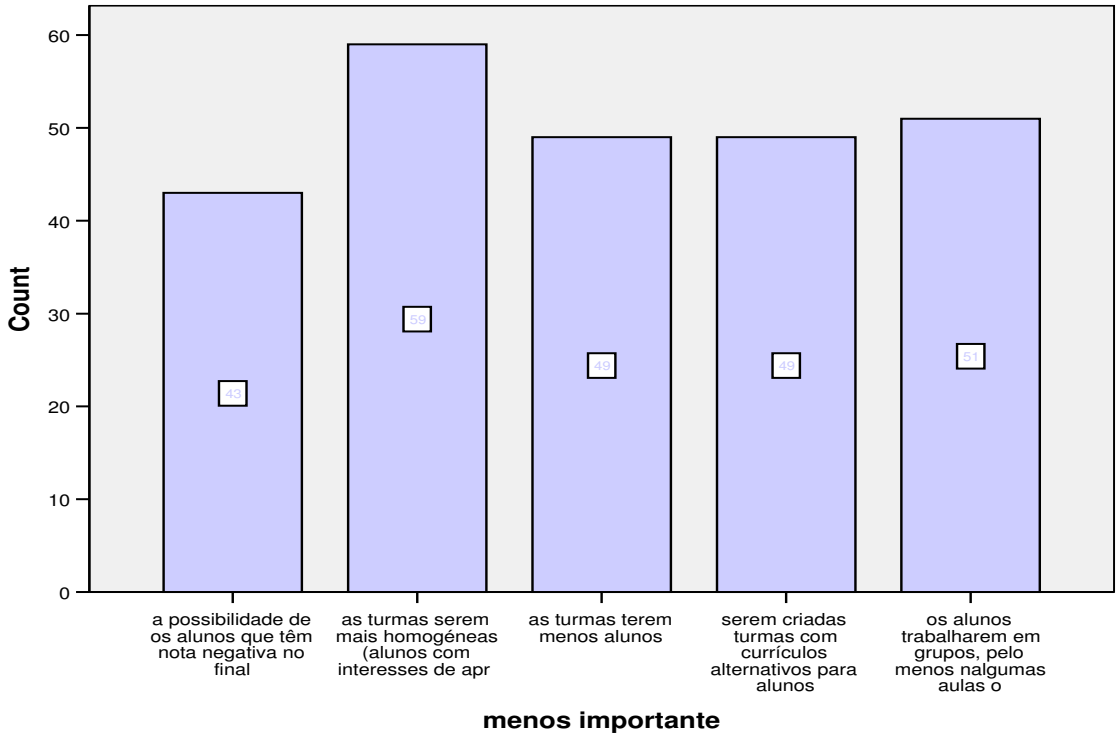
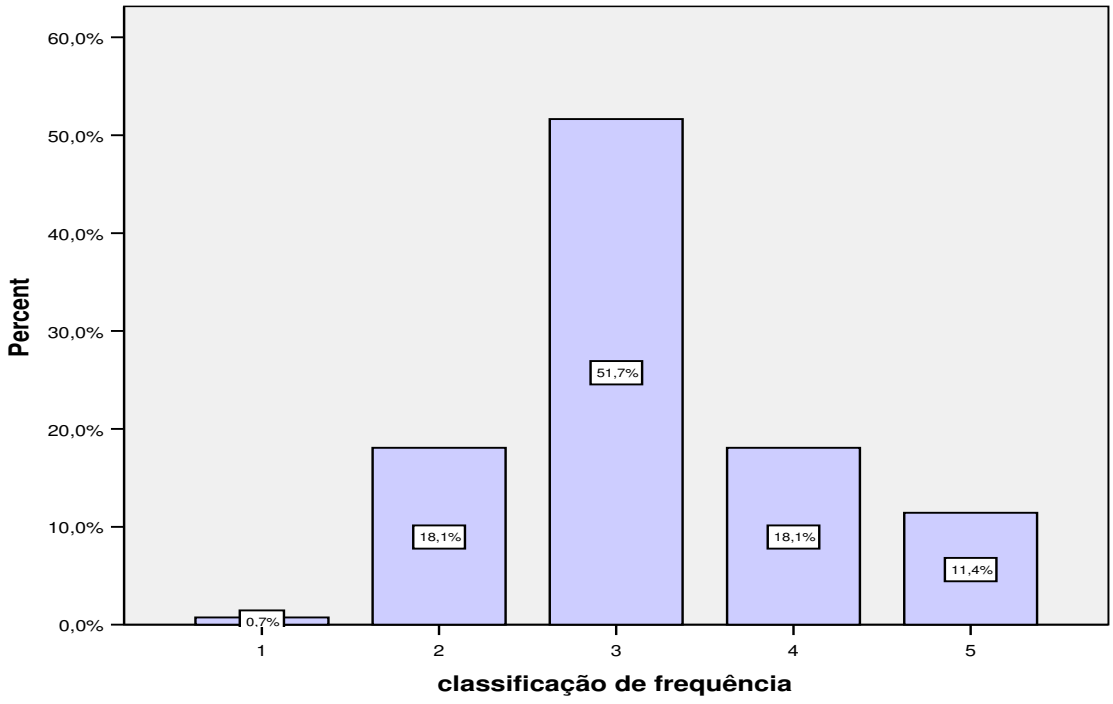


Figura PII.3.22

Com as avaliações que nos foram dadas de cada aluno fizemos os seguintes gráficos:



.

Figura PII.3.23

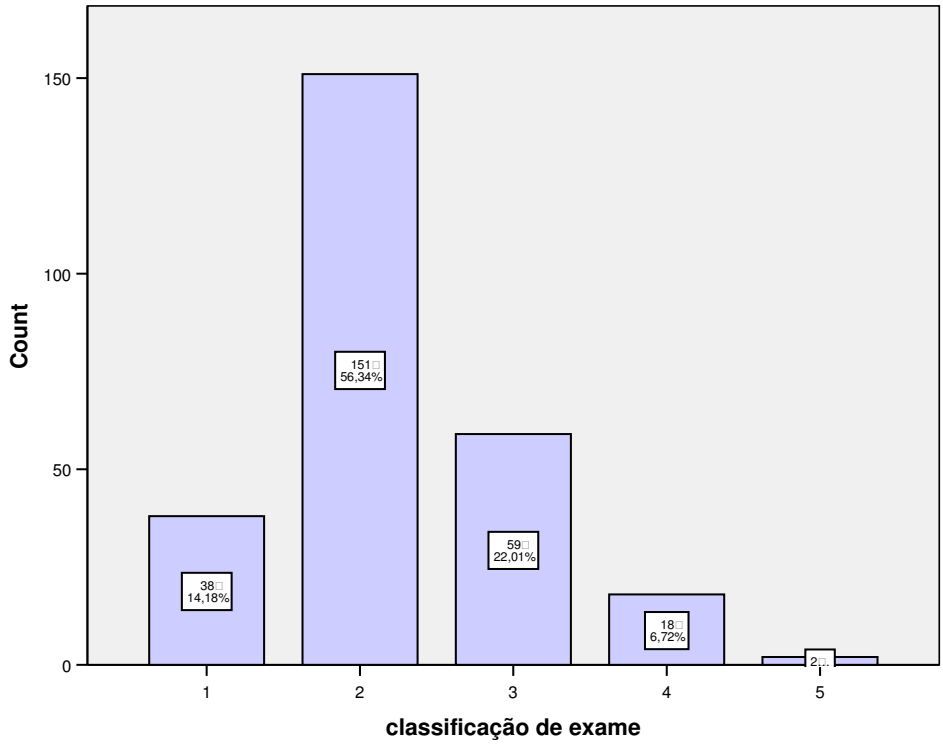


Figura PII.3.24

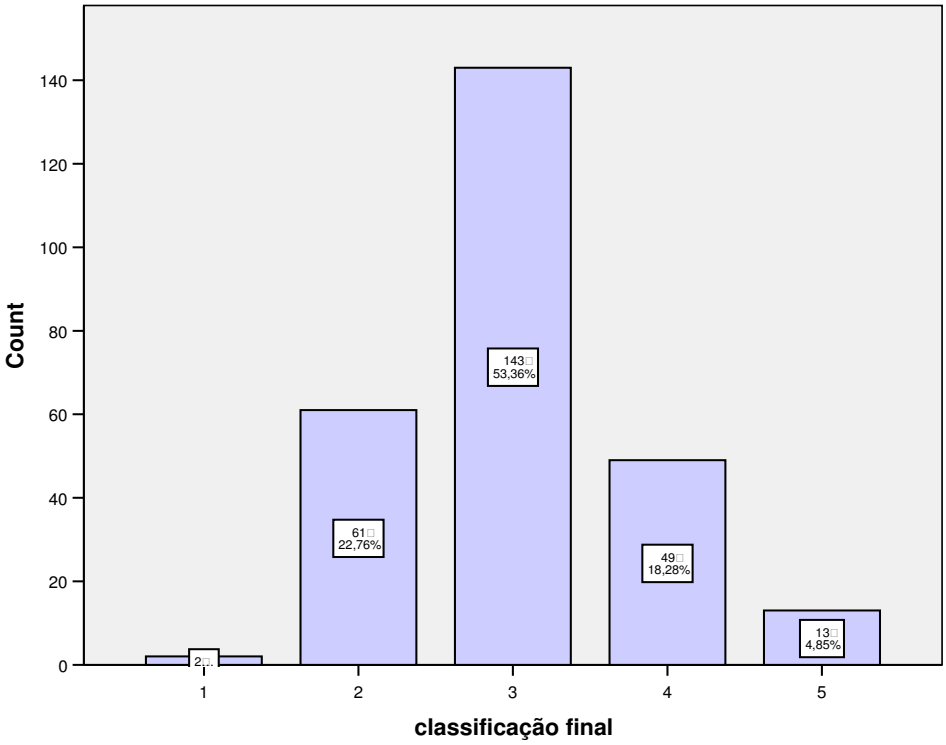


Figura PII.3.25

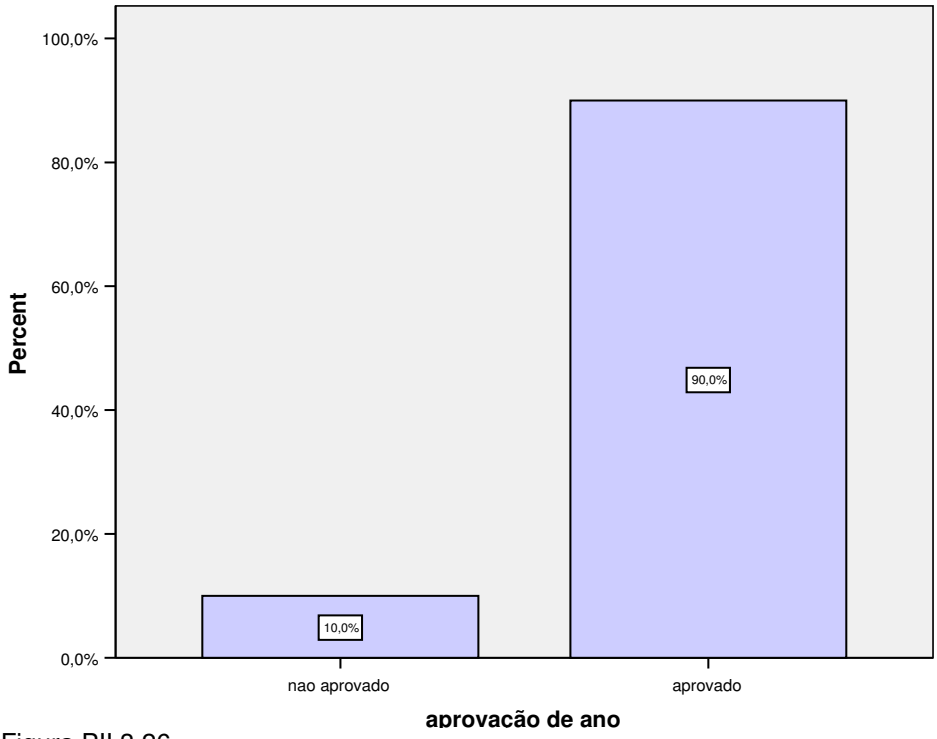


Figura PII.3.26

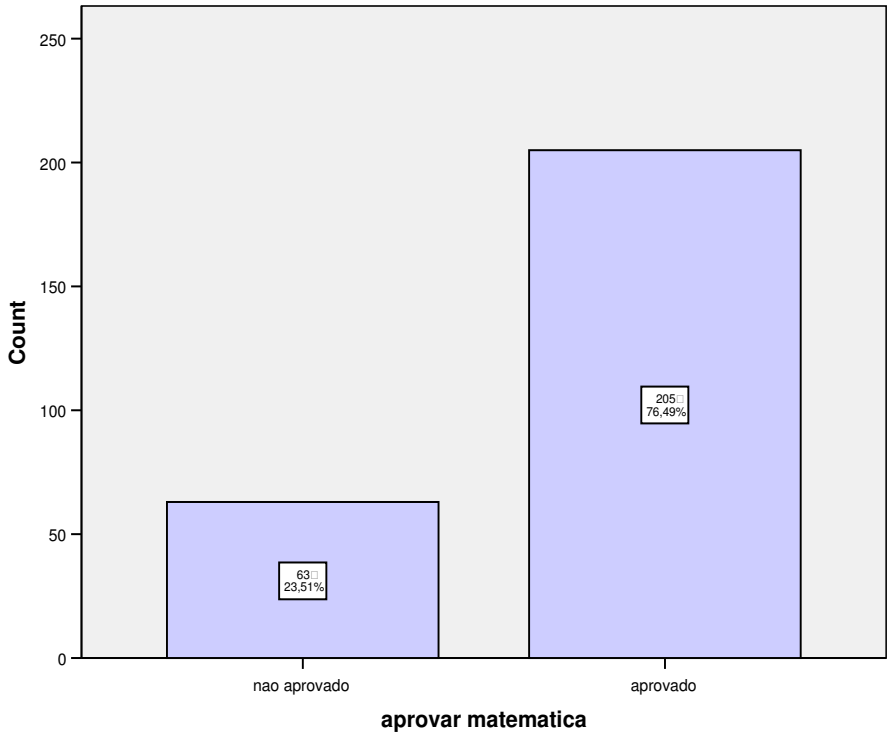


Figura PII.3.27

Da análise dos dados, concluímos que:

- Todos os alunos que tiveram classificação final[18] positiva (maior ou igual a 3), a Matemática foram aprovados no 9º ano de escolaridade.

- Houve alunos que por terem uma classificação de nível 1 no exame de Matemática e uma classificação de frequência de 3, tiveram classificação final de 2, o que fez com que reprovassem, pois já tinham negativa noutras duas disciplinas antes de fazerem exame de Matemática. Assim em alguns casos o resultado do exame aumentou a taxa de insucesso nas escolas (entende-se por sucesso no final do ano lectivo o aluno que foi aprovado no final desse ano).

- Dos alunos que responderam ao inquérito, 0,75% tiveram classificação do exame superior à classificação de frequência, 20,97% tiveram classificação do exame igual à classificação de frequência e 78,28% tiveram classificação do exame inferior à classificação de frequência.

- O aluno que não estuda Matemática tem um risco de sete vezes maior de não ser aprovado do que um aluno que estuda mais de duas horas por semana Matemática, para a amostra observada[19].

- O aluno que estuda 1h por semana Matemática, tem um risco duas vezes maior de não ser aprovado do que um aluno que estuda Matemática mais de duas horas por semana.

- O risco de não transitar de ano vai diminuindo à medida que a habilitação da mãe vai aumentando. Começando com:

[18] Classificação final = 70% da classificação de frequência + 30% da classificação de exame

[19] Estes resultados foram obtidos a partir de uma regressão logística aplicada aos dados da nossa amostra.

.- Um filho de uma mãe com 4ª classe tem um risco de sete vezes maior de não transitar de ano do que um filho de uma mãe licenciada.

4. Utilização de software para a aplicação da Análise de Clusters aos resultados obtidos no inquérito

O *software* utilizado foi o SPSS (Statistical Package for the Social Sciences), versão 13.0.

Houve algumas dificuldades na execução do inquérito, nomeadamente o ter que deixar os inquéritos em certas escolas para posteriormente serem distribuídos aos alunos para responderem, por indicação do Conselho Executivo, e portanto não estar presente na realização dos mesmos; o facto dos alunos estarem no final do ano lectivo e portanto estarem cansados/sobrecarregados de obrigações; por terem que preencher a parte relativa à sua identificação (para que depois conseguíssemos ter acesso às suas classificações de frequência, de exame e final em Matemática) e à minha falta de experiência na execução de inquéritos.

Foram feitos 313 inquéritos. Nas escolas foram escolhidas duas turmas, uma com bons resultados em Matemática no 2º Período, outra com resultados menos bons (na medida do possível).

Escolhemos algumas das variáveis do inquérito baseando-nos nas indicações da Análise Descritiva dos dados para realizarmos a Análise de Clusters.

Não escolhemos todas as variáveis do inquérito porque seriam demasiadas e algumas não se revelaram importantes para a formação de clusters. Poderia dificultar/confundir a interpretação dos mesmos.

4.1 Aplicação de método hierárquico no software SPSS

As variáveis são de diferentes tipos e os softwares de que dispomos, SPSS e Clustan Graphycs8 não aplicam em simultâneo diferentes medidas de semelhança consoante o tipo de variáveis para depois fazer uma combinação destas, como se desejava.

Uma vez que o nosso n era maior do que 200, não é aconselhável a utilização dos métodos hierárquicos, até porque no dendograma que se obtém não se consegue discernir os clusters por estarem sobrepostos.

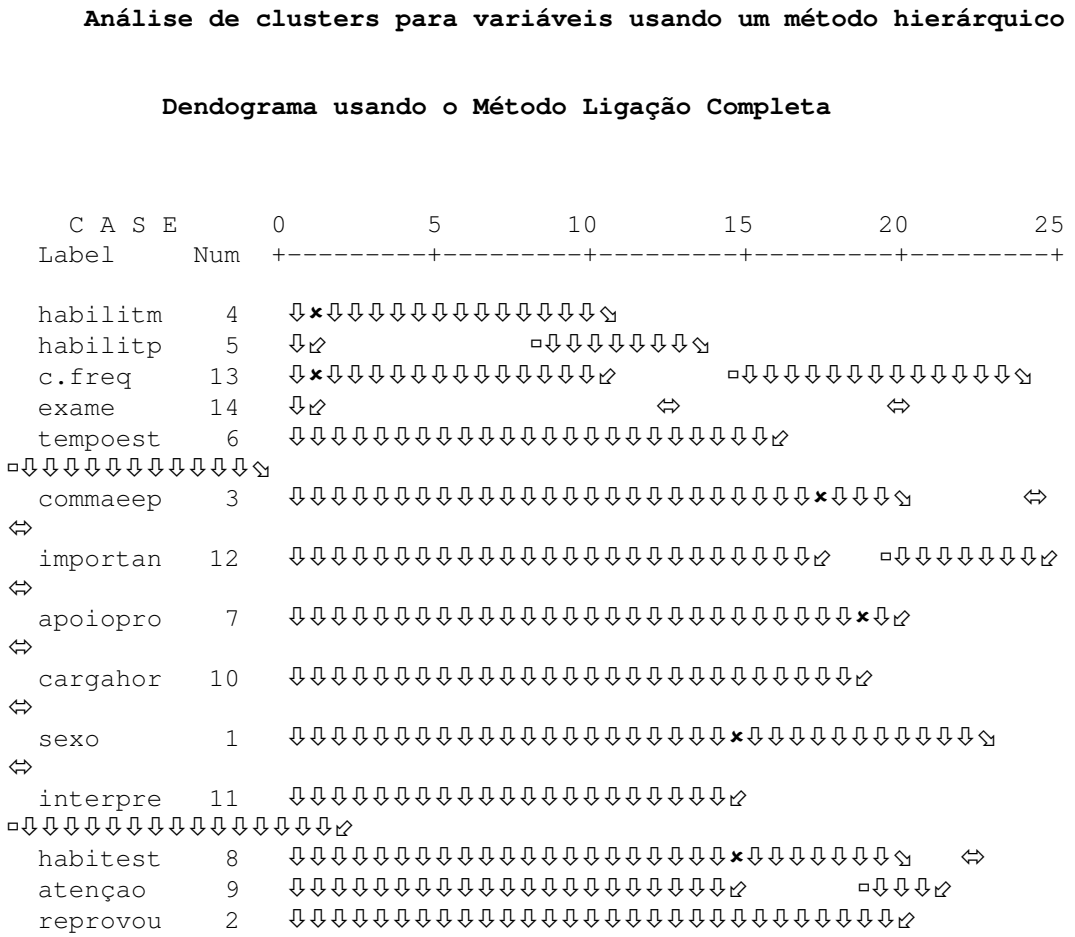
Tendo em conta isto, optou-se por aplicar os métodos hierárquicos mas às variáveis.

A medida de proximidade utilizada foi o coeficiente de correlacção de Pearson, e o método usado foi o de ligação completa.

As variáveis seleccionadas foram: sexo (pergunta 1.2 do inquérito; variável qualitativa binária); idade (pergunta 1.3 do inquérito; variável quantitativa); reprovou (já reprovou?. Pergunta 1.4 do inquérito;variável qualitativa binária); commaeep (reside com mãe e pai?. Pergunta 1.5 do inquérito; variável qualitativa binária); habilitmãe (habilitações da mãe. Pergunta 1.8 do inquérito; variável qualitativa ordinal); habilitpai (habilitações do pai. Pergunta 1.8 do inquérito;variável qualitativa ordinal); tempoestud (tempo semanal de estudo em Matemática. Pergunta 1.9 do inquérito; variável qualitativa ordinal); apoioprof (gostaria que na sua escola existissem aulas de apoio ministradas pelo seu professor de Matemática?. Pergunta 2.2 do inquérito; variável qualitativa ordinal); habitest - falta de hábitos de estudo, atenção - falta de atenção/concentração na sala de aula, interpretaçã - problemas de interpretação, cargaho - carga horária global excessiva, são quatro variáveis qualitativas binárias da pergunta 2.3 do inquérito; importante1 (enumemerar de 1 até 3; as 5 razões apontadas, sendo o 1 o mais importante; o que na opinião do aluno se deverá fazer para aumentar o

sucesso dos alunos em Matemáticamais importante. Pergunta 2.4 do inquérito; variável qualitativa ordinal) e as variáveis: c. freq (classificação de frequência) e exame (classificação de exame*).*

Obtivemos o seguinte dendograma:



Nome abreviado das variáveis	Nome das variáveis
apoiopro	apoio do professor
cargahor	carga horária global excessiva
commaeep	reside com mãe e pai
habilitm	habilitação da mãe
habilitp	habilitação do pai
importan	importante1

```
interpre          dificuldades de interpretação
tempoest          tempo de estudo
atenção           falta de atenção/concentração
habitest          falta de hábitos de estudo
tempoest          tempo semanal de estudo
reprovou          reprovou algum ano
c.freq            classificação de frequência
exame             classificação de exame
```

- As variáveis mais fortemente relacionadas são: a variável classificação de frequência com a variável classificação de exame ; e a variável habilitação da mãe com a variável habilitação do pai.

- Podemos considerar três grupos de variáveis para classificar os jovens:

No **Grupo 1** verificamos que há forte associação entre as habilitações da mãe e do pai assim como entre classificações de frequência e exame. Há uma posterior associação entre estes dois pares ao tempo de estudo.
Verificamos que as habilitações dos pais estão muito relacionadas com as notas que os alunos têm a matemática.

No **Grupo 2** existe associação entre o facto de o aluno viver com a mãe e o pai com aquilo que o aluno considera mais importante fazer para aumentar o sucesso dos alunos em Matemática; assim como, o considerar a carga horária global excessiva e o pretender ter aulas de apoio com o professor, caso tivesse negativa em Matemática. Estes associar-se-ão posteriormente para formar este grupo.

No **Grupo 3** há agregação, ao mesmo nível, entre os alunos pelo sexo (masculino ou feminino) à variável "dificuldades de interpretação"; assim como, "falta de hábitos de estudo" com a "falta de atenção". Este último par associar-se-á à variável " reprovou"; que por sua vez agregar-se-á ao 1º par mencionado neste grupo.

4.2 Aplicação do método não hierárquico k-means, no software SPSS

Foi utilizado o *software* SPSS, versão 13.0 para aplicar o **método não hierárquico "k-means"**, à base de dados obtida do inquérito aos alunos de 9º ano sobre Sucesso/Insucesso em Matemática.
Pretende-se com este estudo encontrar clusters (grupos) dos objectos escolhidos, que foram *"classificação final"* em Matemática. A escolha desta variável tem origem no resultado.

As variáveis seleccionadas foram: sexo; idade; reprovou (já reprovou?); commaeep (reside com mãe e pai?); habilitmãe (habilitações da mãe); habilitpai (habilitações do pai); tempoestud (tempo semanal de estudo em Matemática); apoioprof (gostaria que na sua escola existissem aulas de apoio ministradas pelo seu professor de Matemática?); habitest (falta de hábitos de estudo); atenção (falta de atenção/concentração na sala de aula); interpretaçã (problemas de interpretação); importante1 (a mais importante, das 5 razões apresentadas, se deverá fazer para aumentar o sucesso dos alunos em Matemática?).

Foram retirados os alunos cuja identificação não estava correcta, ou seja os alunos que não tinham classificações devido a não possuirmos identificação destes.

No SPSS, os procedimentos foram: Analyse→Classify→K-Means Cluster … 3 clusters.

Escolhemos 3 clusters, tendo em conta os resultados obtidos nos métodos hierárquicos aplicado às variáveis.

Obtivemos as seguintes tabelas:

Tabela 8 - Iterações

Iteration	Change in Cluster Centers		
	1	2	3
1	2,997	3,738	3,791
2	,136	,286	,313
3	,134	,286	,150
4	,052	,197	,143
5	,064	,189	,073
6	,069	,118	,000
7	,045	,088	,032
8	,000	,101	,089
9	,028	,050	,000
10	,000	,000	,000

Nesta tabela é apresentada a indicação da variação do centro dos clusters em cada passo da iteração. O algoritmo termina no 9º passo no cluster 3; termina no 10º passo nos clusters 1 e 2.

O algoritmo termina quando não há uma variação significativa dos centróides após a atribuição dos objectos pelos 3 clusters pedidos.

A tabela seguinte, permite-nos identificar o cluster a que cada objecto pertence. Podemos verificar para cada caso, isto é, para cada aluno, a classificação final em Matemática assim com o cluster a que pertence e a distância ao centro do cluster, ou quão semelhante é cada observação ao centro do respectivo cluster.

Tabela 9 – Membros dos clusters

Case Number	classificação final em Matemática	Cluster	Distance
1	3	1	3,166
2	3	3	1,494
3	5	3	1,886
4	3	3	2,281
5	4	2	3,191
6	4	3	2,449
7	3	2	2,812

A continuaçaõ desta tabela está no Anexo II deste trabalho.

- Por exemplo, podemos verificar da tabela anterior que o aluno correspondente ao inquérito 7, teve classificação final 3 em Matemática, pertence ao cluster 2 e este objecto está a uma distância de 2,841 do centro do respectivo cluster.

- Como verificamos na tabela anterior os alunos são separados em três clusters.
  - No cluster 1 há 68% de positivas e 32% de negativas
  - No cluster 2 há 79% de positivas e 21% de negativas
  - No cluster 3 há 91% de positivas e 9% de negativas

  - No cluster 1 não existem alunos com nota final 5, no cluster 2 apenas 2% dos alunos têm 5, enquanto que no cluster 3 15% dos alunos tiveram nível 5.

Como podemos verificar, a taxa de sucesso dos alunos, no cluster 3 é muito superior no cluster 3, que corresponde aos alunos que têm

pais com habilitações mais altas, curso médio ou apenas alguns anos de Universidade (ver valores dos centros dos clusters, tabela 10).

Na tabela seguinte podemos observar o centro final de cada cluster em cada variável.

Tabela 10 – Centros finais dos clusters

	Cluster		
	1	2	3
sexo	2	2	1
Já reprovou?	1	1	1
reside com mãe e pai	1	1	1
habilitação da mãe	3	4	6
habilitação do pai	2	4	6
tempo semanal de estudo	2	2	2
gostaria que na sua escola houvessem aulas de apoio ministradas pelo seu prof Mat	1	2	1
falta de hábitos de estudo	1	1	1
falta de atenção/concentração na sala de aula	1	1	1
carga horária global excessiva	0	1	0
problemas de interpretação	1	1	1
o que se deverá fazer em1º lugar para aumentar o sucesso em Mat	2	3	2

➢ Em relação ao centro final dos clusters verificamos que, o centro final dos clusters é idêntico no que respeita a:

- residir com mãe e pai (corresponde ao 1 da tabela);
- tempo semanal de estudo (corresponde ao 2 da tabela, 1h de estudo semanal);
- falta de hábitos de estudo (corresponde ao 1 da tabela);
- falta de atenção/concentração na sala de aula (corresponde ao 1 da tabela);
- problemas de interpretação (corresponde ao 1 da tabela).
- no cluster 2, o centro final do cluster reprovaram (corresponde ao 2 da tabela) e não reprovaram nos restantes clusters(corresponde ao 1 da tabela).

➢ Por outro lado os centros dos clusters diferem no que se refere:

- às habilitações dos pais, pois:
- no cluster 1, os centros finais do cluster, é Básico-2ºciclo (corresponde ao 3 da tabela), para a mãe e Básico-4ª ano, para o pai, (corresponde ao 2 da tabela). (ver esta correspondência, na pergunta 1.8 do inquérito em anexo);
- no cluster 2, o centro final do cluster, é 3º Ciclo (corresponde ao 4 da tabela) para a mãe e para o pai;
- no cluster 3, o centro final do cluster, é superior-curso médio ou apenas alguns anos da Universidade (corresponde ao 6 da tabela) para a mãe e para o pai;

- à carga horária é considerda excessiva, no cluster 2 (corresponde ao 1 da tabela) e não é considerada excessiva nos clusters 1 e 3 (corresponde ao 0 da tabela);
- o que se deve fazer em 1º lugar para aumentar o sucesso, opção 2 do inquérito no cluster 1 e 3 (as turmas serem mais homogéneas (alunos com interesses de aprendizagem idênticos) e opção 3 no cluster 2 (as turmas terem menos alunos)
- sexo, masculino nos clusters 3 (corresponde ao 1 da tabela) e feminino nos clusters 1 e 2 (corresponde ao 2 da tabela);

- Uma vez que os centros dos clusters são idênticos no que se refere à falta de hábitos de estudo, problemas de atenção/concentração, tempo semanal de estudo e dificuldades de interpretação, leva-nos a concluir que as causas principais de insucesso escolar estão associadas a outros factores tais como habilitações dos pais e critérios de avaliação pouco exigentes, entre outrros que não foram abordados neste estudo tais como ambiente familiar e social em que o aluno está inserido. Não esqueçamos que, no Ensino Básico, um aluno transita com três negativas, podendo até transitar com quatro ou mais negativas, se o Conselho de Turma assim o decidir ou se o Encarregado de Educação discordar da decisão do Conselho de Turma em anos não terminais de ciclo, como 5º, 7º e 8º anos. Além disso, o aluno que está na escolaridade obrigatória não perde por faltas (pois se ultrapassar o limite de faltas, a decisão se o aluno perderá ou não por faltas será do Conselho Pedagógico da Escola). Castigos muito brandos para alunos que infringem gravemente as regras e deveres dos alunos (baseadas na legislação em vigor).

Tudo isto tem feito com que os alunos sintam que "podem fazer o que querem" e no final de ano transitarão de ano.

Tabela 11 – Distâncias entre os centróides dos clusters finais

Cluster	1	2	3
1		2,638	5,201
2	2,638		3,122
3	5,201	3,122	

Por exemplo, o centróide do cluster 2 está a uma distância de 2,638, 3,122 dos centróides dos clusters 1 e 3, respectivamente.

A maior distância entre os centróides é verificada entre os centróides dos clusters 1 e 3 enquanto que a menor distância entre estes é verificada entre os centróides dos clusters 1 e 2.

Tabela 12 - ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
sexo	,256	2	,249	234	1,029	,359
Já reprovou?	2,347	2	,213	234	11,025	,000
reside com mãe e pai	,123	2	,138	234	,891	,412
habilitação da mãe	312,894	2	1,122	234	278,845	,000
habilitação do pai	246,968	2	1,116	234	221,305	,000
tempo semanal de estudo	4,103	2	,705	234	5,824	,003
gostaria que na sua escola houvessem aulas de apoio ministradas pelo seu prof Mat	1,496	2	,344	234	4,344	,014
falta de hábitos de estudo	,348	2	,161	234	2,165	,117
falta de atenção/concentração na sala de aula	,662	2	,125	234	5,301	,006
carga horária global excessiva	,799	2	,244	234	3,270	,040
problemas de interpretação	,885	2	,230	234	3,848	,023
o que se deverá fazer em1º lugar para aumentar o sucesso em Mat	30,225	2	1,662	234	18,189	,000

Com a tabela da ANOVA podemos identificar quais as variáveis que permitem a separação dos clusters. Se uma variável descriminar bastante entre os clusters, o quadrado médio, QMC, do cluster há-de ser elevado. Pelo contrário dentro do cluster essa variabilidade (dada pelo Quadrado médio do erro, QME há-

de ser pequena. Portanto as variáveis que mais contribuem para a definição dos clusters são aquelas com maior Quadrado médio do cluter e menor Quadrado médio do erro, ou seja aquela com maior valor de $F = QMC / QME$. Portanto as variáveis que mais contribuiem para a descriminação entre os clusters são as variáveis*: habilitação do pai e habilitação da mãe.*

Tabela 13 - Número de casos em cada cluster

Cluster	1	108,000
	2	61,000
	3	68,000
Valid		237,000
Missing		,000

Dos 237 casos válidos, os alunos estão separados pelos três clusters, sendo o cluster 1 o maior, constituido por 108 classificações finais de alunos. Os clusters 2 e 3 têm 61 e 68 casos, respectivamente.

- Temos que ter em conta a possibilidade da existência de alguns erros, tais como os alunos não terem respondido de forma sincera e correcta, o que poderá dificultar a conclusão.

  No entanto, com os dados que conseguimos apurar, a percentagem de insucesso em Matemática é de 32%, 21% e 9% nos clusters 1,2,3; respectivamente.

  Portanto, a taxa de sucesso é de 68%, 79% e 91% nos clusters 1,2,3; respectivamente.

  No cluster 3 há mais 23% de sucesso do que no cluster 1 e mais 12% do que no cluster 2.

  Como podemos constatar, é no cluster 3 que estão os alunos com pais com mais habilitações literárias, sendo o centro final deste cluster, 6,

correspondente a pais com curso médio (bacharelato) ou alguns anos da Universidade. Aliás, como vimos na tabela 12 –ANOVA, são as habilitações dos pais que mais contribuem para a separação dos clusters.

5. Conclusão

Provamos que faz sentido aplicar uma Análise de Clusters para estudar e interpretar a informação relativa ao sucesso/insucesso a Matemática no 9º ano, contida na base de dados.

A formação de clusters é um desafio interessante e útil. Tem a capacidade de recuperar e filtrar a informação após fazermos escolhas cuidadosas das técnicas a usar.

Com uma amostra que consideramos representativa dos estudantes de 9º ano do Funchal, provámos que uma Análise de Clusters permitiu constatar que de uma maneira geral, os estudantes estudam pouco, têm dificuldades de atenção/concentração, dificuldades de interpretação.

Leva-nos a concluir que, as causas principais do insucesso escolar não estão apenas concentradas nos estudantes, mas estão principalmente associadas a outros factores tais como habilitações dos pais e critérios de avaliação pouco exigentes baseados nas leis em vigor; porque apesar dos alunos estudarem pouco em todos os grupos, alguns deles conseguem obter boas classificações finais em Matemática.

Portanto devia-se investir mais na formação/habilitações das famílias, em horário pós-laboral; porque por um lado aumentava a qualificação dos trabalhadores e aproveitava-se o trabalho dos professores que estão com “horário zero” nas escolas ou até mesmo criando postos de trabalho; por outro lado, o aumento das habilitações dos pais ia desencadear um melhor desempenho profissional por parte destes e faria com que estes dessem apoio mais consistente aos seus filhos, quer a nível científico, quer a nível pedagógico. Com isto, ganhávamos todos nós, ganhava a sociedade em geral, enfim ganhava o país.

Sugestões para futuras investigações:

- Aplicar a Análise de Clusters a bases de dados obtidas em inquéritos a alunos de toda a Região Autónoma da Madeira noutros níveis de ensino, como por exemplo ao 12º ano.
- Estudar/desenvolver/criar *software* que aplique as técnicas de análise de clusters a uma base de dados com medidas de proximidade diferentes (mas combinadas) quando os dados são de natureza diferente.

Bibliografia

- Agresti, 1981; measures of nominal-ordinal associaton. *Journal of the American Statiscal Association*, **76,** 524-529.
- Anderberg Michael, Boris; 1973, *Cluster Analysis for applicatons*, Academic Press, London.
- Arabie P., Hubert L.J. e De Soete G.; 1996, *Clustering and Classification*, Kluwer Academic Publishers, London.
- Baker, F. B. and Hubert, L. J., 1975, Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, **70,** 31-38.
- Belbin, L.1987, The use of non-hierarchical allocation methods of clustering large sets of data. *Australian Computer Journal*, **19,** 32-41.
- Branco, João; 2004, *Uma Introdução à análise de clusters*, Sociedade Portuguesa de Estatística.
- Bryant Peter; 1991, Large-Sample Results for Optimization-Based Clustering Methods. *Journal of Classification*, **8**, 31-44.
- Bryant Peter; 1988, On Characterizing Optimization-Based Clustering Methods. *Journal of Classification*, **5**, 81-84.
- Bryant Peter; 1978, Asymptotic Behaviour of Classification. *Biometrika*, vol 65,nº **2**, 273-281.
- Burbank, F.; 1972, A sequencial space–time cluster analysis of câncer mortality in the United States: etiological implications. *American Journal of Epidemiology*, **95**, 393-417.
- Correia, Luis; 1983, *Escala de Comportamento Escolar*, Porto Editora
- Everitt Brian, Landau, S. e Leese M.; 2001, *Clusters Analysis*, 4th Ed., Arnold, Londres.

- Everitt, Brian S. e Dunn, G.; 1991, *Applied Multivariate Data Analysis*, Arnold, London.
- Everitt, Brian S. e Dunn, G.; 1982, *An introduction to mathematical taxonomy*"Analysis", Arnold, London.
- Fleiss, L. L. aand Zubin, J., 1969, On the methods and Theory of clustering. *Multivariate Behavioral Research*, **4,** 235-250.
- Fraley C. And Raftery A., 1998, How many clusters? Which clustering method? Answers via model-based cluster analysis, *The Computer Journal*, vol. **41**, nº **8**, 578-588.
- Friedman H.P. e Rubin J., 1967, On Some Invariant Criteria for Grouping Data, *Journal of the American statistical Association*, **62,** 1159-1178.
- Gabinete de Avaliação Educacional (gave), 2006, *Reflexão dos Docentes do 3º Ciclo sobre os Resultados do Exame do 9º ano 2005 1ª chamada,* Editorial do Ministério da Educação, Lisboa.
- Gabinete de Avaliação Educacional (gave), 2006, *Resultados do Exame de Matemática do 9º ano 2005 1ª chamada,* Editorial do Ministério da Educação, Lisboa.
- Gnanadesikan, R., 1997, *Methods of Statistical Data Analysis of Multivariate Observations*, John Wiley & Sons
- Gower J. e Legendre P., 1986, Metric and Euclidean properties of dissimilarity coefficients, *Journal of Classification*,**3** ,5-48
- Milligan G e Cooper M., 1985, An examination of procedures for determing the number of clusters in a data set. *Psycometrika*, vol 50, nº **2,** 159-179.
- Milligan G e Cooper M., 1988, A Study of Standardization of variables in Cluster Analysis, *Journal of Classification*, nº **5,** 181-204.

- Jain, A.K. e Dubes, R.C. 1988, *Algoritms for Clustering Data,* Englewood cliffs, NJ, Prentice Hall
- Jobson, J.D.; 1991,*Clusters analysis,* Vol II*, Springer-Verlag*
- Jobson, J.D.; 1991, *Applied MultivariateData Analysis: categorical and multivariate methods,* Vol II, Springer
- K.V. Mardia, J.T. Kent and J.M.Bibby, 1979, *Multivariate Analysis*, Academic Press Limited
- Kaufman Leonard; Rousseeuw, Peter J., 1990, *Finding Groups in Data ,An Introduction to Cluster Analysis*, Wiley Inter-science, Canadá
- Lance G. N. e Williams W.T., 1967, A general Theory of Classificatory sorting strategies: 1. Hierarchical systems. 2 *Computer journal*, **9**, 373-380.
- Maroco, João; 2003, *Análise Estatística com utilização do SPSS,* Edições Sílabo, Lisboa.
- Marriot F.H.C., 1971, Practical Problems in a Method of Cluster Analysis, *Biometrics*, **27,** 501-514.
- Mirkin, Boris; 1996, *Mathematical Classification and Clustering*, Kluwer Academic Publishers, London.
- Pestana Mª Helena, Gageiro N.João; 2000, *Análise de dados para Ciências Sociais - a complementaridade do SPSS*, Edições Sílabo.
- Pestana, Dinis e Velosa, Sílvio; 2002, *Introdução à Probabilidade e Estatistica*, Edição da Fundação Calouste Gulbenkian.
- Petrakis, E.G.M. e Faloutsos, C.; 1997, Similarity searching in medical image databases. *IEEE Transactions on Knowledge and Data Engineering*, **9,** 435-447.
- Price, L. J.; 1993, Identifying cluster overlap with NORMIX population membership probabilities. *Multivariate Behavioral Research*, **28**, 235-262.

- Reis, Elisabete; 2001, *Estatística Multivariada Aplicada*, 2ª edição revista e corrigida, Edições Sílabo.
- Rasmussen and Willet, 1989, Efficiency of hierarchic agglomerative clustering using the ICL distributed Array Processor. *Journal of Documentation*, **45**, (1),1-24.
- Romesburg, H. Charles; 1990, *Cluster Analysis for Researchers*, Krieger Publishing Company, Florida.
- Scott A.J. e Symons M. J.; 1971, Clustering Methods based on Likelihood Ratio Criteria. *Biometrics*, nº **27**, 387-397.
- Smart, R.G., Asbridge, Mann R.E. e Adlaf, E.; 2003, Psychiatric distress among road rage victims and perpetrators. *Canadian Journal of Psychiatrry*, **48**, 681-688.
- Sokal and Sneath, 1963; *Principles of Numeric Taxonomy*. Freeman, London.
- Symons M. J.; 1981, Clustering Criteria and Multivariate Normal Mixtures. *Biometrics*, nº **37**, 35-43.
- Vasconcelos, Rita; 1994, *Contribuição à análise de dados categorizados*, Universidade da Madeira (Tese de Doutoramento).
- Williams and Lambert, 1959, Multivariate methods in plant ecology, 1. Associationanalysis in plant communities, *Journal of Ecology*, **47**, 83-101.
- Wishart, D., 1999, ClustanGraphics3: Interactive graphics for cluster analysis, in *Classification in information Age* (W. Gaul and H. Locarek-Junge, eds), pp. 268-275. Springer-Verlag, Berlin.

ANEXO I - Inquérito

As respostas deste inquérito são totalmente **confidenciais** e destinam-se apenas à realização de um estudo que se está a fazer no âmbito de uma dissertação de Mestrado em Matemática/Ensino na UMa sobre "Análise de clusters e sua aplicação no Sucesso /Insucesso em Matemática".

Assinale apenas com uma cruz, x, em cada pergunta com¤; com excepção da questão 1.10 que pode assinalar mais do que um x.

No início do inquérito é pedido o B.I. para que possamos ter acesso à nota final do 3º Período em Matemática e à nota do exame nacional nesta disciplina.

1- Dados pessoais

1.1. Bilhete de identidade_____

1.2. Sexo: Masculino ☐ Feminino ☐

1.3. Idade :_____anos

1.4. Já reprovou? Não ☐ Sim ☐

Se sim, quantos anos reprovou ?_____

Nos anos que reprovou, teve nota negativa em Matemática?_____

1.1. Com quem reside?

Mãe e pai ☐ Sem a mãe ☐ Sem o pai ☐

Avós ☐ Outros_____

1.2. Tem bom relacionamento com quem reside?

Sim ☐ Não ☐ Às vezes ☐

1.3. Como classificaria o seu ambiente familiar:

Muito Bom ☐ Bom ☐ Razoável ☐

Mau ☐ Muito mau ☐ Outro_____

1.8. Habilitações dos pais (mesmo se já falecidos) ou pessoa com quem reside

.Habilitações	Mãe	Pai	Outro_____
1. nenhum nível de ensino			
2. Básico-1º ciclo/4ªclasse			
3. Básico - 2º ciclo			
4. Básico - 3º ciclo			
5. Secundário			
6. Superior - Curso médio ou apenas alguns anos da Universidade(bacharelato)			
7. Superior - Licenciatura			
8. Superior -mestrado			
9. Superior - doutoramento			
10. Informação desconhecida			

1.1. Tempo semanal de estudo em Matemática (T.P.C. e estudo da matéria dada)

0h ☐ 1h ☐ 2h ☐ mais do que 2h ☐

1.2. Quando tem dúvidas ou dificuldades em Matemática, a quem recorre para esclarecê-las?

às pessoas com quem reside ☐ colegas ☐

ao professor de Matemática ☐ outros_____

1.3. Quantas horas de sono dorme em média por noite?

Menos de 7 horas ☐ Entre 7 h e 8 h ☐ Entre 8 h e 9h ☐

Entre 9h e 10 h ☐ Mais do que 10 h ☐

2. Dados escolares

2.1. Na sua escola existe sala de estudo/oficina de aprendizagem ou aulas de apoio em Matemática?

Sim ☐ Não ☐ Outra_____

2.2. Se na sua escola existissem aulas de Apoio, gostaria que estas fossem ministradas pelo seu professor de Matemática?

Sim ☐ Não ☐ Outra_____

2.3. Assinale com um X os factores principais, que na sua opinião contribuem para o insucesso em Matemática:

a) falta de hábitos de estudo ☐

b) falta de ambiente de estudo em casa ☐

c) falta de atenção/concentração na sala de aula ☐

d) indisciplina existente na sala de aula ☐

e) falta de apoio individualizado por parte do seu professor de Matemática ☐

f) mudança de professor ☐

g) livros adoptados inadequados ☐

h) inadaptação à turma ☐

i) carga horária global excessiva ☐

g) problemas de interpretação ☐

2.1. Enumere de 1 até 3 (1 corresponde ao mais importante e 3 corresponde ao menos importante) o que na sua opinião se deverá fazer para aumentar o sucesso dos alunos em Matemática?

_____ a possibilidade de os alunos que têm nota negativa no final do Período terem apoio com o seu professor de Matemática pelo menos duas vezes por semana (em horário extra-aulas)

_____ as turmas serem mais homogéneas (alunos com interesses de aprendizagem idênticos)

_____ as turmas terem menos alunos

_____ serem criadas turmas com currículos alternativos para alunos que estejam muito desinteressados/desmotivados na escola

_____ os alunos trabalharem em grupos, pelo menos nalgumas aulas ou no 2º tempo de cada bloco de 90minutos

Obrigada pela sua colaboração!

ANEXO II - Tabela de membros dos clusters

Case Number	classificação final em Matemática	Cluster	Distance
1	3	1	3,166
2	3	3	1,494
3	5	3	1,886
4	3	3	2,281
5	4	2	3,191
6	4	3	2,449
7	3	2	2,812
8	4	3	2,121
9	4	2	3,165
10	3	3	2,134
11	4	2	2,555
12	4	3	1,798
13	5	3	1,940
14	3	3	1,413
15	4	3	2,413
16	3	2	2,073
17	5	3	1,562
18	3	1	1,680
19	3	2	2,612
20	2	2	3,718
21	2	1	2,608
22	2	1	2,242
23	1	2	2,159
24	2	1	2,250
25	1	2	3,001
26	2	1	1,377
27	2	1	1,680
28	3	2	2,711
29	3	3	1,697
30	4	1	2,702
31	2	1	1,969
32	4	2	3,092
33	3	2	3,963
34	2	2	2,998
35	3	2	2,323
36	3	1	2,633
37	3	2	3,230
38	3	2	2,389
39	2	1	2,225
40	4	2	2,073

41		4	3	3,686
42		3	1	2,295
43		2	1	3,510
44		2	3	2,514
45		3	1	1,755
46		2	1	2,374
47		2	1	1,960
48		2	1	2,753
49		3	3	2,057
50		3	3	3,572
51		3	1	2,038
52		2	1	1,718
53		3	1	1,988
54		4	3	2,759
55		3	2	2,979
56	2		1	2,551
57		3	2	2,577
58	3		3	2,640
59		4	3	2,107
60	2		3	2,651
61		3	1	1,892
62	3		3	2,461
63		3	2	2,864
64		3	3	3,176
65		2	2	1,989
66	3		3	2,617
67		3	3	2,700
68	3		1	1,858
69		4	3	2,514
70	4		3	3,535
71		5	3	2,345
72		2	3	2,281
73	3		3	4,759
74	3		2	1,680
75		4	1	1,812
76		3	1	3,041
77		3	2	2,143
78		3	2	2,628
79		2	1	1,696
80		2	1	1,173
81		3	3	2,388
82		3	2	2,320
83	4		2	2,500
84		3	1	2,629
85	3		1	1,766
86		3	1	1,739
87		4	2	3,378

88	3	1	1,960
89	3	1	2,213
90	3	3	3,166
91	2	1	2,015
92	3	2	3,209
93	3	3	2,196
94	4	3	3,115
95	5	3	1,757
96	3	3	2,437
97	3	2	2,895
98	4	1	2,303
99	3	2	2,995
100	2	2	2,618
101	2	2	2,085
102	3	3	2,169
103	3	3	1,444
104	4	1	2,366
105	2	2	2,069
106	4	3	2,182
107	3	3	2,216
108	2	2	2,711
109	3	1	2,485
110	3	1	1,787
111	3	2	2,113
112	3	1	1,969
113	4	3	1,688
114	3	3	1,580
115	3	2	4,177
116	2	2	2,376
117	4	2	3,509
118	3	3	2,079
119	4	3	2,029
120	4	3	2,920
121	4	3	3,497
122	5	2	2,696
123	3	3	1,878
124	3	2	3,152
125	4	3	1,909
126	3	1	1,652
127	5	3	4,321
128	3	1	1,707
129	4	3	2,388
130	5	3	3,106
131	4	3	2,014
132	5	3	3,726
133	3	3	2,743
134	3	3	3,737

135	3	1	1,960
136	4	1	2,145
137	3	2	2,747
138	4	1	2,540
139	2	1	1,853
140	2	2	4,247
141	2	1	2,945
142	3	1	2,149
143	2	1	2,200
144	2	2	2,671
145	3	2	1,368
146	3	1	2,709
147	3	1	3,473
148	2	1	1,536
149	2	1	2,836
150	4	1	1,760
151	3	1	2,029
152	3	1	1,462
153	2	1	1,685
154	2	1	2,079
155	3	1	1,499
156	3	1	2,208
157	3	2	3,448
158	3	1	1,945
159	4	1	1,843
160	4	1	2,713
161	3	2	3,500
162	4	1	2,629
163	4	2	2,846
164	3	3	3,272
165	3	3	1,338
166	3	1	3,101
167	3	2	2,889
168	4	1	2,002
169	3	1	2,923
170	3	3	3,617
171	3	3	2,566
172	3	1	1,771
173	4	3	2,209
174	3	1	2,157
175	2	3	2,128
176	4	2	2,116
177	4	3	1,552
178	3	2	3,001
179	5	3	1,814
180	3	2	2,587
181	3	3	4,335

182	3	3	2,235
183	2	1	1,589
184	3	1	1,560
185	2	1	3,287
186	3	1	1,511
187	3	1	3,427
188	2	3	2,363
189	3	1	1,843
190	2	1	3,001
191	3	1	2,661
192	2	1	1,669
193	3	1	3,264
194	3	1	1,227
195	3	1	2,149
196	2	1	2,682
197	3	1	2,826
198	4	1	2,374
199	2	1	2,192
200	3	1	1,978
201	3	2	2,671
202	2	1	1,960
203	4	2	2,618
204	5	3	1,932
205	3	1	2,262
206	4	1	1,462
207	3	1	1,955
208	4	2	2,779
209	3	1	1,474
210	3	2	3,818
211	3	2	2,128
212	3	1	1,945
213	3	2	2,313
214	3	1	2,757
215	3	1	1,149
216	3	1	1,853
217	3	1	1,734
218	3	1	1,635
219	2	2	2,464
220	3	1	2,709
221	2	1	1,921
222	3	1	2,287
223	3	1	2,671
224	2	3	2,262
225	2	1	3,166
226	3	1	1,505
227	2	1	3,074
228	3	1	2,157

229	3	1	1,853
230	2	2	2,995
231	3	2	2,646
232	3	1	2,536
233	3	3	2,287
234	3	1	2,101
235	2	1	1,536
236	3	2	2,631
237	2	1	1,423

Este trabalho de dissertação foi co-financiado por:

CITMA
CENTRO DE CIÊNCIA E TECNOLOGIA DA MADEIRA

PROGRAMA OPERACIONAL
PLURIFUNDOS DA REGIÃO
AUTÓNOMA DA MADEIRA

UNIÃO EUROPEIA
FUNDO SOCIAL EUROPEU