Universidade Aberta

Métodos multivariados para variáveis qualitativas:

aplicação ao estudo de variáveis associadas com a avaliação na disciplina de Matemática de uma escola do Ensino Básico no Concelho de Vila Nova de Gaia

Dina Maria Duarte Cabrita

Mestrado em Estatística, Matemática e Computação

Especialidade em Estatística Computacional

Dissertação orientada pela Professora Doutora Maria do Rosário Ramos

2012

Resumo

Nesta dissertação aprofundou-se o estudo de metodologias de Estatística Multivariada, nomeadamente a Análise de Correspondências Múltiplas e a Análise de *Clusters*. Aplicou-se a Análise de Correspondências Múltiplas sobre os dados de um inquérito realizado em meio escolar, com o objetivo de investigar associações entre o desempenho dos alunos do 3º Ciclo na disciplina de matemática e um conjunto de variáveis referentes aos alunos, aos encarregados de educação e aos professores. A leitura dos resultados da Análise de Correspondências Múltiplas permitiu observar configurações definidas pelas categorias das variáveis. Estas configurações refletem a presença de grupos relativamente homogéneos com perfis distintos. Partindo dos resultados da Análise de Correspondências Múltiplas, com o propósito de definir os grupos e saber o peso de cada um deles na amostra, usou-se a Análise de *Clusters*. Assim, foi efetuada a articulação Análise de Correspondências Múltiplas/Análise de *Clusters* para definir grupos homogéneos de estudantes relativamente à avaliação na disciplina de Matemática e aos hábitos de estudo. A solução foi comparada com os grupos obtidos sobre as variáveis originais. Para validar o número de grupos foram explorados diversos algoritmos de agrupamento.

Palavras chave: *Optimal Scaling*; Análise de Correspondências; Análise de Correspondências Múltiplas; Análise de *Clusters*, Desempenho na disciplina de Matemática

Abstract

In this dissertation the study of methodologies of Multivariate Statistics was deepened, namely Multiple Correspondence Analysis and Cluster Analysis.

Multiple Correspondence Analysis was applied to data of a survey conducted in a school community, with the aim to study relationship between the students´ final results in Mathematics and a set of variables connected with the students themselves, their parents and or tutors and the teachers.

The study of the results of Multiple Correspondence Analysis allowed us to observe configurations defined by categories of variables. These configurations reflect the presence of relatively homogeneous groups with distinct profiles.

Based on the results of Multiple Correspondence Analysis, in order to define the groups and know how important they were in the sample, we made use of Cluster Analysis.

In this way, Multiple Correspondence Analysis and Cluster Analysis were articulated with the purpose to define groups that are homogeneous in terms of success/failure and of study habits. The solution was compared with the one obtained with original variables. To validate the number of groups it was used more than one clustering algorithm.

Key Words: Optimal Scaling*;* Correspondence Analysis; Multiple Correspondence Analysis; Cluster Analysis; Performance in the Mathematics course.

Agradecimentos

À minha orientadora Professora Doutora Maria do Rosário Olaia Duarte Ramos, pelos seus conselhos, opiniões e sugestões.

Aos meus colegas pela ajuda na aplicação dos questionários.

Aos alunos pela colaboração na recolha de dados.

À Teresa pela ajuda na introdução dos dados no SPSS.

Aos meus amigos pela tolerância da minha ausência.

À Raquel pela minha falta de disponibilidade, um pedido de desculpa.

À minha mãe, pelo incentivo, o meu obrigada. A ela dedico esta dissertação.

Índice

Índice de Tabelas

Índice de Figuras

Índice de Gráficos

INTRODUÇÃO

Investigadores de diferentes áreas defrontam-se frequentemente com um problema na análise de dados, quando as variáveis em estudo são de natureza qualitativa ou categórica e se pretende utilizar técnicas estatísticas mais complexas.
Neste sentido, a investigação em técnicas estatísticas e o desenvolvimento de ferramentas (*software*) tem sido constante nas décadas mais recentes.

O enfoque temático deste trabalho é a análise multivariada de dados, nomeadamente os métodos Análise de Correspondências Múltiplas[1] (ACM) e Análise de *Clusters*, com vista a uma aplicação cujos dados são de natureza predominantemente qualitativa.

A ACM tem por objetivo descobrir possíveis associações entre as variáveis de um espaço multidimensional. Esta técnica permite resumir um grande número de variáveis qualitativas, num menor número de variáveis quantitativas, facilitando o estudo das relações entre as diversas caraterísticas existentes num determinado espaço de análise.
Assim, os resultados obtidos com a ACM podem ser utilizados como ponto de partida de técnicas de análise de dados que utilizam variáveis quantitativas, tal como a Análise de *Clusters*.

A Análise de *Clusters* tem como principal objetivo o agrupamento de casos (sujeitos ou variáveis) com base numa ou mais propriedades comuns. Esta técnica tenta formar subgrupos homogéneos (*clusters*) de forma que dentro de um *cluster* a variabilidade dos casos seja mínima, em termos dos seus valores num conjunto de variáveis (propriedades), e que entre *clusters* a variabilidade dos casos seja máxima.

Neste trabalho utilizamos o método de ACM articulado com a Análise de *Clusters* na análise de dados de um inquérito por questionário realizado em meio escolar. Este estudo teve como objetivos:

[1] Também conhecido por Análise de Homogeneidade

a) Identificar as perceções dos alunos sobre o contexto escolar, sobre o seu envolvimento no mesmo e sobre as suas perspetivas de prosseguimento de estudos;
b) Verificar as associações entre o desempenho dos alunos na disciplina de Matemática e um conjunto de variáveis;
c) Procurar grupos de alunos utilizando mais do que um critério e comparar resultados.

Iniciamos a análise multivariada dos dados das respostas aos questionários com a aplicação da Análise de *Clusters*, utilizando a técnica *TwoStep*, tendo como objetivo identificar agrupamentos naturais de indivíduos. Obtivemos uma solução de dois *Clusters* que no contexto do desempenho dos alunos na disciplina de Matemática, pensamos ser pouco informativo e até não fazer sentido na realidade. Assim, realizamos um novo estudo recorrendo a uma outra abordagem que é realizar previamente uma ACM.

Por via da concretização da ACM, procedemos à análise do espaço de desempenho dos alunos, considerando múltiplos indicadores categorizados. Como resultado obtivemos a identificação de diferentes perfis de desempenho.

No sentido de validar a solução sugerida pelo plano da ACM foi aplicado um método hierárquico segundo dois critérios de agregação: *Ward e distância média entre Clusters (average linkage between groups).*

Seguidamente realizamos a formação de grupos de alunos afetos aos vários perfis. Para isso, procedemos ao agrupamento dos indivíduos a partir das duas dimensões encontradas na ACM. Estas dimensões entraram como variáveis originais para a realização da Análise de Clusters, utilizando um método não hierárquico (*k-means cluster*) e novamente o método *TwoStep*.

A articulação ACM/Análise de *Clusters* permitiu, assim, passar da configuração topológica (conhecida a partir dos resultados da ACM) à definição de tipologias.

A estrutura desta dissertação organiza-se em seis capítulos, os quais contemplam a fundamentação teórica deste trabalho e a aplicação das técnicas estudadas aplicadas a dados reais.

No capítulo 1 fazemos uma breve descrição do procedimento *Optimal Scaling*.

No capítulo 2 caraterizamos a Análise de Correspondências, dando enfase à formulação matemática dos conceitos inerentes a esta técnica.

No capítulo 3 abordamos a ACM, esta é apresentada como uma generalização da Análise de Correspondências Simples para o caso de múltiplas variáveis.

No capítulo 4 descrevemos a Análise de *Clusters*, referindo os algoritmos de agrupamento e algumas medidas de distância.

No capítulo 5 desenvolvemos a componente prática deste trabalho, percorrendo as várias etapas desde a caraterização do questionário até à realização da componente de análise estatísstitica.

Por último, no capítulo 6 apresentamos algumas conclusões e notas finais.

1. OPTIMAL SCALING

Na investigação, na área das ciências sociais e comportamentais, a maior parte das variáveis utilizadas são qualitativas (nominais ou ordinais), com medições registadas em escalas com uma unidade de medida incerta. A relação entre as diferentes categorias é muitas vezes desconhecida, e embora com frequência se possa assumir que as categorias são ordenadas, as suas distâncias podem ainda ser desconhecidas. Por outro lado, o número de observações, o número de variáveis e os seus níveis e o grau de complexidade da análise pode dificultar o trabalho do investigador. Este problema pode ser evitado recorrendo à quantificação das variáveis qualitativas.

Nos métodos de análise multidimensionais têm-se verificado um importante desenvolvimento na atribuição de valores (ótimos) quantitativos a escalas qualitativas. Esta forma de quantificação ótima – *Optimal scaling* - é uma abordagem geral para a análise multivariada de dados qualitativos.

O método *Optimal scaling* baseia-se na atribuição de quantificações numéricas para as categorias de cada variável permitindo revelar, de forma visual, bidimensional, através de gráficos (*plots*), as relações entre as variáveis.

Com os procedimentos usados para quantificar os dados, os objetos (indivíduos) e as categorias passam a ter novas coordenadas, designadas respetivamente por *scores* e *centroid coordinates.*

A relação entre as quantificações e as categorias originais pode ser observada graficamente (*Transformation Plots).* Estes gráficos são particularmente adequados para verificar a perfomance do nível de quantificação escolhido.

As referidas quantificações têm a propriedade de guardarem e resumirem as características essenciais das categorias e objetos de análise da matriz de *input* dos dados originais.

Vários métodos de análise de dados, como por exemplo, Análise de Correspondências, ACM e a Análise de Componentes Principais, integram os procedimentos *Optimal scaling*. Estes métodos são geralmente encarados como métodos de redução da complexidade dos dados.
No presente trabalho será abordado o método ACM.

A ACM designa por objeto cada caso ou unidade da análise. A representação gráfica da quantificação dos objetos (*object scores*) permite uma visão sobre a densidade dos agregados obtidos aquando da quantificação das categorias (Pestana e Gageiro, 2008).
A ACM tenta produzir uma solução ótima em que categorias de uma mesma variável são afastadas uma da outra tanto quanto possível. Em termos gráficos, isto implica que os objetos que partilham a mesma categoria estão próximos uns dos outros (agregados). As distâncias entre os objetos traduzem a semelhança ou dissemelhança que carateriza os seus perfis.
No ponto 3.1 será aprofundado o método ACM no contexto da *Optimal scaling.*

2. ANÁLISE DE CORRESPONDÊNCIAS

A Análise de Correspondências (AC) é um método de análise multivariada especialmente delineado para o estudo de variáveis qualitativas (extensível a variáveis quantitativas divididas em classes). Necessita apenas de uma tabela com números positivos que representam frequências observadas de objetos ou indivíduos classificados por uma categoria de linha e uma categoria de coluna; tais categorias devem ser mutuamente exclusivas e exaustivas, ou seja, um indivíduo ou objeto não pode ser classificado em mais de uma categoria de uma mesma variável (Greenacre, 1984).

As primeiras considerações sobre a AC surgem em 1935, com Fisher e Hirschfeld. Mais tarde nos anos 60, esta técnica tomou corpo com Benzécri, na França. A Análise de Correspondências foi desenvolvida com base nas tabelas de contingência. Antes do trabalho de Benzécri, o único tratamento "quantitativo" para dados qualitativos, existente era o teste do qui-quadrado (χ^2), que apenas avalia o grau de independência entre as duas variáveis de partida. A Análise de Correspondências veio permitir detetar relações impercebíveis na análise feita por comparação de variáveis par a par.

O objetivo da AC é descrever as linhas e as colunas de uma tabela de contingência, isto é, estudar a dependência entre os indivíduos e as categorias das variáveis em estudo. Baseia-se na decomposição do qui-quadrado de contingência, sendo o estudo da dependência realizado com base em representações gráficas (em que dois indivíduos ou duas categorias se assemelham tanto mais quanto mais próximos estiverem um do outro) e por parâmetros numéricos que permitem fazer a sua interpretação.

A AC envolve três conceitos básicos (Greenacre,2008):

- a noção de *perfil* das categorias
- o *peso* associado a cada perfil
- a *distância* (a distância qui-quadrado)

Segue-se a apresentação de alguns conceitos, para uma melhor compreensão dos procedimentos utilizados na formulação matemática da Análise de Correspondências (simples) e da Análise de Correspondências Múltiplas (abordada no capítulo 4). Os conceitos são maioritariamente geométricos, o único conceito estatístico envolvido nesta metodologia é a estatística do qui-quadrado.

2.1.Tabela de contingência

Consiste numa tabela cujos valores são frequências de observação de cada uma das possíveis combinações de níveis dos fatores de classificação. Substituindo as frequências absolutas pelas frequências relativas (relativas ao total de observações), obtém-se uma variante da tabela acima indicada. Assim, se **T** indicasse a tabela de frequências absolutas, e *n* o número total de observações associado à tabela, a tabela **F** = **T**/*n* fornece as frequências relativas de cada combinação (*i, j*) de níveis dos fatores. Este tipo de tabela designa-se **tabela de correspondências.** A soma dos elementos duma matriz **F** de correspondências é igual a 1. Outra variante da tabela acima indicada pode ser obtida quando se assinala, não a frequência de observações, mas a presença ou ausência de resposta. Obtemos desta forma, uma *matriz binária de presenças–ausências* constituída por elementos *um* e *zero* que indicam, respetivamente, a presença ou ausência da categoria no indivíduo. Podemos, assim, falar numa **matriz de incidências** ou **matriz indicatriz.**

2.2.Matriz inicial dos dados

Consideremos os dados organizados numa tabela de contingência com *n* linhas (indivíduos) e *p* colunas (categorias de uma variável). A matriz inicial $\boldsymbol{T}$ é definida da seguinte forma:

$$\boldsymbol{T}(n \times p) = \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1j} & \dots & t_{1p} \\ t_{21} & t_{22} & \dots & t_{2j} & \dots & t_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ t_{i1} & t_{i2} & \dots & t_{ij} & \ddots & t_{ip} \\ t_n 1 & t_{n2} & \dots & t_{nj} & \dots & t_{np} \end{bmatrix}$$

$\boldsymbol{T}$ é uma matriz de números não negativos, onde a soma das linhas ou das colunas é superior a zero e cujo *(i, j)* -ésimo elemento $\boldsymbol{t_{ij}}$ com *i= 1, ..., n* e *j=1, ... p* indica o

número de observações (frequência absoluta) efetuadas na combinação *i*-ésima linha (individuo *i*) com a *j*-ésima coluna (categoria *j*).

A soma das frequências na linha *i* da tabela $\boldsymbol{T}$,

$$t_{i\,.} = \sum_{j=1}^{p} t_{ij} \quad (i = 1, \dots n\,),$$

indica o número total de observações (frequência absoluta) associado ao individuo i – *totais marginais de linha*.

Analogamente, a soma das frequências na coluna *j* da tabela,

$$t_{.\,j} = \sum_{i=1}^{n} t_{ij} \quad (j = 1, \dots p\,),$$

indica o número total de observações associado à categoria *j* – *totais marginais de coluna*.

O número *total de observações* (em qualquer combinação de níveis dos dois fatores) é dado por:

$$\boldsymbol{t}_{..} = \sum_{i=1}^{n} t_{i} = \sum_{j=1}^{p} t_{.j} = \sum_{i=1}^{n}\sum_{j=1}^{p} t_{ij} \quad (i = 1, \dots n;\ j = 1, \dots p)$$

A frequência relativa da linha *i* da tabela (nível *i* dos indivíduos, independentemente de quais as categorias correspondentes) é dada por:

$$\boldsymbol{r_i} = \frac{t_{i\,.}}{t_{..}} \quad (i = 1, \dots n)$$

De forma análoga, a frequência relativa da coluna *j* da tabela é dada por:

$$\boldsymbol{c_j} = \frac{t_{.\,j}}{t_{..}} \quad (j = 1, \dots p)$$

Então, dividindo cada elemento da matriz de inicial $\boldsymbol{T}$ pelo número total de observações $\boldsymbol{t}_{..}$ obtém-se matriz de correspondências $\boldsymbol{F}$, cujo termo geral é dado por:

$$\boldsymbol{f_{ij}} = \frac{t_{i\,j}}{t_{..}} \qquad \mathbf{0} \leq \boldsymbol{f_{ij}} \leq \mathbf{1}$$

A **matriz de correspondências** apresenta-se da seguinte forma:

$$\boldsymbol{F} = \frac{T}{t_{..}} = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1j} & \cdots & f_{1p} \\ f_{21} & f_{22} & \cdots & f_{2j} & \cdots & f_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ f_{i1} & f_{i2} & \cdots & f_{ij} & \ddots & f_{ip} \\ f_{n1} & f_{n2} & \cdots & f_{nj} & \cdots & f_{np} \end{bmatrix}$$

cujo *(i, j)* -ésimo elemento $\boldsymbol{f_{ij}}$ é a frequência relativa do individuo *i* na categoria *j*.

Da matriz $\boldsymbol{F}$ podemos definir:

Totais marginais de linha,

$$f_{i\,.} = \sum_{j=1}^{p} f_{ij} \quad (i = 1, \ldots n\,),$$

Totais marginais de coluna,

$$f_{.\,j} = \sum_{i=1}^{n} f_{ij} \quad (j = 1, \ldots p\,),$$

Total de observações,

$$\boldsymbol{f_{..}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{f_{i\,.}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{f_{.\,j}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{p}} \boldsymbol{f_{ij}} \quad (i = 1, \ldots n;\ j = 1, \ldots p)$$

2.3. Massas

A *massa ou peso dos indivíduos* (linhas) da matriz $\boldsymbol{F}$ é dada pelos totais das frequências marginais das linhas, $f_{i\,.}$

$$r_i = f_{i\,.} = \sum_{j=1}^{p} f_{ij}$$

ou seja, $r = \boldsymbol{FI}$ onde $\boldsymbol{I}$ é um vetor unitário.

Analogamente, *a massa ou peso das categorias* (colunas) constituintes da matriz $\boldsymbol{F}$ é dada pelos totais das frequências marginais das colunas, $f_{.j}$

$$c_j = f_{.j} = \sum_{i=1}^{n} f_{ij}$$

ou seja, $c = \boldsymbol{F^t I}$.

Podemos escrever $\boldsymbol{D}_r$ e $\boldsymbol{D}_c$ as matrizes diagonais contendo os vetores das massas das linhas e colunas, *r* e *c*, respetivamente. Assim:

$$\boldsymbol{D}_r = diag(r)$$
$$\boldsymbol{D}_c = diag(c)$$

2.4. Perfis de linha e perfis de coluna

Perfil de linha i é o conjunto de frequências observadas para cada elemento dessa linha, relativas ao total de observações nessa linha. Assim, o perfil da linha *i* é dado pelos *p* valores:

$$pl_j^{(i)} = \frac{f_{ij}}{f_{i\,.}} = \frac{\frac{t_{ij}}{t_{..}}}{\frac{t_{i\,.}}{t_{..}}} = \frac{t_{ij}}{t_{i\,.}} \quad (j = 1, \dots p)$$

Do ponto de vista matricial, a matriz $\boldsymbol{P_L}$ dos perfis de linha calcula-se através do produto

$$\boldsymbol{P_L} = \boldsymbol{D}_r^{-1} F \,,$$

onde $\boldsymbol{D}_r^{-1}$ é a matriz diagonal (*n*×*n*) cuja diagonal é dada pelos recíprocos do vetor de frequências relativas de linha.

De forma semelhante por p*erfil da coluna j* entende-se o conjunto das frequências observadas para cada elemento dessa coluna, relativas ao total de observações nessa coluna. Assim, o perfil da coluna *j* é dado pelos *n* valores:

$$pc_i^{(j)} = \frac{f_{ij}}{f_{.\,j}} = \frac{\frac{t_{ij}}{t_{..}}}{\frac{t_{.\,j}}{t_{..}}} = \frac{t_{ij}}{t_{.\,j}} \quad (i = 1, \dots n)$$

A matriz $\boldsymbol{P_C}$ dos perfis de coluna calcula-se através do produto

$$\boldsymbol{P_C} = \boldsymbol{F}\boldsymbol{D}_c^{-1} \,,$$

onde $\boldsymbol{D}_c^{-1}$ é a matriz diagonal (*p*×p) cuja diagonal é dada pelos recíprocos do vetor de frequências relativas de coluna.

O perfil de uma linha *i* pode ser interpretado como a descrição da relação entre esta linha com todos os elementos de *J* (coluna). O perfil de uma coluna *j* tem interpretação análoga.

Cada perfil de linha ou de coluna pode ser representado como um ponto no espaço onde cada elemento do perfil constitui uma coordenada.
A semelhança entre duas linhas ou duas colunas é medida pela distância entre os seus perfis. Assim, quanto mais semelhantes são os perfis (linhas ou colunas), mais próximos estão os pontos um do outro; de igual modo, dois perfis muito diferentes entre si correspondem a pontos distantes entre si.

2.5. Nuvens de perfis

Cada linha da matriz de perfis de linha, $\boldsymbol{P_L}$, define um ponto no espaço a *n* dimensões, R^p. A nuvem de pontos *N (I) - nuvem de perfis de linha* -, no espaço R^p, é o conjunto dos pontos $i \in I$, cujas coordenadas são dadas pelos perfis f_j^i com massa $f_{i.}$.

Uma representação análoga dos perfis de coluna (colunas de $\boldsymbol{P_C}$) é possível no espaço R^n. A nuvem de pontos *N (J) - nuvem de perfis de coluna* -, no espaço R^n, é o conjunto dos pontos $j \in J$, cujas coordenadas são dadas pelos perfis f_i^j com massa $f_{.j}$.

2.6. Centroide

O centroide pode também ser designado por *centro de gravidade* ou *perfil médio*.
O centroide dos pontos f_j^i, relativos aos indivíduos, com massa $f_{i.}$ é a média ponderada da nuvem *N (I)* e tem coordenadas iguais aos totais marginais de coluna de ***F*.**
Analogamente, o centroide dos pontos f_i^j, relativos aos indivíduos, com massa $f_{.j}$ é a média ponderada da nuvem *N (J)* e tem coordenadas iguais aos totais marginais de linha de **F.**

Perfis que se aproximam do perfil médio, ou seja, do respetivo centro de gravidade serão representados por pontos próximos da origem; contrariamente os perfis que diferem muito do perfil médio, os seus pontos serão representados longe centro de gravidade.

2.7. Distância

Para descrever a diferença entre perfis utiliza-se uma distância ponderada pelo inverso da massa, a distância do qui-quadrado (χ^2).

A distância χ^2entre duas linhas *i* e $i^{'}$, em R^p, é dada por:

$$\sum_{j=1}^{p}\left(\frac{f_{ij}}{r_i}-\frac{f_{i'j}}{r_{i'}}\right)^2/c_j$$

Matricialmente a distância do χ^2 entre linhas é definida por

$$d_{ii'}{}^2=(P_L-P_{L'})^t D_r^{-1}(P_L-P_{L'})$$

A distância χ^2entre duas colunas *j* e $j^{'}$, em R^n, é dada por:

$$\sum_{j=1}^{n}\left(\frac{f_{ij}}{c_j}-\frac{f_{ij'}}{c_{j'}}\right)^2/r_i$$

Por outro lado, matricialmente a distância do χ^2 entre colunas é definida por

$$d_{jj'}{}^2=(P_C-P_{C'})^t D_c^{-1}(P_C-P_{C'})$$

2.8. Inércia

Na AC a variabilidade de uma tabela de dados é medida mediante a inércia, um conceito muito relacionado com a distância do qui-quadrado.

A inércia de uma nuvem de pontos em relação ao seu centro de gravidade é uma medida de variação total, que traduz a dispersão dos pontos da nuvem em torno do centro de gravidade.

A inércia (total) de uma tabela quantifica a variação existente nos perfis de linha e nos perfis de coluna. Cada uma das linhas e cada uma das colunas contribui para a inércia total; denominamos estas contribuições *inércias das linhas* e *inércias das colunas*, respetivamente.

Em R^p a inércia das linhas é dada por

$$\sum_{i=1}^{n} f_{i.}\sum_{j=1}^{p}\left(f_j^i-f_{.j}\right)^t P_c^{-1}\left(f_j^i-f_{.j}\right)$$

Desenvolvendo e fazendo as substituições adequadas, nesta expressão obtemos

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{\left(t_{ij}-\frac{t_{i.}t_{.j}}{t_{..}}\right)^2}{\frac{t_{i.}t_{.j}}{t_{..}}}=\frac{1}{t_{..}}\chi^2$$

sendo χ^2 a estatística do qui-quadrado utilizada nos testes de independência.

Em R^n a inércia das colunas é dada por

$$\sum_{j=1}^{p} f_{.j}\sum_{i=1}^{n}\left(f_i^j-f_{i.}\right)^T P_l^{-1}\left(f_i^j-f_{.j}\right)$$

De igual modo, podemos concluir que a inércia das colunas é igual a $\frac{1}{t_{..}}\chi^2$.

2.9. Algoritmo para obter coordenadas dos perfis de linha e dos perfis de coluna

A Análise de Correspondências baseia-se em resultados diretos da teoria das matrizes, utiliza a decomposição de uma matriz em valores singulares (DVS). Assim, procura-se encontrar o subespaço que melhor represente o conjunto de pontos referentes às linhas e às colunas.

O cálculo das coordenadas dos perfis de linha e dos perfis de coluna em relação aos eixos principais é possível recorrendo a um algoritmo que utiliza a decomposição em valores singulares. O algoritmo desenrola-se nos seguintes passos (Greenacre, 2008):

Passo 1. *Cálculo da matriz* $\boldsymbol{S}$ *dos resíduos estandardizados:*

$$\boldsymbol{S}=\boldsymbol{D}_r^{-1/2}(\boldsymbol{F}-rc^t)\boldsymbol{D}_c^{-1/2}$$

Passo 2. *Cálculo da DVS de* $\boldsymbol{S}$ *:*

$$\boldsymbol{S}=\boldsymbol{U}\boldsymbol{D}_\alpha\boldsymbol{V}^t \text{ onde } \boldsymbol{U}^t\boldsymbol{U}=\boldsymbol{V}^t\boldsymbol{V}=I$$

$\boldsymbol{D}_\alpha$ é a matriz diagonal de valores singulares (positivos) em ordem decrescente:

$\alpha_1 \geq \alpha_2 \geq \cdots$

$\boldsymbol{U}$ *e* $\boldsymbol{V}$ são, respetivamente, os vetores singulares esquerdo e direito.

Passo 3. *Coordenadas estandardizadas das linhas* Φ*:*

$$\Phi=\boldsymbol{D}_r^{-1/2}\,\boldsymbol{U}$$

Passo 4. *Coordenadas estandardizadas das colunas* Γ:

$$\Gamma=\boldsymbol{D}_c^{-1/2}\,\boldsymbol{V}$$

Passo 5*. Coordenadas principais das linhas* R*:*

$$\boldsymbol{R}=\boldsymbol{D}_r^{-1/2}\,\boldsymbol{U}\boldsymbol{D}_\alpha=\Phi\boldsymbol{D}_\alpha$$

Passo 6. *Coordenadas principais das colunas* G*:*

$$\boldsymbol{G}=\boldsymbol{D}_c^{-1/2}\,\boldsymbol{V}\boldsymbol{D}_\alpha=\Gamma\boldsymbol{D}_\alpha$$

Passo 7*. Inercias principais* λ_k*:*

$$\lambda_k=\alpha_k^2,\quad k=1,2,\dots,K \qquad onde\ K=min\{n-1,p-1\}$$

Normalmente as duas primeiras coordenadas principais das linhas e colunas são as mais representativas, sendo estas representadas pelos dois maiores valores próprios da matriz $\boldsymbol{S}$.

A realização da AC tem como objetivo explicar o máximo de inércia possível no primeiro eixo, no segundo eixo explica o máximo de inércia restante, e assim sucessivamente.

Observemos o cálculo dos valores das coordenadas principais e estandardizadas:

$$\boldsymbol{R}\boldsymbol{D}_r\boldsymbol{R}^t=\boldsymbol{G}\boldsymbol{D}_c\boldsymbol{G}^t=\boldsymbol{D}_\lambda$$

$$\Phi\boldsymbol{D}_r\Phi^t=\Gamma\boldsymbol{D}_c\Gamma^t=I$$

A soma ponderada dos quadrados das coordenadas principais na *k-ésima* dimensão é igual à inércia principal (ou valor próprio) $\lambda_k=\alpha_k^2$, o quadrado do *k-ésimo* valor singular. Enquanto a soma ponderada dos quadrados das coordenadas estandardizadas é igual a 1.

Coordenadas principais em função das coordenadas estandardizadas (relações baricêntricas):

$$\boldsymbol{R} = \boldsymbol{D}_r^{-1}\boldsymbol{F}\,\Gamma \qquad\qquad \mathbf{G} = \boldsymbol{D}_c^{-1}\boldsymbol{F}^{\mathrm{t}}\,\Phi$$

Coordenadas principais em função das coordenadas principais (relação entre as linhas e as colunas):

$$\boldsymbol{R} = \boldsymbol{D}_r^{-1}\boldsymbol{F}\boldsymbol{G}\boldsymbol{D}_\lambda^{-1/2} \qquad\qquad \boldsymbol{G} = \boldsymbol{D}_c^{-1}\boldsymbol{F}\boldsymbol{R}\boldsymbol{D}_\lambda^{-1/2}$$

A inércia total da matriz dos dados é igual à soma dos quadrados da matriz **S** (passo 1 do algoritmo anterior):

$$inércia = traço(\boldsymbol{S}\boldsymbol{S}^t) = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{\left(f_{ij} - r_i c_j\right)^2}{r_i c_j}$$

A inércia também é a soma dos quadrados dos valores singulares, ou seja, a soma dos valores próprios:

$$inércia = \sum_{k=1}^{K}\alpha_k^2 = \sum_{k=1}^{K}\lambda_k$$

2.10. Escolha do número de eixos. Proporção de inércia

A interpretação geométrica de um valor próprio λ_k relativamente a um eixo k, é a inércia da nuvem ao longo desse eixo.

A soma da inércia total da nuvem dos eixos é igual ao somatório dos K valores próprios, onde K é o número total de eixos (como indica a expressão anterior).

A inércia do eixo k é λ_k.

Temos então,

$$t^k = \frac{\lambda_k}{\sum_{k=1}^{K}\lambda_k}$$

que representa a contribuição relativa do eixo k para a inércia total da nuvem.

Assim,

$$\frac{\lambda_k}{\sum_{k=1}^{K}\lambda_k} \times 100$$

representa a percentagem de inércia explicada pelo eixo k.

Ao analisar um conjunto de dados, interessa manter o menor número possível de dimensões permanecendo elevada a percentagem de inércia explicada pelos eixos selecionados, para permitir uma análise gráfica com qualidade.

Um equilíbrio entre estes dois pressupostos pode ser encontrado com a ajuda de dois critérios muito comuns:

- representar graficamente, nas abcissas, o número de dimensões e, nas ordenadas, a percentagem de inércia explicada por cada eixo; em seguida, detetar no gráfico uma quebra acentuada da percentagem de inércia explicada e excluir os eixos a partir dela,

- reter um número suficiente de eixos de modo a explicar uma certa proporção de inércia σ (usualmente superior a 50%), ou seja, reter os primeiros q eixos de forma que

$$\frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{K} \lambda_k} \geq \sigma$$

2.11. Representação/Análise gráfica da AC

Inicialmente, o processo gráfico gera uma nuvem de pontos contidos num espaço multidimensional, que torna praticamente impossível a análise visual das relações. A nuvem pode, no entanto, ser projetada em planos escolhidos pela sua capacidade de representar o mais fielmente as distâncias originais dos pontos.

Os pontos distribuem-se, nos planos, segundo a representatividade dos mesmos, de acordo com o valor dos perfis, linha ou coluna, que representam o conjunto de dados. Desta forma, pontos consequentes de perfis semelhantes, localizam-se mais próximos no plano do que pontos resultantes de perfis com características discrepantes, esse fato faz com que a AC desvende modelos de associações entre as variáveis em estudo e as respetivas categorias.

A Análise Classificatória pode ser utilizada como ferramenta de apoio à interpretação da AC, revela-se útil na presença de inúmeras variáveis ou indivíduos quando o objetivo da análise é a classificação.

3. ANÁLISE DE CORRESPONDÊNCIAS MÚLTIPLAS

A Análise de Correspondências Múltiplas (ACM) é uma generalização da AC para o caso de tabelas com dimensão igual ou superior a três.

O recurso à ACM parte do pressuposto implícito na hipótese de estudo de que existem relações preferenciais entre variáveis**.** Esta metodologia é especialmente adaptada ao tratamento de um conjunto de respostas de um inquérito por questionário (no qual as perguntas representam as variáveis). A adoção desta estratégia de análise implica a conjugação de um conjunto de ações preliminares, entre as quais as de codificação das variáveis aquando do desenho dos questionários. Os dados são submetidos a transformações matemáticas, tornando possível visualizar num espaço de menores dimensões a estrutura da matriz dos dados, a partir da qual se realiza a ACM. A aplicação desta técnica, possibilita ao investigador realizar uma abordagem relacional sobre múltiplas variáveis que caraterizam um conjunto de indivíduos, permitindo posteriormente definir diferentes grupos.

Existem diversas ferramentas informáticas para operacionalizar a ACM, designadamente a Linguagem R, o programas ANDAD, SPAD, SAS e o SPSS. No presente trabalho será utilizado o SPSS[2].

A construção dos resultados na lógica das correspondências do SPSS, processa-se através da medição das distâncias entre as categorias das variáveis, considerando estas como um sistema de oposições/associações entre as mesmas. Designam-se, no SPSS, de *object scores* os valores atribuídos aos casos, e de *centroid coordinates* os valores (quantificações) atribuídos às categorias. Os *object scores* das mesmas categorias são projetados próximos uns dos outros (semelhança de *scores*). Ou seja, quando as categorias de diferentes variáveis são projetadas próximas umas das outras, elas pertencem aos mesmos objetos.

[2] Statistical Package for the Social Sciences

3.1. A ACM no contexto da *Optimal Scaling*

Na década de 70 início dos anos 80 vários investigadores, entre os quais se encontram os membros da equipa de Gifi, foram responsáveis pela formalização de métodos adequados à realização de análises multivariadas sobre dados qualitativos. Estes métodos foram acompanhados pelo desenvolvimento de *software* e o aparecimento de programas designados ALSOS (***A**lternating **L**east **S**quares with **O**ptimal **S**caling*).

O algoritmo da ACM é do tipo ALS, a este método está inerente um processo de quantificação (descrito no ponto 3.1.4), que tem por objetivo estimar quantificações ótimas (*optimal scaling*) para os parâmetros em análise.

3.1.1. Organização dos dados

A matriz de *input* submetida a transformação via ACM é por definição uma matriz multidimensional na qual se dispõem *n* objetos (ou indivíduos), que correspondem às linhas da matriz, caraterizados segundo *m* variáveis (de natureza qualitativa) representadas, neste caso, pelas colunas da matriz. A partir desta matriz é necessário construir uma matriz de presenças-ausências, pois é esta a matriz efetivamente usada pelo algoritmo, Carvalho (2008).

Para cada variável j com k_j categorias constrói-se uma matriz binária $\mathbf{G}_\mathrm{j}$ – *matriz indicatriz* – com *n* linhas e com k_j categorias. Fazendo a justaposição das matrizes $\mathbf{G}_\mathrm{j}$ obtém-se a matriz $\boldsymbol{G}$ (Tabela 3.3) do tipo $n \times \sum k_j$.

Cada objeto tem ocorrência unitária por variável, o somatório em linha da matriz *G* coincide com *m* (número de variáveis). Por sua vez, o somatório em coluna corresponde às frequências univariadas (frequência marginal de cada uma das categorias).

A ACM contempla ainda a definição de duas outras matrizes: uma com as frequências das categorias e outra com o número de respostas válidas .

Define-se uma matriz $\boldsymbol{D}_j$ para cada variável j . A diagonal da matriz $\boldsymbol{D}_j$ contém as frequências marginais das k_j categorias de j. Para as *m* variáveis obtém-se a matriz $\boldsymbol{D}$ (Tabela 3.4) cuja diagonal corresponde às somas em coluna da matriz $\boldsymbol{G}$.

A identificação do número de respostas válidas faz-se definindo para cada variável *j* a matriz $\boldsymbol{M}_j$. Do somatório das diversas matrizes $\boldsymbol{M}_j$ obtém-se a matriz $\boldsymbol{M}_*$ (Tabela 3.5). A matriz $\boldsymbol{M}_* = \sum \boldsymbol{M}_j$ contabiliza o número total de respostas válidas para cada objeto.

EXEMPLO

Considere-se uma aplicação com n =10 e m =4, com 4 categorias para as variáveis B e O, 3 categorias para a variável H e 2 categorias para a variável W.

A partir das categorias observadas para os 10 casos construiu-se a matriz de *input* (Tabela 3.2), substituindo as categorias pelos respetivos códigos.

Variáveis	Categorias	Codificação
B	B1	1
	B2	2
	B3	3
	B4	4
H	H1	1
	H2	2
	H3	3
O	O1	1
	O2	2
	O3	3
	O4	4
W	W1	1
	W2	2

Tabela 3. 1 Exemplo de codificação

Nesta matriz é possível observar o perfil dos 10 indivíduos caraterizado por combinações das categorias das 4 variáveis.

Veja-se a correspondência existente entre a matriz de *input* e a matriz binária:

Tabela 3.2 Matriz de Input

Casos	B	H	O	W
1	2	1	4	2
2	1	3	3	2
3	1	2	3	1
4	3	1	1	2
5	1	2	2	1
6	4	1	1	1
7	2	3	4	1
8	3	1	2	2
9	4	3	1	1
10	2	2	4	1

Tabela 3.3 Matriz Binária (***G***)

Casos	B				H			O				W	
1	0	1	0	0	1	0	0	0	0	0	1	0	1
2	1	0	0	0	0	0	1	0	0	1	0	0	1
3	1	0	0	0	0	1	0	0	0	0	1	1	0
4	0	0	1	0	1	0	0	1	0	0	0	0	1
5	1	0	0	0	0	1	0	0	1	0	0	1	0
6	0	0	0	1	1	0	0	1	0	0	0	1	0
7	0	1	0	0	0	0	1	0	1	0	0	1	0
8	0	0	1	0	1	0	0	0	1	0	0	0	1
9	0	0	0	1	0	0	1	1	0	0	0	1	0
10	0	1	0	0	0	1	0	0	0	0	1	1	0

Tabela 3.4 Matriz $\boldsymbol{D}$

$$\boldsymbol{D} = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}$$

Tabela 3.5 Matriz $\boldsymbol{M}_*$

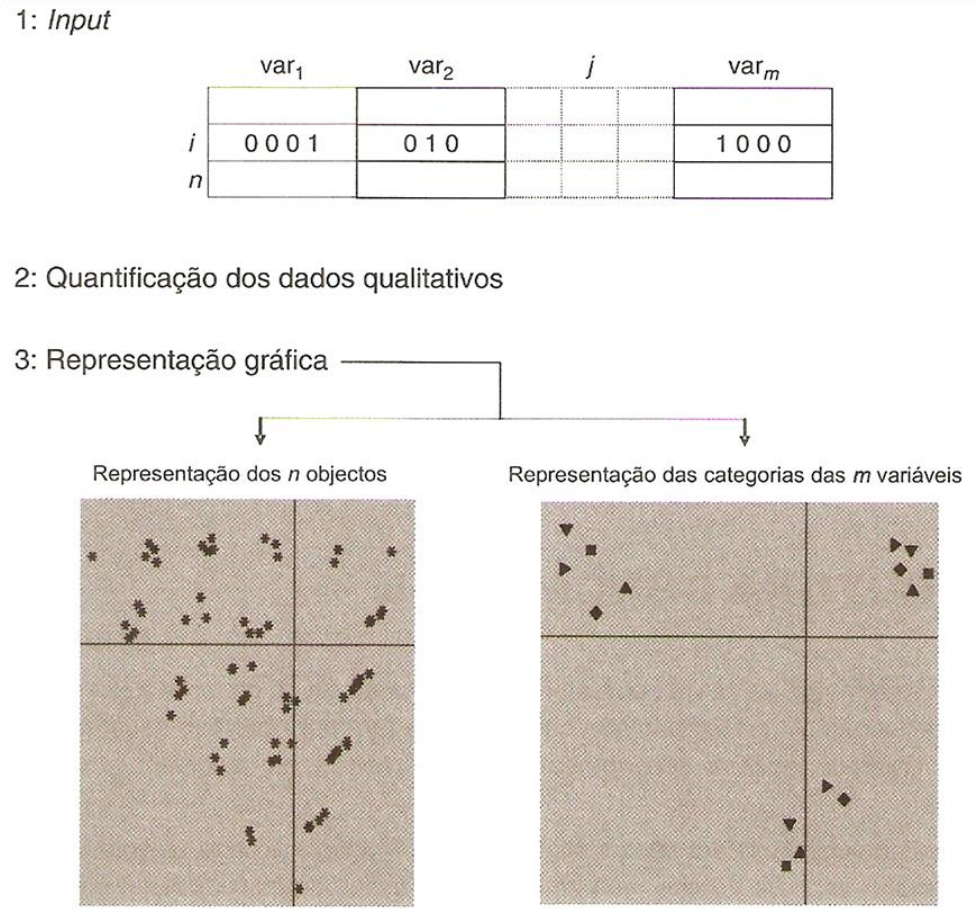
$$\boldsymbol{M}_* = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}$$

A diagonal da matriz $\boldsymbol{M}_*$ regista o valor 4 (número de variáveis) para os 10 casos pois, neste caso, não se registam não-respostas.

Ao método ACM está inerente um procedimento de transformação ótima. Os dados são submetidos a um processo de quantificação, que tem por objetivo estimar quantificações ótimas (*optimal scaling*) para os parâmetros em análise: categorias e objetos. Esta operação decorre segundo um processo que vai estimando de forma alternativa (e iterativa) as quantificações dos parâmetros até ser atingida a solução ótima.

As quantificações (enquanto coordenadas) permitem projetar as categorias ou objetos em planos. A representação das categorias possibilita a análise das associações entre as múltiplas variáveis e a dos objetos permite avaliar o seu posicionamento no espaço.

Figura 1 Trajeto da ACM



Fonte: Carvalho (2008)

3.1.2. Princípio das médias recíprocas

No processo de transformação das categorias e dos objetos está implícito *o princípio das médias reciprocas*, através do qual se relacionam as quantificações das categorias e os *scores* dos objetos, contribuindo de forma recíproca para as suas definições. Assim, define-se a quantificação de uma categoria como sendo a média dos *scores* dos objetos que partilham essa categoria e, por sua vez os *scores* dos objetos são proporcionais à média das quantificações das categorias às quais eles estão associados.

As categorias de cada variável desencadeiam uma partição dos objetos em subgrupos sendo que cada um dos *pontos-categoria*[3] corresponderá ao centroide (centro de gravidade ou ponto médio) das "subnuvens" defindas pelos *pontos-objeto*[4] que

[3] Pontos representativos das categorias.
[4] Pontos representativos dos objetos

partilhem as mesmas categorias. Os objetos associados às mesmas categorias (perfis semelhantes) situam-se próximos, definindo-se subgrupos homogéneos e serão afastados os que se afigurem mais dissemelhantes.

As distâncias entre objetos traduzem graficamente a estrutura das semelhanças/dissemelhanças que carateriza os seus perfis.

Na representação gráfica respeitante à projeção das categorias, temos que as categorias de uma mesma variável tendem a registar projeções distantes. Por outro lado, a proximidade entre categorias de diferentes variáveis corresponde à presença de objetos com perfis semelhantes.

3.1.3. Quantificação das categorias

Considerando a matriz de *input* sabemos que as categorias representam os objetos a ela associados e, por sua vez, os objetos são caraterizados por partilharem algumas categorias. Esta reciprocidade intrínseca à matriz é preservada no procedimento de transformação das categorias e dos objetos com o uso do *princípio das médias reciprocas.*

Para cada categoria de uma variável *j* regista-se a presença de diferentes objetos. As categorias das variáveis efetuam uma partição entre os objetos definindo subgrupos que se caraterizam por partilharem certas categorias das diferentes variáveis em análise.

A quantificação de cada categoria é representativa dos objetos que nela se inserem. Algebricamente a quantificação - Y- das *p* categorias dos m variáveis é dada pela expressão

$$\boldsymbol{Y = D^{-1}\grave{G}X} \qquad (1)$$

sendo Y: matriz de quantificações das categorias

D: matriz de frequências das *p* categorias (número total de categorias em análise)

$\boldsymbol{G}$: matriz de presencias ausências de cada objeto nas categorias

$\boldsymbol{X}$: matriz dos scores dos *n* objetos

A expressão (1) evidencia que a quantificação das categorias (Y) é igual à média dos *scores* dos objetos que nelas se inserem ($\boldsymbol{\grave{G}X}$), ponderada pela frequência de ocorrência das categorias, a qual é retirada da matriz $\boldsymbol{D}$.

Os *scores* dos objetos – matriz X – são, por sua vez, proporcionais à média das quantificações das categorias (Y) associadas a cada objeto.
Para o conjunto das *m* variáveis ter-se-á:

$$\boldsymbol{X} \cong \boldsymbol{G}Y/_{m} \qquad (2)$$

a proporcionalidade pode traduzir-se numa igualdade, basta afetar-lhe um fator de proporcionalidade, que é o valor da inércia (λ) da dimensão para a qual estão a ser determinados os *scores.* Para tal proceder-se-á da seguinte forma:

$$\boldsymbol{X} = \boldsymbol{G}Y/_{m}\ {}^{1}/_{\lambda} \qquad (3)$$

3.1.4. Quantificação ótima

Conforme foi anteriormente referido a ACM usa um algoritmo do tipo ALS, isso significa que as quantificações vão sendo determinadas alternadamente, até ser obtida a solução ótima. Esse procedimento algébrico tem inerente a minimização de uma função perda. O processo iterativo converge quando for mínima esta função perda (Carvalho, 2008) :

$$\sigma(X;Y) = {}^{1}/_{np} \sum_{j} tr\left(\left(\boldsymbol{X} - \boldsymbol{G}_j\, Y_j\right)'\, M_j \left(\boldsymbol{X} - \boldsymbol{G}_j\, Y_j\right)\right)$$

A solução ótima equivale a definir os *scores* dos objetos (X) e as quantificações das categorias (Y_j) garantindo ser mínima a soma do quadrado das distâncias entre os *pontos-objeto* e os correspondentes *pontos-categoria*.

3.1.4.1 Teste de Convergência

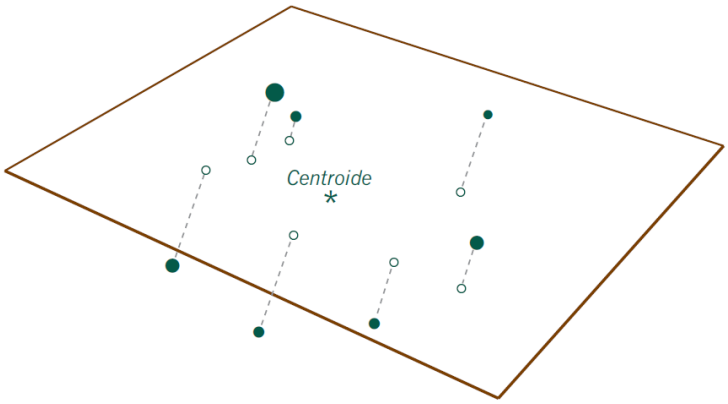
Este teste corresponde à comparação do resultado obtido (valor da função perda) numa iteração com a solução da iteração anterior.
No final de cada iteração a ACM recalcula o valor da função perda comparando o resultado obtido $\sigma(X^{+};Y^{+})$ com o resultado da iteração anterior $\sigma\left(\tilde{X};\tilde{Y}\right)$. Enquanto a diferença $\sigma\left(\tilde{X};\tilde{Y}\right) - \sigma(X^{+};Y^{+})$ for superior ao *critério de convergência* (ε) – o qual é

limitado por zero[5] –, repetir-se-ão iterativamente os passos do algoritmo correspondentes ao cálculo (alternado) das estimativas para as quantificações das categorias e para os *scores* dos objetos. Quando a referida diferença for inferior ao critério de convergência, está concluída a quantificação ótima. O que significa que as quantificações convergem e, como foi referido no ponto 3.1.4, é mínima a soma do quadrado das distâncias entre os *pontos-objeto* e os correspondentes *pontos-categoria.*

3.1.5. Quantificação múltipla

Dado que é muito difícil (ou mesmo impossível) observar e imaginar pontos num espaço com mais de três dimensões, é necessário reduzir a dimensionalidade. Portanto, será interessante poder visualizar os perfis ainda que de forma aproximada num espaço de poucas dimensões. Esta é a essência da Análise de Correspondências: a identificação de subespaços de poucas dimensões (preferencialmente com três ou menos) que contenham os perfis. Assim, é possível projetar no subespaço mencionado os perfis e observar as posições das suas projeções como uma aproximação às suas verdadeiras posições no espaço original de maior dimensionalidade. Na maioria dos casos necessitamos de pelo menos um plano para aproximar ou ajustar o espaço multidimensional da nuvem de perfis.

Figura 2 Perfis num espaço multidimensional e um plano que corta o espaço



Fonte: Greenacre (*2008*)

O plano (Figura 2) que melhor se ajusta deve passar pelo centroide dos pontos (os tamanhos dos pontos indicam que os perfis têm massas diferentes).

[5] No SPSS o valor assumido por defeito é 0,00001.

Neste sentido a ACM é classificada como um método de redução de dados, pois submete os dados a transformações matemáticas permitindo, assim, rever num espaço de menores dimensões a estrutura multifacetada e relacional do espaço de partida.

Como já foi mencionado no processo de transformação iterativo dos dados, a cada categoria é associada uma quantificação e a cada objeto um *score.* Pode estimar-se mais do que uma solução para essas quantificações, ou seja, é possível iterar-se diversas soluções para os *scores* dos objetos e para as quantificações das categorias, correspondendo cada uma das soluções a uma dimensão. Obtém-se assim uma quantificação múltipla para as variáveis nominais (ou categorizadas).

Determinação em cada caso do número de soluções possíveis ou dimensões (*Carvalho, 2008):*

$$r_{max} = \{(n-1); (p - max(m_1; 1))\}$$

sendo:

n: número de indivíduos

p: número de categorias ativas (que entram efetivamente na ACM)

m_1: número de variáveis (ativas) sem não resposta

Como habitualmente o número de indivíduos em análise é superior ao número de categorias, a expressão de uso mais vulgar é $p - max(m_1; 1)$.

Nesse caso para obter o número máximo de dimensões faz-se:

$p - m_1$ se $m_1 \geq 1$ ou

$p - 1$ se, no limite, todas as m variáveis registarem não respostas.

Se $m = m_1$, isto é, não existirem categorias de não resposta, então o número máximo de dimensões virá $p - m$.

3.2. Medidas de análise dos resultados

3.2.1. Medidas de discriminação e contribuições

A partir da quantificação das k_j categorias de cada variável *j*, a ACM prevê que se determine também, uma quantificação para cada uma das *m* variáveis em cada dimensão ***s*** e que se designa por *medida de discriminação*. As medidas indicam a variância de cada variável após concluído o processo de quantificação ótima. A medida de discriminação de uma variável *j* na dimensão ***s*** é dada por (*Carvalho, 2008):*

$$Discr_{js} = {}^{1}/_{n}\, Y'_{(j)}s\, \boldsymbol{D}_j Y_{(j)}s \qquad \begin{array}{l} \boldsymbol{s} = 1, 2, \ldots, r \text{ dimensões} \\ j = 1, 2, \ldots, m \text{ variáveis} \end{array}$$

Como a expressão indica, a medida de discriminação de uma variável é igual à média do quadrado das quantificações $\left(Y'_{js}Y_{js}\right)$ das categorias de uma variável *j*, ponderadas pela respetiva frequência ou peso (matriz D_j).

O valor desta medida varia entre 0 e 1, quanto mais o seu valor se aproximar do limite superior mais as variáveis em questão discriminam os objetos em análise numa dada dimensão. Neste sentido a presença de variáveis com medidas de discriminação elevadas aumentará a possibilidade de definir grupos homogéneos.

3.2.2. Seleção das dimensões

É necessário estudar o número de dimensões a reter para a construção gráfica (e a partir das quais se vai basear a análise) uma vez que, como foi referido na secção 3.1.5, a ACM integra um processo de quantificação múltipla.

A importância de cada uma das dimensões para explicar a variância dos dados de *input*, pode ser analisada através dos valores próprios e da inércia.

Os valores próprios quantificam a variância explicada por dimensão. A inércia varia entre 0 e 1 e quanto mais perto do limite superior mais variância é explicada por dimensão. Assim, a importância das dimensões está hierarquizada e vai diminuindo. É habitual privilegiar 2, 3 dimensões por ser nessas, que se registam os valores de inércia mais elevados. E por consequência, ao explicarem mais variância são mais diferenciadoras dos objetos em estudo.

3.2.3. Representação gráfica

A tradução gráfica das medidas de discriminação possibilita a observação da disposição das múltiplas variáveis nos planos definidos e a avaliação da relevância de cada uma delas nas dimensões.

Quanto mais afastadas estiverem as variáveis da origem do gráfico e mais adjacentes a uma única dimensão, maior é o indício da presença de dimensões que envolvem traços de caraterização distintos. Por outro lado, variáveis próximas da origem, correspondem a variáveis que não são diferenciadoras para as duas dimensões consideradas nesse plano.

Uma variável pode ser relevante em mais do que uma dimensão. Graficamente a variável dispõe-se próximo da diagonal do gráfico "*Discrimination Measures".*

A representação gráfica dos resultados da ACM permite visualizar no plano a nuvem de pontos das categorias das variáveis, assim como a nuvem de pontos dos objetos caraterizados pelas variáveis em causa. A nuvem de pontos dos objetos é formada a partir do conjunto dos perfis de cada linha. Da mesma forma, a nuvem de pontos das categorias é formada a partir do conjunto de perfis de cada coluna.

A partir da leitura do plano é possivel a identificação de associações entre categorias (proximidade das projeções). Para além da disposição relativa das categorias de cada variável é também importante atender à sua localização relativamente à origem. Quanto mais as suas coordenadas se afastarem de zero, maior é a diferenciação que as categorias produzem nos objetos. Dessa observação/leitura do plano, pode resultar a identificação de diferentes configurações. É possível, portanto, realizar a identificação de grupos homogéneos a partir da análise do espaço onde estão representadas as propriedades que os caraterizam.

3.3. Articulação da Análise de Correspondência com a Análise de *Clusters*

Por via da ACM, através da disposição dos indivíduos ou das categorias em planos, é possível visualizar as relações entre as múltiplas variáveis em análise. A existência de diferentes combinações das características em análise (proximidade de categorias de diferentes variáveis) induz a presença de indivíduos que partilham tendencialmente as mesmas características. Estas configurações refletem a presença de grupos com perfis

distintos. Ao ser conhecida a configuração topológica do espaço em análise, se existir interesse, podemos definir os grupos e saber o peso de cada um deles na amostra aplicando a Análise de *Clusters.*

A ACM permite além de todas as combinações gráficas dimensionais possíveis, a disposição das coordenadas das variáveis de cada dimensão. Estes valores das coordenadas são pois utilizados na Análise de *Clusters* como variáveis de *input* (quantitativas).

Neste sentido, Carvalho (2008) refere "para efetuar a articulação ACM/Análise de Clusters usam-se como variáveis de *input* os *scores* dos objetos (indivíduos) nas dimensões que sustentam o plano que, por sua vez, configura os grupos que se pretendem vir a definir" (…) "Pode assim ver-se na articulação destes métodos de Análise de Dados uma estratégia para a partir da configuração topológica se passar à definição de *tipologias".*

Assim, o interesse da combinação dos dois métodos justifica-se uma vez que pode resultar num aperfeiçoamento dos resultados da análise. Embora cada técnica possua particularidades e objetivos específicos de pesquisa, a associação de técnicas diferentes e complementares permite o refinamento das soluções e conclusões encontradas.

4. ANÁLISE DE *CLUSTERS*

Análise Classificatória (*Cluster Analysis*, em inglês), Análise de Agrupamentos, Análise de Grupos, Análise de Aglomerados, Análise de Conglomerados são outras designações em uso na língua portuguesa. Neste trabalho é utilizada a designação Análise de *Clusters* por ser a mais frequente nos trabalhos na área da estatística.

A Análise de *Clusters* é uma técnica multivariada que tem por objetivo facultar uma ou várias partições na massa de dados, em grupos, por algum critério de classificação, de forma que exista homogeneidade dentro do grupo e heterogeneidade entre grupos (Sneath & Sokal, 1973).

Neste tipo de análise os métodos são exploratórios assume-se que não existe dependência entre as variáveis: os grupos definem-se por si mesmo sem que haja uma relação causal entre as variáveis utilizadas. A identificação de grupos de sujeitos ou variáveis permite identificar *outliers* multivariados, e possibilita a geração de hipóteses relativas às relações estruturais entre variáveis. No entanto, trata-se de uma ferramenta de apoio à interpretação, e por isso não deve ser analisada independentemente, mas sim fundamentada com outras técnicas, como por exemplo a Análise de Correspondências.

Na Análise de C*lusters*, os agrupamentos de sujeitos (casos ou itens) ou variáveis é feito a partir de medidas de semelhanças ou de medidas dissemelhança (distância) entre, inicialmente dois sujeitos e mais tarde entre dois *Clusters* de observações usando técnicas hierárquicas ou não-hierárquicas de agrupamento de C*lusters* (Maroco, 2010).

Genericamente, a análise de *clusters* compreende cinco etapas (Reis, 2001):

1. A seleção de indivíduos ou de uma amostra de indivíduos a serem agrupados;
2. A definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos indivíduos;
3. A definição de uma medida de semelhança ou distância entre cada dois indivíduos;

4. A escolha de um critério de agregação ou desagregação dos indivíduos, isto é a definição de um algoritmo de partição / classificação;
5. Por último, a validação dos resultados encontrados.

Os algoritmos de agrupamento operam, geralmente, sobre dois tipos de estrutura de dados:

(i) uma matriz de dimensão *n x p* correspondendo as *n* linhas aos sujeitos e as *p* colunas aos seus atributos ou características;

(ii) quadro de dimensão *n x n* cujos elementos medem as proximidades[6] entre cada par de indivíduos.

4.1. As variáveis

4.1.2. Seleção das variáveis

A escolha das variáveis adequadas para a definição de grupos pode estar relacionada com conhecimento anteriormente adquirido pelo investigador sobre o tema a estudar, o que permitirá, à partida, rejeitar as variáveis irrelevantes.

Variáveis que assumem praticamente o mesmo valor para todos os sujeitos são pouco discriminatórias, e a sua inclusão pouco contribuiria para a determinação da estrutura do agrupamento. Por outro, a inclusão de variáveis com grande poder de descriminação, porém pouco significativas na abordagem do problema, pode mascarar os grupos e levar a resultados equivocados.

Acontece com frequência, o número de variáveis medidas ser grande, dificultando a análise. Respeitando o princípio da parcimônia, devemos tentar diminuir o seu número de forma que a seleção considere tanto a sua relevância como o seu poder de descriminação face ao problema em estudo. Em último caso, pode-se ainda tentar utilizar técnicas estatísticas para redução da dimensionalidade da matriz de dados (*e,g.* a Análise de Componentes Principais).

[6] As proximidades poderão ser semelhanças (medem o grau de similitude entre cada par de sujeitos) ou distâncias (medem o grau de afastamento ou diferença)

4.1.3. Escala das variáveis

Quando as variáveis estão definidas em diferentes escalas de medida e se aplica a análise de *clusters*, qualquer medida de semelhança ou de distância vai refletir sobretudo o peso das variáveis que maiores valores e maior dispersão apresentam. Visando anular este efeito, surgiram várias propostas de estandardização das variáveis. Apresentam-se as mais comuns:

Consideremos as observações originais $x_1, \ldots, x_2$

A transformação mais comum é definida por

$Z_i = \frac{x_i - \bar{x}}{S},\ i = 1, \ldots, n$ onde $\bar{x}$ e S denotam a média e o desvio padrão das observações. Esta transformação faz com que as novas variáveis tenham média nula e variância unitária.

Outra forma de se transformar variáveis é tomar-se os desvios em relação ao menor valor e normalizá-los pela amplitude, ou seja,

$$Z_i = \frac{x_i - x_{(1)}}{x_{(n)} - x_{(1)}}, \qquad i = 1, \ldots, n$$

onde $x_{(1)}$ e $x_{(n)}$ denotam o mínimo e o máximo da amostra, respetivamente.

Podemos ainda tomar a média como fator normalizador,

$$Z_i = \frac{x_i}{\bar{x}},\ \ i = 1, \ldots, n$$

É importante cuidado no processo de estandardização, pois não deve ser tomado como solução ideal para todos os casos. Este processo reduz as diferenças entre os sujeitos anulando os agrupamentos naturais que possam existir nos dados.

4.2. Medidas de semelhança e medidas de dissemelhança

Na análise de Clusters, a escolha da medida de semelhança ou dissemelhança que melhor se adequa ao tipo de dados recolhidos representa um passo crucial.

Segundo Gower e Legender (1986) um coeficiente tem de ser considerado no contexto do estudo estatístico, incluindo a natureza dos dados e do tipo de análise pretendido.

Indicam alguns critérios para a escolha das medidas, no entanto, apesar da sua análise sobre o assunto concluíram que não é possível dar uma resposta definitiva.

As medidas de semelhança (ou dissemelhança) podem ser entre sujeitos ou entre variáveis, de acordo com o objetivo do estudo, respetivamente *clusters* de sujeitos ou *clusters* de variáveis.

4.2.2. Medidas de semelhança e medidas de dissemelhança entre sujeitos

São obtidas de uma matriz multivariada $X_{n\times p}$ resultante da observação de p variáveis em n sujeitos, e são escolhidas de acordo com o tipo de variáveis.
Os números d_{ij} (valor de uma medida de dissemelhança entre o sujeito i e sujeito j) ou s_{ij} (valor de uma medida de semelhança entre o sujeito i e sujeito j) são colocados numa matriz $n \times n$, conhecida por matriz de semelhança (ou dissemelhança).
A proximidade de dois sujeitos $i\ e\ j$ é tanto maior quanto menor é a dissemelhança ou distância entre eles.
O estudo das relações de semelhança é inspirado em modelos geométricos, os sujeitos são representados por pontos no espaço. Deste modo as dissemelhanças observadas entre os sujeitos são visualizadas como a distância entre os respetivos pontos.

4.2.2.1. Dissemelhanças e distâncias – propriedades

Uma medida de dissemelhança d_{ij} entre um sujeito i e um sujeito j deverá satisfazer algumas propriedades:

- $d_{ij} \geq 0,\ \forall\, i,j = 1{:}\,n;$
- $d_{ii} = 0,\ \forall\, i,j = 1{:}\,n;$ (Identidade)
- $d_{ij} = d_{ji}\ ,\ \forall\, i,j = 1{:}\,n$ (Simetria);
- $d_{ij} \leq d_{ik} + d_{kj}\ ,\ \forall\, i,j,k$ (Desigualdade triangular)

No caso de as medidas de dissemelhança verificarem além das três primeiras condições a desigualdade triangular fala-se em **distância.**

4.2.2.2 Semelhanças – propriedades

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

Quando $s_{ij} = 0$ os objetos não são semelhantes

Quando $s_{ij} = 1$ significa que a semelhança é máxima

- $s_{ij} = s_{ji} \quad \forall i, j = 1{:}\, n$ (Simetria)
- $s_{ii} = 1 \quad \forall i = 1{:}\, n$ (Identidade)

As medidas de semelhança (ou dissemelhança) dependem em primeiro lugar, do tipo de variáveis que caraterizam os sujeitos.

4.2.2.3 Medidas de dissemelhança e de semelhança para variáveis quantitativas

São várias as medidas que podem ser utilizadas como medidas de distância ou dissemelhança entre cada par de sujeitos. Assim para dois sujeitos i e j, para as variáveis $v = 1,2, \ldots p$.

- ***Distância Euclidiana***

$$d_{ij} = \left[\sum_{v=1}^{p} \left(X_{iv} - X_{jv}\right)^2\right]^{1/2}$$

- ***Distância Euclidiana ao quadrado***

$$d_{ij} = \sum_{v=1}^{p} \left(X_{iv} - X_{jv}\right)^2$$

- ***Distância de Minkowski***

$$d_{ij} = \left[\sum_{v=1}^{p} \left|X_{iv} - X_{jv}\right|^r\right]^{1/r}$$

para $r = 1$, d_{ij} é o módulo da distância absoluta entre os sujeitos i *e* j relativamente às p- variáveis medidas (conhecida por Distância *city-block*); para $r = 2$, têm-se a distância euclidiana habitual.

- ***Distância absoluta ou de Manhattan***

$$d_{ij} = \sum_{v=1}^{p} \left|X_{iv} - X_{jv}\right|$$

- ***Distância de Chebishev***

$$d_{ij} = \max_{v} |X_{iv} - X_{jv}|$$

- ***Distância de Mahalanobis***

$$d_{ij} = (X_i - X_j)' \sum^{-1} (X_i - X_j)$$

Apenas a distância de Mahalanobis, também chamada distância generalizada, utiliza a matriz de variância e covariância $\sum$ fazendo implicitamente a estandardização das variáveis.

- ***Medida de Semelhança do Cosseno*** – define-se no intervalo [-1, 1]

$$CoSIN(i,j) = \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{jk}^2}} = s_{ij}$$

$CoSIN(i,j)$, cosseno do ângulo formado pelas duas semirretas que unem a origem aos respetivos sujeitos, representados como pontos no espaço.

s_{ij}=1 significa que a semelhança é máxima

s_{ij} = -1 significa que a semelhança é mínima

Coeficiente de correlação de Pearson – medida de semelhança. O seu valor varia entre -1 e +1. Para dois sujeitos $i\ e\ j$, caraterizados por p atributos este coeficiente define-se como

$$r_{ij=} \frac{\sum_{k=1}^{p} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{p} (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{p} (x_{jk} - \bar{x}_j)^2}}$$

com

p = número total de variáveis

x_{ik}= valor da variável $k\ para\ o\ sujeito\ i$ $(k = 1, \dots, p)$

x_{jk}= valor da variável $k\ para\ o\ sujeito\ j$

$\bar{x}_i$= média de todas as variáveis para o sujeito i

$\bar{x}_j$= média de todas as variáveis para o sujeito j

4.2.2.4. Medidas de dissemelhança e de semelhança para variáveis qualitativas

Na procura de elementos semelhantes é frequente o uso de critérios qualitativos, para medir o grau de semelhança entre os sujeitos, segundo variáveis qualitativas.

Estas medidas de semelhança (dissemelhança) geralmente têm valores no intervalo [0, 1].

Se dois sujeitos i e j têm valores iguais para todas as variáveis, então têm coeficiente de semelhança, igual a um, $s_{ij} = 1$.

Se dois sujeitos i e j diferem no máximo para todas as variáveis, então têm coeficiente de semelhança, igual a zero, $s_{ij} = 0$.

Na literatura são muitas são as propostas deste tipo de coeficientes.

4.2.2.4.1. Medidas de semelhança para variáveis nominais binárias. Medidas de associação

Indicadas para definir a semelhança entre os sujeitos de uma amostra multivariada caraterizados por variáveis qualitativas, em especial binárias[7] (as medidas de distância métrica não são aplicáveis). De entre os vários coeficientes de associação (s_{ij}) existentes, destaquem-se: coeficientes de emparelhamento simples (*simple matching*), coeficientes de Jaccard, coeficiente de Russel & Rao, coeficiente de Sorenson e coeficiente de Gower e Legendre.

Considere-se dois sujeitos i *e* j caraterizados por p variáveis nominais dicotómicas onde 1 e 0 significam, respetivamente, a presença e a ausência da característica em questão, as medidas de semelhança entre os dois sujeitos baseiam-se, em geral, nas seguintes quatro quantidades:

a – o número de variáveis para os quais ambos os sujeitos tomam o valor 1 (Presente);

[7] Variáveis que apenas podem tomar dois diferentes valores (1/0, Presente/ Ausente, Sim/Não, Masculino/ Feminino, etc.)

b - o número de variáveis para os quais o sujeito i tomam o valor 1 (Presente) e o sujeito j toma o valor 0 (Ausente);

c – o número de variáveis para os quais o sujeito i tomam o valor 0 (Ausente) e o sujeito j toma o valor 1 (Presente);

d – o número de variáveis em que ambos os sujeitos tomam o valor 0 (Ausente).

O resumo do número de presenças e ausências das características das variáveis sob estudo para cada sujeito i e j pode ser representado na tabela de contingência:

Tabela 4.1 Tabela de contingência1

		Sujeito j		**Totais**
		1	**0**	
Sujeito i	**1**	a	b	a + b
	0	c	d	c + d
Totais		a+c	b+d	p = a+b+c+d

De acordo com a informação da tabela os coeficientes são definidos como:

- ***Coeficientes de emparelhamento simples*** *(simple matching measures)*

$$s_{ij} = \frac{(a+d)}{(a+b+c+d)} \qquad \text{ou} \qquad d_{ij} = \frac{(b+c)}{(a+b+c+d)}$$

s_{ij} – Mede a semelhança entre cada dois sujeitos, varia entre 0 e 1. Representa a razão entre o número de características presentes e ausentes simultaneamente nos dois sujeitos e o número de características totais;

d_{ij} – Mede a distância entre os dois sujeitos, varia entre 0 e 1. Representa a razão entre o número de características presentes num sujeito mas ausentes no outro e o número total de características.

- ***Coeficientes de Jaccard*** **–** Medem a semelhança ou dissemelhança entre dois sujeitos. Não contemplam o número de características ausentes em ambos os sujeitos.

$$s_{ij} = \frac{a}{a+b+c} \;;\quad 0 \leq s_{ij} \leq 1 \qquad ou \qquad d_{ij} = \frac{b+c}{a+b+c} \;;\quad 0 \leq d_{ij} \leq 1$$

- ***Coeficiente de Russel &Rao*** **–** Dá-nos a perfeita semelhança ($s_{ij} = 1$) quando b=c=d=0 e a máxima dissemelhança ($s_{ij} = 0$) quando a=0.

$$s_{ij} = \frac{a}{a+b+c+d};\quad 0 \leq s_{ij} \leq 1$$

- ***Coeficiente de Sorenson*** – Valoriza a ocorrência simultânea da característica presente nos sujeitos.

$$s_{ij} = \frac{2a}{2a+b+c} \;;\; 0 \leq s_{ij} \leq 1$$

- ***Coeficiente de Gower e Legendre*** – Toma a diferença entre concordâncias e discordâncias, relativamente ao número total de variáveis observadas. Ao contrário dos anteriores coeficientes, pode tomar valores negativos, situação que ocorre caso haja mais discordâncias do que concordâncias nos valores das variáveis para os sujeitos $i\ e\ j$. Toma valores entre -1 e 1.

$$s_{ij} = \frac{(a+d)-(b+c)}{a+b+c+d}$$

4.2.2.4.2. Medidas de semelhança para variáveis nominais com mais de dois níveis

Quando a variável qualitativa nominal possui mais do que dois níveis, o artifício usual é a transformação em variáveis binárias através da criação de variáveis fictícias *(dummies*).

Supor o vetor de variáveis qualitativas nominais:

$$y' = (y_1, y_2, \dots y_l)$$

onde a *i*-ésima componente assume l_i níveis, codificados de modo que

$y_i = j$, com $j = 1,2,\dots, l_i$

Supondo também que $\sum l_i = p$. Cada componente irá dar origem a l_i, variáveis binárias $x_k(i)$ tal que

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

Assim, o vetor y de dimensão l é transformado no vetor x de dimensão p, formado por componentes binárias. Esquematicamente tem-se:

$$y' = (y_1, y_2, \dots, y_l) \rightarrow x' = \left(\underbrace{0, \dots, 1, \dots, 0}_{l_1}; \dots; \underbrace{0, \dots 1, \dots, 0}_{l_l} \right)$$

$y_1 \qquad\qquad y_l$

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

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

e tem-se a situação anterior.

Para corrigir o desequilíbrio causado pelo diferente número de níveis de cada, variável, faz-se intervir no cálculo do coeficiente o número de níveis de cada variável. Supondo que há p variáveis, $y_1, \dots y_p$ com $l_1, \dots l_p$ níveis respetivamente, então o coeficiente de semelhança s_{ij} será

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

Sendo I a função indicatória dos níveis dos sujeitos, i e j, na variável k, isto é,

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

$y_k(i)$ e $y_k(j)$ são, respetivamente, os níveis dos sujeitos i e j, na variável k.

4.2.2.4.3. Medidas de semelhança para variáveis ordinais

No caso de variáveis qualitativas do tipo ordinal, uma solução simples é considera-las simplesmente qualitativas e aplicar qualquer um dos coeficientes definidos anteriormente. Este procedimento deixa de considerar a importante propriedade da ordem.

Podemos utilizar uma extensão do conceito de variáveis fictícias para este tipo de variáveis.

Assim, a mesma estratégia usada anteriormente para transformar cada possível realização, numa variável binária, de acordo com a ocorrência do atributo, também pode ser usado nesta situação. Porém deve ser considerada a questão da ordem.

Supor a variável ordinal y, nível de escolaridade, podendo assumir um dos seguintes valores:

1- Primeiro ciclo
2- Segundo ciclo
3- Terceiro Ciclo
4- Secundário
5- Universitário

pode-se criar cinco variáveis binárias, ou seja, $y \rightarrow (x_1, x_2, x_3, x_4, x_5)$.

Uma pessoa com o nível universitário $(x_5 = 1)$ é portadora das características anteriores.

As cinco variáveis binárias permitem definir os vetores associados

Primeiro ciclo $(1,0,0,0,0)$

Segundo ciclo$(1,1,0,0,0)$

Terceiro Ciclo $(1,1,1,0,0)$

Secundário $(1,1,1,1,0)$

Universitário $(1,1,1,1,1)$

A variável ordinal y definida por $y = j,\ j = 1,2,\dots,l$ é transformada em l variáveis dicotómicas x_k, tal que: se $y = k$ então $x_i = 1$, para i=1,2,…,k, e $x_i = 0$, para i=k+1,…,l.

Novamente podem ser usados os coeficientes de semelhança definidos para as variáveis binárias.

EXEMPLO

Suponhamos agora que duas pessoas, A e B, possuem o terceiro ciclo e secundário, respetivamente.

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

Obtemos:

		Pessoa B		Totais
		1	0	
Pessoa A	1	3	0	3
	0	1	1	2
Totais		4	1	5

O coeficiente de emparelhamento simples seria

$$S_{AB}=\frac{3+1}{3+0+1+1}=0,8$$

Enquanto que o coeficiente de Jaccard seria

$$S_{AB}=\frac{3}{3+0+1}=0,75$$

O coeficiente de Jaccard apresenta um valor menor porque não considera as características simultaneamente ausentes.

4.2.2.5 Coeficientes de semelhança para variáveis de diferentes tipos

É frequente a presença de diferentes tipos de variáveis (quantitativas e qualitativas) na procura de sujeitos semelhantes.

A seguir são apresentadas algumas estratégias para aplicar a uma matriz de dados que contém diferentes tipos de variáveis.

Coeficientes combinado de semelhança

Determina-se os coeficientes de semelhança de mesmo sentido (semelhança ou dissemelhança), s_n, s_o e s_q para cada grupo de variáveis (nominais, ordinais e quantitativas), depois constrói-se um único coeficiente ponderado.

Para dois sujeitos A e B esquematicamente tem-se:

$$s_{AB=} w_1 . s_{n_{AB}} + w_2 . s_{o_{AB}} + w_3 . s_{q_{AB}}$$

Onde os w_p, $p = 1,2,3$ são os pesos associados. É comum ponderar pelo número de variáveis envolvidas.

Proposta de Gower

Gower (1971) propõe uma forma mais elaborada do coeficiente de semelhança combinado.

O coeficiente de entre os sujeitos A e B, segundo as *p* variáveis, de qualquer tipo, passa a ser:

$$S_{AB} = \sum I_{i_{AB}} \, S_{i_{AB}} \, / \sum I_{i_{AB}}$$

Para cada variável x_j é definido um coeficiente de semelhança $S_{i,}$ com valores entre 0 e 1. A variável I_i , assume o valor 1 quando a comparação dos sujeitos é possível segundo o critério i, e assume o valor 0 em caso contrário, isto é se o valor da variável é omisso em pelo menos um dos sujeitos A e B. Este coeficiente será indefinido quando todos $I_{i_{AB}} = 0$, ou seja, a comparação dos dois sujeitos não é válida segundo nenhum critério.

Este coeficiente torna-se idêntico ao coeficiente de Jaccard quando as variáveis são todas binárias.

Proposta de Romesburg

Romesburg (1984), sugere esquecer a natureza das variáveis, tratar todas como variáveis quantitativas (todas são codificadas com números) e aplicar a distância euclidiana. A grande desvantagem está na interpretação dos valores dos coeficientes de semelhança, pois estes dependem da codificação das variáveis.

4.2.2.6. Conversão das semelhanças em dissemelhanças

Quando o ponto de partida é uma matriz de semelhanças de um conjunto de sujeitos, pode-se trabalhar diretamente sobre estas medidas de semelhança ou, alternativamente converter as medidas de semelhança em dissemelhança. É possível estabelecer uma relação entre as semelhanças e dissemelhanças dos sujeitos:

$$\boldsymbol{d_{ij} = 1 - s_{ij}}$$

$$\boldsymbol{d_{ij} = 1 - s_{ij}^2}$$

$$\boldsymbol{d_{ij} = \sqrt{1 - s_{ij}^2}}$$

$$\boldsymbol{d_{ij} = \sqrt{1 - s_{ij}}}$$

A primeira destas regras de conversão, no caso de se utilizar o Coeficiente de emparelhamento simples como medida de semelhança, equivale a tomar o coeficiente de dissemelhança.

4.2.3. Medidas de semelhança entre variáveis

É também possível agrupar variáveis através duma Análise de Clusters. Neste caso, o ponto de partida será uma matriz de semelhanças (ou dissemelhanças) entre variáveis.

As variáveis tomam o lugar dos sujeitos e podemos aplicar as medidas de (dis)semelhança utilizadas na análise de sujeitos.
De um modo geral o agrupamento de variáveis é baseado em medidas de correlação ou associação.

Independentemente dos critérios de semelhança e/ ou dissemelhança adotados entre variáveis, o procedimento de classificação que se segue será análogo aos procedimentos de classificação de sujeitos anteriormente apresentados.

4.2.3.1. Medidas de semelhança entre variáveis quantitativas

Seja o comportamento das variáveis X e Y, representados pelos vetores X= $(x_1, \ldots, x_n)'$ e Y= $(y_1, \ldots, y_n)'$. O valor observado da variável no i-ésimo objeto é representado pelo i-ésimo componente de cada vetor, i=1, …, n.

Coeficiente de correlação de Pearson – este coeficiente varia entre [-1,1,] mede a intensidade e a direção da associação de tipo linear entre duas variáveis quantitativas e define-se como

$$r_{xy=} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Poderemos também usar a Medida de Semelhança do Cosseno com variáveis quantitativas.

Medida de Semelhança do Cosseno

$$s_{xy} = \mathrm{CoSIN}(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{k=1}^{p} y_i^2}}$$

$CoSIN(x, y)$, cosseno do ângulo formado pelos dois vetores. À medida que o vetor x "aproxima" de y, o cosseno do ângulo cresce.

4.2.3.2. Medidas de semelhança entre variáveis nominais binárias

Retomemos à Tabela 4.1, nesta situação em que i *e* j representam a i-ésima e a j-ésima variáveis e tomando os valores 1 e 0 para representar as duas categorias das variáveis, os coeficientes de correlação de Pearson e de medida de semelhança do cosseno são dados por

Coeficientes de correlação de Pearson

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

Medida de semelhança do cosseno

$$\mathrm{CoSIN}(\mathrm{i},\mathrm{j}) = \frac{a}{\sqrt{(a+b)(a+c)}}$$

4.2.3.3. Medidas de semelhança entre variáveis nominais com mais de dois níveis

Consideremos as variáveis X e Y com as categorias 1,…,p e 1,…q respetivamente. Suponhamos uma tabela de contingência onde n representa o número total de observações, n_{ij} representa a frequência absoluta do par (X_i,Y_j), e n_i e n_j as frequências marginais de X e Y, respetivamente. Analogamente, definamos as frequências relativas, f_{ij} , f_i e f_j.

Tabela 4.2 Tabela de Contingência2

X	Y						Total
	Y_1	Y_2	…	Y_j	…	Y_q	
X_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1q}	$n_{1.}$
X_2	n_{11}	n_{22}	…	n_{2j}	…	n_{2q}	$n_{2.}$
…	…	…	…	…	…	…	…
X_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{iq}	$n_{i.}$
…	…	…	…	…	…	…	…
X_p	n_{p1}	n_{p2}	…	n_{pj}	…	n_{pq}	$n_{p.}$
Total	$n._1$	$n._2$	…	$n._j$	…	$n._q$	n

Apresenta-se de seguida a medida mais usada *Qui-quadrado* e as variações desta:

Qui-quadrado χ^2de Pearson

$$\boldsymbol{\chi^2 = n \sum_{i=1}^{p} \sum_{j=1}^{q} (f_{ij} - f_{i.} f_{.j})^2 \,/\, f_{i.} f_{.j}}$$

Coeficiente de contingência quadrática média

$$\boldsymbol{\phi^2 = \frac{\chi^2}{n}}$$

Coeficiente de contingência de Pearson

$$\boldsymbol{P = \left(\frac{\phi^2}{1+\phi^2}\right)^{1/2}}$$

Coeficiente de Tschuprow

$$\boldsymbol{T = \left[\frac{\phi^2}{(p-1)(q-1)}\right]^{1/2}}$$

Coeficiente V de Cramer

$$\boldsymbol{V = \sqrt{\frac{\phi^2}{min(p-1, q-1)}}}$$

4.2.3.4. Medida de semelhança entre variáveis ordinais

Como já foi referido o coeficiente de Spearman é uma das medidas de associação entre variáveis ordinais mais usada. Toma valores no intervalo [-1,1,] pode obter-se usando a fórmula do coeficientes de correlação de Pearson, substituindo os valores das observações X e Y pelas respetivas ordens r_1 *e* r_2 *:*

Coeficiente de Correlação de Spearman

$$r_{s=} \frac{\sum_{i=1}^{n}\left(r_{1_i} - \bar{r_1}\right)\left(r_{2_i} - \bar{r_2}\right)}{\sqrt{\sum_{i=1}^{n}\left(r_{1_i} - \bar{r_1}\right)^2 \sum_{i=1}^{n}\left(r_{2_i} - \bar{r_2}\right)^2}}$$

4.3. Métodos hierárquicos

Permitem a obtenção de C*lusters* quer para indivíduos quer para variáveis. Dividem-se em ***aglomerativos*** e ***divisivos***, os mais divulgados e mais utilizados são os hierárquicos aglomerativos. No método aglomerativo, o agrupamento em classes procede por etapas, em geral determinando-se a partir de *n* subgrupos (de um individuo cada) sucessivas

fusões de subgrupos considerados mais "semelhantes", até se encontrar apenas um grupo ou *Cluster* que incluirá a totalidade de *n* indivíduos. No método divisivo o processo é inverso.

O ponto de partida para os métodos hierárquicos é, em geral, uma matriz *n* x *n* cujo elemento genérico (i , j) é uma medida de semelhança (ou dissemelhança) entre o individuo i e o individuo j.

O método aglomerativo desenrola-se de acordo como seguinte algoritmo:

1. Começar com *n Clusters* (um para cada individuo ou variável) e calcular a matriz de dissemelhança (ou de semelhança) $D_{n \text{ x } n;}$

2. Encontrar na matriz os pares *Clusters* (indivíduos ou variáveis) mais semelhantes de acordo com uma medida de distância escolhida;

3. Com os *Cluster* i e j encontrados formar um *Cluster* maior, *Cluster ij* e recalcular a distância deste *Cluster* para os restantes *Clusters* originais;

4. Repetir os passos 2 e 3 até sobrar um único *Cluster*.

4.3.1. Métodos de (des)agregação - Caraterísticas

As medidas de distância entre os indivíduos não representam a única opção a fazer numa Análise de *Clusters.* É necessário também escolher o método de (des)agregação dos indivíduos.

Os métodos hierárquicos de agrupamento diferem no modo como calculam as distâncias entre grupos e os restantes (grupos ou indivíduos). Os métodos mais utilizados são os seguintes:

- ***Menor distância*** (*Single linkage ou Nearest neighbor*) – Consiste em considerar que a distância entre dois grupos é a *menor* distância entre um elemento dum grupo e um elemento do outro grupo. Seja k a distância e dois grupos (i , j):

$$d_{(l,j)k} = min\{d_{lk}\,; d_{jk}\}$$

  Este método tende a produzir classes com indivíduos que podem estar muito distantes entre si, mas pertencendo a uma mesma classe. Este fato resulta de

bastar que exista um elemento duma classe "próximo" de um único elemento de outra classe para que estas sejam atraídas.

- ***Maior distância*** (*Complete linkage ou farthest-neighbor*) – Consiste em considerar que a distância entre dois grupos é a maior distância entre um elemento dum grupo e um elemento do outro grupo. Seja k a distância e dois grupos (i , j):

$$d_{(l,j)k} = max\{d_{lk}\,;\, d_{jk}\}$$

Caso os sujeitos sejam representáveis em R^p , este método tem tendência a produzir classes onde não há grandes diferenças nas distâncias entre pares de elementos mais distantes, ao longo de várias direções - classes *"esféricas"*.

- ***Distância Média entre Clusters*** (*Average linkage between groups*) – Consiste em considerar que a distância entre dois grupos é a média de todas as distâncias entre pares de elementos (um de cada grupo). Seja k a distância e dois grupos (i , j):

$$d_{(l,j)k} = média\,\{d_{lk}\,;\, d_{jk}\}$$

Assim como o método da menor distância este método também tem tendência a produzir classes "esféricas".

- ***Distância Média dentro dos Clusters*** (*Average linkage within groups*) – Semelhante à "Distância média entre *clusters*" mas neste método os *clusters* são unidos de modo a que a soma de quadrados dos erros seja mínima.

- ***Distância Mediana*** (*Median linkage*) – Após formado o primeiro Cluster, a distância deste aos restantes é a mediana das distâncias de cada um dos elementos constituintes deste Cluster a cada um dos restantes indivíduos ou variáveis. Seja k a distância e dois grupos (i , j):

$$d_{(l,j)k} = mediana\{d_{lk}\,;\, d_{jk}\}$$

- ***Método Centroide* –** Toma-se a distância entre dois grupos como sendo a distância entre os centros de gravidade, ou outros pontos considerados "representativos" (centróides), dos grupos. O método do centroide calcula a distância entre dois grupos como a diferença entre as suas médias, para todas as variáveis.

  Este método também tem tendência a produzir classes "esféricas".

- ***Método Ward* (Método da Inércia Mínima) –** Neste método os Clusters são formados de modo a minimizar a soma dos quadrados dos desvios das observações individuais relativamente às médias dos grupos em que são classificadas.

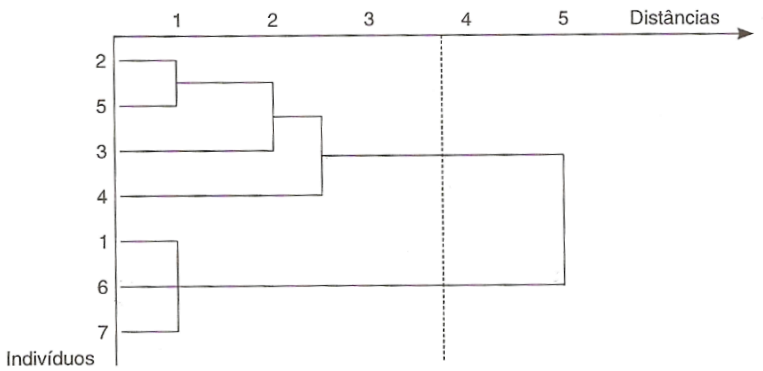
  O método Ward tem tendência a produzir classes com um número aproximado igual de sujeitos.

A escolha do método de (des)agregação condicionará a classificação obtida. A escolha do método deve ser justificada com base na natureza dos dados e no objetivo da análise.

O processo de agrupamento é possível ser visualizado através de uma representação gráfica denominada *dendrograma*. O primeiro mostra todas as fases do processo do agrupamento desde a separação total dos indivíduos até à sua inclusão num grupo apenas. A posição na escala horizontal indica a distância a que os *Clusters* são agrupados. Um corte no dendrograma a qualquer nível de aglomeração produz uma classificação em K grupos ($1 \leq K \leq n$).

Numa classificação segundo uma dada escolha de distâncias entre indivíduos e classes, a representação em dendrograma não é única, uma vez que a ordem dos indivíduos é arbitrária. Reordenações dos indivíduos podem produzir dendrogramas de aspeto diferente, mas a informação neles contida é idêntica.

Figura 3 Dendrograma



A partir do dendrograma anterior, fazendo um corte a uma distância de aproximadamente 3 é possível identificar a existência de dois grupos: (2, 5, 3, 4) e (1, 6, 7).

4.4. Escolha do número de *clusters*

O objetivo da análise de *clusters* é formar grupos homogéneos. Assim, coloca-se o problema da escolha do número apropriado de *clusters*. Existem várias técnicas para se determinar o número adequado de *clusters*.

4.4.1.Análise do dendrograma

O procedimento básico consiste no exame do dendrograma procurando grandes alterações (saltos) na distância para as sucessivas fusões.

Qualquer método produzirá sempre uma classificação em qualquer número de classes, conforme o nível em que decidamos cortar o dendrograma. Por este facto a análise de *clusters* produz classificações mesmo onde elas possam não fazer sentido. Assim, é importante verificar a robustez das classificações obtidas.

Uma boa classificação deverá corresponder a um agrupamento que resulte de cortar o dendrograma numa zona onde as separações entre classes correspondam a grandes distâncias (barras de junção de classes relativamente compridas) - heterogeneidade entre classes. A homogeneidade interna das classes, será tanto maior quanto mais próximo dos indivíduos se fizer o corte.

4.4.2. Coeficiente de fusão

É o valor numérico (distância ou semelhança) para o qual os vários indivíduos se unem para formar um grupo. A comparação gráfica do número de *clusters* com o valor do coeficiente de fusão permite sugerir a escolha do número de clusters. Quando a divisão de um novo grupo não introduz alterações no coeficiente de fusão poderá tornar-se essa partição como sendo ótima (Reis 2001). A escolha ótima coincidirá com uma marcada horizontalidade da curva.

4.5. Métodos não hierárquicos

Os métodos não hierárquicos são válidos apenas para a obtenção de *clusters* de indivíduos (e não de variáveis). Estes métodos são enquadrados como partitivos (dividem os n dados existentes em *K* partições). Fixa-se à partida o número de partições que se pretende constituir e (regra geral) faz-se a classificação dos n indivíduos em *K* Clusters, de modo a otimizar algum critério de homogeneidade interna e heterogeneidade externa. Existem vários métodos não-hierárquicos, o desempenho destes métodos depende da primeira agregação dos indivíduos em *clusters*, e do modo como as novas distâncias entre os centroide[8] dos *clusters* e os indivíduos é calculada.

Um dos métodos partitivos mais frequente nos *softwares* estatísticos é o ***K-means*** que se desenrola nos seguintes passos (Johnson & Wichern,2002):

1. Partição inicial dos sujeitos em *K Clusters* definidos à partida pelo investigador;
2. Calculo dos centróides para cada um dos *K Clusters* e cálculo da distância euclidiana dos centróides a cada sujeito na base de dados;
3. Agrupar os sujeitos aos *Clusters* de cujos centróides se encontram mais próximos, e voltar ao passo 2 até que não ocorra variação significativa na distância mínima de cada sujeito da base de dados a cada um dos centróides dos *K clusters* (ou até que o número máximo de interações ou o critério de convergência – definido pelo analista – seja alcançado).

Cada indivíduo é transferido para o *cluster* que apresenta uma menor distância (por exemplo, distância euclidiana) entre o indivíduo e o centroide do *cluster*. Assim, é

[8] Centroides são os valores médios contidos em cada uma das variáveis do *cluster*.

necessário conhecer os centróides de cada *cluster* ou calculá-los a partir dos dados originais.

4.6. Outros métodos

4.6.1. TwoStep *Cluster* (Análise de *clusters* em duas fases)

Os tradicionais métodos de agrupamento são eficientes e rigorosos quando aplicados a pequenos conjuntos de dados. O mesmo não se verifica no caso de conjuntos de dados muito grandes. Para aplicar os métodos tradicionais é necessário previamente reduzir a dimensão da base de dados, ou seja, o agrupamento é realizado em dois passos como BIRCH (Zhang *et al.* 1996). O método TwoStep *Cluster* usa este procedimento, permitindo dar resposta a conjuntos de dados de enorme dimensão e a utilização de variáveis contínuas, categóricas ou os dois tipos de variável em simultâneo.

Passo 1: formação de uma série de preclusters. O objetivo deste passo é reduzir o tamanho da matriz das distâncias entre todos os pares de casos possíveis. Nesta primeira etapa os dados são percorridos um a um e o algoritmo decide se um determinado individuo deve migrar para um precluster previamente formado ou iniciar um novo precluster. No fim deste procedimento e todos os indivíduos pertencentes ao mesmo precluster são tratados como uma só entidade. Assim, a matriz de distâncias é menor, pois o seu tamanho passa a depender do número de preclusters.

Passo 2: agrupamento dos preclusters. No segundo passo, ocorre o agrupamento hierárquico (dos preclusters formados na etapa anterior) de acordo com o número de *clusters* pretendido.

Medidas de distância: se só existirem variáveis continuas, é possível usar a distância euclidiana entre o centro de dois *clusters*. Quando existem misturadas variáveis contínuas e categóricas, é utilizada a função log-verossimilhança, a distância entre dois *clusters* é expressa pelo decréscimo da função log-verossimilhança. Neste caso o algoritmo fornece melhores resultados quando se verifica a normalidade das variáveis contínuas, e a distribuição multinomial no caso das variáveis categóricas. Estes

pressupostos são difíceis de encontrar em dados reais, no entanto, o algoritmo encontra uma solução razoável mesmo quando os pressupostos são quebrados.

Principais vantagens: utilização de variáveis contínuas e categóricas simultaneamente; agrupamento em duas etapas, aumentando a eficiência do método; o próprio algoritmo encontra um número ótimo de *clusters*, sendo também possível especificar o número de clusters desejado; de fácil interpretação, são disponibilizadas informações sobre a importância de cada variável na formação de cada *cluster* e uma medida de significância estatística (Qui-quadrado para variáveis categóricas e t-test para variáveis contínuas), permitindo a confirmação dos perfis definidos.

4.6.2. Técnicas de densidade

Com aplicação quando se espera observar grupos naturais. Os agrupamentos são formados através da procura de regiões que contenham uma concentração relativamente densa de pontos (casos). Regiões com muitos pontos próximos no espaço, separadas por áreas com poucos pontos (representando ruídos) sugerem *clusters.* Em geral usa-se o método da ligação simples para a obtenção dos *clusters*. Começa-se por escolher um raio r e o número de pontos *P*. Uma região densa – pontos de densidade - é uma região onde uma vizinhança (círculo de raio *r*) de cada ponto contém pelo menos *P* pontos. Os primeiros *clusters* são definidos pelos pontos de densidade. Um ponto cuja distância a todos os pontos de densidade seja superior a *r,* form*a* o seu próprio *cluster*. O algoritmo repete-se e só para quando não houver mais possibilidades de se juntar *clusters* da etapa precedente.

4.6.3. Agrupamentos fuzzy

O agrupamento fuzzy é uma generalização dos métodos partitivos, permitindo que haja sobreposição dos grupos (fuzzy *clusters*). É possível observar o grau de associação de cada elemento em cada grupo, que geralmente de verifica em domínios de dados reais, onde um elemento pertence a diferente grupos, com diferentes graus de associação – vantagem relativamente a outros métodos por partição. No entanto, apresenta a desvantagem do número de coeficientes de associação crescer rapidamente com o aumento do número de elementos e de grupos. Trata-se de uma técnica válida, pois

associa graus de incerteza aos elementos nos grupos, situação que se aproxima das características reais dos dados (Kaufman, 1990).

Um método de agrupamento fuzzy frequentemente utilizado é o *fuzzy c-means.* Este método iterativo inicia-se com *c* valores arbitrários, com base nos quais associa cada elemento ao valor ao qual possui menor distância, formando *c* grupos. Depois determina-se o centro de cada *cluster* formado, e os elementos são reagrupados ao centro mais próximo. Este procedimento termina quando as diferenças entre os centros do passo atual e do anterior sejam mínimas.

4.7. Métodos hierárquicos / métodos não hierárquicos

Enquanto a aplicação dos métodos hierárquicos requer o cálculo de uma matriz de dissemelhanças, os métodos não-hierárquicos aplicam-se diretamente sobre os dados originais, permitindo a sua aplicação a matrizes de dados muito grandes.

Os métodos não-hierárquicos permitem reagrupar os indivíduos num *cluster* diferente daquele em que foram inicialmente incluídos. Nos métodos hierárquicos os indivíduos que sejam incluídos num mesmo *cluster* em qualquer etapa do processo não poderão mais ser separados em etapas posteriores.

A necessária definição, à partida, do número de *clusters*, sem conhecimento da estrutura dos dados, pode representar uma desvantagem nos métodos não-hierárquicos.

Num problema de análise de *clusters*, a validade das soluções encontradas aumenta, se o processo de análise começar com um método hierárquico aglomerativo para determinar o número de grupos, e proceder com o *k-means* para refinar (otimizar) a partição encontrada.

4.8. Escolha da técnica a utilizar

Qualquer um dos métodos de análise de *clusters* impõe um certo grau de estrutura nos dados, e para o investigador assegurar que o resultado obtido não é um artefacto da técnica utilizada, é aconselhável usar diferentes critérios de agrupamento e escolher a estrutura resultante da maior parte deles.

Podemos utilizar outra estratégia para averiguar a estabilidade do agrupamento. Esta estratégia consiste em formar ao acaso, dois subconjuntos do conjunto de observações e aplicar em cada um deles o mesmo critério. A alocação dos sujeitos nas subamostras e na amostra total será semelhante se o agrupamento for estável.

5. ESTUDO DE VARIÁVEIS ASSOCIADAS COM A AVALIAÇÃO NA DISCIPLINA DE MATEMÁTICA

5.1. Identificação dos sujeitos da investigação. Dimensão da amostra

A população-alvo deste estudo é composta pelos alunos do 9º ano de uma escola pública do Ensino Básico. Esta população foi escolhida por nos encontrarmos numa situação privilegiada, no contexto escolar, para inquirir os alunos. Assim, podemos afirmar que se utilizou uma escolha por conveniência ou acessibilidade.

Tendo em conta o número de variáveis e os objetivos desta pesquisa passamos então a determinar a dimensão mínima da amostra para que as análise estatisticas posteriores sejam válidas. Dado que, neste trabalho, pretendemos aplicar várias técnicas de estatística (univariadas e multivariadas), para determinar a dimensão da amostra é necessário utilizar um critério. Seguindo Manuela e Andrew Hill, (2009) podemos utilizar a "Regra do Polegar" que indica que para k variáveis são necessários n=10k sujeitos. Assim, para o nosso estudo, havendo K=13 variáveis necessitamos de um mínimo de n=130 sujeitos.

A nossa amostra é constituída por 142 alunos.

5.2. Instrumento de recolha de dados

O método de recolha de dados utilizado foi o inquérito por questionário.

Construímos um questionário que nos permitiu conhecer a opinião dos alunos em relação ao processo ensino-aprendizagem, nomeadamente no que se refere ao relacionamento entre professor e aluno, o efeito do estatuto socioeconómico no seu rendimento escolar, etc..

O questionário foi anónimo, contendo essa anotação, bem como a da garantia de confidencialidade das respostas no cabeçalho, para a que os inquiridos pudessem responder sem quaisquer tipos de restrições.

A distribuição e aplicação dos questionários teve lugar no primeiro trimestre de 2012, tratando-se de um questionário com questões referentes ao ano letivo anterior (2010/2011). A recolha das respostas ao questionário realizou-se em sala de aula a todos os alunos presentes. Todos os questionários recolhidos foram validados.

5.3. Pré-teste e ajuste do questionário

A investigação principal foi precedida por um pré-teste realizado numa amostra de alunos na mesma situação escolar (a frequentar o nono ano) da amostra sobre a qual recaiu o estudo final. Dos 25 questionários-teste entregues rececionamos 20.

O pré-teste destinou-se a contabilizar o tempo despendido na resposta ao questionário e a detetar eventuais falhas tais como: legibilidade, erros de lógica e/ou perceção das perguntas. Aos respondentes foi solicitada a indicação, no próprio questionário, das principais dificuldades no seu preenchimento.

Após examinadas as dificuldades sentidas, resolvemos, na questão 3, simplificar o quadro relativo à classificação final e nas questões 7 e 8 adicionar uma nota esclarecedora.

Assim, o questionário-teste com as referidas alterações deu lugar ao questionário final da investigação (Anexo II).

5.4. Análise dos questionários. Tratamento dos dados

Depois de recolhidos os dados, as respostas fornecidas pelos alunos às diversas questões foram codificadas, de acordo com as regras do programa informático SPSS (Tabela 1, do Anexo III)

As variáveis "Categoria Profissional da mãe" e "Categoria Profissional do pai" foram definidas com base na Classificação Nacional de profissiões em Portugal (ver Anexo I).

Os dados foram inseridos no *software* SPSS, versão 19. Na Tabela 1 do Anexo III são definidas, em detalhe, as características específicas, de cada uma das 27 variáveis no questionário. Estas caraterísticas são, nomeadamente, o nome (*Name*), o tipo de variável (*Type*), o rótulo (*Label*), os valores da variável (*Values*), e escala de medida (*Measure*), etc..

5.5. Análise exploratória dos dados

Iniciamos com uma análise univariada dos dados, seguindo-se uma análise bivariada de forma a ser possível determinar associações entre alguns pares de variáveis em estudo.

Por fim, passamos a uma análise multivariada dos dados, Análise de Correspondências Múltiplas e Análise de Clusters.

5.5.1. Análise Univariada

Nesta secção, iremos apresentar gráficos, tabelas, medidas de tendência central e de dispersão, entre outras, de forma a ser possível descrever a nossa amostra, evidenciando as principais característica da mesma.

Assim, em primeiro lugar é de salientar que na nossa amostra composta por 142 indivíduos, predomina o sexo masculino, isto é, 67 são do sexo feminino e 75 do masculino (gráfico 5.1 e tabela 5.1).

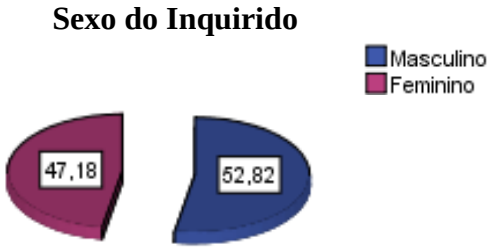


Gráfico 5.1 Distribuição dos inquiridos por sexo

		Frequency	Percent
Valid	Masculino	75	52,8
	Feminino	67	47,2
	Total	142	100,0

Tabela 5.1 Tabela de frequências da variável "Sexo do inquirido"

No que diz respeito ao ano letivo anterior (2010/2011), e ao qual se remetem as questões deste inquérito, 80,28% dos alunos transitaram de ano tendo-se verificado uma percentagem de 19,72% de alunos que ficaram retidos (Gráfico 5.2).

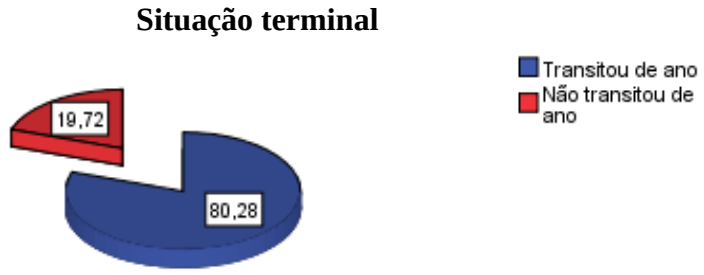


Gráfico 5.2 Situação em que os alunos terminaram o ano letivo anterior

Passemos ao estudo das três variáveis quantitativas nesta análise, "Idade do Inquirido", "Nível a Matemática" e "Nível a Português". É de referir que a idade dos alunos varia entre os 14 e os 16 (Gráfico 5.3) sendo a idade mais observada a de 14 anos.

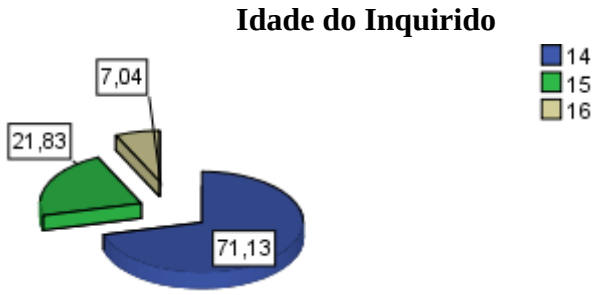


Gráfico 5.3 Distribuição dos inquiridos por idades (Março de 2012)

Observando a tabela 5.2 confirma-se efetivamente que a idade mais frequente (Moda) é de 14 anos; verifica- se também que a idade média destes é 14,36 anos e que 50% dos jovens inquiridos têm idade que varia entre os 14 e os 15 anos (Percentil 25 e Percentil 75). Sabendo que os alunos portugueses, na sua maioria, ingressam no 1º ciclo com 6 anos de idade, seria, por isso, de esperar que ao frequentar o 9º ano tivessem entre 14 e 15 anos, dependendo da data de nascimento destes. Analisando os resultados anteriores constatamos que a nossa amostra está em concordância com estas características, uma vez que a maioria dos jovens inquiridos tem 14 anos. De salientar ainda que a situação dos inquiridos que referem ter 16 anos (7,04%), dever-se-á a situações de retenções ocorridas em anos anteriores (Gráfico 5.3).

		Idade do inquirido	Nível a Matemática	Nível a Português
N	Valid	142	142	142
	Missing	0	0	0
Mean		14,36	3,02	3,18
Median		14,00	3,00	3,00
Mode		14	2	3
Std. Deviation		,611	1,021	,836
Minimum		14	2	2
Maximum		16	5	5
Percentiles	25	14,00	2,00	3,00
	50	14,00	3,00	3,00
	75	15,00	4,00	4,00

Tabela 5.2 Medidas descritivas das variáveis "Idade do inquirido", "Nível a Matemática" e "Nível a Português".

Analisando os resultados para a variável "Nível de Matemática" no ano anterior o nível mais observado é o nível 2 (38,03%), seguido do nível 3 (34,51%), depois do nível 4 (14,79%) e, por último nível 5 (12,68%). Por conseguinte, 38,03% dos alunos inquiridos obtiveram uma classificação inferior a 3 e 61,97% uma nota positiva (Gráfico

5.4). De salientar ainda que a percentagem dos resultados observados decresce do nível 2 (mínimo observado) para o nível 5 (máximo).

Observando a tabela 5.2 confirma-se que o valor modal é o nível 2, assim como o valor mediano. Verifica-se também que 50% dos jovens inquiridos têm classificação a Matemática a variar entre o nível 2 e o nível 4 (Percentil 25 e Percentil 75), o que permite concluir que estamos perante uma amostra com rendimento a Matemática pouco satisfatório.

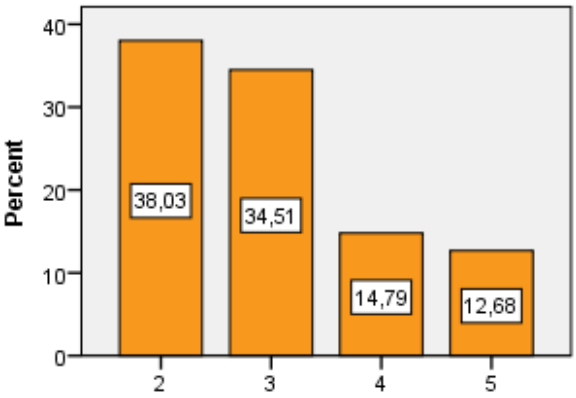


Gráfico 5.4 Nível a Matemática no ano anterior

Na disciplina de Português, comparativamente com Matemática, verificamos menor percentagem de nível 2 (19,01%) e maior de nível 3 (54,23%) Tal como na disciplina de Matemática, os níveis menos observados são níveis 4 e 5 (Gráfico 5.5). Observando a tabela 5.2, confirma-se que o valor modal e o valor mediano é o nível 3. Verifica-se também que 50% dos jovens inquiridos têm classificação a Português a variar entre o nível 3 e o nível 4 (Percentil 25 e Percentil 75), o que permite concluir que estamos perante uma amostra com rendimento satisfatório nesta disciplina.

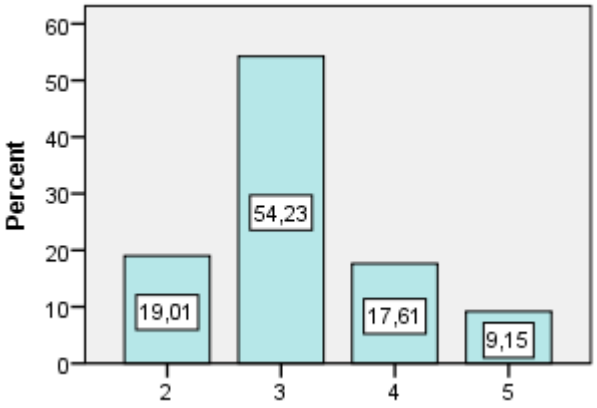


Gráfico 5.5 Nível a Português no ano anterior

Foi questionado aos alunos se gostam de estudar. Os alunos afirmaram: não gostar (18,31%), em parte (54,93%) e apenas 26,76% assumiu gostar de estudar (Gráfico 5.6).

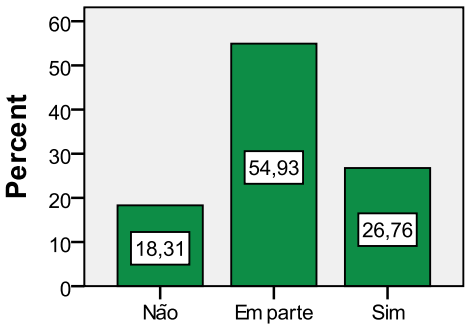


Gráfico 5.6 Gosto de estudar

Relativamente ao agregado familiar, destaca-se o agregado constituído por pai, mãe, irmãos e avós (44,37%), seguido do agregado pai e mãe. Salienta-se ainda que 86,62% dos jovens vive com ambos os pais e que em 13,38% dos agregados familiares está ausente uma das figuras parentais (Gráfico 5.7).

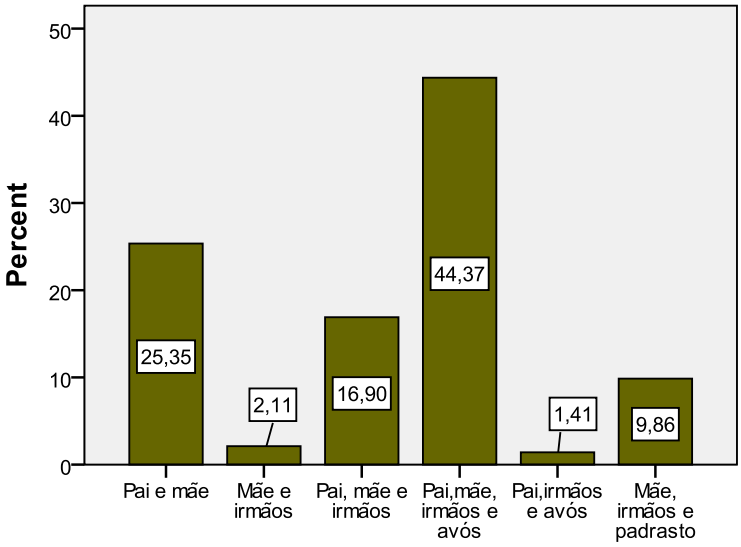


Gráfico 5. 7 Agregado familiar

No Gráfico 5.8 podemos observar a distribuição dos encarregados de educação, destaca-se a mãe com uma percentagem de 78,17% , por outro lado o pai representa 19,72% dos encarregados de educação observados, e apenas 2,11% dos inquiridos têm a avó como encarregada de educação.

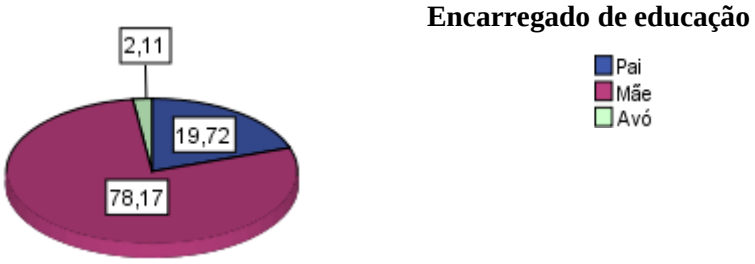


Gráfico 5.8 Encarregado de educação

No contexto socioeconómico dos jovens em análise, observando os Gráficos 5.9 e 5.10, verificamos que maior percentagem observada, tanto de pais (26,06%) como de mães (28,17%), se enquadra na categoria profissional " Trabalhadores Não Qualificados". A a segunda categoria mais observada, também para ambos os pais, é a de "Especialistas /Técnicos", 27,46% para as mães e 23,94% para os pais.

Para o caso das mães, segue-se a categoria profissional "Administrativos/Serviços/Vendedores" (26,06%) logo seguida pela categoria

profissional “Operários/Artífices (11,27%). No caso dos pais a terceira categoria profissional mais observada é a de “Operários/Artífices” (21,13%) seguida da categoria “Administrativos/Serviços/Vendedores” (19,72%).

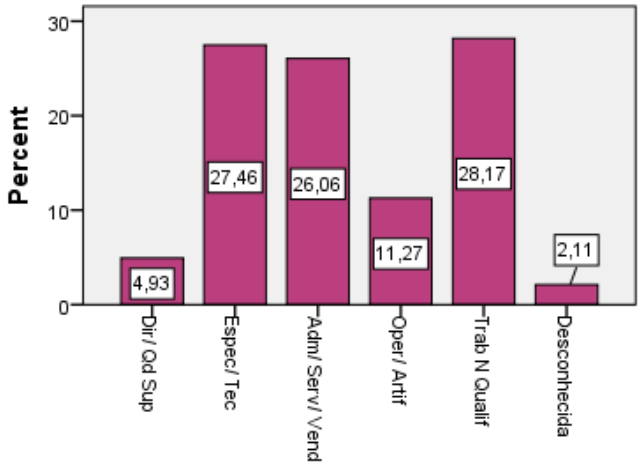


Gráfico 5.9 Distribuição das profissões da mãe do inquirido

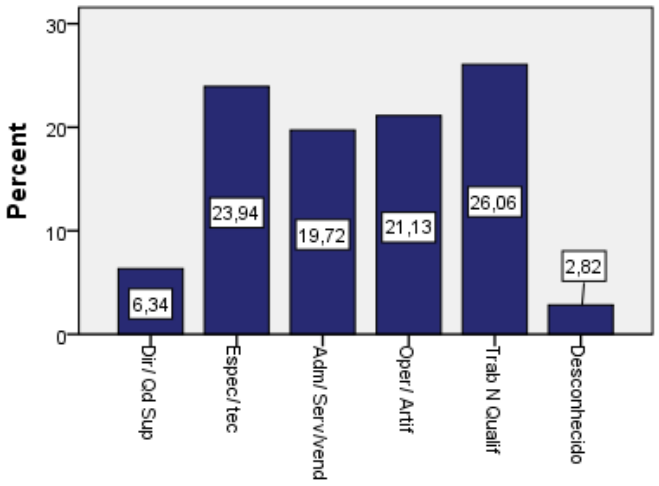


Gráfico 5.10 Distribuição das profissões do pai do inquirido

Ainda no contexto socioeconómico dos jovens, observando os Gráficos 5.11 e 5.12, evidencia-se a percentagem tanto de pais (68,31%) como de mães (69,01%), que se encontram empregados. Em situação de desemprego observamos 16,20% dos pais e, no caso das mães uma percentagem ligeiramente superior, 18,31%. Convém referir, no caso das mães, que alguns alunos indicaram a situação “Em casa (a cuidar do lar)”, a qual foi enquadrada em “Outro/Desconhecido”.

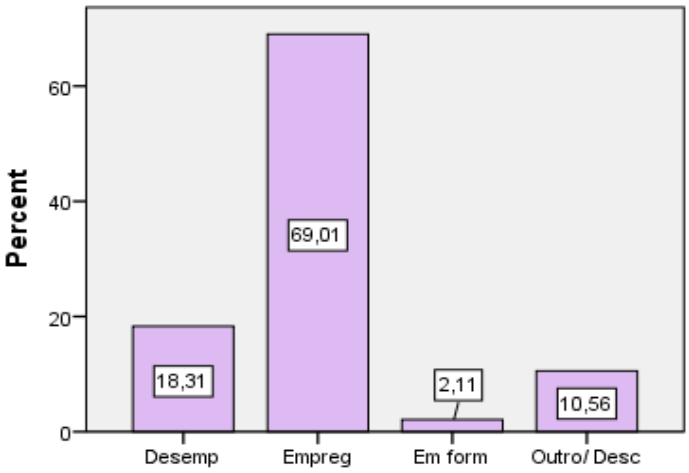


Gráfico 5.11 Situação profissional da mãe do inquirido

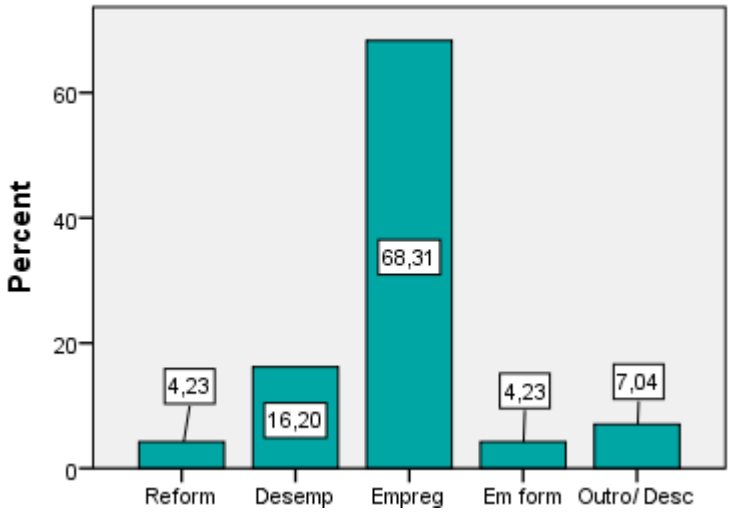


Gráfico 5. 12 Situação profissional do pai do inquirido

No caso dos pais verificamos a situação de Reformado (4,23%), situação não observada nas mães.

Note-se que, no decorrer deste estudo, se assiste a nível mundial a uma crise económica que tem provocado o encerramento de várias empresas e o despedimento de inúmeros funcionários.

Como tal a percentagem de situações de desemprego assinaladas podem incluir situações de desemprego recente e que, consequentemente, ainda não tiveram repercussões a nível socioeconómico.

Questionamos os alunos quanto à existência, em suas casas, dos seguintes recursos: computador, internet, livros não escolares e acesso a canais de TV temáticos. Observando os gráficos 5.12 a 5.15, verificamos que todos os alunos indicam ter computador em casa, 92,96% dos jovens tem acesso a internet, a livros não escolares e a canais TV temáticos.

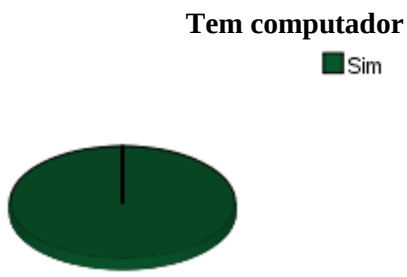


Gráfico 5.12- Tem computador em casa?

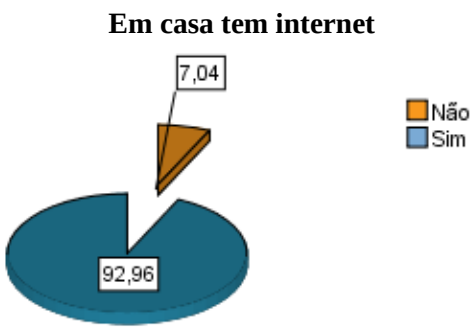


Gráfico 5. 13 Em casa tem acesso internet?

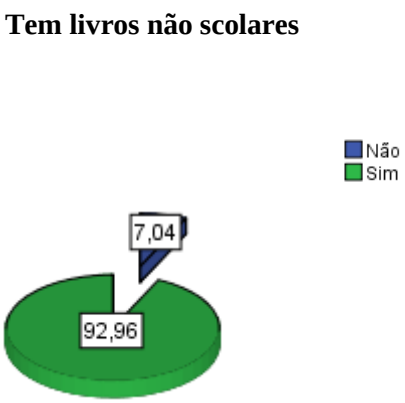


Gráfico 5.14 Em casa tem livros não escolares

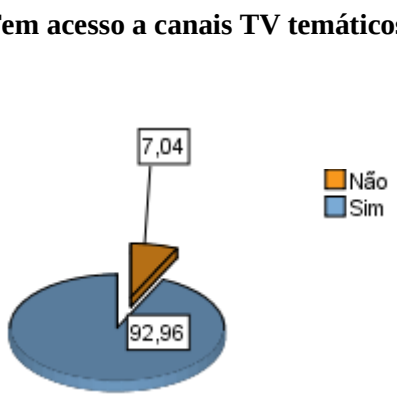


Gráfico 5.15 Em casa tem acesso a canais TV

No que diz respeito à variável "Hora de deitar", podemos verificar pelo Gráfico 5.16 que a hora de deitar entre as 22h e as 23h é a que mais se evidencia (43,66%), seguida da hora de deitar entre as 23h e as 24h com 28,17%. Indicam como hora de deitar depois das 24h, 21,13% dos alunos, o que é preocupante, uma vez que, em consequência deste horário tardio, os alunos não têm o descanso noturno necessário nestas idades (14-16) anos, e à atividade de estudo.

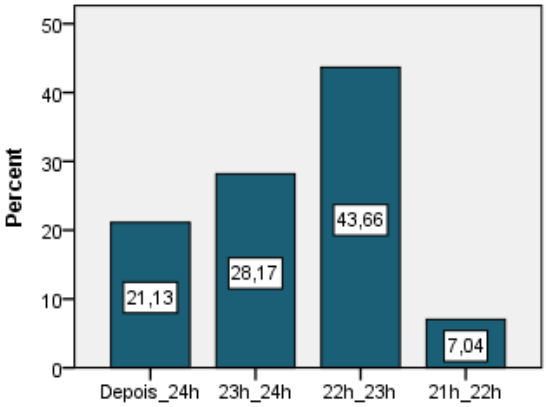


Gráfico 5. 17 Hora de deitar em tempo de aulas

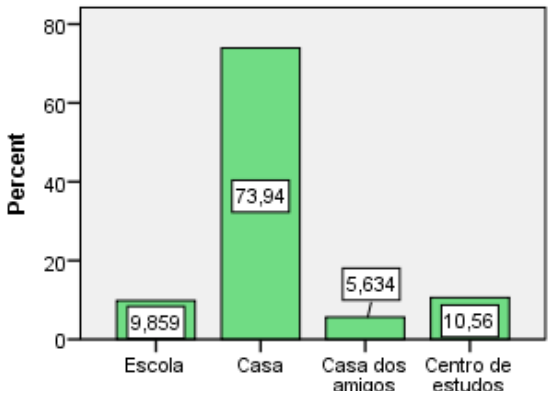


Gráfico 5.16 local habitual de estudo

No gráfico 5.17 verificamos que a casa é o local de estudo que mais se salienta (73,94%), seguido do centro de estudos com 10,56% e da escola (9,86%), por último 5,53% dos alunos indica estudar em casa dos amigos.

A maioria dos inquiridos (61,27%) não frequenta qualquer atividade extracurricular, por outro lado, 38,73% dos alunos frequenta uma atividade. Das atividades praticadas destacam-se as atividades enquadradas nos "Jogos Desportivos Coletivos" (21,83%), seguidas das "Atividades Rítmicas Expressivas" com uma pequena percentagem (3,52%), como se pode ler no Gráfico 5.18.

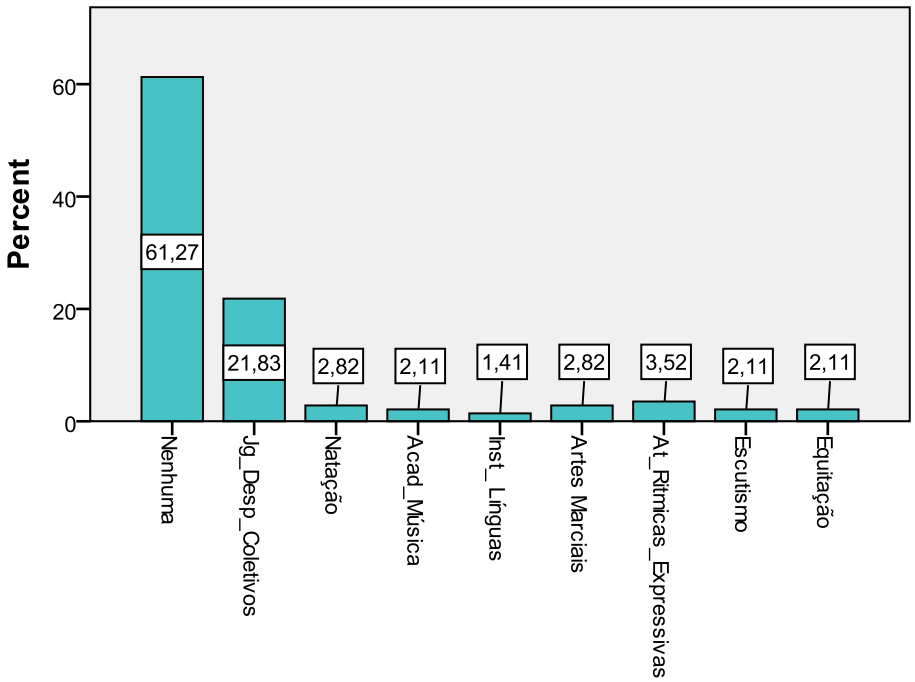


Gráfico 5.18 Atividade extracurricular praticada

De seguida, pedimos aos jovens para se situarem numa das três posições, “Sim”, “Não”, Às vezes”, em relação à assiduidade, pontualidade, participação, empenho, iniciativa e distração. Pela observação dos Gráficos 5.19 a 5.20 verificamos que a maioria dos inquiridos considera-se assídua, pontual, empenhada. Quanto à participação e distração a maioria dos jovens indica ser participativo e distraído às vezes.

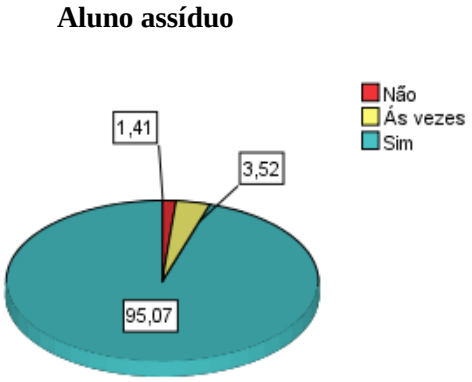


Gráfico 5.20 Consideras-te um aluno assíduo

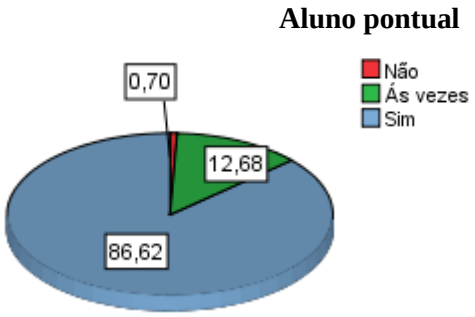


Gráfico 5.19 Consideras-te um aluno pontual

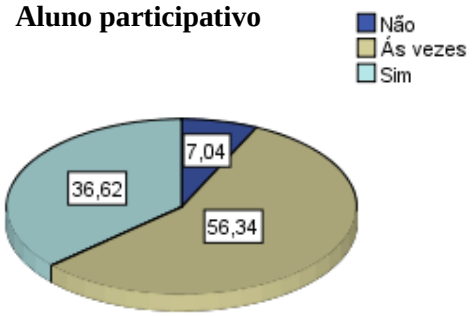


Gráfico 5. 21 Consideras-te um aluno articipativo

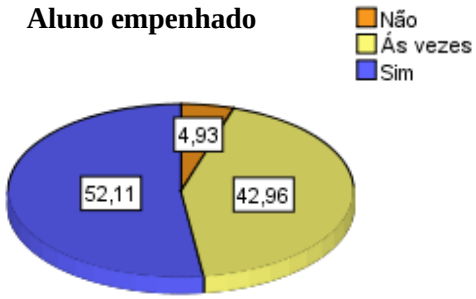


Gráfico 5.22 Consideras-te um aluno empenhado

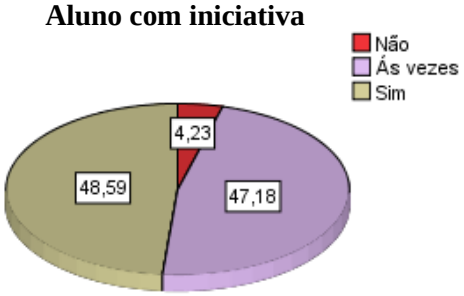


Gráfico 5. 23 Consideras-te um aluno com iniciativa

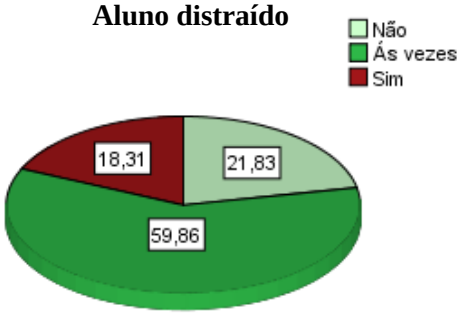


Gráfico 5.24 Consideras-te um aluno distraído

Relativamente às variáveis "Primeira disciplina preferida" e "Primeira disciplina não preferida", (Gráficos 5.25 e 5.26), salienta-se a Educação Física como disciplina preferida (28,87%), seguida de Ciências Naturais (16,90%). A disciplina não preferida mais observada é o Inglês (24,65%), logo seguida da Matemática (23,24%)

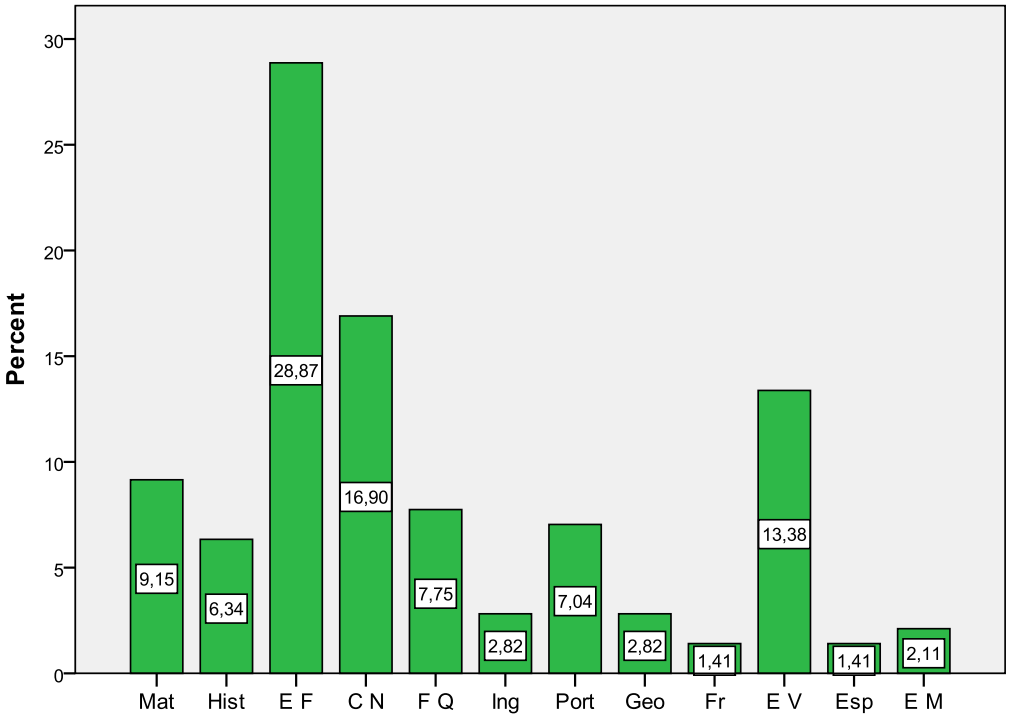


Gráfico 5.25 Primeira disciplina preferida

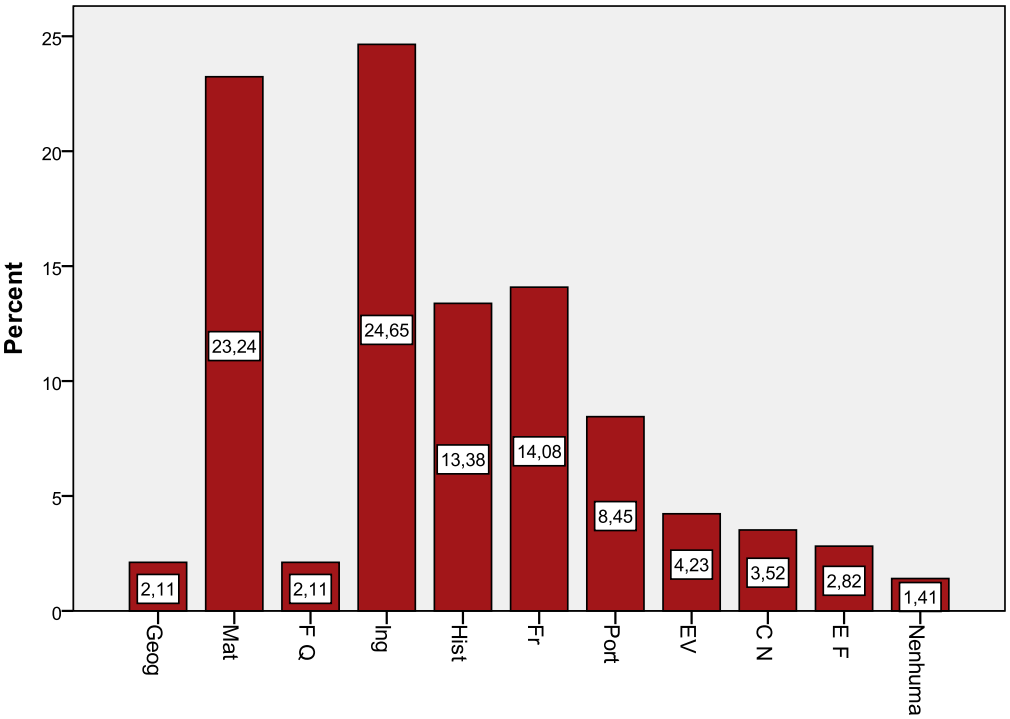


Gráfico 5. 26 Primeira disciplina não preferida

Para a variável "Dificuldades a Matemática" verificamos que, a maioria (57,04%) dos inquiridos indica "Às vezes". Por outro lado, observam-se percentagens quase idênticas para as respostas "Não" e "Sim bastantes", respetivamente 21,83% e 21,13% (Gráfico 5.26).

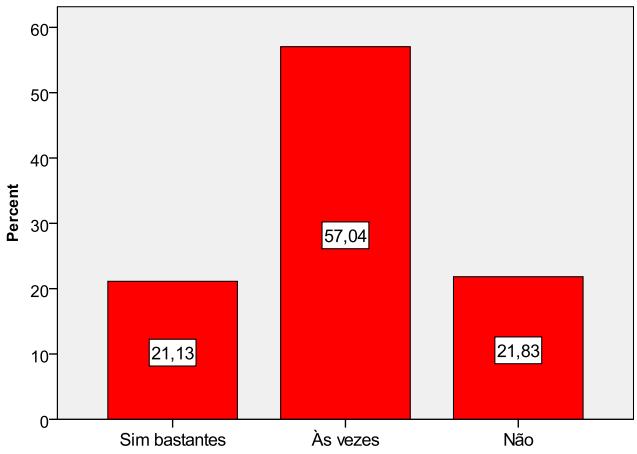


Gráfico 5.27 Dificuldades na disciplina de Matemática

No contexto do trabalho escolar, foram colocadas questões relativas aos hábitos de trabalho e à opinião dos alunos sobre a utilidade do Trabalho de casa (TPC):

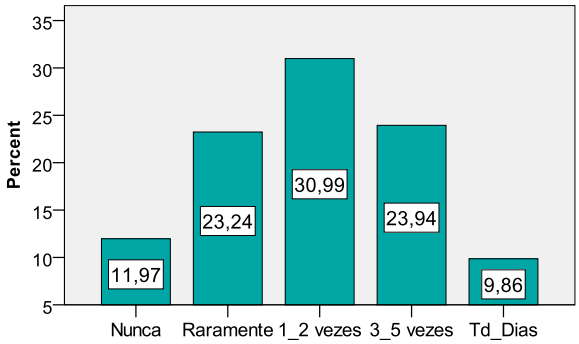


Gráfico 5.28 Frequência do estudo de Matemática na semana

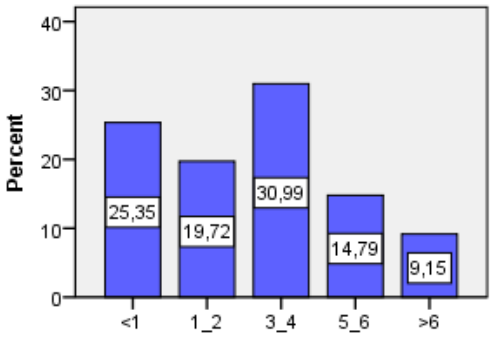


Gráfico 5. 29 Horas de estudo de Matemática na semana

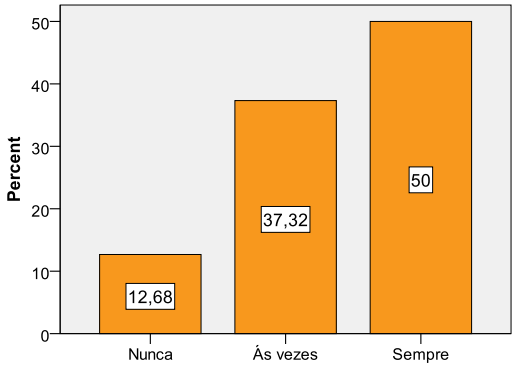


Gráfico 5. 30- Realiza os TPC de Matemática

Analisando os gráficos 5.28 a 5.31, constatamos que a maior percentagem dos inquiridos indica estudar (por semana), uma a duas vezes, no total três a quatro horas (30,99%). Note-se que 23,24% dos alunos raramente estuda e, 23,35% estuda em média, por semana, menos de uma hora. Observamos também que 50% dos jovens realiza sempre o TPC, no entanto, 37,32% realiza o TPC só às vezes. Estes valores demonstram que um número significativo dos alunos tem falta de hábitos de trabalho.

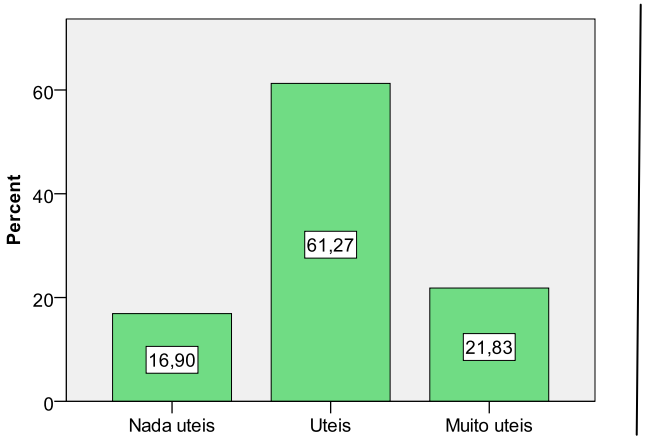


Gráfico 5.32 Utilidade do TPC

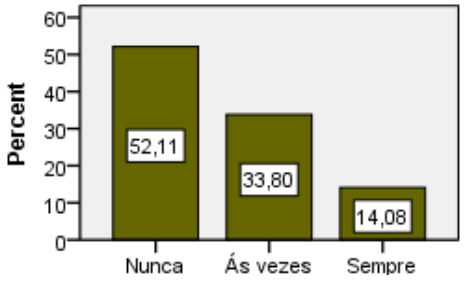


Gráfico 5.31 Ajuda nos TPC de Matemática

Apesar desta situação relativa aos nossos jovens ser preocupante, no gráfico 5.31 podemos visualizar que, 61,27% dos inquiridos considera úteis os TPCs.
No gráfico 5.32 verificamos que a maioria dos alunos indica não ter ajuda na realização do TPC, apenas 14,08% recebe sempre apoio na realização do TPC.

Ainda no contexto do TPC, foi pedido aos alunos que indicassem se, os TPCs servem para: "Praticar", "Ajudar a memorizar as matérias" e "Tomar consciência das dúvidas" (Gráficos 5.33- 5.35). Nas três situações observadas a percentagem do "Sim" prevalece, o que demonstra que a maioria dos jovens inquiridos tem opinião positiva relativamente ao contributo do TPC no seu processo de aprendizagem.

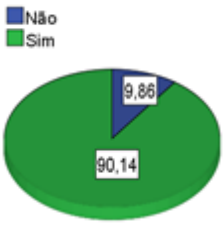


Gráfico 5.35 O TPC ajuda a praticar

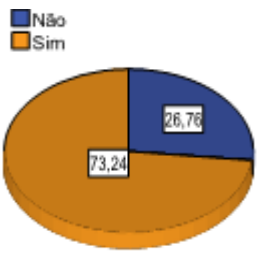


Gráfico 5. 34 O TPC ajuda a memorizar as matérias

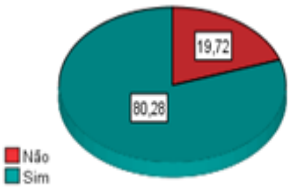


Gráfico 5. 33 O TPC ajuda a tomar consciência das minhas dúvidas

Outro aspeto analisado no questionário referia-se à opinião dos jovens sobre o apoio recebido na escola por parte dos professores. Para tal, foi-lhes pedido para indicarem se concordavam, ou em parte, ou não, com as seguintes questões: "Os professores ajudaram-me a compreender as matérias", "Os professores ouviram as minhas ideias e opiniões", "Os professores ouviram os meus problemas", Os professores ajudaram-me a ultrapassar as minhas dificuldades. Em todas as questões abordadas, a maioria dos inquiridos apresenta uma ideia positiva acerca do apoio recebido na escola por parte dos professores. No entanto, também em todas as questões, observamos uma percentagem razoável de alunos indica receber apoio apenas em parte (Gráficos 5.36- 5.39).

Professores ajudaram a compreender as matérias

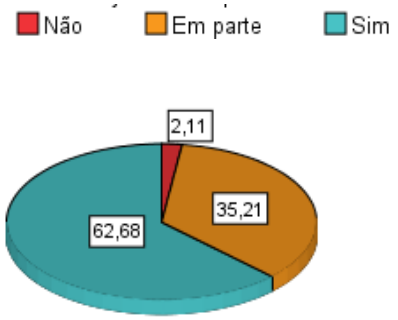


Gráfico 5.37 Os professores ajudaram-me a compreender as matérias

Professores ouviram as minhas opiniões

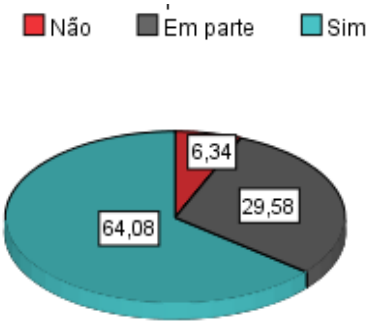


Gráfico 5. 36 Professores ouviram as minhas opiniões

Professores ouviram os meus problemas

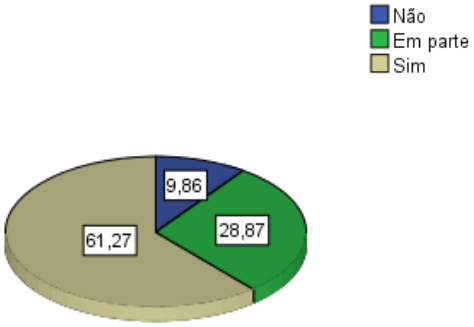


Gráfico 5.38 Professores ouviram os meus problemas

Professores ajudaram a ultrapassar as minhas dificuldades

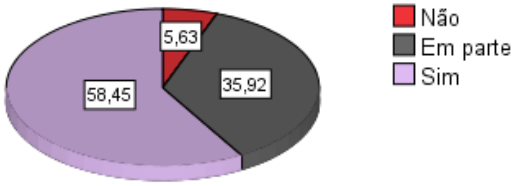


Gráfico 5.39 Professores ajudaram a ultrapassar as minhas dificuldades

Colocaram-se ainda duas questões referentes à ordem/disciplina dentro da sala de aula: "Havia elementos problemáticos?" e "Os professores conseguiram impor a ordem?" Uma percentagem elevada (66,20%) de alunos indica a existência de elementos problemáticos (Gráfico 5.40 e 5.41). Relativamente à questão "Os professores conseguiram impor a ordem?" a maioria (66,20%) dos alunos respondeu "Sim".

Existência de elementos problemáticos

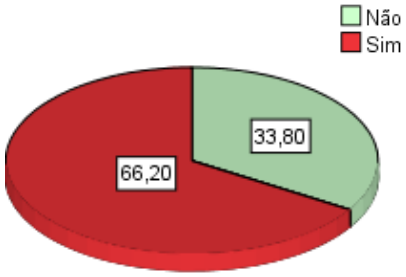


Gráfico 5.41 Havia elementos problemáticos?

Professores conseguiam impor ordem

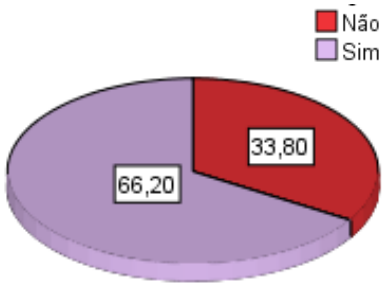


Gráfico 5.40 Professores conseguiam impor ordem?

Relativamente à questão "O que pensas estar a fazer daqui a 5 anos"(Gráfico 5.42), a maioria dos alunos perspetiva tirar um Curso superior" (57,04%), um número menor de alunos pensa em formação de nível médio ou curso não superior (16,20%) . Alguns alunos respondem "A trabalhar com o 12ºano" (9,15%) ou "A trabalhar sem 12ºano" (7,04%). Por último 10,56% dos jovens ainda não sabe ou não pensou.

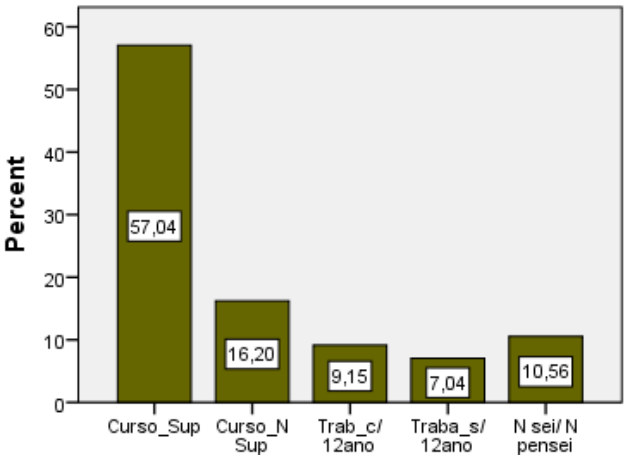


Gráfico 5.42-Perspetivas do aluno **(**O que pensas estar a fazer daqui a 5 anos?)

Verificamos que 73.24% dos inquiridos perspetiva obter formação diferenciada não superior ou superior.

Na última questão pretende-se saber, na opinião dos alunos, qual a importância do trabalho desenvolvido na escola na sua formação. Atribui "Grande" importância 54,93% dos alunos (Gráfico 5.43), este valor está próximo da percentagem de alunos que

perspetiva frequentar um curso superior (Gráfico 5.42). "Alguma" importância é indicada por 30,28% dos inquiridos e, por último 14,79% considera que o trabalho desenvolvido na escola irá ter "Pequena" a importância na sua formação. Note-se que, no questionário, existia também a opção de resposta "Nenhuma", no entanto, esta resposta não foi observada (Gráfico 5.43).

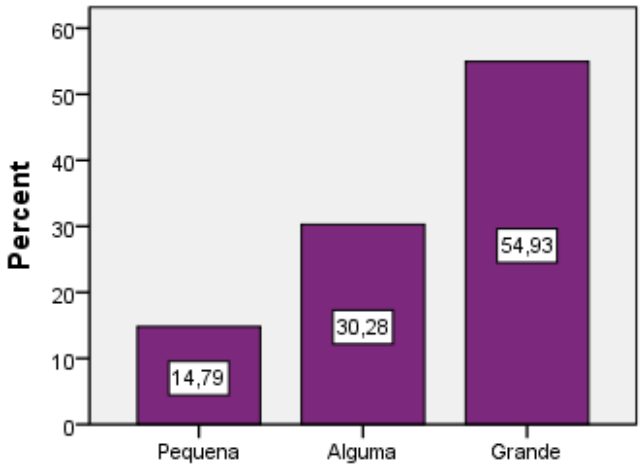


Gráfico 5.43 Importância do trabalho escolar na formação

5.5.2. Análise Bivariada

Após termos estudado as variáveis individualmente, avançamos para um estudo das eventuais relações existentes entre pares de variáveis.

Dada a complexidade do problema do insucesso escolar a Matemática, iremos analisar a possível associação entre a variável "Nível a Matemática" com uma seleção de variáveis:

"Perspetivas do aluno", "Categoria profissional da mãe", "Categoria profissional do pai", "Situação profissional da mãe", "Situação profissional do pai", "Atividade extracurricular", "Frequência do estudo de Matemática", "Horas de estudo de Matemática", "Realiza o TPC de Matemática" e "Importância da formação recebida".

A análise entre estas variáveis será apresentada em tabelas de contingência.

Para avaliar se a independência das variáveis é aceitável, ou para medir o afastamento à independência recorreu-se ao Qui-Quadrado de Pearson (χ^2). Quando as variáveis estão numa situação próxima da independência, o valor do Qui-Quadrado é próximo de zero e, quanto maior for afastamento à independência, maior será o seu valor (Silvestre, A. 2007).

A medida do Qui-Quadrado tem como principal requisito o fato de não mais de 20% das células da tabela de contingência terem frequências esperadas menores que 5 e de que nenhuma célula da tabela de contingência tenha frequências esperadas inferiores a 1. No caso das tabelas de 2x2 alguns investigadores consideram ainda que é necessário não existir nenhuma célula com frequência esperada inferior a 5 (Pestana, 2008). Caso estas condições não sejam cumpridas é possível agregar categorias adjacentes de modo a aumentar as frequências esperadas de cada célula (Cochran, 1954). No caso de não independência (Silvestre, A. 2007) propõe, para avaliar o grau de associação entre as variáveis, algumas medidas que se baseiam no valor do Qui-Quadrado, nomeadamente: Coeficiente de Contingência quadrática média ($\boldsymbol{\phi^2}$), Coeficiente de Contingência de Pearson[9] e o Coeficiente V de Cramer[10]. Estes coeficientes encontram-se definidos no ponto 4.2.3.3.

Para avaliar a intensidade de associação entre duas variáveis ordinais podemos determinar o Coeficiente de correlação de Spearman[11] (Reis, 2009), definido no ponto 4.2.3.4.

5.5.2.1. Nível Matemática/ Perspetivas do aluno

Começamos por averiguar se o Nível de Matemática está de algum modo associado às "Perspetivas do aluno". Observando a Tabela 5.3 parece existir uma associação entre estas duas variáveis. Verificamos, essencialmente, que os inquiridos que perspetivam adquirir formação superior os que obtêm os melhores resultados à disciplina de Matemática.

		Níve lMatemática				
		2	3	4	5	Total
Perspetivas do aluno	Curso Sup	11	35	19	16	81
	Curso N Sup	17	4	1	1	23
	Trab. c/ 12ano	8	4	1	0	13
	Trab. s/ 12ano	8	2	0	0	10
	N sei/ N pensei	10	4	0	1	15
Total		54	49	21	18	142

Tabela 5.3 Tabela de contingência: "Nível a Matemática" e " Perspetivas do aluno"

Dado se tratar da análise entre uma variável quantitativa (Nível a Matemática) e uma variável qualitativa (Perspetivas do aluno), para se verificar a existênciade associação entre estas duas variáveis,

[9] O seu valor máximo depende da dimensionalidade da tabela de contingência.
[10] Assume o valor 1 no caso de associação perfeita, qualquer que seja o número de linhas e de colunas da tabela.
[11] Quanto mais próximo o seu valor estiver dos extremos (-1 e 1), maior a correlação entre as ordenações das variáveis, quanto mais próximo de 0, menor essa correlação.

calculamos a medida do Qui-Quadrado de Pearson (Tabela 5.4).

Dado que mais de 20% das células da tabela de contingência têm frequências esperadas menores que 5, teremos de recodificar as variáveis em estudo, de modo a aumentar as frequências esperadas em cada célula.

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	52,689[a]	12	,000
N of Valid Cases	142		

a. 12 cells (60,0%) have expected count less than 5. The minimum expected count is 1,27.

Tabela 5.4 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis "Nível a Matemática" e "Perspetivas"

Como já referimos, da análise da Tabela 5.3, ressalta desta uma possível associação entre "Nível a Matemática" que o aluno obtém e o fato de pensar tirar, no futuro, um "Curso Superior.

Procedemos, por isso, à recodificação das variáveis "Nível a Matemática" e "Perspetivas do aluno", da seguinte forma:

A variável "Nível a Matemática" é recodificada com o nome "NotaMat1" (variável ordinal) de acordo com as seguintes condições:

1. N<3; 2. N_3 3. N 4_ 5

A variável "Perspetivas do aluno" é recodificada com o nome "PerpetivasFuturas" (variável nominal) de acordo com as seguintes condições:

1. C_sup; 2. Outra_Ns/Np

Obtemos uma nova tabela de contingência (Tabela 5.5), constituída pelas variáveis recodificadas "NotaMat1" e "PerpetivasFuturas".

		NotaMat1			Total
		N<3	N_3	N_4_5	
PerpetivasFuturas	C_sup	11	35	35	81
	Outra_Ns/ Np	43	14	4	61
Total		54	49	39	142

Tabela 5.5 Tabela de contingência: **"**NotaMat1" (recodificada) e "PerpetivasFuturas" (recodificada)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	50,795[a]	2	,000
N of Valid Cases	142		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 16,75.

Tabela 5.6 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis **"**NotaMat1" e "PerpetivasFuturas"

Para se poder verificar a existência de associação entre estas duas variáveis calculamos novamente o Qui-Quadrado de Pearson. Atendendo ao valor obtido para esta medida, χ^2 = 50,795 (Tabela 5.6), fica comprovado que, de fato, não existe independência entre o Nível que o aluno obtém à disciplina de Matemática e o fato de o aluno pensar, ou não, tirar um Curso Superior.

No entanto, para ser possível quantificar o grau de associação entre estas duas variáveis temos de calcular as seguintes medidas de associação:

Como já foi referido o coeficiente V de Cramer assume Valores no intervalo 0 a 1, sendo a associação entre as variáveis tanto mais elevada quanto mais próximo de 1 estiver o coeficiente. Logo, dado que o valor de V é 0,508 (Tabela 5.7) podemos afirmar que existe associação entre as perspetivas do aluno e o nível que o mesmo obtém a Matemática.

		Value	Approx. Sig.
Nominal by Nominal	Phi	,598	,000
	Cramer's V	,598	,000
	Contingency Coefficient	,513	,000
N of Valid Cases		142	

Tabela 5.7 Medidas de associação

5.5.2.2. Nível Matemática/ Categoria profissional da mãe

Passamos a averiguar se a variável "Nível a Matemática" está, de alguma forma, associada à variável "Categoria profissional da mãe". Observando a Tabela 5.8, verificamos uma possível associação entre estas variáveis.

		Nível Matemática				
		2	3	4	5	Total
Cat_Prof_Mãe	Dir/ Qd Sup	1	1	2	3	7
	Espec/ Tec	3	14	12	10	39
	Adm/ Serv/ Vend	11	19	3	4	37
	Oper/ Artif	6	8	2	0	16
	Trab N Qualif	32	5	2	1	40
	Desconhecida	1	2	0	0	3
Total		54	49	21	18	142

Tabela 5.8 Tabela de contingência: "Nível a Matemática" e "Categoria profissional da mãe"

Observa-se que os inquiridos provenientes de um contexto socioeconómico de "Especialistas/Técnicos" são os que conseguem os melhores resultados à disciplina de Matemática. Por outro lado, que os inquiridos provenientes de um contexto socioeconómico de "Trabalhador não qualificado" são os que conseguem os piores resultados à disciplina de Matemática.

Para verificarmos a existência de associação entre estas duas variáveis teríamos de aplicar o Qui-Quadrado de Pearson. No entanto, dado que existem células da tabela de contingência com frequências esperadas inferiores a 1, teremos de recodificar as variáveis em estudo.

A variável "Categoria profissional da mãe" é recodificada com o nome "CatProfMãe1" (variável nominal) de acordo com as seguintes condições:

1. Especialista/Técnicos; 2.Outras categorias

A variável "Nível a Matemática" é substituída pela variável "NotaMat1".
Obtemos uma nova tabela de contingência (Tabela 5.9), constituída pelas variáveis recodificadas "CatProfMãe1" e "NotaMat1".

		NotaMat1			Total
		N<3	N_3	N_4_5	
CatProfMãe1	Espec/Técnicos	3	14	22	39
	Outras categorias	51	35	17	103
Total		54	49	39	142

Tabela 5.9 Tabela de contingência: "Categoria profissional da mãe" (recodificada1) e "Nível a Matemática" (recodificada2)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	29,444[a]	2	,000
N of Valid Cases	142		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 10,71.

Tabela 5.10 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis e **"**NotaMat1" e "CatProfMãe1

A existência de associação entre estas duas variáveis pode ser verificada calculando novamente o Qui-Quadrado de Pearson. Considerando o valor obtido para esta medida, $\chi^2 = 29{,}444$ (Tabela 5.10), fica comprovado que, não existe independência entre a nota que o inquirido atinge à disciplina de Matemática e o fato de ter, ou não, uma mãe pertencente à categoria "Especialista/Técnicos".

Para ser possível quantificar o grau de associação entre estas duas variáveis "CatProfMãe1" e "NotaMat1" temos de calcular as seguintes medidas de associação:

		Value	Approx. Sig.
Nominal by Nominal	Phi	,455	,000
	Cramer's V	,455	,000
	Contingency Coefficient	,414	,000
N of Valid Cases		142	

Tabela 5.11 Medidas de associação

Dado o valor de $V = 0{,}455$ (Tabela 5.11) podemos afirmar que o fato de a mãe pertencer à categoria "Especialista/Técnicos" condiciona o nível que o aluno obtém à disciplina de Matemática.

Ao codificarmos a variável "Categoria profissional da mãe" da forma "CatProfMãe1", acabamos por perder uma informação mais específica, de uma possível associação entre o nível a Matemática e o fato de a mãe pertencer à categoria profissional "Trabalhador não qualificado".

Para ser possível proceder à análise de uma associação entre o nível obtido a Matemática quando a mãe pertence, ou não, à categoria "Trabalhador não qualificado", teremos de recodificar a variável "Categoria profissional da mãe" com o nome "CatProfMãe2", da seguinte maneira:

1. Trab Não Qualif (Trabalhador não qualificado)
2. Trab Qualif/Desc (Trabalhador qualificado/Desconhecido)

A variável "Nível a Matemática" é substituída pela variável "NotaMat1".

		NotaMat1			Total
		N<3	N_3	N_4_5	
CatProfMãe2	Trab Não Qualif	32	5	3	40
	Trab Qualif/Desc	22	44	36	102
Total		54	49	39	142

Tabela 5.12 Tabela de contingência: "Categoria profissional da mãe" (recodificada2) e "Nível a Matemática" (recodificada1)

Na nova tabela de contingência (Tabela 5.12) constituída pelas variáveis qualitativas "CatProfMãe2" (nominal) e "NotaMat1 (ordinal), podemos observar que os jovens inquiridos com mãe pertencente à categoria "Trabalhador não qualificado" são, de fato os que obtém os piores níveis.

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	41,694[a]	2	,000
N of Valid Cases	142		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 10,99.

Tabela 5.13 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis e "NotaMat1" e "CatProfMãe2

Para analisar a associação entre estas duas variáveis recorremos, uma vez mais ao Qui-Quadrado de Pearson. O valor do χ^2 = 41,694 (Tabela 5.13) comprova que também existe associação entre o nível à disciplina de Matemática e aluno ter, ou não, a mãe pertencente à categoria profissional "Trabalhador não qualificado".

A determinação da medida da intensidade da associação entre estas duas variáveis, requere, uma vez mais, o calculo dos valores das seguintes medidas de associação:

Dado o valor de V = 0,542 (Tabela 5.14) podemos afirmar que o fato de a mãe pertencer à categoria "Trabalhador não qualificado" condiciona de forma mais relevante o nível que o jovem inquirido obtém a Matemática.

		Value	Approx. Sig.
Nominal by Nominal	Phi	,542	,000
	Cramer's V	,542	,000
	Contingency Coefficient	,476	,000
N of Valid Cases		142	

Tabela 5.14 Medidas de associação

Tal como suponhamos, comprovamos que, na nossa amostra existe uma associação forte entre a categoria profissional da mãe e o nível a Matemática que o inquirido obtém.

5.5.2.3. Nível Matemática/ Categoria profissional do pai

Neste ponto, iremos analisar a associação entre a variável "Nível a Matemática" e "Categoria profissional do pai".

Observando a Tabela 5.15, tal como no ponto 5.5.2.2, verificamos que, os inquiridos provenientes de um contexto socioeconómico de "Especialistas/Técnicos" são os que conseguem os melhores resultados à disciplina de Matemática. Por outro lado, inquiridos provenientes de um contexto socioeconómico de "Trabalhador não qualificado" são os que conseguem os piores resultados à disciplina de Matemática.

		Nível Matemática				
		2	3	4	5	Total
Cat_Prof_Pai	Dir/ Qd Sup	1	2	3	3	9
	Espec/ tec	3	13	9	9	34
	Adm/ Serv/vend	7	13	3	5	28
	Oper/ Artif	12	14	3	1	30
	Trab N Qualif	29	5	3	0	37
	Desconhecido	2	2	0	0	4
Total		54	49	21	18	142

Tabela 5.15 Tabela de contingência "Nível a Matemática" e "Categoria Profissional"

Dado que existem células da tabela de contingência com frequências esperadas inferiores a 1, teremos de recodificar as variáveis em estudo, de modo a aumentar as frequências esperadas em cada célula.

A variável "Categoria profissional do pai" é recodificada com o nome "CatProfPai1" (variável nominal) de acordo com as seguintes condições:

1. Especialista/Técnicos
2. Outras categorias

A variável "Nível a Matemática" é substituída pela variável "NotaMat1".

Obtemos uma nova tabela de contingência (Tabela 5.16), constituída pelas variáveis recodificadas "CatProfPai1" e "NotaMat1".

		NotaMat1			
		N<3	N_3	N_4_5	Total
CatProfPai1	Espec/Técnicos	3	13	18	34
	Outras categorias	51	36	21	108
Total		54	49	39	142

Tabela 5.16 Tabela de contingência: "Categoria profissional do Pai (recodificada1) e "Nível a Matemática" (recodificada 1)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	20,771[a]	2	,000
N of Valid Cases	142		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,34.

Tabela 5.17 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis e "NotaMat1" e "CatProfPai1

Para se puder verificar a existência de associação entre estas duas variáveis calculamos o Qui-Quadrado de Pearson. Atendendo ao valor obtido para esta medida, χ^2 = 20,771 (Tabela 5.17), fica comprovado que, de fato, não existe independência entre o Nível que o aluno obtém à disciplina de Matemática e o fato de o pai do mesmo pertencer, ou não, à categoria "Especialista/Técnicos".

Para aferirmos o grau de associação entre estas duas variáveis determinamos as medidas de associação:

Dado o valor de V = 0,382 (Tabela 5.18) podemos afirmar que o fato de o pai pertencer à categoria "Especialista/Técnicos" condiciona o nível que o jovem inquirido obtém a Matemática.

		Value	Approx. Sig.
Nominal by Nominal	Phi	,382	,000
	Cramer's V	,382	,000
	Contingency Coefficient	,357	,000
N of Valid Cases		142	

Tabela 5.18 Medidas de associação

Assim, como foi referido no ponto 5.5.2.2, ao codificarmos a variável "Categoria profissional do pai" da forma "CatProfPai1", acabamos por perder informação. Por esta razão, vamos averiguar a associação entre o nível a Matemática e o fato de pai pertencer à categoria profissional "Trabalhador não qualificado".

Teremos de recodificar a variável "Categoria profissional do pai" com o nome "CatProfPai2", da seguinte maneira:

1. Trab Não Qualif (Trabalhador não qualificado)
2. Trab Qualif/Desc (Trabalhador qualificado/Desconhecido)

A variável "Nível a Matemática" é substituída pela variável "NotaMat1".

A nova tabela de contingência (Tabela 5.19) constituída pelas variáveis "CatProfPai2" e "NotaMat1, mostra que os jovens inquiridos com pai pertencente categoria Trabalhador não qualificado" são os que maioritariamente, obtém os piores níveis à disciplina de Matemática.

		NotaMat1			Total
		N<3	N_3	N_4_5	
CatProfPai2	Trab Não Qualif	29	5	3	37
	Trab Qualif/Desc	25	44	36	105
Total		54	49	39	142

Tabela 5.19 Tabela de contingência: "Categoria profissional do Pai" (recodificada2) e "Nível a Matemática" (recodificada1)

Para analisar a associação entre estas duas variáveis recorremos, uma vez mais ao Qui-Quadrado de Pearson.

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	34,641[a]	2	,000
N of Valid Cases	142		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 10,16.

Tabela 5.20 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis e "NotaMat1" e "CatProfPai2

O valor do χ^2 = 34,641 (Tabela 5.20) comprova que também existe associação entre o nível à disciplina de Matemática e aluno ter, ou não, o pai pertencente à categoria profissional "Trabalhador não qualificado".

Dado que o valor de *V* é 0,494 (Tabela 5.21) podemos afirmar que o facto de o pai pertencer, ou não, à categoria "Trabalhador não qualificado" condiciona a nota que o aluno obtém a Matemática.

		Value	Approx. Sig.
Nominal by	Phi	,494	,000
Nominal	Cramer's V	,494	,000
	Contingency Coefficient	,443	,000
N of Valid Cases		142	

Tabela 5.21 Medidas de associação

5.5.2.3. Nível Matemática/ Situação profissional da mãe

		Nível_Matemática				Total
		2	3	4	5	
Situ_Prof_Mãe	Desemp	18	6	2	0	26
	Empreg	30	38	16	14	98
	Em form	1	1	1	0	3
	Outro/ Desc	5	4	2	4	15
Total		54	49	21	18	142

Tabela 5.22 Tabela de contingência: "Nível a Matemática" e "Situação profissional da mãe"

De seguida, pretendemos averiguar se a variável "Nível a Matemática" está de alguma forma associada à variável "Situação profissional da mãe". Observando a Tabela 5.22, que resulta do cruzamento destas variáveis, verificamos a existência de células com frequências esperadas

inferiores a 1, teremos de recodificar as variáveis em estudo.

A variável "Situação profissional da mãe" é recodificada com o nome "SituaProfMãe1" (variável nominal) de acordo com as seguintes condições:

1. Desempregada
2. Outra Situ Prof

A variável "Nível a Matemática" é substituída pela variável "NotaMat1".

		NotaMat1			
		N<3	N_3	N_4_5	Total
SituaProfMãe1	Desempregada	18	7	3	28
	Outra_Situ_Prof	36	42	36	114
Total		54	49	39	142

Tabela 5.23 Tabela de contingência: "Situação profissional da mãe" (recodificada1) e "Nível a Matemática" (recodificada1)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10,800[a]	2	,005
N of Valid Cases	142		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 7,69.

Tabela 5.24 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis "NotaMat1"e SituaProfMãe1"

O valor do Qui-Quadrado de Pearson, $\chi^2 = 10{,}800$ (Tabela 5.24) comprova que não existe independência entre a nota que o aluno obtém à disciplina de Matemática e a Situação profissional da mãe.

A observação da Tabela 5.25 permite aferir o grau de associação das duas variáveis. Verificamos o valor de 0,276 para o coeficiente *V* de Cramer, podemos afirmar que o facto de a mãe se encontrar desempregada condiciona a nota que o aluno obtém a Matemática.

		Value	Approx. Sig.
Nominal by Nominal	Phi	,276	,005
	Cramer's V	,276	,005
	Contingency Coeffici	,266	,005
N of Valid Cases		142	

Tabela 5.25 Medidas de associação

5.5.2.4. Nível Matemática/ Situação profissional do pai

De seguida, pretendemos averiguar se a variável "Nível a Matemática" está de alguma forma associada à variável "Situação profissional do pai". Observando a Tabela 5.26, que resulta do cruzamento destas variáveis, verificamos a existência de células com frequências esperadas inferiores a 1, teremos de recodificar as variáveis em estudo.

		Nível_Matemática				
		2	3	4	5	Total
Sit_Prof_Pai	Reform	3	2	0	1	6
	Desemp	16	5	2	0	23
	Empreg	26	41	16	14	97
	Em form	4	0	1	1	6
	Outro/ Desc	5	1	2	2	10
Total		54	49	21	18	142

Tabela 5.26 Tabela de contingência: "Nível a Matemática" e "Situação profissional do pai".

A variável "Situação profissional do pai" é recodificada com o nome "Situ_Prof_Pai1" (variável nominal) de acordo com as seguintes condições:

1. Desempregado
2. Outra Situ Prof

A variável "Nível a Matemática" é substituída pela variável "NotaMat1".

		NotaMat1			
		N<3	N_3	N_4_5	Total
SituaProfPai1	Desempregado	16	5	2	23
	Outra_Sit_Prof	38	44	37	119
Total		54	49	39	142

Tabela 5.27 Tabela de contingência: "Situação profissional do pai" (recodificada1) e "Nível a Matemática" (recodificada1)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	11,995[a]	2	,002
N of Valid Cases	142		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,32.

Tabela 5.28 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis "NotaMat1"e SituaProfPai1"

O valor do Qui-Quadrado de Pearson, $\chi^2 = 11{,}995$ (Tabela 5.28) comprova que não existe independência entre a nota que o aluno obtém à disciplina de Matemática e a Situação profissional do pai.

Dado que o valor da medida de associação, V de Cramer, é 0,291(Tabela 5.29) podemos afirmar que o facto de o pai se encontrar desempregado condiciona a nota que o aluno obtém a Matemática.

		Value	Approx. Sig.
Nominal by Nominal	Phi	,291	,002
	Cramer's V	,291	,002
	Contingency Coefficient	,279	,002
N of Valid Cases		142	

Tabela 5.29 Medidas de associação

5.5.2.5. Nível Matemática/ Atividade extracurricular

De seguida, pretendemos averiguar se a variável "Nível a Matemática" está de alguma forma associada à variável "Atividade extracurricular".

		Nível_Matemática				Total
		2	3	4	5	
Atividade extracurricurricular	Não	38	28	14	7	87
	Sim	16	21	7	11	55
Total		54	49	21	18	142

Tabela 5.30 Tabela de contingência: "Nível a Matemática" e "Atividade extracurricular".

Para se puder verificar a existência de associação entre estas duas variáveis calculamos o Qui-Quadrado de Pearson.

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6,294[a]	3	,098
N of Valid Cases	142		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,97.

Tabela 5.31 Qui-Quadrado relativo ao cruzamento das variáveis "Nível_Matemática" e "Atividade extracurricular".

Sendo *p-value* = 0,098 > α = 0.05 e Qui-Quadrado de Pearson, $\chi^2 = 6{,}294$ (Tabela 5.31) não rejeitamos a hipótese de o nível que o aluno obtém à disciplina de Matemática ser independente do mesmo praticar, ou não, uma Atividade extracurricular.

5.5.2.6. Nível Matemática/Frequência do estudo de Matemática (na semana)

Vamos agora, averiguar se a variável "Nível a Matemática" está de alguma forma associada à variável "Frequência do estudo de Matemática".

		Nível_Matemática				
		2	3	4	5	Total
Frequência_Estudo_Mat	Nunca	9	6	0	2	17
	Raramente	23	5	3	2	33
	1_2 vezes	15	18	7	4	44
	3_5 vezes	5	14	7	8	34
	Td_Dias	2	6	4	2	14
Total		54	49	21	18	142

Tabela 5.32 Tabela de contingência: "Nível a Matemática" e "Frequência do estudo de Matemática"

Observando a Tabela 5.32, que resulta do cruzamento das variáveis ordinais "Nível a Matemática" e "Frequência do estudo de Matemática", verificamos a existência de células com frequências esperadas inferiores a 1. Assim, é necessário recodificar as variáveis em estudo.

A variável "Nível a Matemática" é substituída pela variável "NotaMat1".

		NotaMat1			
		N<3	N_3	N_4_5	Total
Frequência_Estudo_Mat	Nunca	9	6	2	17
	Raramente	23	5	5	33
	1_2 vezes	15	18	11	44
	3_5 vezes	5	14	15	34
	Td_Dias	2	6	6	14
Total		54	49	39	142

Tabela 5.33 Tabela de contingência: "Nível a Matemática" e "Frequência do estudo de Matemática"

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	29,730[a]	8	,000
N of Valid Cases	142		

a. 3 cells (20,0%) have expected count less than 5. The minimum expected count is 3,85.

Tabela 5.34 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis "NotaMat1" e "Frequência do estudo de Matemática"

Para se puder verificar a existência de associação entre estas duas variáveis calculamos o Qui-Quadrado de Pearson. Atendendo ao valor obtido para esta medida, $\chi^2 = 29{,}730$ (Tabela 5.34), fica comprovado que, de fato, não existe independência entre o Nível que

o aluno obtém à disciplina de Matemática e a Frequência do estudo de Matemática (na semana).

Para aferirmos o grau de associação entre estas duas variáveis determinamos as medidas de associação:

		Value	Approx. Sig.
Nominal by Nominal	Phi	,458	,000
	Cramer's V	,324	,000
	Contingency Coefficient	,416	,000
Ordinal by Ordinal	Spearman Correlation	,394	,000
N of Valid Cases		142	

Tabela 5.35 Medidas de associação

Como o coeficiente de correlação de Spearman é 0,394 (Tabela 5.35), podemos afrmar que existe associação (positiva) entre as ordenações das duas variáveis.

5.5.2.7. Nível Matemática/Horas de estudo de Matemática (na semana)

De seguida, pretendemos averiguar se a variável "Nível a Matemática" está de alguma forma associada à variável "Horas de estudo de Matemática".

		Nível_Matemática				Total
		2	3	4	5	
Horas_Estudo Mat	<1	24	9	1	2	36
	1_2	15	7	4	2	28
	3_4	8	22	6	8	44
	5_6	5	6	6	4	21
	>6	2	5	4	2	13
Total		54	49	21	18	142

Tabela 5.36 Tabela de contingência: "Nível a Matemática" e "Horas de estudo de Matemática"

Dado que mais de 20% das células da tabela de contingência têm frequências esperadas menores que 5, teremos de recodificar as variáveis em estudo, de modo a aumentar as frequências esperadas em cada célula.

A variável "Nível a Matemática" é substituída pela variável "NotaMat1".

		NotaMat1			Total
		N<3	N_3	N_4_5	
Horas Estudo Mat	<1	24	9	3	36
	1_2	15	7	6	28
	3_4	8	22	14	44
	5_6	5	6	10	21
	>6	2	5	6	13
Total		54	49	39	142

Tabela 5.37 Tabela de contingência: "Horas de estudo de Matemática" e "Nível a Matemática" (recodificada1).

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	32,212[a]	8	,000
N of Valid Cases	142		

a. 3 cells (20,0%) have expected count less than 5. The minimum expected count is 3,57.

Tabela 5.38 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis "NotaMat1" e "Horas de estudo de Matemática".

O valor de Qui-Quadrado de Pearson, $\chi^2 = 32{,}212$ (Tabela 5.38) comprova que não existe independência entre a nota que o aluno obtém à disciplina de Matemática e o número de horas de estudo (na semana).

Dado que o valor do coeficiente de correlação de Spearman é 0,419 (Tabela 5.39), podemos afrmar que existe associação (positiva) entre as ordenações das duas variáveis. O número de horas de estudo condiciona a nota que o aluno obtém a Matemática.

		Value	Approx. Sig.
Nominal by Nominal	Phi	,476	,000
	Cramer's V	,337	,000
	Contingency Coefficient	,430	,000
Ordinal by Ordinal	Spearman Correlation	,419	,000
N of Valid Cases		142	

Tabela 5.39 Medidas de associação

5.5.2.8. Nível Matemática/Realiza o TPC de Matemática

Neste ponto, pretendemos averiguar se a variável "Nível a Matemática" está de alguma forma associada à variável "Realiza o TPC de Matemática".

		Nível_Matemática				
		2	3	4	5	Total
RealizaTPC Mat	Nunca	10	6	1	1	18
	Às vezes	38	9	2	4	53
	Sempre	6	34	18	13	71
Total		54	49	21	18	142

Tabela 5.40 Tabela de contingência: "Nível Matemática" e "Realiza o TPC de Matemática"

Observando a Tabela 5.40, verificamos que a maioria dos alunos que obtém nível igual ou superior a 3 realiza sempre o TPC.

Dado que mais de 20% das células da tabela de contingência têm frequências esperadas menores que 5, teremos de recodificar as variáveis em estudo, de modo a aumentar as frequências esperadas em cada célula.

A variável “Nível a Matemática” é substituída pela variável “NotaMat1”.

		NotaMat1			
		N<3	N 3	N 4 5	Total
Realiza TPC Mat	Nunca	10	6	2	18
	Às vezes	38	9	6	53
	Sempre	6	34	31	71
Total		54	49	39	142

Tabela 5.41 Tabela de contingência: e “Nível a Matemática” (recodificada1) e “Realiza o TPC de Matemática”.

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	55,560[a]	4	,000
N of Valid Cases	142		

a. 1 cells (11,1%) have expected count less than 5. The minimum expected count is 4,94.

Tabela 5.42 Resultados de Qui-Quadrado relativo ao Cruzamento das variáveis “NotaMat1” e “Realiza o TPC de Matemática”.

Atendendo ao valor do Qui-Quadrado $\chi^2 = 55{,}560$ (Tabela 5.42), fica comprovado que, de fato, não existe independência entre a nota que o inquirido obtém à disciplina de Matemática e o fato de ralizar, ou não, o TPC de Matemática.

Para aferirmos o grau de associação entre estas duas variáveis determinamos as medidas de associação.

Dado que o coeficiente de correlação de Spearman é 0,535 (Tabela 5.43), podemos afrmar que existe associação (positiva) entre as ordenações as duas variáveis. Concluimos que a realização, ou não, do TPC condiciona o nível que o aluno obtém à disciplina de Matemática.

		Value	Approx. Sig.
Nominal by Nominal	Phi	,626	,000
	Cramer's V	,442	,000
	Contingency Coefficient	,530	,000
Ordinal by Ordinal	Spearman Correlation	,535	,000
N of Valid Cases		142	

Tabela 5.43 Medidas de associação

5.5.2.9. Nível Matemática/Utilidade do TPC

De seguida, pretendemos averiguar se a variável "Nível a Matemática" está de alguma forma associada à variável "Utilidade do TPC".

		Nível_Matemática				
		2	3	4	5	Total
Utilidade TPC	Nada úteis	16	6	1	1	24
	Úteis	34	28	11	14	87
	Muito úteis	4	15	9	3	31
Total		54	49	21	18	142

Tabela 5.44 Tabela de contingência: "Nível amatemática" e "Utilidade do TPC"

Dado que mais de 20% das células da tabela de contingência têm frequências esperadas menores que 5, teremos de recodificar as variáveis em estudo, de modo a aumentar as frequências esperadas em cada célula.

A variável "Nível a Matemática" é substituída pela variável "NotaMat1".

		NotaMat1			
		N<3	N_3	N_4_5	Total
Utilidade TPC	Nada úteis	16	6	2	24
	Úteis	34	28	25	87
	Muito úteis	4	15	12	31
Total		54	49	39	142

Tabela 5.45 Tabela de contingência: "Nível amatemática"(recodificada1) e "Utilidade do TPC"

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	17,520[a]	4	,002
N of Valid Cases	142		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,59.

Tabela 5.46 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis "NotaMat1" e "Utilidade do TPC ".

Atendendo ao valor do Qui-Quadrado $\chi^2 =$ 17,520 Tabela 5.46), fica comprovado que, não existe independência entre a nota que o inquirido obtém à disciplina de Matemática e a utilidade que o aluno atribui à realização do TPC.

Dado que o coeficiente de correlação de Spearman é 0,326 (Tabela 5.47), podemos afrmar que existe associação (positiva) entre o nível que o inquirido obtém a Matemática e a utilidade que o aluno atribui à realização do TPC.

		Value	Approx. Sig.
Nominal by Nominal	Phi	,351	,002
	Cramer's V	,248	,002
	Contingency Coefficient	,331	,002
Ordinal by Ordinal	Spearman Correlation	,326	,000
N of Valid Cases		142	

Tabela 5.47 Medidas de associação

5.5.2.10. Nível Matemática/Importância da formação recebida

Por último, pretendemos averiguar se a variável "Nível a Matemática" está de alguma forma associada à variável "Importância da formação recebida".

		Nível_Matemática				
		2	3	4	5	Total
Importância da formação	Pequena	17	2	2	0	21
	Alguma	26	10	4	3	43
	Grande	11	37	15	15	78
Total		54	49	21	18	142

Tabela 5.48 Tabela de contingência: "Nível a Matemática" e "Importância da formação".

Dado que mais de 20% das células da tabela de contingência têm frequências esperadas menores que 5, teremos de recodificar as variáveis em estudo, de modo a aumentar as frequências esperadas em cada célula.

A variável "Nível a Matemática" é substituída pela variável "NotaMat1".

		NotaMat1			
		N<3	N_3	N_4_5	Total
Importância da formação	Pequena	17	2	2	21
	Alguma	26	10	7	43
	Grande	11	37	30	78
Total		54	49	39	142

Tabela 5.49 Tabela de contingência: e "Nível a Matemática" (recodificada1) e "Importância da formação".

Observando a Tabela 5.49, esta sugere uma associação entre o nível obtido à disciplina de Matemática e a importância que o aluno atribui à formação recebida na escola: os

alunos com melhores níveis a Matemática são os que obtêm menos casos de "Pequena" e "Alguma".

Para avaliar a independência entre estas duas variáveis recorremos, uma vez mais ao valor do Qui-Quadrado de Pearson.

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	44,617[a]	4	,000
N of Valid Cases	142		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 5,77.

Tabela 5.50 Resultados de Qui-Quadrado relativo ao cruzamento das variáveis "NotaMat1" e "Importância da formação"

Verificamos na Tabela 5.50, o valor de Qui-Quadrado $\chi^2 =$ 44,617, comprova-se que de fato existe associação entre o nível obtido à disciplina de Matemática e a importância que o aluno atribui à formação recebida na escola.

Para aferirmos o grau de associação entre estas duas variáveis determinamos as medidas de associação:

		Value	Approx. Sig.
Nominal by Nominal	Phi	,561	,000
	Cramer's V	,396	,000
	Contingency Coefficient	,489	,000
Ordinal by Ordinal	Spearman Correlation	,503	,000
N of Valid Cases		142	

Tabela 5.51 Medidas de associação.

Como o coeficiente de correlação de Spearman é 0,503 (Tabela 5.51), podemos afirmar que existe associação (positiva) entre as ordenações das duas variáveis. Concluimos que a importância que o aluno atribui à formação recebida na escola condiciona o nível que o aluno obtém à disciplina de Matemática.

5.5.3. Análise Multivariada

5.5.3.1. Análise de *Clusters* com variáveis originais

Iniciamos a análise multivariada com a aplicação da Análise de *Clusters*, utilizando a técnica *TwoStep*, com o objetivo de identificar agrupamentos naturais de indivíduos. Das variáveis iniciais[12], optamos por selecionar as variáveis que consideramos com

[12] As variáveis do espaço de análise deste estudo encontram-se descritas na tabela 1 (anexo III).

interesse no contexto do problema do insucesso escolar a Matemática. Atendemos, ainda, aos resultados da análise bivariada que comprovaram existir associação entre estas variáveis e o Nível a Matemática obtido pelo aluno. Assim, foram selecionadas as seguintes variáveis originais: "Nível a Matemática", "Realiza TPC de Matemática", "Horas de estudo na semana", "Frequência de estudo na semana", "Importância da formação recebida", "Consideras-te um aluno empenhado", "Categoria profissional da mãe", "Categoria profissional do pai" e "Utilidade do TPC".

Foram realizados três estudos, tendo sido utilizada a medida de distância Log-likelihood, pois estamos a trabalhar com variáveis qualitativas. No primeiro estudo não foi fixado o número de *clusters* a reter (apenas por defeito o número máximo, 15), no segundo solicitaram-se três *clusters* e, por último foram requeridos quatro *clusters.* No primeiro estudo foram realizados dois ensaios, um utilizando o critério *Akaike's Information Criterion* (AIC) e outro utilizando o critério *Bayes* Information Criterion (BIC), em ambos obtivemos uma solução com dois clusters.

As soluções de três e quatro clusters obtidas são de fraca qualidade como se pode verificar no Anexo IV, Figura 1 e 2. No primeiro estudo, obtivemos uma solução de dois clusters de qualidade razoável que passamos a expor.

Podemos observar, na Figura 4, na parte inferior da barra horizontal a "*silhouette measure of cohesion and separation",* trata-se de uma medida da qualidade do agrupamento. Essencialmente baseia-se nas distâncias médias entre os objetos e pode variar entre -1 e +1. Um valor de medida inferior a 0,20 indica uma má solução, entre 0,20 e 0,50 indica uma solução razoável e superior a 0,50 indica boa solução. No nosso caso, como já foi referido, indica uma solução razoável.

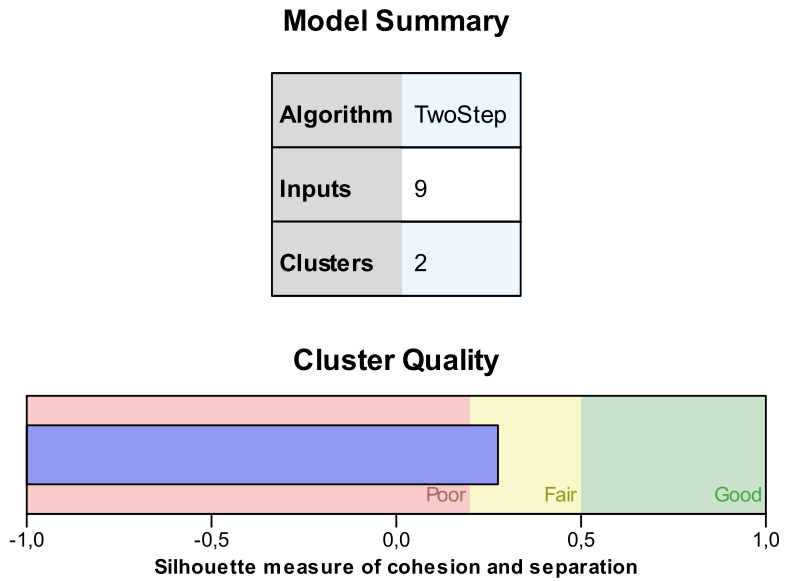


Figura 4 Qualidade do Agrupamento

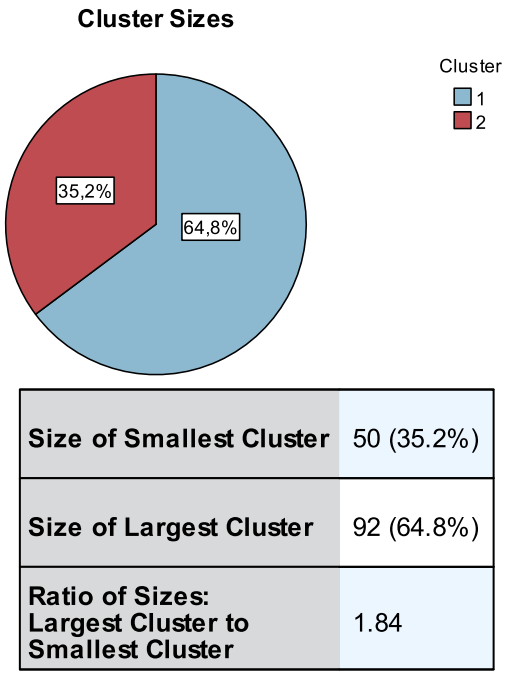


Figura 5 Distribuição por *Clusters*

A Figura 5 informa-nos sobre a dimensão dos *clusters.*

O *cluster* 1 é o maior, sendo constituido por noventa e dois alunos (64,8%), o *cluster* 2 é formado por cinquenta alunos (35,2%).

Na Figura 6 podemos ver uma descrição dos dois *clusters*/grupos. Salientamos nesta descrição o seguinte: no *cluster* 1, o grupo é composto por alunos que tendencialmente conseguem um desempenho satisfatório a Matemática (nível 3; 45,7%) e cujo pai e mãe pertencem à categoria profissional "Especialistas/Técnicos" (respetivamente 34,8% e 39,1%); por outro lado no *cluster* 2, o grupo apresenta um desempenho não

satisfatório (nível 2; 86,0%); o pai e a mãe dos alunos pertencem à categoria profissional "Trabalhadores não Qualificados" (respetivamente 60,0% e 54,0%); nos dois grupos os alunos consideram úteis os trabalhos de casa. No Anexo IV na Tabela 2 podemos consultar o cruzamento entre as variáveis de *input* da Análise de *Clusters* e os dois *clusters.*

Cluster	1	2
Inputs	RealizaTPC Mat Sempre (77.2%)	RealizaTPC Mat Às vezes (68.0%)
	Nível_Matemática 3 (45.7%)	Nível_Matemática 2 (86.0%)
	Frequência Estudo Mat 1_2 vezes (39.1%)	Frequência Estudo Mat Raramente (50.0%)
	Horas Estudo Mat 3_4 (41.3%)	Horas Estudo Mat <1 (60.0%)
	Categ. Profi. Pai Espec/ tec (34.8%)	Categ. Profi. Pai Trab N Qualif (60.0%)
	Utilidade TPC Uteis (67.4%)	Utilidade TPC Uteis (50.0%)
	Importância da formação Grande (72.8%)	Importância da formação Alguma (42.0%)
	Aluno empenhado Sim (70.7%)	Aluno empenhado Às vezes (74.0%)
	Categ. Profi. Mãe Espec/ Tec (39.1%)	Categ. Profi. Mãe Trab N Qualif (54.0%)

Figura 6 Descrição dos C*lusters*

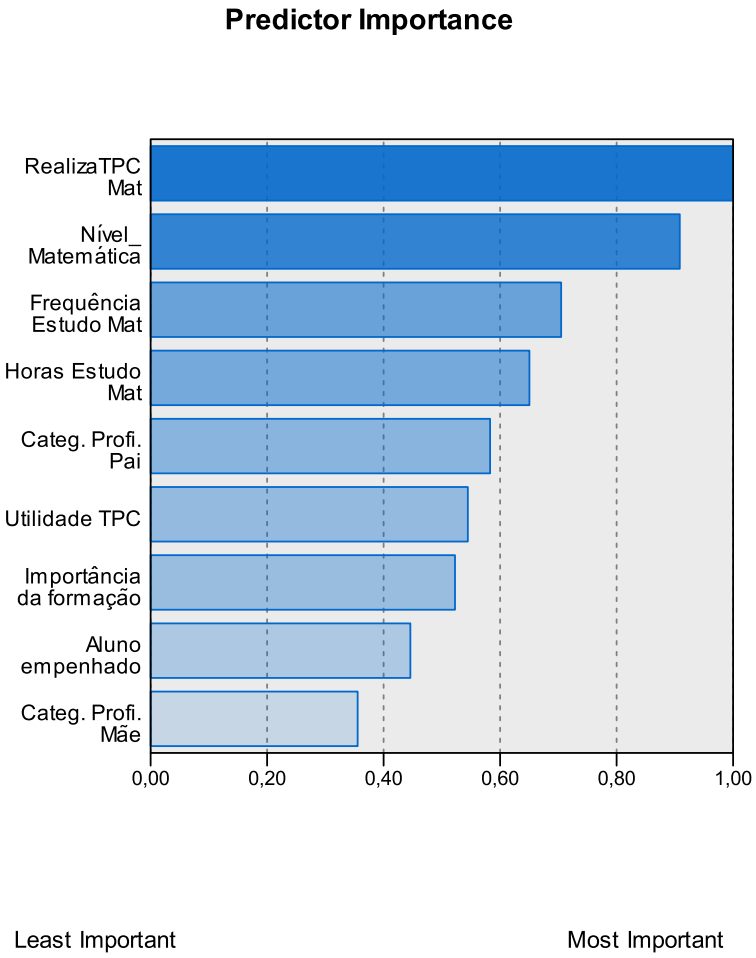


Gráfico 5. 44 Importância das variáveis

Do Gráfico 5.44 obtemos informação relativa à importância das variáveis para a solução do agrupamento. Podemos ver que a variável mais importante é a "Realiza o TPC de Matemática", seguida por "Nível a Matemática".

No contexto do desempenho dos alunos na disciplina de Matemática, pensamos que não faz muito sentido o agrupamento dos alunos em apenas dois *Clusters*/Grupos. Assim, para averiguar outra solução e fundamentar a mesma, realizou-se outro estudo aplicando previamente a Análise de Correspondências Múltiplas (ACM), com se passa a axplicar na secção seguinte.

5.5.3.2. Aplicação da ACM

A ACM foi aplicada com o objetivo de identificar possíveis perfis no espaço de partida. Para a aplicação desta técnica recorreu-se, uma vez mais, ao software SPSS versão 19, tendo-se seguido as seguintes fases:

- Seleção e interpretação das dimensões
- Interpretação do plano

5.5.3.2.1. Seleção das dimensões mais representativas

Foi usado um número elevado de dimensões a fim de ser possível analisar o comportamento dos valores próprios e da inércia nas múltiplas dimenões (Anexo V, Tabela 3). Solicitou-se uma solução com 29 dimensões que no caso corresponde ao número máximo (número total de categorias menos o número de variáveis da ACM) que é possível definir neste estudo.

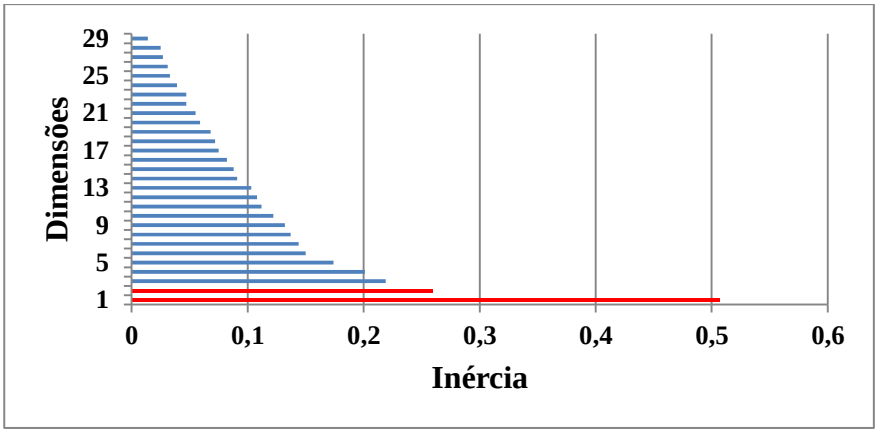


Gráfico 5.45 Representação da variância das dimensões[13]

[13] Gráfico de barras em Excel, construído a partir da informação da Tabela 3 do Anexo V.

A partir da análise da distribuição dos valores próprios e da inércia (Anexo V, Tabela 3) e da análise da variância das dimensões (Gráfico 5.45), verifica-se que:
- as duas primeiras dimensões são as mais representativas em termos de inércia,
- a partir da dimensão 3 o acréscimo de variância explicada tende a ser mínimo.

Assim, avançamos para uma solução de duas dimensões (privilegiando as duas primeiras).

5.5.3.2.2. Distribuição das variáveis nas duas primeiras dimensões

O valor do Alpha de Cronbach- permite avaliar a qualidade do ajustamento do modelo, dimensão por dimensão- mostra que ambas as dimensões têm um bom ajustamento: 0,879 para dimensão 1 e 0,644 para a dimensão 2 (Tabela 5.52).
As duas primeiras dimensões explicam juntas 23,8% da variância total. Este cálculo[14] pressupõe o número máximo de dimensões (Anexo V, Tabela 3).
Como se distribuem as nove variáveis nestas dimensões é o que se passa a analisar por via da leitura das medidas de discriminação (Tabela 5.53). Deve proceder-se à seleção das variáveis mais importantes para cada dimensão. Para tal pode tomar-se como referência o valor médio das medidas de discriminação (Tabela 5.52) em cada dimensão (Carvalho, 2008). As variáveis que mais discriminam numa dada dimensão são aquelas cuja medida de discriminação é pelo menos próximo do valor da inércia nessa dimensão.

Dimension	Cronbach's Alpha	Variance Accounted For		
		Total (Eigenvalue)	Inertia	% of Variance
1	,879	4,566	,507	50,737
2	,644	2,337	,260	25,962
Total		6,903	,767	
Mean	,799[a]	3,451	,383	38,349

a. Mean Cronbach's Alpha is based on the mean Eigenvalue.

Tabela 5.52 Valores próprios e inércias

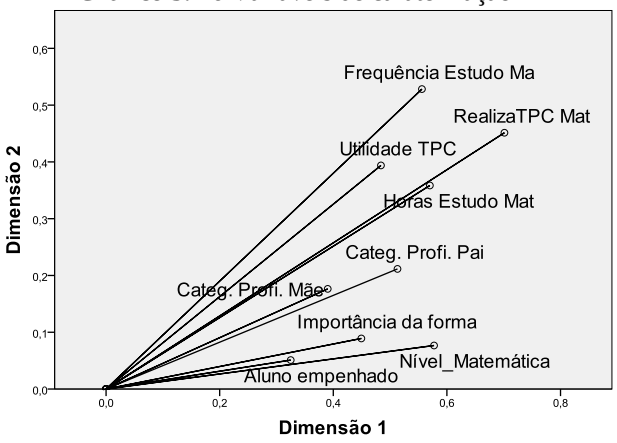
[14] A percentagem de variância explicada por cada dimensão =(inércia / variância total) x 100.

As variáveis assinaladas na Tabela 5.53 são as que mais discriminam em cada uma das dimensões, uma vez que os seus valores são próximos da inércia. No caso das variáveis "Categ. Profi. Mãe", "Importância da formação" e "Aluno empenhado", apesar de discriminarem menos que as outras 4 assinaladas na dimensão1, foram incluídas na dimensão 1, pois apresentam maiores medidas de discriminação nesta dimensão. A seleção da variável "Categ. Profi. Mãe" é justificada, ainda, por um motivo de natureza qualitativa, pois a variável "Categ. Profi. Pai" também foi escolhida.

Tabela 5.53 Variáveis de caraterização

	Dimension		Mean
	1	2	
Nível_Matemática	,577	,077	,327
Frequência Estudo Mat	,556	,528	,542
RealizaTPC Mat	,701	,451	,576
Utilidade TPC	,484	,394	,439
Horas Estudo Mat	,570	,358	,464
Categ. Profi. Mãe	**,390**	,176	,283
Categ. Profi. Pai	,513	,212	,362
Importância da formação	,449	,089	,269
Aluno empenhado	,325	,051	,188
Active Total	4,566	2,337	3,451
% of Variance	50,737	25,962	38,349

Gráfico 5. 46 Variáveis de caraterização



O Gráfico 5.46 corrobora a informação da Tabela 5.53. As variáveis que mais discriminam em cada dimensão são aquelas mais próximas de cada uma delas.

A leitura das medidas de discriminação permite concluir que as variáveis "Frequência Estudo Mat", "RealizaTPC Mat" e "Horas Estudo Mat", são relevantes nas duas dimensões. Em termos gráficos podemos ter esta perceção pela disposição destas variáveis na proximidade da diagonal do Gráfico 5.46.

5.5.3.2.3. Interpretação/Nomeação das dimensões

Para a dimensão 1 são particularmente determinantes as variáveis; "Nível_Matemática", "Frequência Estudo Mat", "RealizaTPC Mat", "Horas Estudo Mat", "Categ. Profi. Mãe", "Categ. Profi. Pai", "Importância da formação" e "Aluno empenhado". Esta dimensão combina a profissão dos pais do aluno com o desempenho do mesmo na disciplina de Matemática. Podemos pensar que esta dimensão aponta para o *"contexto socioeconómico* e *desempenho do aluno".*

A dimensão 2 contempla as variáveis : "Frequência Estudo Mat", "RealizaTPC Mat", "Horas Estudo Mat" e "Utilidade TPC". Esta dimensão refere-se ao trabalho individual do aluno e ao reconhecimento da utilidade do TPC por parte do mesmo, de certa forma remete para os *"hábitos de trabalho do aluno".*

O "contexto socioeconómico e desempenho do aluno" e os *"hábitos de trabalho do aluno",* são, portanto, os dois eixos temáticos do espaço em análise.

De seguida, com o objetivo de identificar as categorias mais diferenciadoras entre os objetos em análise, passamos à leitura das quantificações das categorias, em cada dimensão. Assim, a partir das quantificações e das contribuições determinadas pela ACM (ver Anexo V, Tabelas 3A a 11) e tendo como referência as variáveis que mais discriminam nas duas dimensões, identificaram-se: as coordenadas das categorias, e as contribuições mais elevadas. Desta forma foi-nos possível encontrar associações e oposições entre as categorias das variáveis selecionadas em cada dimensão[15]. Os resultados desta análise são apresentados nas Tabelas 5.54 e 5.55.

Pela leitura da Tabela 5.54 podemos afirmar que, na dimensão 1, temos tendencialmente em associação entre categorias que refletem baixos desempenhos e as categorias profissionais de menor qualificação. Em oposição, aparecem categorias que apontam para um melhor desempenho e categorias que refletem qualificação profissional. Assim, a dimensão 1 separa os alunos em função do seu desempenho e da categoria profissional dos respetivos pais.

[15] Fala-se em associações quando as categorias apresentam coordenadas com o mesmo sinal, por outro lado as oposições são avaliadas a partir das coordenadas com sinais opostos.

DIMENSÃO 1		
Variável	**Dim1 < 0**	**Dim1 > 0**
Nível a Matmática	• 2	• 3 • 4 • 5
Realiza TPC de Matemática	• Nunca • Às vezes	• Sempre
Horas de estudo na semana	• < 1	• 3_4 • 5_6 • > 6
Frequência de estudo na semana	• Nunca • Raramente	• 3_5 vezes • Td_dias
Importância da formação recebida	• Pequena • Alguma	• Grande
Consideras-te um aluno empenhado	• Às vezes	• Sim
Categoria profissional da mãe	• Trab N Qualif	• Espec/tec
Categoria profissional do pai	• Trab N Qualif	• Espec/tec

Tabela 5.54 Agregação das categorias na dimensão 1

DIMENSÃO 2		
Variável	**Dim2 < 0**	**Dim2 > 0**
Realiza TPC de Matemática	• Às vezes	• Nunca
Horas de estudo na semana	• 1_2 • 3_4	• < 1 • > 6
Frequência de estudo na semana	• 1_2 vezes	• Nunca • Td dias
Utilidade do TPC	• Úteis	• Nada úteis • Muito úteis

Tabela 5.55 Agregação das categorias na dimensão 2

Quanto à dimensão 2 (Tabela 5.55), salienta-se para as quatro variáveis, a associação das suas categorias intermédias, em oposição às categorias dos extremos.

Concluída a leitura/interpretação das dimensões identificamos as variáveis e as categorias que devem, sobremaneira, orientar a leitura dos planos, definidos a partir das dimensões que acabamos de interpretar.

5.5.3.2.4. Leitura dos planos: identificação de configurações

O objetivo desta fase é definir grupos de indivíduos que partilhem as mesmas caraterísticas. Para tal, pretendemos identificar no plano, com as projeções das categorias, as distâncias e oposições entre categorias face aos eixos definidos.

O Gráfico 5.45 representa o plano que cruza as duas primeiras dimensões. Temos assim o espaço de análise definido segundo os dois eixos estruturantes:
"Contexto socioeconómico e classificação a Matemática" e Hábitos de trabalho.

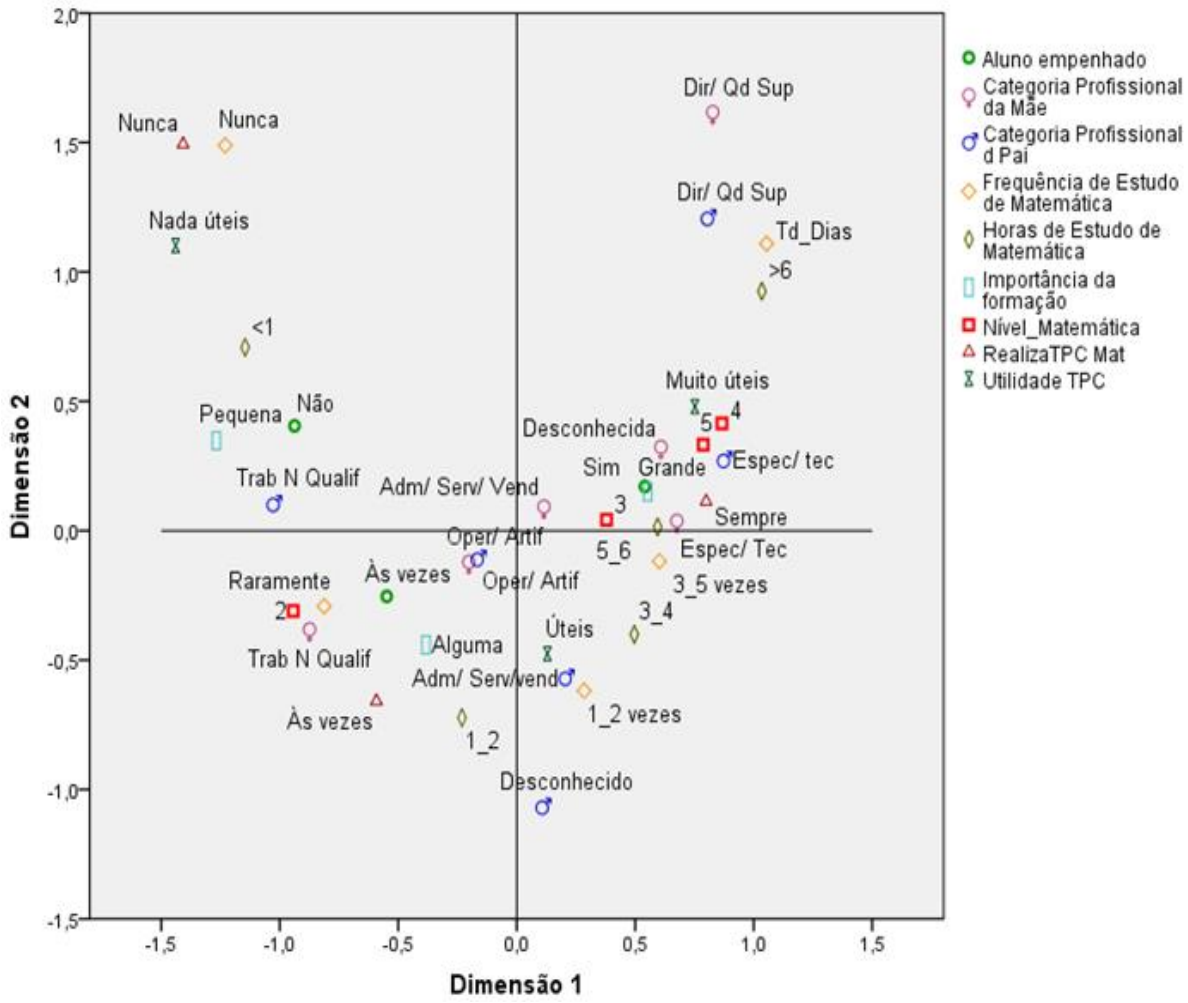


Gráfico 5.47 Representação das categorias

Pela leitura da distribuição das categorias no plano é possível identificar 4 perfis de alunos.

Perfil 1: grupo de alunos que tendencialmente conseguem os níveis superiores de classificação a Matemática (4 e 5), estudam todos os dias, consideram o TPC muito útil e realizam-no sempre, totalizam mais de 6 horas de estudo por semana, consideram-se empenhados no estudo , atribuem grande importância à formação recebida na escola. Os pais dos alunos são trabalhadores qualificados.

Perfil 2: grupo de alunos que tendencialmente atingem o nível 3 a Matemática, estudam 3 a 5 vezes por semana, consideram o TPC útil e realizam-no sempre, totalizando entre 3 a 4 horas de estudo por semana, consideram-se empenhados no estudo , atribuem grande importância à formação recebida na escola. Os pais dos alunos são trabalhadores qualificados.

Perfil 3: grupo de alunos que tendencialmente atingem o nível 2 a Matemática, raramente estudam, no entanto, reconhecem utilidade ao TPC e realizam-no às vezes, totalizando entre 1 a 2 horas de estudo por semana, consideram-se empenhados no estudo apenas às vezes , atribuem alguma importância à formação recebida na escola. Ambos os pais são trabalhdores não qualificados.

Perfil 4: grupo de alunos que tendencialmente atingem o nível 2 a Matemática, nunca estudam, não reconhecem qualquer utilidade ao TPC e nunca o realizam, totalizam menos de uma hora de estudo por semana, consideram-se empenhados no estudo apenas às vezes , atribuem pequena importância à formação recebida na escola. Ambos os pais são trabalhdores não qualificados.

Da análise do plano destaca-se ainda a forma aproximadamente parabólica da distribuição das categorias - oposição, na dimensão 2, das categorias dos extremos às categorias intermédias - conhecida por *Efeito de Guttman*.

Nesta fase da análise vamos procurar definir grupos de alunos a partir das duas dimensões encontradas na ACM que vão entrar como variáveis originais para a realização de uma Análise de *Clusters*, com o objetivo de podermos classificar os grupos segundo as suas carateristicas.

5.5.3.3. Aplicação da Análise de *Clusters* após a ACM

O objetivo deste ponto é a definição dos grupos de alunos a partir das duas dimensões encontradas na ACM. Estas dimensões vão entrar como variáveis originais para a realização de uma Análise de Clusters, com o objetivo de podermos classificar os grupos segundo as suas carateristicas.

A Análise de *Clusters* foi aplicada em duas fases:

- Aplicação de um método hierárquico, para validar a solução sugerida pelo plano da ACM;
- Utilização do método não hierárquico (*k-means*) e do método *TwoStep*, para a construção propriamente dita dos grupos de alunos afetos aos vários perfis.

5.5.3.3.1. Validação da solução sugerida pelo plano da ACM

Realizaram-se dois ensaios segundo dois critérios de agregação: *Ward e distância média entre Clusters (average linkage between groups)*. Foram selecionamos quatro clusters. Esta solução de Clusters foi avaliada recorrendo aos coeficientes de fusão[16] (Anexo VI, Tabelas 12 e 13) e ao critério do R-quadrado (*R-squared*)[17].

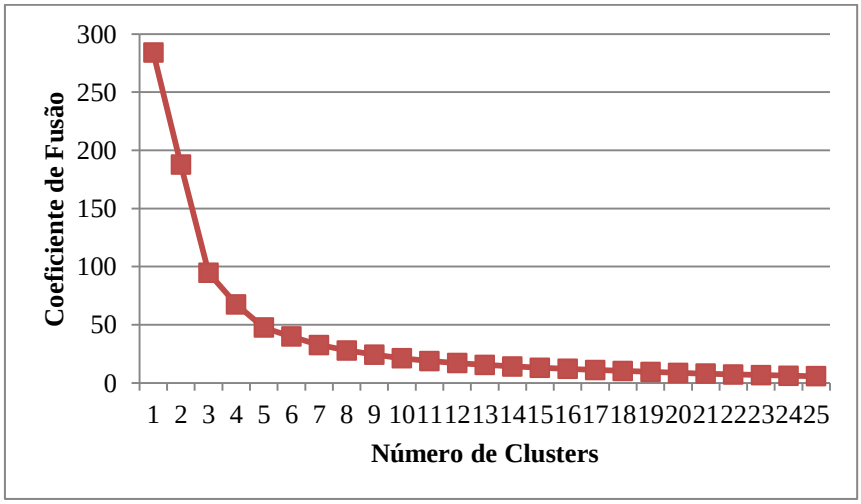


Gráfico 5.48 Coeficiente de fusão - critério de Ward

[16] Para facilitar a leitura dos gráficos selecionaram-se apenas os coeficientes de fusão referentes às últimas 25 agregações.

[17] O *R-squared* é uma medida da percentagem da variância total que é retida em cada uma das soluções dos Clusters.Os cálculos foram realizados com auxílio da ANOVA *one-way* do SPSS.

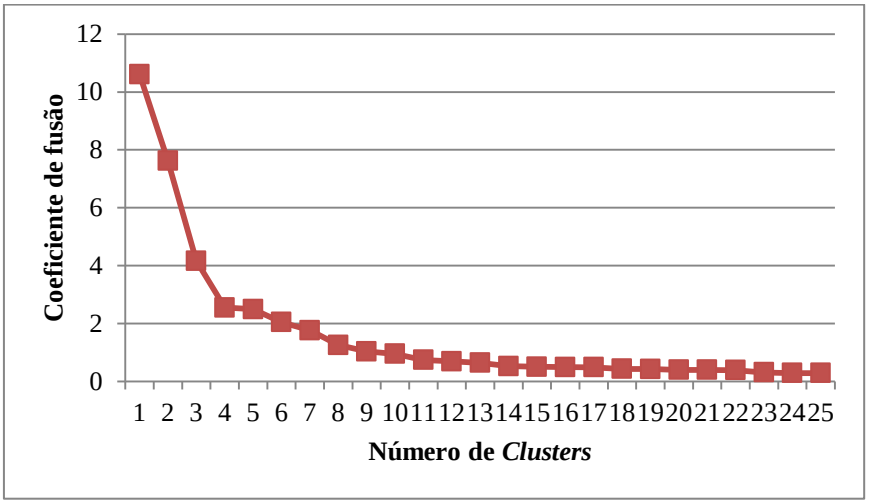


Gráfico 5.49 Coeficientes de fusão- critério da distância média entre *clusters*

Como se pode observar nos gráficos acima, registam-se declives acentuados até à solução com quatro *clusters* indicando-nos a possibilidade de selecionar quatro *clusters* como os mais pertinentes para a análise.

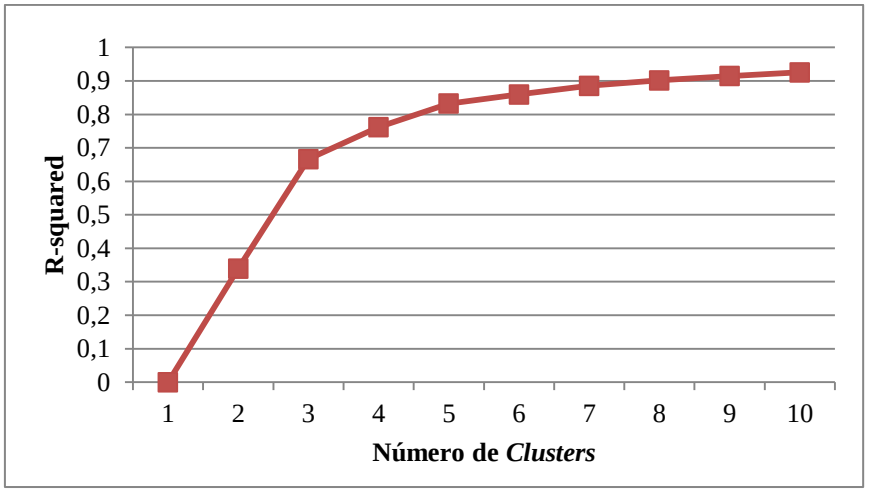


Gráfico 5.50 *R-Squared*

Podemos verificar no gráfico 5.50 que os ganhos de variabilidade retida por mais do que quatro *clusters* é relativamente pequeno quando comparada com a evolução de um para quatro *clusters*. Também este resultado indica que podemos reter quatro *clusters*.

Podemos concluir que os resultados do agrupamento hierárquico validam assim a solução sugerida pelo plano da ACM.

5.5.3.3.2. Construção dos grupos

Método de agrupamento não hierárquico *(k-means*)

Avançando para a segunda fase da Análise de *Clusters* utilizou-se a técnica de agrupamento não hierárquico *(k-means)* segundo o método iterar e classificar. No SPSS este procedimento está incluido no método *iterate and classify*. A interrupção do algoritmo ocorre quando a variação dos centroides deixa de ser significativa, ou quando é atingido o número máximo de iterações (definido pelo investigador) seja alcançado.

No nosso caso verifica-se (ver Anexo VI –Tabela 14) que o algoritmo termina no quarto passo uma vez que não ocorre variação dos centroides dos *clusters* 1,2 e 3 após o terceiro passo, o centroide do C*luster* 4 estabiliza logo depois do segundo passo.

A partição final pode ver-se na Tabela 5.56.

	N	%
Cluster/grupo1	17	12,0
Cluster/grupo2	61	42,9
Cluster/grupo3	44	31,0
Cluster/grupo4	20	14,1
Total	142	100,0

Tabela 5.56 Distribuição por *Cluster*

No Gráfico 5.51 pode observar-se a disposição dos alunos (objetos) segundo o *cluster* de pertença[18], verificamos que os quatro grupos estão bem delimitados.

No Gráfico 5.52 podemos ver os centroides dos *clusters*/grupos encontrados. A distância entre os centroides dos c*lusters* pode ser consultada no Anexo VI- Tabela 16.

[18] Usaram-se como coordenadas os scores determinados pela ACM nas duas dimensões em análise.

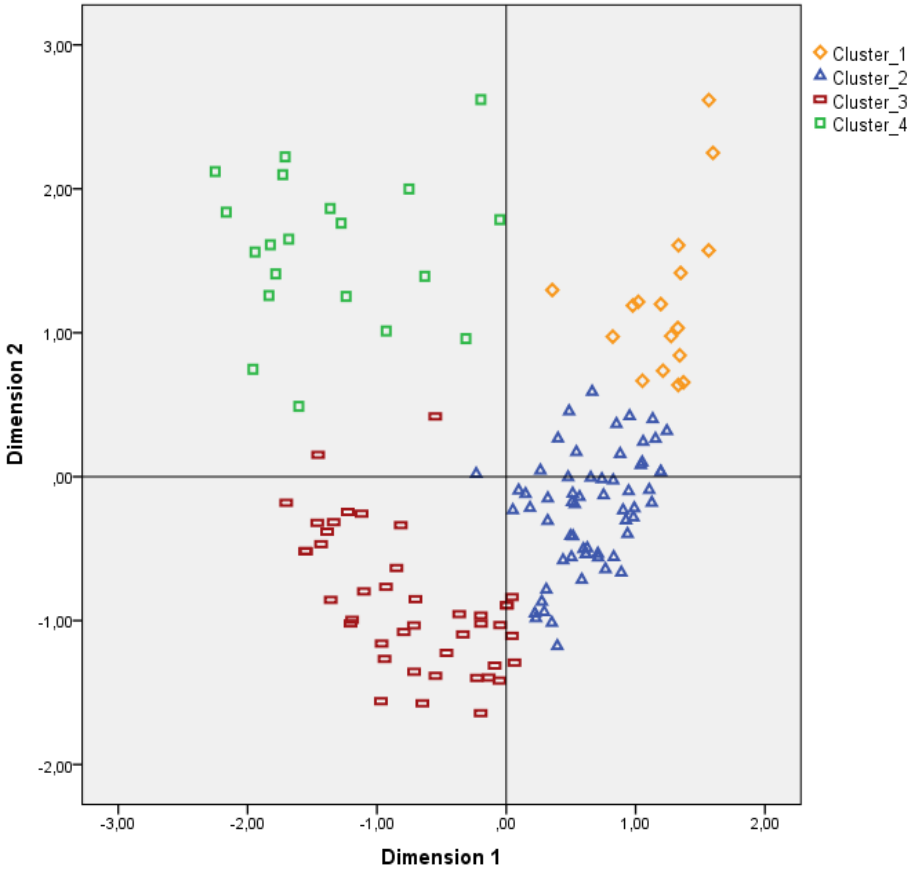


Gráfico 5. 51 Representação dos objetos

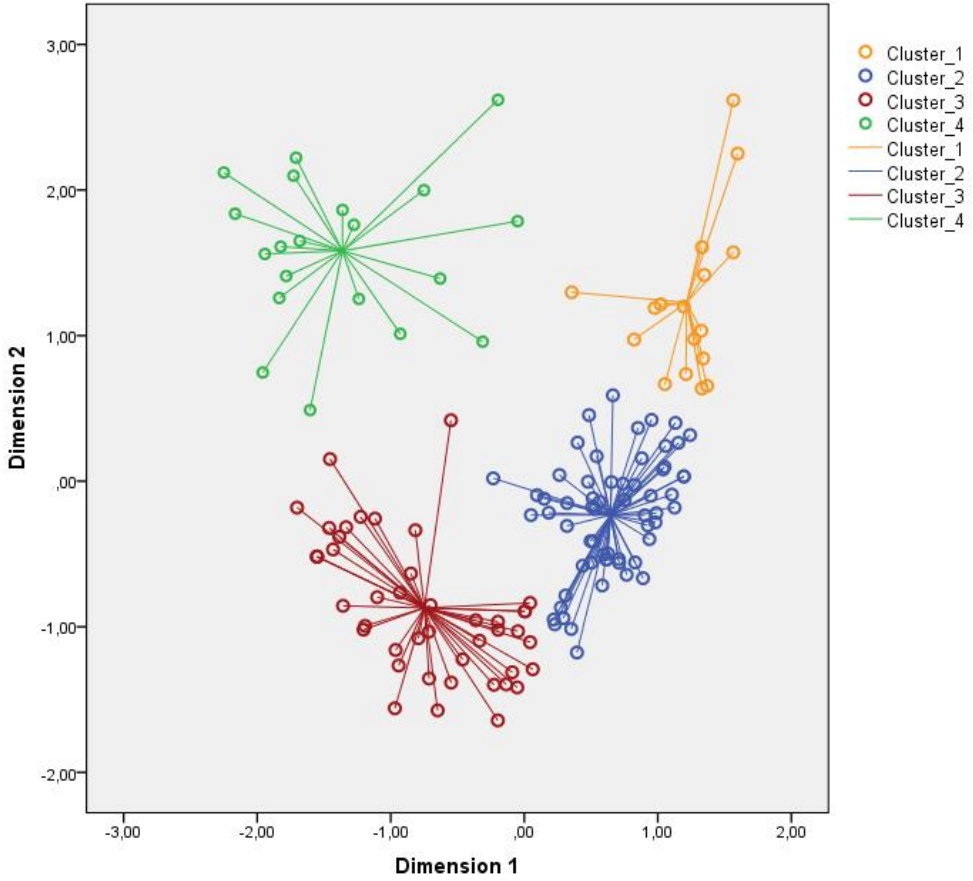


Gráfico 5.52 Centroides

Método *TwoStep*, com utilização a medida de distância Log-likelihood.

Model Summary

Algorithm	TwoStep
Inputs	2
Clusters	4

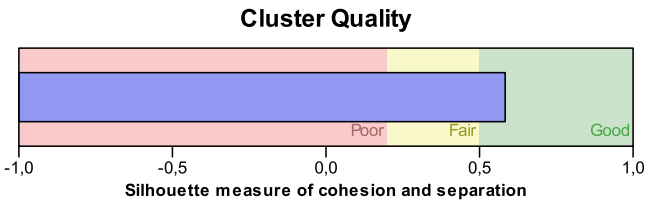
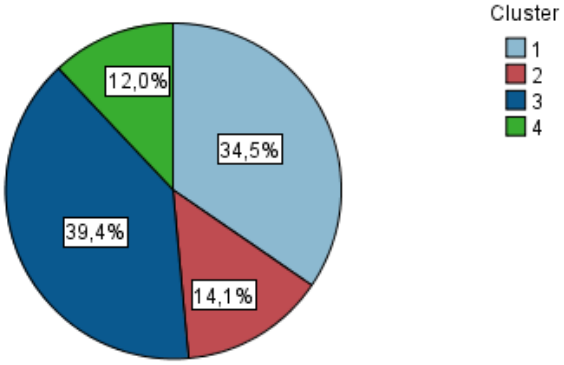


Figura 7 Qualidade do agrupamento



Size of Smallest Cluster	17 (12%)
Size of Largest Cluster	56 (39.4%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	3.29

Figura 8 Distribuição por *Clusters*

A aplicação do metodo *TwoStep* indica-nos uma boa solução para quatro *Clusters* (Figura 7). Verificamos que a distribuição por Cluster obtida neste último ensaio (Figura 8) é muito aproximada da distribuição obtida no ensaio anterior (Tabela 5.56).

5.6. Resultados

Pretendemos agora fazer corresponder cada um dos quatro C*lusters*/grupos a cada um dos quatro perfis configurados pelo plano da ACM (Gráfico 5.47). Para tal realizamos o

cruzamento entre as variáveis da ACM e os quatro c*lusters* obtidos nos dois ensaios (*K-meams* e *TwoStep).*

O resultado do cruzamento das variáveis com os quatro *Clusters* obtidos com a aplicação do método *TwoStep* pode ser consultado no Anexo VI na Tabela 18.

De seguida apresentamos o cruzamento das variáveis com os quatro *Clusters* resultantes da aplicação do método não hierárquico (*k-means*).

Considerando os valores assinalados nas Tabelas 5.57 e 5.58 obteve-se a seguinte correspondência *clusters/*perfis:

- Cluster 1 com 12,0% - grupo com boa classificação na disciplina de Matemática (nível 4, 47,1%). No estudo semanal destacam-se os níveis máximos, os alunos estudam todos os dias (64,7%), consideram o TPC muito útil (64,7%) e realizam-no sempre (100%), em média estudam mais de seis por semana horas (47,1%). Estes alunos atribuem grande importância à formação recebida na escola (94,1%) e consideram-se empenhados no estudo (88,2%). Relativamente à profissão do pai sobressai a categoria Especialistas/Técnicos (52,9%), no caso da mãe destacam-se as categorias Diretor/Quadro Superior e Pessoal administrativo/Serviços (29,4%) logo seguidas da categoria Especialistas/Técnicos (23,5%). Este grupo revela o melhor desempenho no estudo da disciplina de Matemática.
- Cluster 2 com 42,9% - grupo com classificação satisfatória na disciplina de Matemática (nível 3, 50,8%). Os alunos estudam entre uma a duas vezes por semana (44,3%), consideram o TPC útil (75,4%) e realizam-no sempre (82,0%), em média estudam três a quatro horas por semana (52,5%). Estes alunos atribuem grande importância à formação recebida na escola e consideram-se empenhados no estudo (73,8%). Relativamente à profissão do pai e da mãe salienta-se a categoria Especialistas/Técnicos (respetivamente 34,4% e 47,5%).
- Cluster 3 com 31,0% - grupo com classificação não satisfatória na disciplina de Matemática (nível 2, 86,4%). Os alunos raramente estudam (45,5%), consideram o TPC útil (77,3%) e realizam-no às vezes (88,6%), em média

estudam uma a duas horas por semana (34,1%). Estes alunos atribuem alguma importância à formação recebida na escola (54,5%), e consideram-se empenhados no estudo apenas às vezes (79,5%). Relativamente à profissão do pai e da mãe destaca-se a categoria Trabalhador Não Qualificado (respetivamente 43,2% e 56,8%).

- Cluster 4 com 14,1% - grupo com classificação não satisfatória na disciplina de Matemática (nível 2, 55,0%). No estudo semanal destacam-se os níveis minimos, os alunos nunca estudam (70,0%), consideram o TPC nada útil (90,0%) e nunca o realizam (85,0%), em média estudam menos de uma hora por semana (90,0%). Estes alunos atribuem pequena importância à formação recebida na escola (40,0%), e consideram-se empenhados no estudo apenas às vezes (50,0%). Relativamente à profissão do pai e da mãe destaca-se a categoria Trabalhador Não Qualificado (respetivamente 55,0% e 35,0%).

		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
		N	N %	N	N %	N	N %	N	N %
Nível a Matemática	2	0	,0%	5	8,2%	38	86,4%	11	55,0%
	3	6	35,3%	31	50,8%	6	13,6%	6	30,0%
	4	8	47,1%	12	19,7%	0	,0%	1	5,0%
	5	3	17,6%	13	21,3%	0	,0%	2	10,0%
	Total	17	100,0%	61	100,0%	44	100,0%	20	100,0%
Frequência de Estudo de Matemática	Nunca	0	,0%	1	1,6%	2	4,5%	14	70,0%
	Raramente	1	5,9%	6	9,8%	20	45,5%	6	30,0%
	1_2 vezes	0	,0%	27	44,3%	17	38,6%	0	,0%
	3_5 vezes	5	29,4%	24	39,3%	5	11,4%	0	,0%
	Td_Dias	11	64,7%	3	4,9%	0	,0%	0	,0
	Total	17	100,0%	61	100,0%	44	100,0%	20	100,0%
Realiza TPC de Matemática	Nunca	0	,0%	0	,0%	1	2,3%	17	85,0%
	Às vezes	0	,0%	11	18,0%	39	88,6%	3	15,0%
	Sempre	17	100,0%	50	82,0%	4	9,1%	0	,0%
	Total	17	100,0%	61	100,0%	44	100,0%	20	100,0%
Horas de Estudo de Matemática	<1	1	5,9%	3	4,9%	14	31,8%	18	90,0%
	1_2	0	,0%	11	18,0%	15	34,1%	2	10,0%
	3_4	2	11,8%	32	52,5%	10	22,7%	0	,0%
	5_6	6	35,3%	10	16,4%	5	11,4%	0	,0%
	>6	8	47,1%	5	8,2%	0	,0%	0	,0%
	Total	17	100,0%	61	100,0%	44	100,0%	20	100,0%

Tabela5.57 Caraterização dos *Clusters*

		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
		N	N %	N	N %	N	N %	N	N %
Utilidade do TPC	Nada úteis	0	,0%	0	,0%	6	13,6%	18	90,0%
	Úteis	6	35,3%	46	75,4%	34	77,3%	1	5,0%
	Muito Úteis	11	64,7%	15	24,6%	4	9,1%	1	5,0%
	Total	17	100,0%	61	100,0%	44	100,0%	20	100,0%
Importância da formação	Pequena	0	,0%	3	4,9%	10	22,7%	8	40,0%
	Alguma	1	5,9%	13	21,3%	24	54,5%	5	25,0%
	Grande	16	94,1%	45	73,8%	10	22,7%	7	35,0%
	Total	17	100,0%	61	100,0%	44	100,0%	20	100,0%
Categoria Profissinal da Mãe	Dir/ Qd Sup	5	29,4%	1	1,6%	0	,0%	1	5,0%
	Espec/ Tec	4	23,5%	29	47,5%	2	4,5%	4	20,0%
	Adm/ Serv/ Vend	5	29,4%	18	29,5%	9	20,5%	5	25,0%
	Oper/ Artif	1	5,9%	5	8,2%	7	15,9%	3	15,0%
	Trab N Qualif	0	,0%	8	13,1%	25	56,8%	7	35,0%
	Desconhecida	2	11,8%	0	,0%	1	2,3%	0	,0%
	Total	17	100,0%	61	100,0%	44	100,0%	20	100,0%
Categoria Profissinal do pai	Dir/ Qd Sup	5	29,4%	3	4,9%	0	,0%	1	5,0%
	Espec/ tec	9	52,9%	21	34,4%	2	4,5%	2	10,0%
	Adm/ Serv/vend	1	5,9%	17	27,9%	9	20,5%	1	5,0%
	Oper/ Artif	1	5,9%	12	19,7%	12	27,3%	5	25,0%
	Trab N Qualif	1	5,9%	6	9,8%	19	43,2%	11	55,0%
	Desconhecido	0	,0%	2	3,3%	2	4,5%	0	,0%
	Total	17	100,0%	61	100,0%	44	100,0%	20	100,0%
Aluno empenhado	Não	0	,0%	2	3,3%	2	4,5%	3	15,0%
	Às vezes	2	11,8%	14	23,0%	35	79,5%	10	50,0%
	Sim	15	88,2%	45	73,8%	7	15,9%	7	35,0%
	Total	17	100,0%	61	100,0%	44	100,0%	20	100,0%

Tabela 5.58 Caraterização dos *Clusters*

Os alunos pertencentes aos grupos três e quatro revelam um desempenho no estudo da disciplina de Matemática bastante preocupante, certamente reflexo da falta de hábitos de trabalho e da pequena importância atribuida à formação recebida na escola. Outro fator que aparece associado a este desempenho não satisfatório é a ausência de qualificação profissional dos pais dos alunos, como já foi observado na secção 5.5.2. deste trabalho.

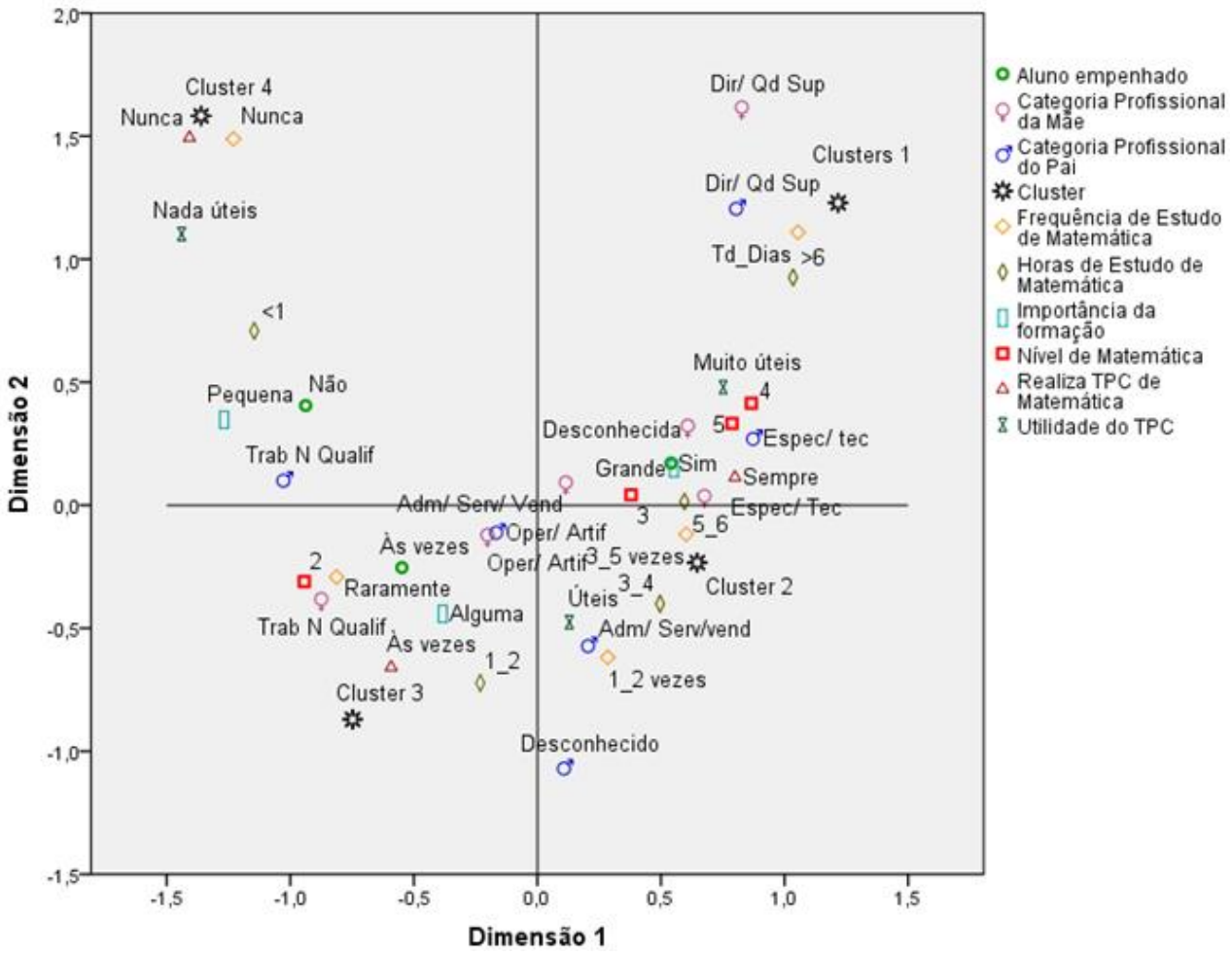


Gráfico 5.53 Disposição dos *Clusters* no espaço de análise

A projeção dos quatro *clusters* no plano da ACM permite validar a consistência desta classificação.

No Gráfico 5.53 observamos que a posição de cada um dos Clusters está próxima das categorias que caraterizam os respetivos perfis.

A projeção dos quatro *clusters (TwoStep)* no plano da ACM é possível observar no Anexo VI- Gráfico1. Também neste caso verificamos a proximidade de cada Clusters das categorias que caraterizam os respetivos perfis.

6. CONCLUSÕES E NOTAS FINAIS

Atingida a reta final do nosso trabalho, é o momento de tecer algumas considerações. Não pretendemos neste ponto repetir todos os resultados obtidos, mas apenas salientar aqueles que nos parecem mais importantes tendo em conta os três objetivos iniciais: identificar as perceções dos alunos sobre o contexto escolar, sobre o seu envolvimento no mesmo e sobre as suas perspetivas de prosseguimento de estudos; verificar as associações entre o desempenho dos alunos na disciplina de Matemática e um conjunto de variáveis; e procurar grupos de alunos utilizando mais do que um critério e comparar resultados.

Relativamente ao primeiro objetivo, constatamos que a maioria dos alunos gosta de estudar apenas em parte, no entanto, atribui grande importância ao trabalho desenvolvido na escola, apresenta uma ideia positiva acerca do apoio recebido por parte dos professores e espera estar a frequentar um curso superior daqui a cinco anos.

Quanto ao segundo e terceiro objetivos, identificamos, por via da concretização da ACM, quatro perfis. Posteriormente, recorrendo à Análise de *Clusters* foi possível a construção dos grupos afetos aos vários perfis. Nos grupos um e dois os alunos apresentam tendencialmente classificação igual ou superior ao nível três, dão grande importância ao trabalho desenvolvido na escola e apresentam hábitos/rotinas de estudo. Nestes dois grupos a maioria dos alunos tem pais qualificados profissionalmente. Por outro lado, nos grupos três e quatro, os alunos apresentam tendencialmente classificação inferior ao nível três, dão alguma ou pequena importância ao trabalho desenvolvido na escola e não manifestam hábitos de estudo. Nestes dois grupos a maioria dos pais dos alunos pertence à categoria profissional trabalhador não qualificado.

Apresentamos de seguida algumas considerações sobre a combinação das técnicas estatísticas utilizadas neste trabalho.
A ACM é uma técnica útil para disponibilizar associação entre variáveis qualitativas (ou categorizadas). Analisando o resultado gráfico da ACM, observamos as combinações

entre as categorias das variáveis e podemos identificar (caso existam) diferentes grupos, mas não é possível a definição efetiva da tipologia e daí justificar-se a combinação com outra técnica multivariada, neste caso a Análise de *Clusters.* A Análise de *Clusters* melhorou, de forma significativa a interpretação gráfica da ACM e permitiu a definição de quatro grupos de alunos.

A articulação da ACM com a Análise de *Clusters* validou os resultados da ACM, uma vez que os *clusters* gerados pela última técnica usada se apresentam geometricamente próximos dos perfis ACM.

Assim, concluímos que mediante os objetivos da análise/estudo é interessante a combinação de ambos os métodos, uma vez que resulta num aperfeiçoamento dos resultados de uma análise. Embora cada técnica possua particularidades e objetivos específicos de pesquisa, a associação de técnicas diferentes e complementares neutraliza os aspetos menos robustos de cada uma das técnicas e permite o refinamento das soluções e conclusões encontradas.

Por último, como sugestões para pesquisas futuras, pesamos que seria útil:

- a criação de uma nova variável que contemple a qualificação e a situação profissional dos pais, evitando a observação de categorias com um número muito reduzido de inquiridos o que pode baralhar os resultados. Propomos uma variável com cinco categorias agrupando em quatro categorias as profissões ativas (Trabalhadores com qualificação superior, Trabalhadores com qualificação intermédia, Trabalhadores sem qualificação e Outros ativos) e numa categoria as situações não ativas;
- alargar este tipo de estudo a outras escolas do ensino básico e/ou secundárias de outros concelhos do país e comparar os resultados obtidos.

BIBLIOGRAFIA

AGRESTI, A. (2002); *Categorial Data Analysis*, 2nd ed. John Wiley&Sons

CADIMA, JORGE (2010), *Apontamentos de Estatística Multivariada.* Disponível em: http://www.isa.utl.pt/dm/mestrado/mmacb/UCs/em/em.html

CARVALHO, HELENA (2000); *Homogeneidade e Correspondências Múltiplas: Comparação de dois métodos de análise* . Temas em Métodos Quantitativos, 1. Elizabeth Reis e Manuel Alberto Ferreira (editores), Lisboa, Edições. Sílabo, pp. 239-269.

CARVALHO, HELENA (2001); *Análise de Homogeneidade (HOMALS) – Quantificação Óptima e Múltipla de Dados Qualitativos*. Temas em Métodos Quantitativos, 2. Manuel Alberto Ferreira, Rui Menezes e Margarida Cardoso (editores), Lisboa, Edições Sílabo, pp.41-134.

CARVALHO, HELENA (2008); *Análise Multivariada de Dados Qualitativos.* 1ªed., Lisboa, Edições Sílabo.

COCHRAN, W.G. (1954); *Some methods for strengthening the common chi-squared tests.* Biometrics.

EVERITT, B. (1980); *Cluster Analysis.* 2a. ed., Halsted Press.

GREENACRE, M.J. (1984); *Theory and Applications of Correspondence Analysis*. London, Academic Press.

GREENACRE, M.J. (2008); *La práctica del análisis de correspondencias*. Fundación BBVA. Disponível em: http://www.fbbva.es/TLFU/tlfu/esp/publicaciones/libros/fichalibro/index.jsp?codigo=300

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W., (1998); *Multivariate Data Analysis*. 5th Edition, Prentice-Hall.

HILL, M. HILL, A. (2009); *Investigação por Questionário*. Lisboa, Edições Sílabo.

IEFP; *Classificação Nacional de Profissões em Portugal.* Disponível em: http://www.iefp.pt/formacao/CNP/Paginas/CNP.aspx

JONHSON, R.A.,WICHEN D.W (2002); *Applied Multivariate Statistical Methods*. Prentice Hall.

KAUFMAN, L.; ROUSSEEUW, P. J. (1990); *Finding Groups in Data.* Wiley Interscience

LIMA, TERESA (2010); *Lições de Álgebra Linear*. Imprensa da Universidade de Coimbra

MAROCO, JOÃO (2010); *Análise Estatística com utilização do SPSS*. Lisboa, Edições Sílabo.

MURTEIRA, BENTO J. F. (1990); *Probabilidades e Estatística, Vol. I e II*. Lisboa, McGraw-Hill.

MURTEIRA, BENTO. J. F. (1993); *Análise Exploratória de Dados – Estatística descritiva*. Lisboa, McGraw-Hill.

PEREIRA, H.G., SOUSA, AJ. (2002), *Análise de Dados para o Tratamento de Quadros Multidimensionais*. Disponível em: http://biomonitor.ist.utl.pt/~ajsousa/AnalDadosTratQuadMult.html

PESTANA, M. H. E GAGEIRO, J. N. (2008), *Análise de Dados para Ciências Sociais – A complementaridade do SPSS*. 5ª ed., Edições Sílabo.

REIS, ELIZABETH (2000); *A Análise de Clusters e as Aplicações às Ciências Empresariais: uma visão crítica da teoria dos grupos estratégicos* – Temas em Métodos Quantitativos 1. Elizabeth Reis e Manuel Alberto Ferreira (editores), Lisboa, Edições. Sílabo, pp. 205-238.

REIS, ELIZABETH (2001); *Estatística Multivariada Aplicada*. 2ª ed., Edições Sílabo.

REIS, ELIZABETH (2009); *Estatística Descritiva*. , 7ª ed., Edições Sílabo.

SILVESTRE, A. (2007); *Análise de Dados e Estatística Descritiva*. Escolar Editora.

SNEATH, P. H. ; SOKAL, R.R. (1973); *Numerical Taxonomy*. W. H. Freeman

SPSS Inc, (1998); *Optimal Scaling Methods for Multivariate Categorical Data Analysis*, White paper, USA.

SPSS Inc, (2001); *The SPSS Twostep Cluster Component. A scalable component enabling more efficient customer segmentation*, White paper – technical report, USA.

SPSS Inc, (2004); *SPSS Categories® 13.0*, USA.

SPSS Inc, (2007); *SPSS Categories™ 16.0*, USA.

SPSS Inc, (2007); *SPSS Statistics Base 17.0. User's Guide*, USA.

SPSS *Tutorials*. Disponível em: http://www.spsstools.net/spss.htm

ZHANG, T., R. RAMAKRISHNON, M. LIVNY (1996); *BIRCH: An efficient data clustering method for very large databases.* Disponível em: http://www.cs.sfu.ca/CourseCentral/459/han/papers/zhang96.pdf

ANEXO I - Classificação Nacional de Profissões em Portugal

Detalhes sobre a **Classificação Nacional de Profissões em Portugal,** segundo o Instituto do Emprego e Formação Profissional (IEFP)**:**

- Grande Grupo 1 – Quadros Superiores da Administração Pública, Dirigentes e Quadros Superiores de Empresa
- Grande Grupo 2 – Especialistas das Profissões Intelectuais e Científicas
- Grande Grupo 3 – Técnicos e Profissionais de Nível Intermédio
- Grande Grupo 4 – Pessoal Administrativo e Similares
- Grande Grupo 5 – Pessoal dos Serviços e Vendedores
- Grande Grupo 6 – Agricultores e Trabalhadores Qualificados da Agricultura e Pescas
- Grande Grupo 7 – Operários, Artífices e Trabalhadores Similares
- Grande Grupo 8 – Operários de Instalações e Máquinas e Trabalhadores de Montagem
- Grande Grupo 9 – Trabalhadores Não Qualificados

➢ Grande Grupo 1 – Quadros Superiores da Administração Pública, Dirigentes e Quadros Superiores de Empresa

Os trabalhadores classificam-se nos seguintes Sub Grandes Grupos:
1.1 - Quadros Superiores da Administração Pública
1.2 - Diretores de Empresa
1.3 - Diretores e Gerentes de Pequenas Empresas

➢ Grande Grupo 2 - Especialistas das Profissões Intelectuais e Científicas

Os trabalhadores classificam-se nos seguintes Sub Grandes Grupos:
2.1 - Especialistas das Ciências Físicas, Matemáticas e Engenharia
  2.1.1 - Físicos, Químicos e Especialistas Similares
  2.1.2 - Matemáticos, Estaticistas e Especialistas Similares
  2.1.3 - Especialistas da Informática
  2.1.4 - Arquitetos, Engenheiros e Especialistas Similares
2.2 - Especialistas das Ciências da Vida e Profissionais da Saúde.

2.2.1 - Especialistas das Ciências da Vida

2.2.2 - Médicos e Profissões Similares - à exceção dos Enfermeiros
2.2.3 - Enfermeiros
2.3 - Docentes do Ensino Secundário, Superior e Profissões Similares.
2.4 - Outros Especialistas das Profissões Intelectuais e Científicas.

- **Grande Grupo 3 - Técnicos e Profissionais de Nível Intermédio**

Os trabalhadores classificam-se nos seguintes Sub Grandes Grupos:
3.1 - Técnicos e Profissionais de Nível Intermédio das Ciências Físicas e Químicas, da
Engenharia e Trabalhadores Similares
3.2 - Profissionais de Nível Intermédio das Ciências da Vida e da Saúde
3.3 - Profissionais de Nível Intermédio do Ensino
3.4 - Outros Técnicos e Profissionais de Nível Intermédio

- **Grande Grupo 4 – Pessoal Administrativo e Similares**

Os trabalhadores classificam-se nos seguintes Sub Grandes Grupos:
4.1 - Empregados de Escritório
4.1.1 - Secretários e Operadores de Equipamentos de Tratamento de Informação
4.1.2 - Empregados dos Serviços de Contabilidade e dos Serviços Financeiros
4.1.3 - Empregados de Aprovisionamento, de Planeamento e dos Transportes
4.1.4 - Empregados de Biblioteca, Carteiros e Trabalhadores Similares
4.1.5 - Empregados de Escritório Não Classificados em Outra Parte
4.2 - Empregados de Receção, Caixas, Bilheteiros e Similares

- **Grande Grupo 5 – Pessoal dos Serviços e Vendedores**

Os trabalhadores classificam-se nos seguintes Sub Grandes Grupos:
5.1 - Pessoal dos Serviços Diretos e Particulares, de Proteção e Segurança
5.1.1 - Assistentes, Cobradores, Guias e Trabalhadores Similares
5.1.2 - Ecónomos e Pessoal do Serviço de Restauração
5.1.3 - Vigilantes, Assistentes Médicos e Trabalhadores Similares
5.1.4 - Outro Pessoal dos Serviços Diretos e Particulares
5.1.5 - Astrólogos e Trabalhadores Similares
5.1.6 - Pessoal dos Serviços de Proteção e Segurança
5.2 - Manequins, Vendedores e Demonstradores
5.2.1 - Manequins e Outros Modelos
5.2.2 - Vendedores e Demonstradores
5.2.3 - Vendedores de Quiosque e de Mercados

- **Grande Grupo 6 – Agricultores e Trabalhadores Qualificados da Agricultura e Pescas**

Os trabalhadores classificam-se nos seguintes Sub Grandes Grupos:
6.1 - Agricultores e Trabalhadores Qualificados da Agricultura, Criação de Animais e Pescas
- 6.1.1 - Agricultores e Trabalhadores Qualificados de Culturas Agrícolas
- 6.1.2 - Criadores e Trabalhadores Qualificados do Tratamento de Animais
- 6.1.3 - Agricultores e Trabalhadores Qualificados da Policultura, Criação e Tratamento de Animais
- 6.1.4 - Trabalhadores Florestais e Similares
- 6.1.5 - Trabalhadores da Aquacultura e Pescas

6.2 - Agricultores e Pescadores - Agricultura e Pesca de Subsistência

- **Grande Grupo 7 – Operários, Artífices e Trabalhadores Similares**

Os trabalhadores classificam-se nos seguintes Sub Grandes Grupos:
7.1 - Operários, Artífices e Trabalhadores Similares das Indústrias Extrativas e da Construção Civil
7.2 - Trabalhadores da Metalurgia e da Metalomecânica e Trabalhadores Similares
7.3 - Mecânicos de Precisão, Oleiros e Vidreiros, Artesãos, Trabalhadores das Artes Gráficas e Trabalhadores Similares
7.4 - Outros Operários, Artífices e Trabalhadores Similares

- **Grande Grupo 8 – Operários de Instalações e Máquinas e Trabalhadores de Montagem**

Os trabalhadores classificam-se nos seguintes Sub Grandes Grupos:
8.1 - Operadores de Instalações Fixas e Similares
8.2 - Operadores de Máquinas e Trabalhadores da Montagem
8.3 - Condutores de Veículos e Embarcações e Operadores de Equipamentos Pesados Móveis

- **Grande Grupo 9 – Trabalhadores Não Qualificados**

9.1 - Trabalhadores Não Qualificados dos Serviços e Comércio
9.2 - Trabalhadores Não Qualificados da Agricultura e Pescas
9.3 - Trabalhadores Não qualificados das Minas, da Construção e Obras Públicas, da Indústria Transformadora e dos Transportes

ANEXO II - Questionário

Questionário aos alunos

Este questionário faz parte de uma pesquisa sobre o rendimento dos alunos a Matemática. A tua colaboração é fundamental para o sucesso da mesma.

Não escrevas o teu nome em nenhuma parte do questionário, pois está assegurada a confidencialidade da informação facultada.

Lê com atenção cada questão antes de responderes. Se tiveres alguma dúvida coloca o dedo no ar para seres esclarecido.

Nota: Nenhuma resposta deverá ficar em branco

1) Qual é a tua Idade? _____anos.

Para cada uma das situações, assinala **com um X** nos quadrados correspondentes de acordo com a tua situação/opinião.

2) Qual é o teu sexo: Feminino ☐ Masculino ☐

3) Quais foram as notas com que terminaste o ano letivo anterior (2010/2011)?

Disciplinas / Nível	P	I	F	H	G	M	FQ	CN	EV	EM	ET	EF	AP	EA	FC
Nível 1															
Nível 2															
Nível 3															
Nível 4															
Nível 5															
NS															
S															
SB															

4) Indica qual a situação em que terminaste o ano letivo anterior?

Transitei de ano ☐ Não transitei de ano ☐

5) O que pensas estar a fazer daqui a 5 anos?

A tirar um curso superior	
A tirar um curso não superior	
A trabalhar com o 12º ano concluído	
A trabalhar sem o 12º ano concluído	
Ainda não sei/não pensei nisso	

6) Com quem vives?

Pai e mãe		Pai, mãe, irmãos e avós	
Pai e irmãos		Pai, irmãos e avós	
Mãe e irmãos		Mãe, irmãos e avós	
Pai, mãe e irmãos		Avós e irmãos	
Pai		Mãe	
Avós		Tios	

Outros. Quem?

Nas perguntas 7 e 8, indica a profissão dos teus pais, mesmo que não se encontrem a exercer a profissão neste momento.

7) Indica qual é a profissão do teu pai: _____
(apresenta só a profissão principal)

8) Indica qual é a profissão da tua mãe: _____
(apresenta só a profissão principal)

9) Indica qual é a situação de vida dos teus pais:

	Pai	Mãe
Reformado(a)		
Desempregado(a)		
Empregado(a)		
Emigrante (a trabalhar no estrangeiro)		
Em formação		
Inválido(a) para trabalhar		
Em casa (a cuidar do lar)		
Outra. Qual?		

10) O teu Encarregado de Educação é:

O teu pai ☐ A tua mãe ☐

Outro ☐ Quem? _____

11) Em tua casa, tens:

	Sim	Não
Computador		
Acesso a Internet		
Livros não escolares		
Acesso a Canais de TV temáticos		

12) A que horas te deitas em tempo de aulas?

21-22h ☐ 22-23h ☐ 23-24h ☐ Depois das 24h ☐

13) No global, gostas de estudar?

Sim ☐ Em parte ☐ Não ☐

14) Estudas habitualmente:

Na escola	
Em casa	
Em casa dos amigos	

Noutro local. Qual?

15) Frequentas alguma atividade extracurricular? Não ☐

Sim ☐ Qual?_____

16) Consideras-te um aluno(a):

	Sim	**Não**	**Às vezes**
Assíduo			
Pontual			
Participativo			
Empenhado			
Com iniciativa			
Distraído			

17) Completa as frases:

17.1) As minhas disciplinas preferidas são: (indica as disciplinas, por ordem de preferência) 1º lugar _____ 2º lugar _____	17.2) As disciplinas que menos gosto são: (indica as disciplinas por ordem de menor preferência) 1º lugar _____ 2º lugar _____

18) Sentes dificuldades na disciplina de Matemática?

Sim, bastantes ☐ Às vezes, depende dos assuntos ☐ Não ☐

19) Indica com que frequência estudas Matemática:

Nunca	Raramente (menos de 1 vez por semana)	1 a 2 vezes por semana	3 a 5 vezes por semana	Todos os dias

20) Em média, quantas horas por semana dedicas ao estudo da disciplina de Matemática?

_____ (horas)

21) Realizas os trabalhos de casa propostos pelo professor de Matemática:

Sempre ☐ Às vezes ☐ Nunca ☐

22) Tens ajuda na realização dos trabalhos de casa de Matemática?

Sim, sempre ou quase ☐ Às vezes ☐ Nunca ☐

23) Consideras que os trabalhos de casa são:

Nada úteis ☐ Úteis ☐ Muito úteis ☐

24) Na tua opinião os trabalhos de casa:

	Sim	Não
24.1. Servem para tomar consciência das minhas dúvidas		
24.2. Servem para me ajudar a memorizar as matérias		
24.3. Servem para praticar o que demos nas aulas		

25) Na tua opinião acerca do apoio que recebeste dos teus professores:

	Sim	Não	Em parte (alguns professores)
25.1. Os professores ajudaram-me a compreender as matérias			
25.2. Os professores ouviram as minhas ideias e opiniões			
25.3. Os professores ouviram os meus problemas			
25.4. Os professores ajudaram-me a ultrapassar as minhas dificuldades			

26) Qual é a tua opinião acerca da ordem/disciplina que existia dentro da sala de aula?

	Sim	Não
26.1. Havia elementos problemáticos		
26.2. Os professores conseguiam impor a ordem		

27) Consideras que o trabalho que desenvolves na escola irá ter importância na tua formação?

Grande ☐ Alguma ☐ Pequena ☐ Nenhuma ☐

Terminaste o questionário.
Obrigada pela tua colaboração.

ANEXO III -Variáveis

	Variável	Designação	Tipo de variável	Valores assumidos
1	Idade do inquirido	Idade	Scale	Anos completos
2	Sexo do inquirido	Sexo	Nominal	1-Masculino 2-Feminino
3	Níveis do ano letivo anterior			
3.1	Português	NívelPort	Ordinal	2, 3, 4, 5
3.2	Inglês	NívelIng	Ordinal	2, 3, 4, 5
3.3	Francês	NívelFr	Ordinal	2, 3, 4, 5
3.4	História	NívelHist	Ordinal	2, 3, 4, 5
3.5	Geografia	NívelGeo	Ordinal	2, 3, 4, 5
3.6	Matemática	NívelMat	Ordinal	2, 3, 4, 5
3.7	Físico-Química	NívelFQ	Ordinal	2, 3, 4, 5
3.8	Ciências Naturais	NívelCN	Ordinal	2, 3, 4, 5
4	Situação Terminal	Situação Terminal	Nominal	1-Não Transitou de ano 2- Transitou de ano
5	Perspetivas do aluno (O que pensas estar a fazer daqui a 5 anos?)	Perspetivas	Nominal	1-Curso Superior 2-Curso Não Superior 3-Trabalhar com o12ºano 4- Trabalhar sem o 12ºano 5-Não sei/Não pensei
6	Agregado familiar	Agregado	Nominal	1-Pai e mãe 2-Mãe e irmãos 3-Pai, mãe e irmãos 4-Pai, mãe, irmãos e avós 5-Pai, irmãos e avós 7-Mãe, irmãos e padrasto
7	Categoria Profissional da mãe	Cat_Prof_Mãe	Nominal	1-Diretor/Quadro Superior 2-Especialistas/Técnicos 3-Pessoal administrativo/Serviços e Vendedores 4-Operários/Artífices 5-Trabalhador não qualificado 6-Desconhecido
8	Categoria Profissional do pai	Cat_Prof_Pai	Nominal	1-Diretor/Quadro Superior 2-Especialistas/Técnicos 3-Pessoal administrativo/Serviços e Vendedores 4-Operários/Artífices 5-Trabalhador não qualificado 6-Desconhecido
9	Situação Profissional dos pais			
9.1	Situação Profissional do pai	Si_Prof_Pai	Nominal	1-Reformado 2-Desempregado 3-Empregado 4-Em formação 5-Outro/desconhecido

	Variável	Designação	Tipo de variável	Valores assumidos
9.2	Situação Profissional do mãe	Si_Prof_Mãe	Nominal	1-Reformada 2-Desempregada 3-Empregada 4-Em formação 5-Outro/desconhecido
10	Encarregado de Educação	EE	Nominal	1-Pai 2-Mãe 3-Avó
11	Recursos em casa			
11.1	Em casa tem computador	RecurComputador	Nominal	1-Não 2-Sim
11.2	Em casa tem internet	RecurInternet	Nominal	1-Não 2-Sim
11.3	Em casa tem livros não escolares	RecurLivros	Nominal	1-Não 2-Sim
11.4	Em casa tem canais de TV temáticos	RecurCanaisTemáticos	Nominal	1-Não 2-Sim
12	Hora de Deitar	HoraDeitar	Nominal	1-Depois das 24h 2-Entre as 23h e 24h 3- Entre as 22h e 23h 4- Entre as 21h e 22h
13	Gostas de Estudar	GostoEstudar		1-Não 2-Em parte 3-Sim
14	Local de Estudo	LocalEstudo	Nominal	1-Escola 2-Casa 3-Casa dos amigos 4-Centro de estudos 5-Outros
15	Atividade Extracurricular			
15.1	Pratica atividade Extracurricular	AtividadeExtra	Nominal	1-Não 2-Sim
15.2	Atividade Extracurricular Praticada	Ativ_Ext_Prat	Nominal	1-Nenhuma 2-Jogos desportivos coletivos 3-Natação 4-Academia Música 5-Instituto de línguas 6-Artes Marciais 7-Atividades Rítmicas Expressivas 8-Escutismo 9-Equitação
16	Aluno			
16.1	Consideras-te um aluno assíduo	AlunoAssiduo	Ordinal	1-Não 2-Às vezes 3-Sim
16.2	Consideras-te um aluno pontual	AlunoPontual	Ordinal	1-Não 2- Às vezes 3-Sim
16.3	Consideras-te um aluno participativo	AlunoParticipativo	Ordinal	1-Não 2- Às vezes 3-Sim
16.4	Consideras-te um aluno empenhado	AlunoEmpenhado	Ordinal	1-Não 2- Às vezes 3-Sim
16.5	Consideras-te um aluno com iniciativa	AlunoIniciativa	Ordinal	1-Não 2- Às vezes 3-Sim

	Variável	Designação	Tipo de variável	Valores assumidos
16.6	Consideras-te um aluno distraído	AlunoDistraído	Ordinal	1-Não 2- Às vezes 3-Sim
17	Disciplinas			
17.1	Primeira disciplina preferida	PrimeiraPreferida	Nominal	P, I, F, H, G, M, FQ, CN, EV, EM, ET, EF, AP, EA, FC
17.2	Segunda disciplina preferida	SegundaPreferida	Nominal	P, I, F, H, G, M, FQ, CN, EV, EM, ET, EF, AP, EA, FC
17.3	Primeira disciplina não preferida	PrimeiraNãoPreferida	Nominal	P, I, F, H, G, M, FQ, CN, EV, EM, ET, EF, AP, EA, FC
17.4	Segunda disciplina não preferida	SegundaNãoPreferida	Nominal	P, I, F, H, G, M, FQ, CN, EV, EM, ET, EF, AP, EA, FC
18	Dificuldades Mat	Dificuldades	Ordinal	1-Sim bastantes 2- Às vezes 3-Não
19	Frequência de estudo Matemática na semana	FreqEstudoMat	Ordinal	1-Nunca 2-Raramente 3-1 a 2 vezes 4-3 a 5 vezes 5-Todos os dias
20	Horas de estudo na semana	HorasEstudo	Ordinal	1-<1 2-1 a 2 3-3 a 4 4-5 a 6 5->6
21	Realiza o TPC de Matemática	TPCMat	Ordinal	1-Nunca 2- Às vezes 3-Sempre
22	Ajuda no TPC de Matemática	AjudaTPC Mat	Ordinal	1-Nunca 2- Às vezes 3-Sempre
23	Utilidade do TPC	UtilidadeTPC	Ordinal	1-Nada úteis 2-Úteis 3-Muito úteis
24	Os TPCs servem para:			
24.1	tomar consciência das dúvidas	TPCConsDúvidas	Nominal	1-Não 2-Sim
24.2	ajudar a memorizar	TPCAjudaMemo	Nominal	1-Não 2-Sim
24.3	praticar	TPCPraticar	Nominal	1-Não 2-Sim
25	Apoio recebido por parte dos professores:			
25.1	Ajudaram a compreender as matérias	ProfAjudaCompreend	Ordinal	1-Não 2-Em parte 3- Sim
25.2	Ouviram as minhas opiniões	ProfOuveOpiniões	Ordinal	1-Não 2-Em parte 3- Sim
25.3	Ouviram os meus problemas	ProfOuveProblemas	Ordinal	1-Não 2-Em parte 3- Sim
25.4	Ajudaram a ultrapassar as dificuldades	ProfAjudaDificuldade	Ordinal	1-Não 2-Em parte 3- Sim

	Variável	Designação	Tipo de variável	Valores assumidos
26	Ordem/disciplina dentro da sala de aula			
26.1	Existência de elementos Problemáticos	ElemProbl	Nominal	1-Não 2-Sim
26.2	Professores conseguiam impor ordem	Ordem	Nominal	1-Não 2-Sim
27	Importância da formação recebida	ImpFormação	Ordinal	2-Pequena 3-Alguma 4-Grande

Tabela 1- Caraterísticas das variáveis

ANEXO IV – Análise de *Clusters* (com variáveis originais)

Model Summary

Algorithm	TwoStep
Inputs	9
Clusters	3

Cluster Quality

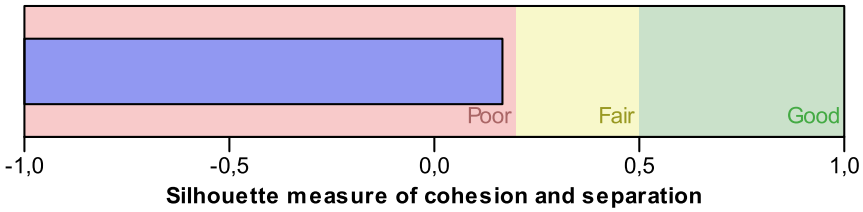


Figura 1- Qualidade do Agrupamento 2

Model Summary

Algorithm	TwoStep
Inputs	9
Clusters	4

Cluster Quality

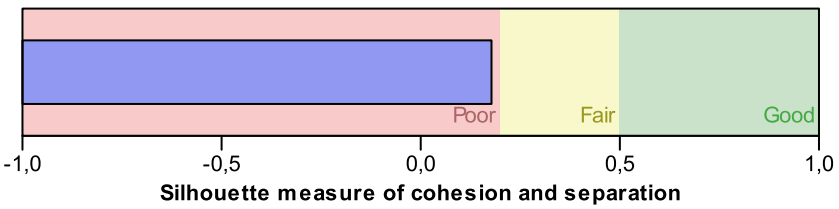


Figura 2- Qualidade do Agrupamento 3

		TwoStep Cluster Number					
		Outlier Cluster		1		2	
		Count	Column N %	Count	Column N %	Count	Column N %
Nível_Matemática	2	0	,0%	11	12,0%	43	86,0%
	3	0	,0%	42	45,7%	7	14,0%
	4	0	,0%	21	22,8%	0	,0%
	5	0	,0%	18	19,6%	0	,0%
	Total	0	,0%	92	100,0%	50	100,0%
RealizaTPC Mat	Nunca	0	,0%	2	2,2%	16	32,0%
	Às vezes	0	,0%	19	20,7%	34	68,0%
	Sempre	0	,0%	71	77,2%	0	,0%
	Total	0	,0%	92	100,0%	50	100,0%
Frequência Estudo Mat	Nunca	0	,0%	3	3,3%	14	28,0%
	Raramente	0	,0%	8	8,7%	25	50,0%
	1_2 vezes	0	,0%	36	39,1%	8	16,0%
	3_5 vezes	0	,0%	31	33,7%	3	6,0%
	Td_Dias	0	,0%	14	15,2%	0	,0%
	Total	0	,0%	92	100,0%	50	100,0%
Horas Estudo Mat	<1	0	,0%	6	6,5%	30	60,0%
	1_2	0	,0%	16	17,4%	12	24,0%
	3_4	0	,0%	38	41,3%	6	12,0%
	5_6	0	,0%	19	20,7%	2	4,0%
	>6	0	,0%	13	14,1%	0	,0%
	Total	0	,0%	92	100,0%	50	100,0%
Utilidade TPC	Nada úteis	0	,0%	2	2,2%	22	44,0%
	Úteis	0	,0%	62	67,4%	25	50,0%
	Muito úteis	0	,0%	28	30,4%	3	6,0%
	Total	0	,0%	92	100,0%	50	100,0%
Categ. Profi. Mãe	Dir/ Qd Sup	0	,0%	7	7,6%	0	,0%
	Espec/ Tec	0	,0%	36	39,1%	3	6,0%
	Adm/ Serv/ Vend	0	,0%	26	28,3%	11	22,0%
	Oper/ Artif	0	,0%	8	8,7%	8	16,0%
	Trab N Qualif	0	,0%	13	14,1%	27	54,0%
	Desconhecida	0	,0%	2	2,2%	1	2,0%
	Total	0	,0%	92	100,0%	50	100,0%
Categ. Profi. Pai	Dir/ Qd Sup	0	,0%	9	9,8%	0	,0%
	Espec/ tec	0	,0%	32	34,8%	2	4,0%
	Adm/ Serv/vend	0	,0%	22	23,9%	6	12,0%
	Oper/ Artif	0	,0%	18	19,6%	12	24,0%
	Trab N Qualif	0	,0%	7	7,6%	30	60,0%

	Desconhecido	0	,0%	4	4,3%	0	,0%
	Total	0	,0%	92	100,0%	50	100,0%
Importância da formação	Pequena	0	,0%	3	3,3%	18	36,0%
	Alguma	0	,0%	22	23,9%	21	42,0%
	Grande	0	,0%	67	72,8%	11	22,0%
	Total	0	,0%	92	100,0%	50	100,0%

Tabela 2- Caraterização dos *Clusters*

ANEXO V – ACM

Dimension	Variance Accounted For		
	Total (Eigenvalue)	Inertia	% of Variance
1	4,566	,507	50,737
2	2,337	,260	25,962
3	1,973	,219	21,928
4	1,813	,201	20,148
5	1,569	,174	17,432
6	1,350	,150	14,995
7	1,295	,144	14,384
8	1,234	,137	13,709
9	1,184	,132	13,151
10	1,099	,122	12,206
11	1,010	,112	11,221
12	,971	,108	10,788
13	,931	,103	10,346
14	,820	,091	9,116
15	,789	,088	8,762
16	,734	,082	8,159
17	,673	,075	7,477
18	,652	,072	7,242
19	,615	,068	6,830
20	,531	,059	5,895
21	,493	,055	5,480
22	,423	,047	4,699
23	,421	,047	4,682
24	,352	,039	3,914
25	,295	,033	3,281
26	,277	,031	3,080
27	,245	,027	2,720
28	,221	,025	2,456
29	,128	,014	1,423
Total	29,000	3,222	
Mean	1,000	,111	11,111

Tabela 3- Distribuição dos valores próprios e da inércia

Points:Coordinates

Category	Frequency	Centroid Coordinates	
		Dimension	
		1	2
2	54	-,943	-,310
3	49	,379	,042
4	21	,866	,415
5	18	,787	,332

Category	Mass	Inertia	Contribution			
			Of Point to Inertia of Dimension		Of Dimension to Inertia of Point	
			1	2	1	2
2	,042	,069	,074	,016	,546	,059
3	,038	,073	,011	,000	,076	,001
4	,016	,095	,024	,011	,130	,030
5	,014	,097	,017	,006	,090	,016
Active Total	,111	,333	,126	,033		

Tabela 3A - Variável "Nível a Matemática"

Points:Coordinates

Category	Frequency	Centroid Coordinates	
		Dimension	
		1	2
Nunca	17	-1,230	1,489
Raramente	33	-,813	-,292
1_2 vezes	44	,284	-,618
3_5 vezes	34	,602	-,118
Td_Dias	14	1,054	1,109

Category	Mass	Inertia	Contribution			
			Of Point to Inertia of Dimension		Of Dimension to Inertia of Point	
			1	2	1	2
Nunca	,013	,098	,040	,114	,206	,302
Raramente	,026	,085	,034	,008	,200	,026
1_2 vezes	,034	,077	,005	,051	,036	,172
3_5 vezes	,027	,085	,019	,001	,114	,004
Td_Dias	,011	,100	,024	,052	,122	,135
Active Total	,111	,444	,122	,226		

Tabela 4-Variável **"**Frequência do estudo"

Points:Coordinates

Category	Frequency	Centroid Coordinates	
		Dimension	
		1	2
Não	7	-,938	,405
Às vezes	61	-,549	-,254
Sim	74	,541	,171

Category	Mass	Inertia	Contribution			
			Of Point to Inertia of Dimension		Of Dimension to Inertia of Point	
			1	2	1	2
Não	,005	,106	,009	,003	,046	,008
Às vezes	,048	,063	,028	,012	,227	,049
Sim	,058	,053	,033	,007	,318	,032
Active Total	,111	,222	,071	,022		

Tabela 5- Variável "Aluno empenhado"

Points:Coordinates

Category	Frequency	Centroid Coordinates	
		Dimension	
		1	2
Nunca	18	-1,408	1,494
Às vezes	53	-,592	-,659
Sempre	71	,799	,113

Category	Mass	Inertia	Contribution			
			Of Point to Inertia of Dimension		Of Dimension to Inertia of Point	
			1	2	1	2
Nunca	,014	,097	,055	,121	,288	,324
Às vezes	,041	,070	,029	,069	,209	,258
Sempre	,056	,056	,070	,003	,638	,013
Active	,111	,222	,154	,193		
Total						

Tabela 6- Variável "Realiza o TPC"

Points:Coordinates

Category	Frequency	Centroid Coordinates	
		Dimension	
		1	2
Nada úteis	24	-1,440	1,102
Úteis	87	,129	-,475
Muito úteis	31	,752	,480

Category	Mass	Inertia	Contribution			
			Of Point to Inertia of Dimension		Of Dimension to Inertia of Point	
			1	2	1	2
Nada úteis	,019	,092	,077	,088	,421	,247
Úteis	,068	,043	,002	,059	,026	,357
Muito úteis	,024	,087	,027	,022	,158	,064
Active	,111	,222	,106	,169		
Total						

Tabela 7- Variável "Utilidade do TPC"

Points:Coordinates

Category	Frequency	Centroid Coordinates	
		Dimension	
		1	2
Pequena	21	-1,267	,348
Alguma	43	-,383	-,441
Grande	78	,552	,150

Category	Mass	Inertia	Contribution			
			Of Point to Inertia of Dimension		Of Dimension to Inertia of Point	
			1	2	1	2
Pequena	,016	,095	,052	,008	,279	,021
Alguma	,034	,077	,010	,025	,064	,084
Grande	,061	,050	,037	,005	,372	,027
Active	,111	,222	,098	,038		
Total						

Tabela 8- Variável "Importância da formação"

Points:Coordinates

Category	Frequency	Centroid Coordinates	
		Dimension	
		1	2
Dir/ Qd Sup	7	,827	1,617
Espec/ Tec	39	,676	,038
Adm/ Serv/ Vend	37	,115	,092
Oper/ Artif	16	-,202	-,121
Trab N Qualif	40	-,874	-,381
Desconhecida	3	,608	,323

Category	Mass	Inertia	Contribution			
			Of Point to Inertia of Dimension		Of Dimension to Inertia of Point	
			1	2	1	2
Dir/ Qd Sup	,005	,106	,007	,055	,035	,136
Espec/ Tec	,031	,081	,027	,000	,173	,001
Adm/ Serv/ Vend	,029	,082	,001	,001	,005	,003
Oper/ Artif	,013	,099	,001	,001	,005	,002
Trab N Qualif	,031	,080	,047	,018	,300	,057
Desconhecida	,002	,109	,002	,001	,008	,002
Active Total	,111	,556	,085	,075		

Tabela 9- Variável "Categoria profissional da mãe"

Points:Coordinates

Category	Frequency	Centroid Coordinates	
		Dimension	
		1	2
Dir/ Qd Sup	9	,804	1,205
Espec/ tec	34	,872	,269
Adm/ Serv/vend	28	,204	-,573
Oper/ Artif	30	-,167	-,112
Trab N Qualif	37	-1,028	,100
Desconhecido	4	,107	-1,071

Category	Mass	Inertia	Contribution			
			Of Point to Inertia of Dimension		Of Dimension to Inertia of Point	
			1	2	1	2
Dir/ Qd Sup	,007	,104	,009	,039	,044	,098
Espec/ tec	,027	,085	,040	,007	,239	,023
Adm/ Serv/vend	,022	,089	,002	,028	,010	,081
Oper/ Artif	,023	,088	,001	,001	,007	,003
Trab N Qualif	,029	,082	,060	,001	,373	,004
Desconhecido	,003	,108	,000	,014	,000	,033
Active Total	,111	,556	,112	,091		

Tabela 10- Variável "Categoria profissional do pai"

Points:Coordinates

Category	Frequency	Centroid Coordinates	
		Dimension	
		1	2
<1	36	-1,146	,709
1_2	28	-,232	-,723
3_4	44	,496	-,400
5_6	21	,594	,015
>6	13	1,034	,925

Category	Mass	Inertia	Contribution			
			Of Point to Inertia of Dimension		Of Dimension to Inertia of Point	
			1	2	1	2
<1	,028	,083	,073	,055	,446	,171
1_2	,022	,089	,002	,044	,013	,128
3_4	,034	,077	,017	,021	,110	,072
5_6	,016	,095	,011	,000	,061	,000
>6	,010	,101	,021	,034	,108	,086
Active	,111	,444	,125	,153		
Total						

Tabela 11- Variável "Horas de estudo"

ANEXO VI – Análise de *Clusters* (a partir dos resultados da ACM)

Tabela 12- Aglomeração (Between Groups)

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	74	138	,000	0	0	4
2	44	105	,000	0	0	13
3	62	96	,000	0	0	60
4	36	74	,000	0	1	48
5	8	47	,000	0	0	22
6	14	39	,000	0	0	63
7	104	121	,000	0	0	60
8	35	113	,001	0	0	20
9	82	87	,001	0	0	35
10	9	12	,001	0	0	75
11	7	116	,001	0	0	13
12	30	71	,001	0	0	23
13	7	44	,002	11	2	35
14	10	90	,002	0	0	64
15	127	142	,003	0	0	72
16	16	93	,003	0	0	68
17	50	130	,003	0	0	20
18	17	102	,003	0	0	47
19	5	51	,004	0	0	26
20	35	50	,004	8	17	62
21	69	125	,005	0	0	63
22	8	110	,005	5	0	73
23	30	133	,006	12	0	43
24	22	29	,006	0	0	67
25	78	140	,007	0	0	39
26	5	120	,007	19	0	29
27	20	112	,007	0	0	34
28	34	60	,008	0	0	51
29	5	88	,008	26	0	54
30	40	95	,008	0	0	61
31	18	21	,008	0	0	74
32	70	117	,008	0	0	69
33	1	79	,009	0	0	58
34	20	37	,011	27	0	55
35	7	82	,011	13	9	70

36	6	115	,011	0	0	42
37	41	109	,012	0	0	106
38	38	101	,012	0	0	100
39	56	78	,013	0	25	80
40	32	67	,013	0	0	85
41	2	26	,014	0	0	73
42	6	103	,015	36	0	92
43	30	46	,015	23	0	61
44	43	84	,016	0	0	118
45	11	124	,016	0	0	51
46	33	61	,016	0	0	76
47	17	83	,017	18	0	52
48	36	100	,017	4	0	80
49	4	76	,018	0	0	118
50	72	131	,019	0	0	58
51	11	34	,019	45	28	87
52	17	64	,020	47	0	83
53	57	98	,021	0	0	68
54	5	86	,022	29	0	82
55	20	137	,022	34	0	79
56	114	135	,024	0	0	108
57	55	106	,025	0	0	83
58	1	72	,025	33	50	90
59	52	73	,025	0	0	95
60	62	104	,026	3	7	90
61	30	40	,027	43	30	91
62	35	65	,027	20	0	87
63	14	69	,027	6	21	70
64	10	13	,028	14	0	88
65	23	85	,028	0	0	77
66	27	111	,029	0	0	102
67	22	68	,029	24	0	98
68	16	57	,030	16	53	97
69	63	70	,032	0	32	82
70	7	14	,036	35	63	81
71	25	97	,037	0	0	86
72	107	127	,039	0	15	101
73	2	8	,042	41	22	97
74	18	128	,047	31	0	100
75	9	66	,049	10	0	84
76	33	108	,049	46	0	95
77	23	92	,049	65	0	94
78	75	123	,050	0	0	102

79	20	99	,050	55	0	96
80	36	56	,051	48	39	103
81	7	48	,052	70	0	92
82	5	63	,052	54	69	109
83	17	55	,056	52	57	89
84	9	42	,059	75	0	108
85	32	126	,060	40	0	99
86	25	119	,064	71	0	128
87	11	35	,068	51	62	109
88	10	122	,069	64	0	98
89	17	24	,075	83	0	113
90	1	62	,075	58	60	99
91	3	30	,085	0	61	112
92	6	7	,086	42	81	115
93	139	141	,087	0	0	126
94	23	49	,090	77	0	110
95	33	52	,094	76	59	123
96	20	28	,102	79	0	119
97	2	16	,103	73	68	112
98	10	22	,106	88	67	114
99	1	32	,111	90	85	120
100	18	38	,111	74	38	111
101	107	136	,119	72	0	114
102	27	75	,123	66	78	120
103	36	134	,133	80	0	116
104	80	118	,136	0	0	141
105	31	58	,136	0	0	121
106	41	53	,137	37	0	116
107	45	54	,154	0	0	125
108	9	114	,175	84	56	111
109	5	11	,191	82	87	115
110	23	94	,197	94	0	124
111	9	18	,213	108	100	124
112	2	3	,217	97	91	119
113	15	17	,236	0	89	127
114	10	107	,255	98	101	128
115	5	6	,262	109	92	127
116	36	41	,268	103	106	132
117	59	132	,289	0	0	125
118	4	43	,290	49	44	123
119	2	20	,311	112	96	133
120	1	27	,391	99	102	129
121	31	77	,400	105	0	132

122	19	81	,403	0	0	134
123	4	33	,428	118	95	126
124	9	23	,436	111	110	133
125	45	59	,488	107	117	131
126	4	139	,496	123	93	138
127	5	15	,503	115	113	129
128	10	25	,526	114	86	136
129	1	5	,643	120	127	136
130	89	91	,695	0	0	134
131	45	129	,742	125	0	135
132	31	36	,953	121	116	137
133	2	9	1,038	119	124	137
134	19	89	1,257	122	130	135
135	19	45	1,768	134	131	138
136	1	10	2,047	129	128	139
137	2	31	2,493	133	132	139
138	4	19	2,550	126	135	140
139	1	2	4,167	136	137	140
140	1	4	7,628	139	138	141
141	1	80	10,607	140	104	0

Tabela 13- Aglomeração (Ward)

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	74	138	,000	0	0	4
2	44	105	,000	0	0	15
3	62	96	,000	0	0	84
4	36	74	,000	0	1	52
5	8	47	,000	0	0	23
6	14	39	,000	0	0	82
7	104	121	,000	0	0	66
8	35	113	,001	0	0	24
9	82	87	,001	0	0	55
10	9	12	,002	0	0	76
11	7	116	,002	0	0	15
12	30	71	,003	0	0	39
13	10	90	,004	0	0	72
14	127	142	,005	0	0	71
15	7	44	,006	11	2	55
16	16	93	,008	0	0	69

17	50	130	,009	0	0	24
18	17	102	,011	0	0	56
19	5	51	,013	0	0	31
20	69	125	,015	0	0	67
21	40	133	,018	0	0	39
22	22	29	,021	0	0	103
23	8	110	,024	5	0	81
24	35	50	,028	8	17	65
25	78	140	,031	0	0	40
26	20	112	,035	0	0	35
27	34	60	,039	0	0	54
28	88	120	,042	0	0	31
29	18	21	,047	0	0	74
30	70	117	,051	0	0	64
31	5	88	,055	19	28	61
32	1	79	,060	0	0	62
33	6	115	,066	0	0	44
34	41	109	,071	0	0	92
35	20	37	,077	26	0	58
36	38	101	,083	0	0	95
37	32	67	,090	0	0	98
38	2	26	,097	0	0	81
39	30	40	,104	12	21	51
40	56	78	,111	0	25	92
41	43	84	,119	0	0	114
42	11	124	,127	0	0	54
43	33	61	,136	0	0	75
44	6	103	,144	33	0	106
45	4	76	,153	0	0	112
46	64	83	,162	0	0	56
47	72	131	,171	0	0	62
48	57	98	,181	0	0	69
49	114	135	,193	0	0	100
50	55	106	,206	0	0	88
51	30	46	,218	39	0	68
52	36	100	,231	4	0	101
53	52	73	,244	0	0	99
54	11	34	,257	42	27	102
55	7	82	,270	15	9	82
56	17	64	,284	18	46	88
57	23	85	,298	0	0	73
58	20	137	,312	35	0	77
59	13	68	,326	0	0	72

60	27	111	,340	0	0	83
61	5	86	,356	31	0	116
62	1	72	,373	32	47	98
63	25	97	,392	0	0	79
64	63	70	,412	0	30	84
65	35	65	,433	24	0	102
66	104	126	,454	7	0	97
67	48	69	,476	0	20	85
68	30	95	,501	51	0	89
69	16	57	,525	16	48	107
70	75	123	,550	0	0	108
71	107	127	,575	0	14	96
72	10	13	,602	13	59	86
73	23	92	,630	57	0	87
74	18	128	,659	29	0	115
75	33	108	,689	43	0	99
76	9	66	,721	10	0	78
77	20	99	,757	58	0	93
78	9	42	,793	76	0	118
79	25	119	,829	63	0	127
80	139	141	,872	0	0	114
81	2	8	,918	38	23	107
82	7	14	,967	55	6	85
83	24	27	1,016	0	60	108
84	62	63	1,068	3	64	97
85	7	48	1,122	82	67	106
86	10	122	1,178	72	0	103
87	23	49	1,235	73	0	111
88	17	55	1,294	56	50	122
89	3	30	1,360	0	68	121
90	80	118	1,428	0	0	127
91	31	58	1,496	0	0	113
92	41	56	1,564	34	40	110
93	20	28	1,640	77	0	126
94	45	54	1,716	0	0	117
95	38	94	1,797	36	0	111
96	107	136	1,879	71	0	119
97	62	104	1,965	84	66	116
98	1	32	2,054	62	37	124
99	33	52	2,144	75	53	123
100	53	114	2,235	0	49	115
101	36	134	2,346	52	0	110
102	11	35	2,471	54	65	120

103	10	22	2,600	86	22	119
104	15	129	2,730	0	0	122
105	59	132	2,874	0	0	117
106	6	7	3,036	44	85	133
107	2	16	3,220	81	69	121
108	24	75	3,414	83	70	120
109	19	81	3,615	0	0	125
110	36	41	3,822	101	92	131
111	23	38	4,055	87	95	129
112	4	89	4,293	45	0	128
113	31	77	4,537	91	0	131
114	43	139	4,809	41	80	123
115	18	53	5,085	74	100	118
116	5	62	5,390	61	97	124
117	45	59	5,767	94	105	128
118	9	18	6,219	78	115	129
119	10	107	6,716	103	96	134
120	11	24	7,286	102	108	130
121	2	3	7,955	107	89	126
122	15	17	8,703	104	88	130
123	33	43	9,479	99	114	136
124	1	5	10,294	98	116	135
125	19	91	11,151	109	0	132
126	2	20	12,123	121	93	137
127	25	80	13,132	79	90	134
128	4	45	14,235	112	117	132
129	9	23	15,559	118	111	137
130	11	15	17,050	120	122	133
131	31	36	18,793	113	110	138
132	4	19	21,151	128	125	136
133	6	11	24,346	106	130	135
134	10	25	27,919	119	127	139
135	1	6	32,545	124	133	139
136	4	33	39,958	132	123	140
137	2	9	47,641	126	129	138
138	2	31	67,530	137	131	140
139	1	10	94,701	135	134	141
140	2	4	187,624	138	136	141
141	1	2	284,000	139	140	0

Iteration History[a]				
Iteration	Change in Cluster Centers			
	1	2	3	4
1	,729	,423	,687	,594
2	,517	,074	,177	,151
3	,208	,070	,018	,000
4	,000	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 4. The minimum distance between initial centers is 2,590.

Tabela 14- Variação do centro dos *Clusters* em cada interação

Final Cluster Centers				
	Cluster			
	1	2	3	4
Object scores dimension 1	1,22	,65	-,75	-1,36
Object scores dimension 2	1,23	-,23	-,87	1,58

Tabela 15- Média das variáveis em cada um dos *Clusters*

Distances between Final Cluster Centers				
Cluster	1	2	3	4
1		1,569	2,875	2,602
2	1,569		1,533	2,707
3	2,875	1,533		2,529
4	2,602	2,707	2,529	

Tabela 16- Distâncias entre os centroide dos *Clusters*

Number of Cases in each Cluster		
Cluster	1	17,000
	2	61,000
	3	44,000
	4	20,000
Valid		142,000
Missing		,000

Tabela 17- Distribuição por *Cluster*

		TwoStep Cluster Number									
		Outlier Cluster		1		2		3		4	
		N	N %	N	N %	N	N %	N	N %		N %
Nível a Matemática	2	0	,0%	38	77,6%	11	55,0%	5	8,9%	0	,0%
	3	0	,0%	9	18,4%	6	30,0%	28	50,0%	6	35,3%
	4	0	,0%	1	2,0%	1	5,0%	11	19,6%	8	47,1%
	5	0	,0%	1	2,0%	2	10,0%	12	21,4%	3	17,6%
	Total	0	,0%	49	100,0%	20	100,0%	56	100,0%	17	100,0%
Realiza TPC de Matemática	Nunca	0	,0%	1	2,0%	17	85,0%	0	,0%	0	,0%
	Às vezes	0	,0%	43	87,8%	3	15,0%	7	12,5%	0	,0%
	Sempre	0	,0%	5	10,2%	0	,0%	49	87,5%	17	100,0%
	Total	0	,0%	49	100,0%	20	100,0%	56	100,0%	17	100,0%
Frequência do Estudo de Matemática	Nunca	0	,0%	2	4,1%	14	70,0%	1	1,8%	0	,0%
	Raramente	0	,0%	21	42,9%	6	30,0%	5	8,9%	1	5,9%
	1_2 vezes	0	,0%	18	36,7%	0	,0%	26	46,4%	0	,0%
	3_5 vezes	0	,0%	8	16,3%	0	,0%	21	37,5%	5	29,4%
	Td_Dias	0	,0%	0	,0%	0	,0%	3	5,4%	11	64,7%
	Total	0	,0%	49	100,0%	20	100,0%	56	100,0%	17	100,0%
Utilidade do TPC	Nada úteis	0	,0%	6	12,2%	18	90,0%	0	,0%	0	,0%
	Úteis	0	,0%	39	79,6%	1	5,0%	41	73,2%	6	35,3%
	Muito úteis	0	,0%	4	8,2%	1	5,0%	15	26,8%	11	64,7%
	Total	0	,0%	49	100,0%	20	100,0%	56	100,0%	17	100,0%
Horas de Estudo de Matemática	<1	0	,0%	14	28,6%	18	90,0%	3	5,4%	1	5,9%
	1_2	0	,0%	16	32,7%	2	10,0%	10	17,9%	0	,0%
	3_4	0	,0%	14	28,6%	0	,0%	28	50,0%	2	11,8%
	5_6	0	,0%	5	10,2%	0	,0%	10	17,9%	6	35,3%
	>6	0	,0%	0	,0%	0	,0%	5	8,9%	8	47,1%
	Total	0	,0%	49	100,0%	20	100,0%	56	100,0%	17	100,0%
Importância da formação	Pequena	0	,0%	10	20,4%	8	40,0%	3	5,4%	0	,0%
	Alguma	0	,0%	25	51,0%	5	25,0%	12	21,4%	1	5,9%
	Grande	0	,0%	14	28,6%	7	35,0%	41	73,2%	16	94,1%
	Total	0	,0%	49	100,0%	20	100,0%	56	100,0%	17	100,0%
Categoria Profissional da Mãe	Dir/ Qd Sup	0	,0%	0	,0%	1	5,0%	1	1,8%	5	29,4%
	Espec/ Tec	0	,0%	4	8,2%	4	20,0%	27	48,2%	4	23,5%
	Adm/ Serv/ Vend	0	,0%	10	20,4%	5	25,0%	17	30,4%	5	29,4%
	Oper/ Artif	0	,0%	7	14,3%	3	15,0%	5	8,9%	1	5,9%
	Trab N Qualif	0	,0%	27	55,1%	7	35,0%	6	10,7%	0	,0%
	Desconhecida	0	,0%	1	2,0%	0	,0%	0	,0%	2	11,8%

	Total	0	,0%	49	100,0%	20	100,0%	56	100,0%	17	100,0%
Categoria Profissional da pai	Dir/ Qd Sup	0	,0%	0	,0%	1	5,0%	3	5,4%	5	29,4%
	Espec/ tec	0	,0%	2	4,1%	2	10,0%	21	37,5%	9	52,9%
	Adm/ Serv/vend	0	,0%	13	26,5%	1	5,0%	13	23,2%	1	5,9%
	Oper/ Artif	0	,0%	12	24,5%	5	25,0%	12	21,4%	1	5,9%
	Trab N Qualif	0	,0%	19	38,8%	11	55,0%	6	10,7%	1	5,9%
	Desconhecido	0	,0%	3	6,1%	0	,0%	1	1,8%	0	,0%
	Total	0	,0%	49	100,0%	20	100,0%	56	100,0%	17	100,0%
Aluno empenhado	Não	0	,0%	2	4,1%	3	15,0%	2	3,6%	0	,0%
	Às vezes	0	,0%	38	77,6%	10	50,0%	11	19,6%	2	11,8%
	Sim	0	,0%	9	18,4%	7	35,0%	43	76,8%	15	88,2%
	Total	0	,0%	49	100,0%	20	100,0%	56	100,0%	17	100,0%

Tabela 18- Caraterização dos *Clusters*

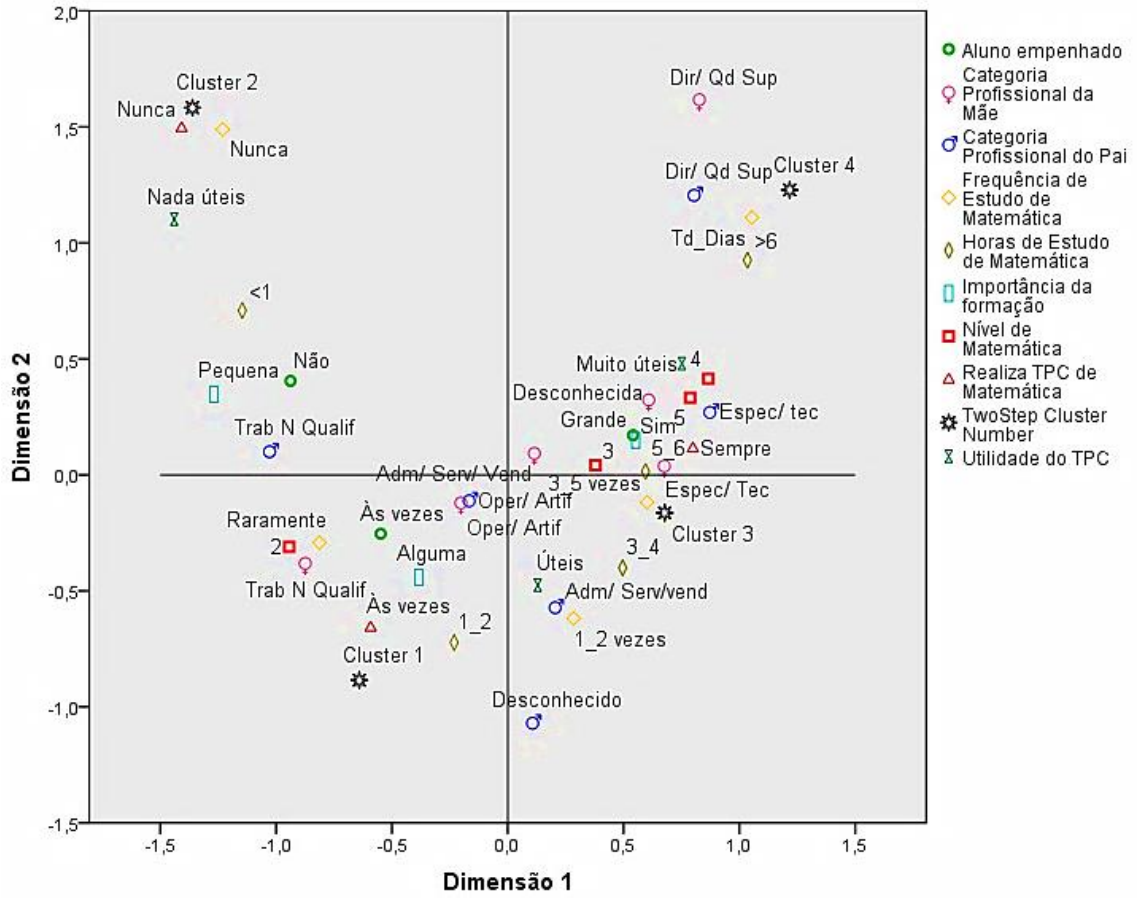


Gráfico 1- Disposição dos *Clusters* (*TwoStep*) no espaço de análise