Universidade do Minho
Escola de Ciências

Ana Rita Vieira da Silva

Análise Estatística Multivariada no estudo da relação de variáveis de um solo residual granítico com a cultura da vinha. Caso da casta Vinhão.

Outubro de 2011

Análise Estatística Multivariada no estudo da relação de variáveis de um solo residual granítico com a cultura da vinha. Caso da casta Vinhão.
Ana Rita Vieira da Silva
UMinho | 2011

Universidade do Minho
Escola de Ciências

Ana Rita Vieira da Silva

Análise Estatística Multivariada no estudo da relação de variáveis de um solo residual granítico com a cultura da vinha. Caso da casta Vinhão.

Relatório de Mestrado
Mestrado em Estatística de Sistemas
Área de Especilaização em Engenharia e Estatística

Trabalho efetuado sob a orientação da
Professora Doutora Ana Cristina Braga

Outubro de 2011

AGRADECIMENTOS

No decorrer deste trabalho tive a oportunidade de contactar com um conjunto de pessoas que me apoiaram e ajudaram, e sem as quais não seria possível a realização deste trabalho. A todos eles deixo uma mensagem de agradecimento:

À Professora Doutora Ana Cristina Braga, minha orientadora científica, pelo apoio, carinho, amizade, compreensão e dedicação que sempre demonstrou. Ao longo deste percurso sempre me transmitiu segurança e confiança que me permitiram seguir em frente.

Ao meu orientador da empresa, Dr. Jorge Oliveira, à Dr.ª Isabel Araújo e à Dr.ª Teresa Mota, pela oportunidade de realização deste estágio, pela forma como me acolheram, pela amizade e carinho que sempre demonstraram e pelo ambiente e experiências que me proporcionaram.

Ao Waylson Quartezani, aluno de doutoramento do Instituto Superior Técnico (IST) de Lisboa, pelas longas conversas e esclarecimentos sobre a possibilidade de implementação da geoestatística neste trabalho. Agradeço também o Professor Doutor Amílcar Soares do IST e à Professora Doutora Cecília Azevedo do Departamento de Matemática da Universidade do Minho pela possibilidade deste contacto.

A todos os que de alguma forma contribuíram para a concretização deste trabalho agradeço com amizade.

Agradeço ainda à minha família e amigos, em especial ao meu marido Alberto, pelo apoio e compreensão que sempre me dispensaram.

RESUMO

Com uma extensa tradição na produção e exploração de vinho, Portugal tem neste elemento um fator cultural basilar da sua identidade nacional que é reconhecido internacionalmente. De modo a proteger e melhorar este importante valor é fundamental melhorar a rentabilidade dos sistemas produtivos e minimizar a degradação do recurso solo para garantir a sustentabilidade dos sistemas de produção vitivinícola. Nesta medida, a informação detalhada sobre os solos de aptidão vitícola deve ser um instrumento de gestão essencial para a rentabilização dos investimentos e aumento da qualidade do produto final.

O trabalho desenvolvido no âmbito do Projeto Agrocontrol teve como principal objetivo o estudo e identificação das características físico-químicas do solo que influenciam o crescimento das videiras, a qualidade das uvas e por conseguinte a qualidade dos vinhos. Com este conhecimento pretende-se otimizar a produção, de forma a conseguir um produto com características mais uniformes, em consonância com adequadas tecnologias de vinificação e assim fazer face a um mercado cada vez mais competitivo.

A parcela em estudo situa-se na Estação Vitivinícola Amândio Galhano (EVAG), no concelho dos Arcos de Valdevez pertencente à Comissão de Viticultura da Região dos *Vinhos Verdes* (CVRVV). A vinha em estudo compreende apenas a casta tinta Vinhão.

Recorrendo aos dados obtidos em campo e no laboratório e fazendo uso de técnicas de Estatística Multivariada e da determinação de correlações entre as variáveis do solo e aquelas que determinaram a qualidade do vinho, foi possível identificar quais as características do solo que interferiram no desenvolvimento das videiras e na qualidade das uvas e respetivos vinhos em 2010. O estudo efetuado pretende contribuir para a fonte de informação do sistema de produção, a qual permitirá uma tomada de decisão mais robusta e orientada, efetuando-se deste modo, uma agricultura de forma sustentada.

ABSTRACT

With a long tradition in wine production and exploration, Portugal has in this element an important cultural factor that is one of the foundations of its national identity and international recognition. In order to protect and improve this important value, it is essential to increase the efficiency of production systems and minimize the degradation of natural soil recourses, preserving the sustainability of the wine production. In this context, the detailed information about the properties of soils that are suitable for wine production should constitute an important management instrument, towards the optimization of investments and the improvement of the quality of the final product.

As part of Agrocontrol Project, the main objective of this research is to study and identify the physical and chemical properties of the soil that influence the vineyard growth, the quality of grapes and therefore the quality of the wine. This information is used to optimize the wine production, achieving a more homogenous product and ensuring the proper production techniques required by the wine highly competitive market.

The parcel of land studied is located in Estação Vitivinícola Amândio Galhano (EVAG), in the municipality of Arcos de Valdevez, belonging to the commission of wine production of the "*Vinhos Verdes*" region (Comissão de Viticultura da Região dos *Vinhos Verdes* - CVRVV). The vineyard is composed of "Vinhão" grapes exclusively.

Using the data collected "in situ" and in the lab, it was applied multivariate statistic techniques and determined the correlations between the variables of the soil and the variables characterizing the wine quality. The results obtained allowed the identification of the soil characteristics that influence the development of vineyard, the quality of grapes and consequently the quality of wine in 2010. This study intends to contribute to the improvement of the information resources used by the production systems, leading to a more robust and objective decision process and making wine production more sustainable.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE VARIÁVEIS

Acidez_T: Acidez total (mosto)

Acido_malico: Ácido málico (mosto)

Acido_tartárico: Ácido tartárico (mosto)

Açucares: Açucares (mosto)

Avaliacao_Global: Avaliação global (vinho)

AzT: Azoto Total (solo)

B: Boro (solo)

Ca: Cálcio assimilável (solo)

Cd: Cádmio (solo)

cor_v: Cor (vinho)

Cr: Crómio (solo)

CTC: Capacidade de troca catiónica (solo)

DA: Densidade aparente (solo)

FF: Fração fina (solo)

FG: Fração grosseira (solo)

FG1: Família de compostos em C_6 do aroma das uvas na forma glicosilada

FG2: Família de álcoois do aroma das uvas na forma glicosilada

FG3: Família de álcoois monoterpénicos do aroma das uvas na forma glicosilada

FG4: Família de óxidos e dióis monoterpénicos do aroma das uvas na forma glicosilada

FG5: Família de norisoprenóides em C_{13} do aroma das uvas na forma glicosilada

FG6: Família de fenóis voláteis do aroma das uvas na forma glicosilada

FG7: Família de compostos carbonilados do aroma das uvas na forma glicosilada

FL1: Família de compostos em C_6 do aroma das uvas na forma livre

FL2: Família de álcoois do aroma das uvas na forma livre

FL3: Família de álcoois monoterpénicos do aroma das uvas na forma livre

FL4: Família de fenóis voláteis do aroma das uvas na forma livre

FL5: Família de compostos carbonilados do aroma das uvas na forma livre

intensidade_g: intensidade gustativa (vinho)

intensidade_o: intensidade olfativa (vinho)

K_2O: Potássio assimilável (solo)

limpidez_g: limpidez gustativa (vinho)

limpidez_o: limpidez olfativa (vinho)

limpidez_v: limpidez visual (vinho)

Mg: Magnésio assimilável (solo)

MO: Matéria orgânica (solo)

N: Nitratos (solo)

Ncachos: Número de cachos por videira

Ni: Níquel (solo)

Nota_Final: Nota final (vinho)

Nvaras: Número de varas por videira

P_2O_5: Fósforo assimilável (solo)

Pcacho_kg: Peso médio do cacho por videira

persistência_g: persistência gustativa (vinho)

pH: pH (solo)

pH_mosto: pH (mosto)

Pvara_g: Peso médio da vara por videira

Pvaras_kg: Peso das varas por videira

qualidade_g: qualidade gustativa (vinho)

qualidade_o: qualidade olfativa (vinho)

TAP: Teor de álcool provável (mosto)

Uvas_kg_vid: Peso de uvas por videira

ABREVIATURAS

AC: Análise de Clusters

ACP: Análise de Componentes Principais

ANOVA: Análise da Variância

CP: Componente Principal

CVRVV: Comissão de Viticultura da Região dos *Vinhos Verdes*

DOC: Denominação de Origem Controlada

EVAG: Estação Vitivinícola Amândio Galhano

FEDER: Fundo Europeu de Desenvolvimento Regional

GC-MS: *Gas-chromatography-mass spectrometry*

I&DT: Investigação e Desenvolvimento Tecnológico

MPB: Modo de Produção Biológico

QREN: Quadro de Referência Estratégico Nacional

CAPÍTULO 1: INTRODUÇÃO

No primeiro capítulo é apresentado um enquadramento do tema deste projeto de investigação descrevendo a importância da influência do solo e suas características na qualidade dos vinhos. É também exposta uma breve descrição do projeto Agrocontrol e da empresa onde decorreu o estágio. Finalmente apresentam-se os objetivos orientadores do estudo desenvolvido e a estrutura global do relatório.

1.1. Enquadramento

1.1.1. Influência do solo na qualidade do vinho

A qualidade de um vinho depende de inúmeros fatores. Genericamente começa na planta (casta e porta-enxerto) a qual é influenciada pelos fatores ambientais (clima e solo) e pela tecnologia vitícola (sistemas de condução, fertilizações, entre outros). Por fim, a tecnologia enológica irá dar origem ao produto final, o vinho, que manifestará o efeito de todos estes fatores (Araújo, 2004).

Os fatores ambientais, nomeadamente o clima e o solo, intervêm de uma forma inequívoca na qualidade das uvas e consequentemente na do vinho. Estes fatores abióticos, além de poderem ser limitantes ou mesmo impeditivos do estabelecimento da cultura, são responsáveis por uma grande diversidade de situações (Clímaco, 1997).

É fundamental que as variedades de uma dada região estejam perfeitamente adaptadas às condições edafoclimáticas da sua região, permitindo na grande maioria dos anos, condições de maturação perfeitas a fim de conseguir regularmente uma produção de qualidade (Clímaco e Castro, 1991).

A ligação da geologia à viticultura surge da aplicação da cartografia geológica e de solos, climatologia, hidrologia e medição de parâmetros pontuais e globais do solo. Tem como objetivo identificar e estudar múltiplas variáveis que determinam o comportamento quer físico quer químico dos solos e que influenciam o crescimento da planta e a qualidade final do fruto. O conhecimento de algumas destas variáveis e a sua influência no solo, bem como o efeito do tipo e estado dos solos sobre as plantas, pode ser utilizado para implementar medidas que permitam otimizar a produção, de forma a conseguir um produto com características mais uniformes e que influenciem significativamente o processo de vinificação.

Existe uma opinião generalizada que a escolha de um solo apropriado, a preparação do terreno e todas as melhorias possíveis de serem realizadas, são os fatores chave para o sucesso da viticultura e o primeiro passo para a obtenção de uvas com qualidade superior (Layon, 2004).

O solo é um elemento indispensável para a cultura da vinha, sendo a sua influência complexa. Esta depende da estrutura física, da composição química, da água e da temperatura. Por outro lado, quer a estrutura física quer a composição química são dependentes da origem geológica do solo (Galet, 1993). Em casos extremos, o solo pode ser impeditivo à cultura da vinha ao não permitir a adequada penetração das raízes da planta. A vocação vitícola de uma dada região é determinada também pelo solo.

A natureza do solo e consequentemente a sua textura e estrutura condicionam a razão entre as folhas e as raízes da planta (Leme e Malheiro, 1998; Tomasi *et al.,* 1998). Em solos pobres em água, a planta desenvolve as raízes mais importantes em detrimento das folhas influenciando deste modo o comportamento da videira no que diz respeito ao vigor e à produção.

A implementação de técnicas culturais que fomentem a atividade biológica dos solos (mobilização mínima dos solos com a racional gestão de culturas de cobertura) tem revelado uma extrema importância para a preservação da estrutura do solo. Desta forma, as propriedades físicas do solo, tais como infiltração, trocas gasosas, dureza do solo, manter--se-ão em proporções ótimas para a produção agrícola (Layon, 2004).

Os elementos que a vinha necessita dividem-se em três categorias: elementos principais (N, P, K), elementos secundários (Ca, Mg, S) e micronutrientes ou oligoelementos (Fe, Cu, Zn, Mn, B, Mb, Cl). Estes últimos são extremamente importantes, embora sejam absorvidos pela vinha em pequeníssimas quantidades. A sua carência provoca doenças e causam fitotoxicidade quando absorvidos em excesso (Araújo, 2004).

A avaliação das necessidades da vinha é feita através de análise da terra complementada com análise foliar, uma vez que a primeira pode não ser suficiente para indicar se o teor de micronutrientes é suficiente ou está em excesso. Por outro lado, só uma parte dos elementos são assimiláveis pelas raízes, o resto é retido pelo poder absorvente do solo (Araújo, 2004).

O padrão de variabilidade económica da vinha tem-se verificado relativamente estável ao longo do tempo. Este facto pode ser explicado pelo caráter perene das videiras. Por outro lado, a variabilidade a nível do vigor da videira ou outro índice de vegetação é normalmente atribuída à variação de água disponível. Estando esta última intimamente

ligada à profundidade do solo, sendo este um fator com elevada estabilidade temporal (Layon, 2004).

1.1.2. O Projeto Agrocontrol

O Projeto Agrocontrol cofinanciado pelo "ON.2 – O Novo Norte" e pelo Quadro de Referência Estratégico Nacional (QREN) através do Fundo Europeu de Desenvolvimento Regional (FEDER), surgiu da necessidade de se identificar e estudar múltiplas variáveis que determinam o comportamento físico e químico dos solos, que por sua vez influenciam o desenvolvimento da videira, a qualidade das uvas e por conseguinte a qualidade dos vinhos.

Este projeto consolida a estratégia empresarial e científica das empresas Sinergeo e Vinalia no desenvolvimento de processos e metodologias inovadoras na sua área de atuação. No sentido de reforçar as capacidades de Investigação e Desenvolvimento Tecnológico (I&DT), o projeto é apoiado pela Universidade do Minho, entidade que através do estatuto de *spinoff* já se encontra ligada a estas empresas, e pela Estação Vitivinícola Amândio Galhano (EVAG), pertencente à Comissão de Viticultura da Região dos *Vinhos Verdes* (CVRVV).

A parcela em estudo, no âmbito deste projeto, situa-se na EVAG, na Quinta Campos de Lima, no concelho dos Arcos de Valdevez, e está inserida numa região apta à produção do vinho com Denominação de Origem Controlada (DOC) - *Vinho Verde*. A vinha em estudo compreende apenas a casta Vinhão, encontrando-se explorada em Modo de Produção Biológico (MPB), certificada pela EcoCert Portugal.

1.1.3. A empresa

O estágio foi desenvolvido na empresa Sinergeo – Soluções Aplicadas em Geologia, Hidrogeologia e Ambiente Lda. Esta empresa foi fundada em 2006 por profissionais licenciados em Geologia. Dedica-se desde então à prestação de serviços, consultoria e execução de projetos nas áreas da geologia, hidrogeologia, geofísica e geotecnia.

Esta empresa aposta na melhoria das capacidades técnicas e inovadoras da equipa e na aquisição dos mais modernos equipamentos e *software* como forma de aumentar a sua eficácia e eficiência e reforçar a imagem de profissionalismo e dinamismo. Tem como

missão o desenvolvimento e implementação de soluções de promoção e proteção dos recursos geológicos com vista à valorização do território.

A Figura 1 expressa o modo como desenvolvem as suas áreas de negócio.

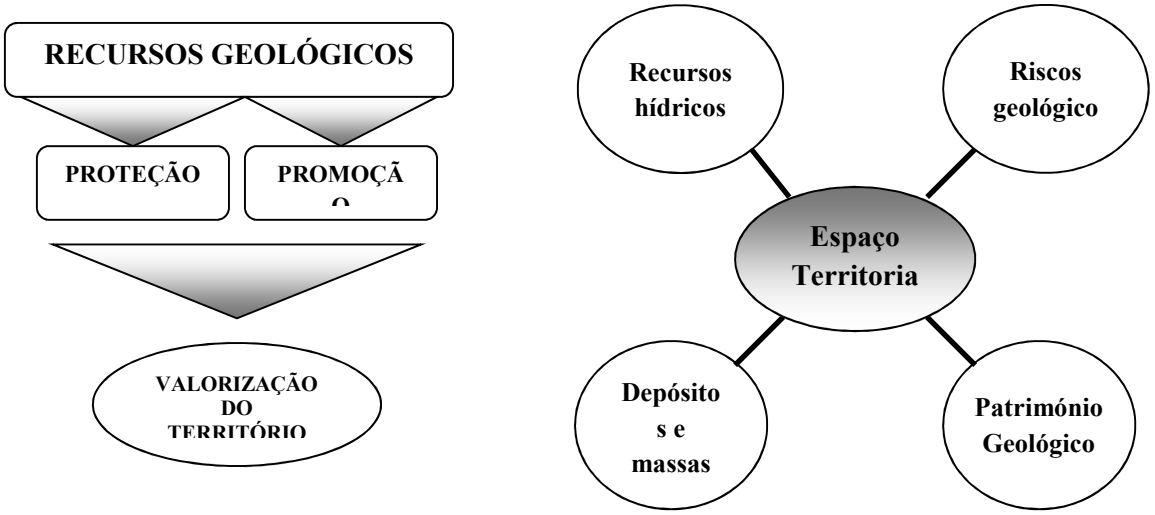


Figura 1 - Áreas de intervenção da Sinergeo

A Sinergeo pretende que a estratégia económica seja uma consequência da estratégia científica, ou seja, que a maior percentagem do negócio seja resultante de serviços gerados a partir das atividades de I&DT. É desta forma, com investigação, inovação e criatividade, que se destacam num mercado cada vez mais global e cada vez mais competitivo e criam valor para a organização, seus parceiros, colaboradores e sociedade em geral.

1.2. Objetivos

O objetivo geral desta investigação é identificar de entre as múltiplas variáveis que determinam o comportamento físico e químico dos solos, as que influenciam o desenvolvimento da videira, a qualidade final da uva e por conseguinte a qualidade dos vinhos, para o ano 2010, na parcela em estudo. Este objetivo concretizou-se através dos seguintes objetivos específicos:

- descrever o comportamento de cada uma das variáveis e detetar a possível existência de *outliers*;
- verificar se existe um pequeno número de variáveis que seja responsável por explicar uma proporção elevada da variação total associada ao conjunto original de dados do solo e reduzir a dimensionalidade do problema;

- detetar na parcela a existência de zonas que se distinguem ao nível das características do solo;
- identificar quais as variáveis do solo que influenciam a qualidade do vinho e quais as zonas da parcela que apresentam défice/excesso nas concentrações dessas variáveis para assim se aplicarem corretivos.

1.3. Estrutura do Relatório

Este relatório é constituído, para além desta introdução, por mais três capítulos.

No capítulo dois faz-se uma abordagem teórica das técnicas estatísticas selecionadas para o desenvolvimento e concretização dos objetivos definidos.

Assim, começa-se por definir os principais conceitos relacionados com a técnica de redução da dimensionalidade dos dados: Análise de Componentes Principais (ACP). É apresentada a definição desta técnica e sua abordagem do ponto de vista geométrico e algébrico, faz-se uma descrição das propriedades das Componentes Principais (CPs), apresentam-se a descrição e implicações da utilização da matriz de correlações na ACP, descrevem-se alguns critérios de seleção das CPs e sua interpretação e, por fim, definem-se *scores* das CPs e suas aplicações noutras técnicas de Estatística Multivariada.

Neste segundo capítulo é feita ainda uma abordagem teórica à técnica de agrupamento de dados: Análise de *Clusters* (AC). Apresenta-se uma breve descrição desta técnica, definem-se as principais medidas de proximidade e os principais métodos de agrupamento, realizando-se uma abordagem mais pormenorizada dos métodos hierárquicos aglomerativos por serem os aplicados no âmbito deste trabalho. Apresentam-se formas de validação dos resultados obtidos e ainda formas de interpretação dos *clusters*, nomeadamente a existência de diferenças significativas entre os *clusters* obtidos detetadas com a aplicação do teste não paramétrico de Kruskal-Wallis.

Finalmente é feita, neste capítulo, uma abordagem ao coeficiente de correlação de Spearman e seu teste de significância para avaliação da existência e grau de associação entre as variáveis do solo e as variáveis que determinam a qualidade do vinho.

No capítulo três começa-se por apresentar uma breve descrição da parcela em estudo e dos dados em análise, de seguida aplicam-se todas as técnicas estatísticas descritas no capítulo 2 e faz-se a análise e discussão dos resultados obtidos.

Por fim, no quarto capítulo, são apresentadas as principais conclusões obtidas no estudo do capítulo anterior e ainda algumas sugestões para trabalhos futuros.

CAPÍTULO 2: ENQUADRAMENTO TEÓRICO

Neste capítulo é apresentado o enquadramento teórico das técnicas estatísticas implementadas neste trabalho. Faz-se uma abordagem detalhada das técnicas de Estatística Multivariada: análise de componentes principais e análise de *clusters*, e apresenta-se também uma descrição do coeficiente de correlação de Spearman e seu teste de significância.

2.1. Análise de Componentes Principais

2.1.1. Introdução

A Análise de Componentes Principais (ACP) é possivelmente uma das mais antigas e utilizadas técnicas de Estatística Multivariada. As primeiras descrições desta técnica foram apresentadas por Pearson (1901) e Hotelling (1933) e a revisão dos seus trabalhos pode ser encontrada na coleção de artigos e revistas Bryant e Atchley (1975), citado por Jolliffe (2002).

Apesar de apresentarem abordagens distintas, ambos chegam à derivação das componentes principais. Pearson fá-lo através de problemas de otimização geométrica, procurando encontrar retas e planos que melhor se ajustem a um conjunto de pontos num espaço *p*--dimensional e Hotelling apresenta uma abordagem mais algébrica, procurando o menor "conjunto fundamental de variáveis independentes" que determinem o conjunto das *p* variáveis originais e introduz o conceito de Componentes Principais (Jolliffe, 2002).

A ACP é um método estatístico multivariado que permite transformar um conjunto original de *p* variáveis correlacionadas num novo conjunto de *p* variáveis não correlacionadas, denominadas Componentes Principais (CPs). Estas novas variáveis independentes são combinações lineares das variáveis originais e são calculadas por ordem decrescente de importância, ou seja, a primeira explica o máximo possível da variância dos dados originais, a segunda o máximo possível da variância ainda não explicada e assim sucessivamente (Reis, 2001).

De acordo com Chatfield e Collins (1995), com a aplicação desta técnica espera-se que as primeiras m $(<< p)$ CPs expliquem grande parte da variabilidade dos dados iniciais. Se algumas das variáveis originais se encontram muito correlacionadas, elas estão efetivamente a dar a mesma informação e pode-se ter uma relação muito próxima da linear entre essas variáveis.

A transformação ocorrida é, na verdade, uma rotação no espaço *p*-dimensional. O espaço gerado pelas primeiras *m* componentes principais é, de facto, um subespaço vetorial *m*--dimensional do espaço *p*-dimensional original. Quando o valor *m* é pequeno, por exemplo dois, é possível uma representação gráfica direta dos *n* indivíduos que ajudará a interpretar as semelhanças entre eles. A ACP pode entender-se também como a procura do subespaço de melhor ajustamento (Villardón, 2011).

É de esperar então que poucas das primeiras CPs nos ajudem intuitivamente a perceber melhor os dados, e que sejam úteis em análises posteriores onde se possa operar com um menor número de variáveis. Na prática, nem sempre é fácil rotular as componentes principais, portanto a sua principal utilização reside na redução da dimensionalidade dos dados originais por forma a simplificar análises subsequentes. Por exemplo, a representação gráfica dos *scores* das duas primeiras CPs para cada indivíduo é uma forma eficaz de tentar encontrar agrupamentos nos dados e reduzir a dimensionalidade para dois (Chatfield e Collins, 1995).

2.1.2. Abordagem ao Problema

Dispõe-se de valores de *p* variáveis avaliadas em *n* indivíduos dispostos numa matriz **X** de dimensão $n \times p$:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

onde x_{ij} é o valor da *j*-ésima variável no indivíduo *i*.

O objetivo de uma ACP pode ser formulado do seguinte modo: aproximar a nuvem de *n* pontos em IR^p por outra nuvem de *n* pontos num subespaço de menor dimensão através de projeções ortogonais, da forma mais fidedigna possível.

Na consideração deste problema é vantajoso começar por centrar as colunas da matriz de dados. Subtraindo a cada coluna de **X** a sua média, estas colunas passam a ter média nula. A centragem das colunas da matriz de dados traduz-se numa translação da nuvem de *n* pontos de forma a fazer coincidir o seu centro de gravidade com a origem do sistema de eixos, permitindo que os subespaços a considerar para eventuais projeções estejam mais próximos da nuvem de pontos. Note-se ainda que qualquer subespaço tem que conter a origem – o elemento nulo do espaço (Cadima, 2010).

Segundo Peña (2002) pretende-se encontrar um subespaço de dimensão menor que p tal que ao projetar sobre ele os pontos, estes conservem a sua estrutura com a menor distorção possível.

Considere-se o caso particular bidimensional $(p=2)$ como o representado na Figura 2.

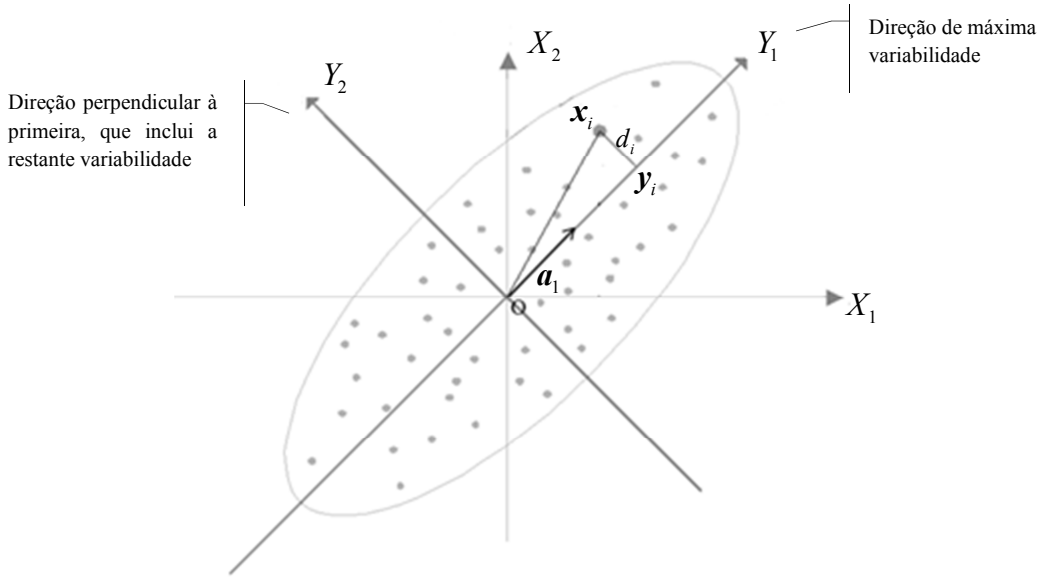


Figura 2 - Projeção dos pontos na 1ª CP (adaptada de Gonzálvez, 1999 e Villardón, 2011)

Pretende-se encontrar um subespaço de dimensão 1, uma reta, com direção definida por um vetor unitário $\boldsymbol{a}_1$ de tal forma que esta reta passe perto da maioria dos pontos. Para tal pretende-se encontrar a direção $\boldsymbol{a}_1$ de modo que as distâncias entre os pontos originais, $\boldsymbol{x}_i$, e as suas projeções na reta, $\boldsymbol{y}_i$, sejam tão pequenas quanto possível.

Designando por d_i a distância entre o ponto $\boldsymbol{x}_i$ e a sua projeção sobre a referida reta, pretende-se:

$$\min \sum_{i=1}^{n} d_i^2 \tag{1}$$

A Figura 2 mostra que ao projetar cada ponto sobre a reta se forma um triângulo retângulo cuja hipotenusa é a distância do ponto à origem e, pelo Teorema de Pitágoras, tem-se:

$$\|\boldsymbol{x}_i\|^2 = \|\boldsymbol{y}_i\|^2 + \|d_i\|^2 \tag{2}$$

Somando esta expressão para todos os pontos obtém-se:

$$\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T = \sum_{i=1}^{n} \boldsymbol{y}_i^2 + \sum_{i=1}^{n} d_i^2 \tag{3}$$

Como o primeiro membro é constante, minimizar $\sum_{i=1}^{n} d_i^2$, a soma das distâncias à reta de todos os pontos, equivale a maximizar $\sum_{i=1}^{n} \boldsymbol{y}_i^2$, a soma dos quadrados dos valores das projeções. Como as projeções $\boldsymbol{y}_i = \boldsymbol{a}_i^T \boldsymbol{x}_i$ são variáveis de média zero, maximizar a soma dos seus quadrados equivale a maximizar a sua variância. Obtém-se assim como critério encontrar a direção da projeção que maximiza a variância dos dados projetados.

Estendendo a procura a um subespaço de dimensão 2 que melhor resuma os dados, procura-se o plano que melhor se aproxima dos pontos. O problema consiste agora em encontrar uma nova direção definida por um vetor $\boldsymbol{a}_2$ que, sem perda de generalidade, seja ortogonal a $\boldsymbol{a}_1$ e que verifique a condição de que a projeção de cada ponto sobre este eixo maximize as distâncias entre os pontos projetados. Isto equivale a encontrar uma segunda variável $\boldsymbol{y}_2$, não correlacionada com $\boldsymbol{y}_1$ e que tenha variância máxima (Peña, 2002).

Se se considerar um qualquer subespaço *m*-dimensional, a aplicação sucessiva da ideia acima descrita permite obter as direções principais (González, 1999).

Este problema pode também ser abordado do ponto de vista geométrico com o mesmo resultado final. Observando a Figura 2 vemos ainda que os pontos se situam segundo uma elipse e podem ser descritos pela sua projeção na direção do eixo maior da elipse. Considerando dimensões maiores tem-se elipsoides e a melhor aproximação dos dados é a proporcionada pela sua projeção sobre o eixo maior da elipsoide (Peña, 2002).

2.1.3. Derivação das Componentes Principais

Considere-se a variável aleatória *p*-dimensional $\boldsymbol{X}^T = \begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}$ com média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$. A matriz de covariâncias $\boldsymbol{\Sigma}$ é semidefinida positiva, ou seja, $\forall \boldsymbol{a} \in R^p, \boldsymbol{a}^T \boldsymbol{\Sigma} \boldsymbol{a} \geq \boldsymbol{0}$ e é simétrica, estas propriedades em conjunto garantem a existência de valores próprios reais não negativos (Azevedo, 2010).

Pretende-se encontrar um novo conjunto de variáveis $Y_1, Y_2, ..., Y_p$ não correlacionadas e cujas variâncias decresçam da primeira até à última.

Cada nova variável Y_j, $j \in \{1,2,...,p\}$ é obtida como uma combinação linear das variáveis originais $X_1, X_2, ..., X_p$, ou seja:

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + ... + a_{pj}X_p = \boldsymbol{a}_j^T \boldsymbol{X} \tag{4}$$

em que $\boldsymbol{a}_j^T = \left[a_{1j}\ a_{2j}\ ...\ a_{pj} \right]$ é um vetor de constantes (Chatfield e Collins, 1995).

A derivação das várias CPs faz-se, segundo Chatfield e Collins (1995) e Jolliffe (2002), seguindo o mesmo procedimento, o qual se passa a descrever:

A primeira CP, Y_1, é a combinação linear das variáveis originais que tem variância máxima.

A média de Y_1 é:

$$\mathrm{E}\left(Y_1\right) = \mathrm{E}\left(\boldsymbol{a}_1^T \boldsymbol{X}\right) = \boldsymbol{a}_1^T \mathrm{E}\left(\boldsymbol{X}\right) = \boldsymbol{a}_1^T \boldsymbol{\mu} \tag{5}$$

e a sua variância é dada por:

$$\begin{aligned} Var\left(Y_1\right) &= \mathrm{E}\left[\left(Y_1 - \mathrm{E}\left(Y_1\right)\right)^2 \right] = \mathrm{E}\left[\left(\boldsymbol{a}_1^T \boldsymbol{X} - \boldsymbol{a}_1^T \boldsymbol{\mu}\right)^2 \right] = \mathrm{E}\left[\left(\boldsymbol{a}_1^T \left(\boldsymbol{X} - \boldsymbol{\mu}\right)\right)^2 \right] = \\ &= \mathrm{E}\left(\boldsymbol{a}_1^T \left(\boldsymbol{X} - \boldsymbol{\mu}\right)\left(\boldsymbol{X} - \boldsymbol{\mu}\right)^T \boldsymbol{a}_1\right) = \boldsymbol{a}_1^T \mathrm{E}\left[\left(\boldsymbol{X} - \boldsymbol{\mu}\right)\left(\boldsymbol{X} - \boldsymbol{\mu}\right)^T \right] \boldsymbol{a}_1 = \boldsymbol{a}_1^T \boldsymbol{\Sigma}\, \boldsymbol{a}_1 \end{aligned} \tag{6}$$

Para maximizar a variância de Y_1 basta aumentar os coeficientes do vetor $\boldsymbol{a}_1$ tanto quanto se queira. Para que esta maximização tenha solução única é necessário impor uma restrição de normalização ao vetor $\boldsymbol{a}_1$, e, sem perda de generalidade, impõe-se que $\boldsymbol{a}_1$ seja unitário, ou seja, $\boldsymbol{a}_1^T \boldsymbol{a}_1 = \sum_{i=1}^{p} a_{i1}^2 = 1$ (Jolliffe, 2002; Peña, 2002).

A primeira CP é então determinada pela escolha do vetor $\boldsymbol{a}_1$ que maximiza a função objetivo $Var\left(Y_1\right) = \boldsymbol{a}_1^T \boldsymbol{\Sigma}\, \boldsymbol{a}_1$ sujeita à restrição $\boldsymbol{a}_1^T \boldsymbol{a}_1 = 1$.

O procedimento usual para maximizar uma função de várias variáveis $f\left(x_1, x_2, ..., x_p\right)$ sujeita a uma restrição, neste caso do tipo $g\left(x_1, x_2, ..., x_p\right) = c$, é o método dos multiplicadores de Lagrange, no qual se começa por definir a função Lagrangeana $L\left(\mathbf{x}\right) = f\left(\mathbf{x}\right) + \lambda\left[g\left(\mathbf{x}\right) - c \right]$, em que λ é o multiplicador de Lagrange associado à restrição de igualdade.

Neste problema tem-se:

$$L(\boldsymbol{a}_1) = \boldsymbol{a}_1^T \boldsymbol{\Sigma} \boldsymbol{a}_1 + \lambda\left(\boldsymbol{a}_1^T \boldsymbol{a}_1 - 1\right) \tag{7}$$

Derivando *L* em ordem a $\boldsymbol{a}_1$ e igualando a zero, vem:

$$\frac{\partial L}{\partial \boldsymbol{a}_1} = \mathbf{0} \Leftrightarrow 2\boldsymbol{\Sigma}\boldsymbol{a}_1 - 2\lambda\boldsymbol{a}_1 = \mathbf{0} \Leftrightarrow \boldsymbol{a}_1\left(\boldsymbol{\Sigma} - \lambda\boldsymbol{I}\right) = \mathbf{0}, \tag{8}$$

onde $\boldsymbol{I}$ é a matriz identidade de dimensão $p \times p$.

Para que esta equação tenha solução para $\boldsymbol{a}_1$ que não seja a solução nula, é necessário que a matriz $\left(\boldsymbol{\Sigma} - \lambda\boldsymbol{I}\right)$ seja uma matriz singular, ou seja, $\left|\boldsymbol{\Sigma} - \lambda\boldsymbol{I}\right| = 0$, o que significa que λ é um valor próprio de $\boldsymbol{\Sigma}$. Mas $\boldsymbol{\Sigma}$, por ser semidefinida positiva, pode ter até *p* valores próprios não negativos $\lambda_1, \lambda_2, ..., \lambda_p$. Para se decidir qual dos *p* valores próprios deve ser escolhido para a determinação da primeira CP, deve-se ter em conta que se pretende maximizar a variância desta componente, ou seja, pretende-se maximizar:

$$Var\left(Y_1\right) = \boldsymbol{a}_1^T \boldsymbol{\Sigma} \boldsymbol{a}_1 \underbrace{=}_{\text{por (8)}} \boldsymbol{a}_1^T \lambda \boldsymbol{I}\, \boldsymbol{a}_1 = \lambda\, \boldsymbol{a}_1^T \boldsymbol{a}_1 = \lambda \tag{9}$$

pelo que λ deve ser o maior dos valores próprios, seja λ_1, e o seu vetor próprio associado, $\boldsymbol{a}_1$, define os coeficientes da primeira CP.

A segunda CP, $Y_2 = \boldsymbol{a}_2^T \boldsymbol{X}$, será derivada por um processo idêntico ao anterior, contudo além da restrição $\boldsymbol{a}_2^T \boldsymbol{a}_2 = 1$ deverá ser considerada a restrição de que Y_1 e Y_2 não poderão estar correlacionadas. Tem-se:

$$\begin{aligned} Cov\left(Y_2, Y_1\right) = Cov\left(\boldsymbol{a}_2^T \boldsymbol{X}, \boldsymbol{a}_1^T \boldsymbol{X}\right) = \mathrm{E}\left[\boldsymbol{a}_2^T \left(\boldsymbol{X} - \boldsymbol{\mu}\right)\left(\boldsymbol{X} - \boldsymbol{\mu}\right)^T \boldsymbol{a}_1\right] = \\ = \boldsymbol{a}_2^T\, \mathrm{E}\left[\left(\boldsymbol{X} - \boldsymbol{\mu}\right)\left(\boldsymbol{X} - \boldsymbol{\mu}\right)^T\right] \boldsymbol{a}_1 = \boldsymbol{a}_2^T \boldsymbol{\Sigma} \boldsymbol{a}_1 \end{aligned} \tag{10}$$

A igualdade (10) deve ser zero, como $\boldsymbol{\Sigma}\boldsymbol{a}_1 = \lambda_1 \boldsymbol{a}_1$, então uma condição simplificada dessa equação é $\boldsymbol{a}_2^T \boldsymbol{a}_1 = 0$, ou seja, $\boldsymbol{a}_1$ e $\boldsymbol{a}_2$ devem ser ortogonais.

Uma vez mais, para maximizar $Var(Y_2) = \boldsymbol{a}_2^T \boldsymbol{\Sigma} \boldsymbol{a}_2$ sujeita às duas restrições anteriormente descritas, é necessário introduzir dois multiplicadores de Lagrange, λ e δ, e considerar a função Lagrangeana Aumentada:

$$L\left(\boldsymbol{a}_2\right) = \boldsymbol{a}_2^T \boldsymbol{\Sigma} \boldsymbol{a}_2 - \lambda\left(\boldsymbol{a}_2^T \boldsymbol{a}_2 - 1\right) - \delta \boldsymbol{a}_2^T \boldsymbol{a}_1 \tag{11}$$

Derivando *L* em ordem a $\boldsymbol{a}_2$ e igualando a zero, vem:

$$\frac{\partial L}{\partial \boldsymbol{a}_2} = \mathbf{0} \Leftrightarrow 2\boldsymbol{\Sigma a}_2 - 2\lambda \boldsymbol{a}_2 - \delta \boldsymbol{a}_1 = \mathbf{0} \Leftrightarrow 2\boldsymbol{a}_2 \left(\boldsymbol{\Sigma} - \lambda \boldsymbol{I}\right) - \delta \boldsymbol{a}_1 = \mathbf{0} \tag{12}$$

Multiplicando à esquerda a equação (12) por $\boldsymbol{a}_1^T$ e simplificando tem-se:

$$2\boldsymbol{a}_1^T \boldsymbol{\Sigma a}_2 - \delta = \mathbf{0} \tag{13}$$

Como $Cov\left(Y_2, Y_1\right) = 0 \Leftrightarrow \boldsymbol{a}_2^T \boldsymbol{\Sigma a}_1 = 0$ então $\delta = 0$ no(s) ponto(s) estacionário(s) e a equação (12) torna-se:

$$\boldsymbol{a}_2 \left(\boldsymbol{\Sigma} - \lambda \boldsymbol{I}\right) = \mathbf{0} \tag{14}$$

cuja solução corresponde a λ_2, segundo maior valor próprio de $\boldsymbol{\Sigma}$, e $\boldsymbol{a}_2$ o correspondente vetor próprio que define os coeficientes da segunda CP.

Continuando este tipo de argumento, tem-se que a *j*-ésima CP é definida pelo vetor próprio unitário correspondente ao *j*-ésimo maior valor próprio da matriz $\boldsymbol{\Sigma}$.

2.1.4. Propriedades das Componentes Principais

- A soma das variâncias das p CPs é igual à soma das variâncias das p variáveis originais.

Sejam $\boldsymbol{A}$ a matriz $p \times p$ dos vetores próprios de $\boldsymbol{\Sigma}$, $\boldsymbol{A} = \left[\boldsymbol{a}_1 \ \boldsymbol{a}_2 \ \cdots \ \boldsymbol{a}_p\right]$, e $\boldsymbol{Y}$ o vetor de ordem $p \times 1$ das CPs, $\boldsymbol{Y} = \left[Y_1 \ Y_2 \ \cdots \ Y_p\right]$. Então:

$$\boldsymbol{Y} = \boldsymbol{A}^T \boldsymbol{X}. \tag{15}$$

Tem-se que:

$$Var\left(\boldsymbol{Y}\right) = \boldsymbol{A}^T \boldsymbol{\Sigma A} = \boldsymbol{\Lambda}, \tag{16}$$

sendo $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$.

Ou ainda,

$$\boldsymbol{\Sigma} = \boldsymbol{A \Lambda A}^T \tag{17}$$

uma vez que $\boldsymbol{A}$ é uma matriz ortogonal tal que $\boldsymbol{A}^T \boldsymbol{A} = \boldsymbol{I}$.

Viu-se já que a variância de cada CP é o valor próprio da matriz $\boldsymbol{\Sigma}$ que lhe está associado, logo a soma dessas variâncias é:

$$\sum_{j=1}^{p} Var\left(Y_j\right) = \sum_{j=1}^{p} \lambda_j = tr\left(\boldsymbol{\Lambda}\right) \tag{18}$$

Mas isso é também o traço da matriz $\boldsymbol{\Sigma}$ que é a soma dos elementos diagonais de $\boldsymbol{\Sigma}$, ou seja, a soma das variâncias das p variáveis originais:

$$tr\left(\boldsymbol{\Lambda}\right) = tr\left(\boldsymbol{A}^T \boldsymbol{\Sigma}\, \boldsymbol{A}\right) = tr\left(\boldsymbol{\Sigma}\, \boldsymbol{A}\, \boldsymbol{A}^T\right) = tr\left(\boldsymbol{\Sigma}\right) = \sum_{i=1}^{p} Var\left(X_i\right) \tag{19}$$

Logo, a soma das variâncias das p CPs, $tr\left(\boldsymbol{\Lambda}\right)$, é igual à soma das variâncias das p variáveis originais, $tr\left(\boldsymbol{\Sigma}\right)$ (Cadima, 2010; Chatfield e Collins, 1995).

- Dado que a variância de cada componente é dada pelo respetivo valor próprio e a variância total é dada pela soma de todos os valores próprios, pode então concluir-se que a proporção de variância total original explicada pela *j*-ésima CP é dada por:

$$\frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} \tag{20}$$

E ainda que a proporção de variabilidade total dos dados originais explicada pelas primeiras $m < p$ CPs é:

$$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \tag{21}$$

A componente com maior valor próprio será a que explica melhor a variância total e assim sucessivamente (Cadima, 2010; Reis, 2001).

- A covariância entre a *i*-ésima variável original e a *j*-ésima CP é dada por:

$$Cov\left(Y_j, X_i\right) = Cov\left(Y_j, \sum_{k=1}^{p} a_{ik} Y_k\right) = a_{ij}\, Var\left(Y_j\right) = a_{ij}\, \lambda_j \tag{22}$$

onde:

a_{ij} - é o coeficiente de X_i na combinação linear que define a *j*-ésima CP (Y_j);

λ_j - é a variância da CP Y_j (*j*-ésimo valor próprio de $\boldsymbol{\Sigma}$) (Reis, 2001).

- A correlação entre a *i*-ésima variável original e a *j*-ésima CP é dada por:

$$Cor\left(Y_j, X_i\right) = \frac{Cov\left(Y_j, X_i\right)}{\sqrt{Var\left(Y_j\right)Var\left(X_i\right)}} = \frac{\lambda_j a_{ij}}{\sqrt{\lambda_j \sigma_i^2}} = a_{ij}\frac{\sqrt{\lambda_j}}{\sigma_i} \tag{23}$$

onde:

σ_i - é o desvio-padrão da variável X_i (Cadima 2010; Peña, 2002).

2.1.5. ACP sobre a Matriz de Correlações

Uma das limitações da ACP é a sensibilidade das componentes principais quando as variáveis estão expressas em diferentes unidades de medida.

Quando as unidades de medida das variáveis são muito diferentes elas contribuem com pesos diferentes na análise. Se, por exemplo, se diminuir a unidade de medida de uma variável qualquer por forma a que aumentem em magnitude os seus valores numéricos (passar uma variável de *km* para *m*, por exemplo), o peso dessa variável na análise aumentará pois aumenta a sua variância e covariância com as restantes variáveis (Peña, 2002). Desta forma, as variáveis com maior variância tenderão a dominar as primeiras CPs (González, 1999).

Para tentar solucionar o problema da inconstância das CPs perante alterações das unidades de medida das variáveis originais, sugere-se frequentemente que a ACP seja efetuada sobre as variáveis centradas e reduzidas, ou seja, que aos valores observados de cada variável seja subtraído o seu valor médio e que se divida pelo seu desvio-padrão. Considerando x_{ij} o *i*-ésimo valor observado na *j*-ésima variável, esta abordagem consiste em trabalhar a transformação (Cadima, 2010):

$$x_{ij} \rightarrow z_{ij} = \frac{x_{ij} - \overline{x}_{\bullet j}}{\sigma_j} \tag{24}$$

Derivar as CPs a partir das variáveis estandardizadas com variância unitária corresponde exatamente a aplicar a ACP à matriz de correlações $\boldsymbol{P}$ (Chatfield e Collins, 1995; Reis, 2001). O procedimento matemático é o mesmo e as CPs são agora combinações lineares das variáveis estandardizadas, em que os coeficientes são dados pelos vetores próprios da matriz $\boldsymbol{P}$ (Cadima, 2010).

Contudo, os valores e vetores próprios de $\boldsymbol{P}$ e de $\boldsymbol{\Sigma}$ não têm uma relação muito simples. Em particular, se as CPs obtidas através da matriz $\boldsymbol{P}$ são expressas em função de $\mathbf{x}$, através de uma nova transformação de $\mathbf{z}$ em $\mathbf{x}$, estas novas CPs não são as mesmas que

as derivadas a partir da matriz $\boldsymbol{\Sigma}$, exceto em circunstâncias muito especiais (Chatfield e Collins, 1989, citado por Jolliffe, 2002). Isto acontece porque as CPs são invariantes para transformações ortogonais das variáveis, mas não são, em geral, para outras transformações, como é o caso da transformação de $\mathbf{x}$ em $\mathbf{z}$ (von Storch and Zwiers, 1999, citado por Jolliffe, 2002). Assim, as CPs obtidas a partir da matriz de covariância ou da matriz de correlação não dão a mesma informação, nem podem ser derivadas diretamente umas das outras (Jolliffe, 2002).

Quando as CPs são derivadas a partir de $\boldsymbol{P}$, a variância total é dada por:

$$tr(\boldsymbol{P}) = \underbrace{1+1+\ldots+1}_{\text{p vezes}} = p \tag{25}$$

Pelo que, a proporção de variância total explicada pela *j*-ésima CP é:

$$\frac{\lambda_j}{p} \tag{26}$$

Uma das propriedades das CPs baseadas na matriz de correlações é que, a correlação entre a *i*-ésima variável original e a *j*-ésima CP apresentada na equação (23) é dada por (Cadima, 2010):

$$Cor(Y_j, X_i) = a_{ij}\sqrt{\lambda_j} \tag{27}$$

2.1.6. Critérios de seleção do número de Componentes Principais

A variância explicada pela *j*-ésima CP é igual *j*-ésimo maior valor próprio da matriz de covariâncias (ou correlações).

A existência de dependência linear aproximada entre algumas das variáveis originais faz com que os valores próprios menores sejam muito próximos de zero, pelo que, a sua contribuição para a explicação da variância total será mínima. Portanto, retirar essas componentes da análise não implica uma perda significativa de informação, permitindo uma redução da dimensão dos dados e tornando os resultados mais simples e de interpretação mais clara (Reis, 2001).

Existem vários critérios práticos para determinar o número de CPs a reter:

1. Selecionar as CPs que permitam explicar mais de 70% da variância total. Este critério é subjetivo, havendo divergências relativamente ao limiar mínimo de variância explicada (Reis, 2001). Jolliffe (2002) sugere um corte na faixa dos 70% a 90%, contudo ressalva

que este intervalo poderá ser maior ou menor dependendo dos detalhes práticos da base de dados.

2. Excluir as CPs cujos valores próprios são inferiores à média dos valores próprios. No caso da análise ser feita a partir da matriz de correlações, reter as CPs cujos valores próprios são maiores que 1 – Critério de Kaiser (Reis, 2001).
3. Utilizar a representação gráfica da variância explicada por cada componente principal $\left(\lambda_j\right)$ em função do número de ordem de cada CP, o *scree-plot* proposto por Cattell (1966), citado por Jolliffe (2002). Unindo estes pontos obtemos uma linha poligonal e a abcissa do ponto onde se dá uma mudança brusca de declive corresponde ao número de CPs a reter. Uma alternativa ao *scree-plot*, desenvolvida em ciências atmosféricas, é representar $\log\left(\lambda_j\right)$, em vez de λ_j, em função de j, esta representação é conhecida como o diagrama *log-eigenvalue* (LEV) (Jolliffe, 2002).
4. Um quarto critério, mais formal que os anteriores, consiste em reter as CPs cuja variância é significativamente diferente de zero e apenas se pode aplicar quando estas são derivadas a partir da matriz de covariâncias amostral ($\boldsymbol{S}$). Bartlett desenvolveu um procedimento para testar a hipótese de que os $p-k$ valores próprios de $\boldsymbol{\Sigma}$ são iguais. Não rejeitar a hipótese nula corresponde a reter as k primeiras CPs.

   A estatística de teste de Bartlett é dada por:

$$M\left[-\ln\left|\boldsymbol{S}\right|+\sum_{j=1}^{k}\ln\lambda_j+\left(p-k\right)\ln l\right] \tag{28}$$

com $M=n-k-\frac{1}{6}\left[2\left(p-k\right)+1+\frac{2}{p-k}\right]$ e $l=\frac{1}{p-k}\left[tr\left(\boldsymbol{S}\right)-\sum_{j=1}^{k}\lambda_j\right]$ e segue uma distribuição de χ^2 com $\left[\frac{1}{2}\left(p-k-1\right)\left(p-k+2\right)\right]$ graus de liberdade (Reis, 2001).

2.1.7. Interpretação das Componentes Principais e Rotação

Os resultados computacionais de uma ACP são habitualmente apresentados através dos transformados dos vetores próprios $\boldsymbol{a}_j$:

$$\boldsymbol{a}_j^{*}=\lambda_j^{1/2}\boldsymbol{a}_j,\ j\in\left\{1,2,...,p\right\} \tag{29}$$

Estes vetores são tais que a soma dos quadrados dos seus elementos é igual ao correspondente valor próprio λ_j, em vez de 1, uma vez que:

$$\boldsymbol{a}_j^{*T}\,\boldsymbol{a}_j^{*} = \left(\lambda_j^{1/2}\,\boldsymbol{a}_j\right)^T \left(\lambda_j^{1/2}\,\boldsymbol{a}_j\right) = \lambda_j\,\boldsymbol{a}_j^T\,\boldsymbol{a}_j = \lambda_j \tag{30}$$

Tomando $\boldsymbol{C} = \left[\boldsymbol{a}_1^{*}\ \boldsymbol{a}_2^{*} \ldots \boldsymbol{a}_p^{*}\right]$, matriz dos pesos (*loadings*), temos que $\boldsymbol{C} = \boldsymbol{A}\,\boldsymbol{\Lambda}^{1/2}$ e da equação (17) vem que $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^T$.

Deste modo, as CPs refletem exatamente a proporção de variância explicada pelos dados correspondentes (Chatfield e Collins, 1995).

Os vetores $\boldsymbol{a}_j^{*}$ têm, segundo estes autores, duas interpretações diretas:

1. São vetores de pesos das variáveis iniciais nas componentes.

Para que as CPs tenham variância unitária toma-se $\boldsymbol{Y}^{*} = \boldsymbol{\Lambda}^{1/2}\,\boldsymbol{Y}$. A transformação inversa, $\boldsymbol{X} = \boldsymbol{A}\,\boldsymbol{Y}$ (supondo que $\boldsymbol{X}$ tem média zero) resulta em $\boldsymbol{X} = \boldsymbol{A}\,\boldsymbol{\Lambda}^{1/2}\,\boldsymbol{Y}^{*} = \boldsymbol{C}\boldsymbol{Y}^{*}$. À semelhança do que acontece na Análise Fatorial, os elementos de $\boldsymbol{C}$ podem ser considerados como *loadings* das CPs.

2. A segunda interpretação de $\boldsymbol{C}$ surge quando se analisa a matriz de correlação $\boldsymbol{P}$ de $\boldsymbol{X}$ de forma que $\boldsymbol{P} = \boldsymbol{C}\boldsymbol{C}^T$.

Neste caso, a covariância entre a *i*-ésima variável original e a *j*-ésima CP é dada pela equação (22), uma vez que $Var\left(Y_j\right) = \lambda_j$.

As variáveis X_i foram estandardizadas de modo a terem variância unitária, então a correlação entre a *i*-ésima variável original e a *j*-ésima CP dada pela equação (27) e a matriz de correlações por:

$$Cor\left(\boldsymbol{Y}, \boldsymbol{X}\right) = \Lambda^{1/2}\,A^T = C^T \tag{31}$$

Assim, quando C^T é calculada a partir da matriz de correlação $\boldsymbol{P}$, os seus elementos medem as correlações entre as CPs e as variáveis originais estandardizadas permitindo a sua interpretação.

Segundo Reis (2001), a soma dos quadrados dos *loadings* das variáveis para cada componente (soma em coluna) dá-nos o valor próprio correspondente:

$$\sum_{i=1}^{p} a_{ij}^{*2} = \lambda_j \tag{32}$$

e a soma dos quadrados dos *loadings* das CPs para cada variável (soma em linha) dá-nos a proporção da variância de cada variável explicada pelas CPs retidas:

$$h_j = \sum_{i=1}^{m} a_{ij}^{*2}, \quad m \leq p \tag{33}$$

a h_j chama-se comunalidade e será igual a 1 se forem consideradas todas as CPs.

De acordo com esta autora, a interpretação de cada CP é feita com base nos *loadings* das variáveis e estaria simplificada se cada variável tivesse um peso elevado para uma das CPs e baixo ou próximo de zero para as restantes, o que pode ser conseguido através da rotação das CPs. Jolliffe (2002) também refere que uma forma simples de ajudar a interpretação das CPs é através da rotação das mesmas. Contudo, segundo este autor, esta prática apresenta alguns inconvenientes tornando-se a sua implementação pouco consensual.

A aplicação de um processo de rotação tem como principal objetivo transformar os coeficientes das CPs numa estrutura simplificada (Thurstone, 1947, citado por Reis, 2001). Sejam $\boldsymbol{C}$ e $\boldsymbol{B}$ matrizes de dimensão $p \times m$, dos vetores $\boldsymbol{a}_j^*$ antes da rotação e $\boldsymbol{b}_j$ depois da rotação, respetivamente:

$$\boldsymbol{C} = \left[\boldsymbol{a}_1^* \ \boldsymbol{a}_2^* \ \cdots \ \boldsymbol{a}_m^* \right] \text{ e } \boldsymbol{B} = \left[\boldsymbol{b}_1 \ \boldsymbol{b}_2 \ \cdots \ \boldsymbol{b}_m \right],$$

onde m é o número de CPs retidas e $\boldsymbol{B} = \boldsymbol{CG}$, sendo $\boldsymbol{G}$ uma matriz de dimensão $m \times m$.

Para que $\boldsymbol{B}$ seja uma estrutura simplificada é necessário que:

- Cada linha da matriz dos *loadings* tenha pelo menos um zero;
- Cada coluna da matriz dos *loadings* tenha pelo menos m zeros;
- Para cada par de colunas da matriz dos *loadings* deve haver variáveis cujos *loadings* sejam próximos de zero numa das colunas, mas não na outra e quando $m \geq 4$, cada coluna deve ter um maior número de *loadings* nulos do que não nulos. Estas condições permitem garantir a independência dos vetores após a rotação.

Uma estrutura simplificada é muito difícil de obter e não existem garantias de que as CPs se mantenham independentes depois da rotação.

Existem vários métodos de rotação que podemos agrupar em dois tipos: rotações ortogonais e oblíquas.

Relativamente aos métodos ortogonais, o problema consiste em encontrar uma matriz $\boldsymbol{G}$ ortogonal que maximize:

$$\sum_{j=1}^{m} \left\{ \sum_{i=1}^{p} b_{ij}^4 - \frac{c}{p} \left(\sum_{i=1}^{p} b_{ij}^2 \right)^2 \right\} \tag{34}$$

sendo c uma constante que varia consoante o método.

O método de rotação ortogonal mais popular e melhor, já demostrado por alguns autores (Gebhardt (1968), Deleeuw e Pruzanski (1978) e TenBerge (1984), citados por Reis (2001)), é o VARIMAX onde $c = 1$. Este método pretende minimizar o número de variáveis em cada CP.

A rotação ortogonal QUARTIMAX, $c = 0$, tem como objetivo que cada variável tenha um peso elevado para um número reduzido de CPs e quase nulo para as restantes.

O método EQUIMAX, $c = m/2$, tem como objetivo simplificar simultaneamente as linhas e as colunas da matriz dos *loadings*.

Os métodos de rotação oblíquos fazem com que se perca o pressuposto de independência entre as CPs, contudo permitem que estas rodem livremente de maneira a simplificarem o agrupamento das variáveis e a interpretação das CPs.

Depois de aplicado um método de rotação, cabe ao investigador interpretar, com alguma subjetividade, as combinações lineares ponderadas das variáveis e qual o rótulo a atribuir a cada CP (Reis, 2001).

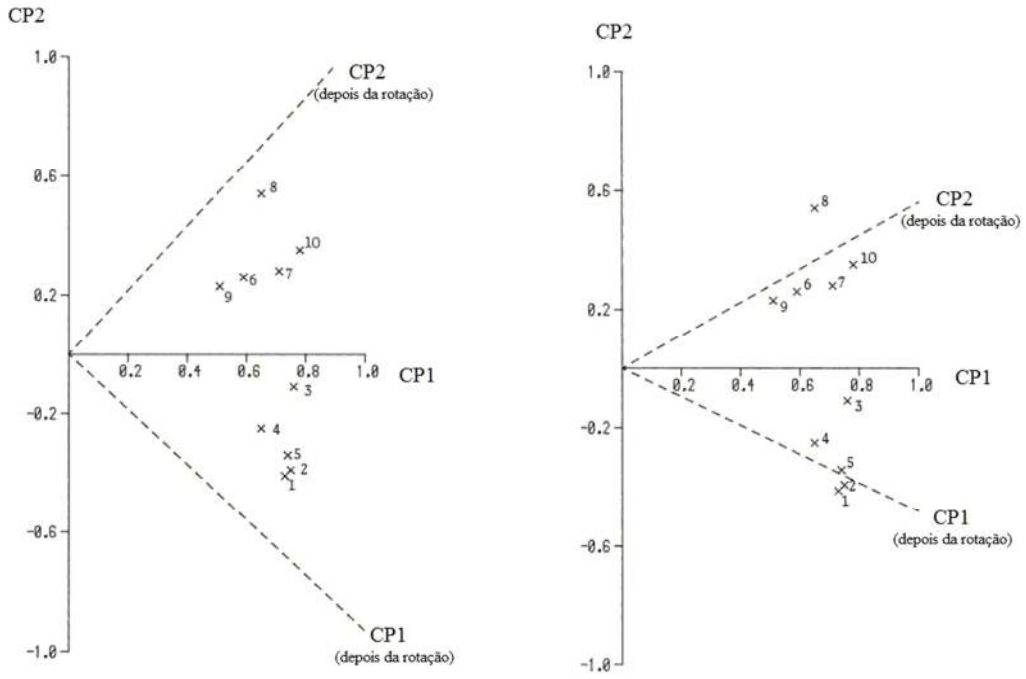


Figura 3 - Rotação ortogonal (à esquerda) e oblíqua (à direita) das CPs (adaptada de Jolliffe, 2002 e Reis, 2001)

2.1.8. *Scores* das Componentes Principais e suas aplicações noutras técnicas de Estatística Multivariada

A representação do *i*-ésimo indivíduo (*i*-ésima linha de $\boldsymbol{X}$) na componente Y_j é o *i*-ésimo coeficiente desse vetor. Considerando que *m* CPs são suficientes para representar adequadamente o conjunto original de dados, o *i*-ésimo indivíduo é representado por um ponto cujas coordenadas são dadas pela *i*-ésima linha de $\boldsymbol{A}_m^T\boldsymbol{X}$, sendo $\boldsymbol{A}_m$ a matriz de dimensão $p \times m$. Usualmente designam-se os coeficientes de cada indivíduo numa CP (coeficientes dos vetores $\boldsymbol{Y}_j$) por *scores* (Cadima, 2010).

Os *m scores* das CPs do *i*-ésimo indivíduo são dados por (Silva, 2007):

$$\begin{aligned} y_{i1} &= \boldsymbol{a}_1^T\boldsymbol{x}_i = a_{11}x_{i1} + a_{12}x_{i2} + \ldots + a_{1p}x_{ip} \\ y_{i2} &= \boldsymbol{a}_2^T\boldsymbol{x}_i = a_{21}x_{i1} + a_{22}x_{i2} + \ldots + a_{2p}x_{ip} \\ &\vdots \qquad\qquad\qquad\qquad\qquad\qquad , i \in \{1,2,\ldots,n\} \\ y_{im} &= \boldsymbol{a}_m^T\boldsymbol{x}_i = a_{m1}x_{i1} + a_{m2}x_{i2} + \ldots + a_{mp}x_{ip} \end{aligned} \tag{35}$$

Se as primeiras CPs explicam grande parte da variabilidade total dos dados originais então é boa ideia usar os *scores* destas primeiras CPs em análises posteriores (Chatfield e Collins, 1995).

A representação gráfica dos dados deve ser feita sempre que possível. Se as duas primeiras CPs explicam grande proporção da variação total, é vantajoso representar os valores dos *scores* das duas primeiras CPs para cada indivíduo e assim detetar a presença de *outliers* ou de agrupamentos ou *clusters* de indivíduos (Chatfield e Collins, 1995). Contudo, muitas vezes não há uma estrutura de grupos clara nos dados e nestes casos pode usar-se a Análise de *Clusters*.

De acordo com Jolliffe (2002), a Análise de *Clusters* é a técnica multivariada onde mais frequentemente se necessita de uma redução da dimensionalidade preliminar.

Os *scores* das CPs também são usados para a representação de *biplots*.

O *biplot* trata-se de uma generalização do gráfico de dispersão que contém marcadores para as variáveis e marcadores para os indivíduos observados. É especialmente utilizado na análise de componentes principais, onde este gráfico pode indicar distâncias entre indivíduos e *clusters* de indivíduos, bem como mostrar as variâncias e correlações entre variáveis (Gabriel, 1971).

Segundo Ferreira (2010), o método *biplot* gera uma representação procurando que:

- as distâncias entre os indivíduos projetados num espaço de dimensão reduzida sejam próximas das originais;

- a projeção dos indivíduos nos eixos sejam as mais próximas das originais;
- os vetores que representam as variáveis originais ajudem a verificar quanto peso cada novo eixo dá a cada uma das variáveis originais;
- o cosseno do ângulo entre os vetores que representam as variáveis originais se aproxime da correlação entre essas variáveis.

Uma vez que nos *biplots* de Galindo (1985) temos indivíduos e variáveis representados na mesma escala, Vairilinhos e Galindo (2004) consideram fazer sentido interpretar distâncias entre indivíduos e variáveis como preponderância de uma variável para explicar um indivíduo ou como a contribuição desse indivíduo para os valores dessa variável.

2.2. Análise de *Clusters*

2.2.1. Introdução

A necessidade de agrupar objetos semelhantes e classificá-los tem acompanhado a evolução da humanidade desde os seus primórdios até aos dias de hoje. Constituir grupos é uma característica da atividade humana e um suporte essencial do método de aprendizagem e do próprio método científico em geral (Branco, 2004).

A Análise de *Clusters* (AC) integra uma série de procedimentos sofisticados de Estatística Multivariada que podem ser usados para classificar objetos observando apenas as (dis)semelhanças entre eles, mais concretamente, para tentar organizar um conjunto de objetos em grupos relativamente homogéneos (*clusters*).

Dado um conjunto de *n* indivíduos avaliados em relação a *p* variáveis, este método agrupa os indivíduos em função da informação existente, de tal modo que indivíduos pertencentes a um mesmo grupo sejam tão semelhantes quanto possível e sempre mais semelhantes aos elementos do mesmo grupo do que a elementos dos restantes grupos (Reis, 2001).

Muitas vezes esta análise tem por objetivo o agrupamento de variáveis e não de indivíduos, nesse caso os objetos da análise são as próprias variáveis (Branco, 2004).

De acordo com Johnson (1982) o principal objetivo da AC é descobrir agrupamentos naturais de indivíduos ou variáveis. Para isso é necessário definir uma escala quantitativa para medir (dis)semelhanças entre eles.

Para o agrupamento de indivíduos usam-se habitualmente coeficientes de dissemelhança, muitos deles baseados em distâncias, e para o agrupamento de variáveis usam-se geralmente medidas de correlação ou associação (semelhanças). Contudo, é possível, de um modo geral, construir uma medida de dissemelhança a partir de uma semelhança e vice-versa, como se verá mais à frente (Branco, 2004).

Segundo Branco (2004), a AC opera essencialmente sobre dois tipos de estruturas de dados identificadas por dois formatos de matrizes:

1. A matriz dos dados, $\mathbf{X} = \left[x_{ij} \right]$, $i \in \{1,...,n\}$ e $j \in \{1,...,p\}$, em que x_{ij} representa o valor da variável *j* observada no indivíduo *i*. Esta matriz pode incluir variáveis quantitativas (contínuas e discretas) e variáveis qualitativas (nominais e ordinais).
2. A matriz de dissemelhanças (semelhanças), $\mathbf{D} = \left[d_{ij} \right]$ $\left(\mathbf{S} = \left[s_{ij} \right] \right)$, $i, j \in \{1,...,n\}$ quadrada e em geral simétrica em que d_{ij} $\left(s_{ij} \right)$ representa o valor da dissemelhança (semelhança) entre os objetos *i* e *j*.

Para se escolher a medida de proximidade e o algoritmo de agrupamento adequado é necessário ter em consideração a natureza das variáveis (contínuas, nominais, ordinais, binárias). Se, adicionalmente, as variáveis se apresentam com unidades de medida diferentes e se aplica a AC sem uma estandardização prévia, qualquer medida de (dis)semelhança vai refletir sobretudo o peso das variáveis que maiores valores e maior dispersão apresentam (Reis, 2001). Contudo, Branco (2004) salienta que a estandardização levanta alguma polémica, havendo autores que consideram tratar-se de um comodismo visto que existem formas apropriadas de tratar dados heterogéneos, tais como atribuir pesos diferentes às variáveis de forma a homogeneizar a sua contribuição na construção dos índices de semelhança.

Em síntese, esta metodologia (AC) compreende duas etapas fundamentais (Azevedo, 2010):

1. Definir uma medida de semelhança ou de dissemelhança/distância;
2. Escolher um método de agrupamento, ou seja, definir um algoritmo de partição/classificação.

2.2.2. Medidas de Proximidade

De acordo com Branco (2004), o processo de construção de *Clusters* tem por base as ideias de semelhança e dissemelhança, conhecidas por proximidades. Dois objetos pertencem ao mesmo *cluster* se são semelhantes e pertencem a *clusters* diferentes se não são semelhantes, também se pode dizer neste último caso se são dissemelhantes.

A semelhança mede o grau de parecença ou proximidade entre dois objetos.

Em muitas situações a medida de proximidade mais fácil de obter é a semelhança s_{ij} entre dois objetos *i* e *j*. Esta medida de proximidade deve satisfazer as seguintes propriedades:

1. $s_{ij} \geq 0, \forall i, j$
2. $s_{ij} = s_{ji}, \forall i, j$
3. s_{ij} é tanto maior quanto maior for a semelhança entre os objetos.

A dissemelhança reflete o grau de diferença, afastamento ou divergência entre os objetos.

Dissemelhança entre dois objetos *i* e *j* de uma dada coleção define-se como a função d_{ij} dos objetos cujos valores verificam as seguintes propriedades:

1. $d_{ij} \geq 0, \quad \forall i, j$
2. $d_{ii} = 0, \quad \forall i, j$

3. $d_{ij} = d_{ji}, \quad \forall i, j$

Se além destas propriedades a dissemelhança satisfaz também a desigualdade triangular:

4. $d_{ij} \leq d_{ik} + d_{kj}, \quad \forall i, j, k$

diz-se que satisfaz as propriedades de uma semimétrica ou semidistância (embora muitas dissemelhanças não a satisfaçam) e se verifica ainda a propriedade:

5. $d_{ij} = 0$ se e só se $i = j$

diz-se que é uma métrica ou distância.

No caso de verificarem a propriedade ultramétrica (mais forte que a desigualdade triangular):

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

diz-se que a dissemelhança é ultamétrica.

Na maioria das situações práticas é suficiente que se satisfaçam as propriedades 1, 2 e 3.

É possível estabelecer uma relação entre semelhanças e dissemelhanças dos mesmos objetos. A dissemelhança d_{ij} pode obter-se da semelhança s_{ij} usando uma função decrescente, por exemplo, $d_{ij} = k - s_{ij}$, por sua vez s_{ij} pode obter-se de d_{ij} através da transformação $s_{ij} = \dfrac{k}{k + d_{ij}}$, onde k é uma constante adequada.

As medidas de proximidade dependem da natureza das características que são observadas nos objetos. Assim, no caso das variáveis quantitativas a medida de dissemelhança mais conhecida é a distância euclidiana entre os objetos i e j e é definida por:

$$d_{ij} = \left[\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2\right]^{\frac{1}{2}} \tag{36}$$

ou, na forma vetorial:

$$d_{ij} = \left[\left(\mathbf{x}_i - \mathbf{x}_j\right)^T\left(\mathbf{x}_i - \mathbf{x}_j\right)\right]^{\frac{1}{2}} \tag{37}$$

onde $\mathbf{x}_i$ e $\mathbf{x}_j$ são vetores linha da matriz $\mathbf{X}$, ou seja, vetores de observações relativas aos objetos i e j, respetivamente.

No caso das variáveis estarem medidas em unidades diferentes, terem variâncias muito diferentes ou estarem correlacionadas, a distância euclidiana nem sempre é satisfatória pois as variáveis intervêm com diferentes pesos na determinação das dissemelhanças. Nestes

casos podem usar-se outras distâncias dela derivadas. De um modo geral, introduz-se uma matriz de pesos **A** e constrói-se a distância euclidiana ponderada entre os objetos *i* e *j*:

$$d_{ij} = \left[\left(\mathbf{x}_i - \mathbf{x}_j \right)^T \mathbf{A} \left(\mathbf{x}_i - \mathbf{x}_j \right) \right]^{\frac{1}{2}} \tag{38}$$

onde:

- Se $\mathbf{A} = \mathbf{I}$ tem-se a distância euclidiana;
- Se $\mathbf{A} = \frac{1}{p}\mathbf{I}$ tem-se a distância euclidiana média;
- Se $\mathbf{A} = \mathbf{D}^{-1} = \left[diag\left(s_1^2, s_2^2, ..., s_p^2 \right) \right]^{-1}$ tem-se a distância euclidiana estandardizada, onde $\mathbf{D}$ é a matriz diagonal das variâncias das colunas de **X**.
- Se $\mathbf{A} = \mathbf{S}^{-1}$ tem-se a distância de Mahalanobis, onde **S** é a estimativa da matriz de covariâncias das *p* variáveis. Esta distância além de reduzir a dependência das unidades de medida, reduz também a influência da correlação entre variáveis, o que pode mascarar ainda mais os resultados de uma AC (Hartigan, citado por Branco, 2004).

Segundo Branco (2004) também se podem construir dissemelhanças com base na família de métricas de Minkowski:

$$d_{ij} = \left[\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^r \right]^{\frac{1}{r}}, r \geq 1 \tag{39}$$

onde:

- Se $r = 1$ tem-se a métrica do quarteirão também designada por distância *city-block*, conhecida pelo seu comportamento robusto relativamente a *outliers*;
- Se $r = 2$ tem-se a distância euclidiana;
- Se $r \to \infty$ obtém-se $d_{ij} = \max_k \left| x_{ik} - x_{jk} \right|$ e tem-se a métrica do máximo ou distância de *Chebychev*.

Geralmente, quanto maior for o valor de *r*, maior será o peso relativo de indivíduos muito dissemelhantes dos restantes (Cadima, 2010).

Com o mesmo objetivo da distância euclidiana ponderada também se pode pensar em aplicar pesos às métricas de Minkovski (Branco, 2004).

Muitas vezes interessa agrupar variáveis em vez de indivíduos. Para isso basta efetuar a AC sobre as linhas da transposta da matriz de dados. As variáveis tomam o lugar dos

indivíduos e as medidas de proximidade entre indivíduos podem ser agora usadas. No entanto, para análise de variáveis as medidas mais adequadas são as de correlação e associação. Para duas variáveis *i* e *j*, encontrada a semelhança pode obter-se a dissemelhança fazendo, por exemplo, $d_{ij} = \sqrt{1 - s_{ij}}$ (Branco, 2004).

2.2.3. Métodos de Agrupamento

Os dois principais grupos de métodos de agrupamento são os hierárquicos e os não hierárquicos.

Nos métodos não-hierárquicos considera-se à partida um número fixo de grupos que se pretende construir e faz-se uma classificação inicial dos *n* indivíduos nesses grupos. Através de transferências de indivíduos de um grupo para outro procura-se determinar uma "boa" classificação, no sentido de tornar os grupos mais internamente homogéneos e externamente heterogéneos (Cadima, 2010).

Nos métodos hierárquicos, o agrupamento efetua-se por etapas, partindo-se de *n* grupos (de um único indivíduo cada) procedem a sucessivas fusões de grupos considerados mais "semelhantes". Cada fusão reduz o número de grupos em uma unidade (Cadima, 2010). Segundo Branco (2004), nestes métodos sempre que um indivíduo é atribuído a um *cluster* nunca mais o abandona. Para aplicar os métodos hierárquicos existem dois tipos de algoritmos:

1. Os algoritmos aglomerativos (ou ascendentes) partem de *n* grupos de um indivíduo cada e vão formando novos grupos por aglutinação sucessiva de grupos anteriores;
2. Os algoritmos divisivos (ou descendentes) partem de um grupo inicial de *n* indivíduos e vão formando novos grupos por divisão sucessiva dos grupos formados anteriormente até se obter *n* grupos singulares.

A estrutura hierárquica proveniente destes procedimentos pode representar-se por um gráfico bidimensional ao qual se dá o nome de **dendograma**. Este gráfico tem a forma de uma árvore invertida, com a raiz para cima e os ramos para baixo, embora dependa do *software* que o produz (o SPSS, por exemplo, fornece o dendograma na horizontal). Os nós internos representam os *clusters* e a altura dos troncos indica a distância a que os *clusters* se ligam, alturas pequenas indicam que a aglutinação é feita entre *clusters* razoavelmente homogéneos (Branco, 2004).

Os métodos de AC mais divulgados e utilizados são os hierárquicos aglomerativos pois os divisivos são computacionalmente muito pesados (Reis, 2001). Neste trabalho apenas se descrevem os métodos hierárquicos aglomerativos.

O ponto de partida comum a todos os métodos hierárquicos é a construção de uma matriz de semelhanças ou de dissemelhanças (distâncias) (Reis, 2001).

2.2.3.1. Métodos Hierárquicos Aglomerativos

Escolhida uma medida de (dis)semelhança segue-se a escolha de um método de agrupamento. De acordo com Branco (2004), os algoritmos hierárquicos aglomerativos têm sido os mais populares e descrevem-se em poucos passos:

1. Começa-se por considerar os n indivíduos iniciais como n grupos singulares. Neste caso, a dissemelhança entre os grupos coincide com a matriz de dissemelhanças $\boldsymbol{D} = \left[d_{ij} \right]$, onde d_{ij} é a dissemelhança entre os indivíduos i e j.
2. Identificam-se os dois grupos mais próximos, ou seja, procura-se na matriz de dissemelhanças o elemento mais pequeno.
3. Unem-se esses dois grupos, ou seja, forma-se um único *cluster* com os dois elementos. Atualiza-se a matriz $\boldsymbol{D}$ eliminando as linhas e colunas correspondentes a esses dois grupos e introduz-se uma nova linha e coluna com as dissemelhanças calculadas entre o novo grupo e cada um dos restantes.
4. Repetir os passos 2 e 3 até que todos os indivíduos estejam contidos num único *cluster*.

O passo 3 exige que se defina a dissemelhança entre dois grupos, em que pelo menos um deles tem mais do que um indivíduo. Existem várias formas de definir esta dissemelhança entre dois grupos e a cada uma delas está associado um método hierárquico aglomerativo.

Os métodos hierárquicos aglomerativos mais comuns, segundo Branco (2004), são:

1. Método da ligação simples (ou método do vizinho mais próximo)

A dissemelhança entre dois grupos A e B é dada pela menor das dissemelhanças entre cada elemento de A e cada elemento de B:

$$d_{AB} = \min \left\{ d_{ij} : i \in A,\ j \in B \right\} \tag{40}$$

A dissemelhança entre dois grupos é determinada pelos indivíduos mais próximos (os vizinhos mais próximos), o que significa que uma única ligação é suficiente para juntar os

grupos. De cada vez que se adiciona um indivíduo a um grupo, as distâncias do novo grupo aos restantes são menores ou não se alteram. Há assim uma tendência para que os grupos cresçam e se juntem para formar grupos maiores deixando os indivíduos isolados firmes na sua posição, o que revela a capacidade deste método detetar *outliers*.

Este método tende então a construir grupos alongados, se dois grupos não estiverem bem separados podem ser classificados como sendo o mesmo grupo desde que tenham dois indivíduos próximos.

O método da ligação simples apresenta uma propriedade única, a sua indiferença a casos de empate. Se houver duas dissemelhanças iguais e menores que as restantes, o resultado final não se altera caso se escolha uma ou outra para produzir um novo grupo e prosseguir a análise.

2. Método da ligação completa (ou método do vizinho mais afastado)

A dissemelhança entre dois grupos A e B é dada pela maior das dissemelhanças entre cada elemento de A e cada elemento de B:

$$d_{AB} = \max\left\{d_{ij} : i \in A,\ j \in B\right\} \tag{41}$$

Ao contrário do método anterior, este método serve-se dos indivíduos mais afastados para derivar a medida de proximidade entre grupos. Ao acrescentar um indivíduo a um grupo, a distância do novo grupo aos restantes aumenta ou não se altera, há assim uma tendência para que grupos grandes não cresçam mais.

3. Método da ligação média

A dissemelhança entre os grupos A e B é a média das dissemelhanças entre todos os pares de indivíduos, formados com um indivíduo de cada grupo:

$$d_{AB} = \frac{\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d_{ij}}{n_A\, n_B} \tag{42}$$

Este método traduz um compromisso entre as duas situações extremas dos dois métodos anteriores e, segundo Reis (2001), além da vantagem de evitar valores extremos possui ainda a vantagem de tomar toda a informação dos grupos. É recomendado por muitos autores que o consideram superior aos métodos da ligação simples e da ligação completa.

4. Método do centróide

A dissemelhança entre os grupos A e B é a distância entre os seus centróides, ou seja,

$$d_{AB} = d\left(\overline{\mathbf{x}}_A, \overline{\mathbf{x}}_B\right) \tag{43}$$

onde $\overline{\mathbf{x}}_A$ e $\overline{\mathbf{x}}_B$ são os centróides dos grupos *A* e *B* respetivamente, isto é, $\overline{\mathbf{x}}_A = \frac{\sum_{i \in A} \mathbf{x}_i}{n_A}$ e $\overline{\mathbf{x}}_B = \frac{\sum_{i \in B} \mathbf{x}_i}{n_B}$ e $\mathbf{x}_i$ é o vetor das *p* observações do indivíduo *i*.

Em cada passo, os grupos a aglutinar são os que têm os centróides mais próximos. Um inconveniente deste método é a dificuldade de interpretação provocada pelo facto da distância de fusão entre dois grupos poder aumentar ou diminuir de passo para passo. A distância entre *clusters* pode ser qualquer mas a medida com mais sucesso em termos de facilidade de aplicação e clareza dos resultados produzidos é, de acordo com este autor, o quadrado da distância euclidiana.

5. Método da mediana

Este método é semelhante ao anterior. Contudo, ao aglutinar dois grupos *A* e *B*, os seus centróides, $\overline{\mathbf{x}}_A$ e $\overline{\mathbf{x}}_B$, recebem pesos iguais antes de produzirem o centróide do novo *cluster*, $\overline{\mathbf{x}}$. Este novo centróide fica a meio dos centróides dos grupos aglutinados e desta forma evita-se que o grupo com mais indivíduos absorva o grupo mais pequeno.

A mediana aqui referida não se trata da mediana estatística mas da mediana de um triângulo, ou seja, o segmento de reta que liga o vértice de um triângulo ao ponto médio do lado oposto.

6. Método de Ward

De acordo com o método de Ward (1963) o critério de fusão de dois grupos *A* e *B* é baseado no incremento da soma dos quadrados que ocorre quando estes grupos são aglutinados. Este incremento é:

$$SSW_{A \cup B} - \left(SSW_A + SSW_B\right) \tag{44}$$

onde $SSW_A = \sum_{i \in A} \sum_{j=1}^{p} \left(x_{ijA} - \overline{x}_{jA}\right)^2$ é a soma dos quadrados dentro do grupo *A*, x_{ijA} é a observação do indivíduo *i* do grupo *A* na variável *j* e $\overline{x}_{jA}$ é a média da variável *j* no grupo *A*. Da mesma forma se obtém SSW_B e $SSW_{A \cup B}$.

O incremento da soma dos quadrados corresponde a uma perda de informação. Em cada passo do algoritmo são formados todos os pares de *clusters* possíveis e para cada par é

calculado o incremento resultante da reunião dos *clusters*. Os *clusters* escolhidos para formar um novo *cluster* são aqueles a que corresponde o menor incremento, isto é, a menor perda de informação resultante da aglutinação.

Segundo Reis (2001) esta técnica pode ainda ser resumida nas seguintes etapas:

1. Calcular as médias das variáveis para cada grupo;
2. Calcular o quadrado da distância euclidiana entre essas médias e os valores das variáveis para cada indivíduo;
3. Somar as distâncias para todos os indivíduos;
4. Minimizar a variância dentro dos grupos.

Não existe um melhor método de agrupamento em AC. É prática comum utilizar vários critérios e para cada um destes experimentar várias (dis)semelhanças e comparar os resultados obtidos. Se estes forem semelhantes, é possível concluir que os resultados são fidedignos (Reis, 2001).

2.2.4. Validação dos resultados obtidos

Como já foi referido, a AC tem como principal objetivo criar grupos homogéneos e a aplicação dos métodos hierárquicos permite a apresentação dos resultados sob a forma de um dendograma. A questão que se coloca é por onde cortar o dendograma por forma a obter um número de grupos ótimo. Infelizmente esta importante decisão ainda não está completamente resolvida.

Um método simples é, de acordo com Reis (2001), a comparação gráfica do número de *clusters* com o coeficiente de fusão, ou seja, o valor numérico (distância ou semelhança) para o qual vários objetos se unem para formar um grupo. Maroco (2007) salienta também que este método sugere que o número de *clusters* naturais a reter se efetue quando o declive da reta que une a distância entre dois *clusters* for relativamente pequeno. Reis (2001) refere que a partição ótima poderá ser considerada quando a divisão de um novo grupo não introduz alterações significativas no coeficiente de fusão.

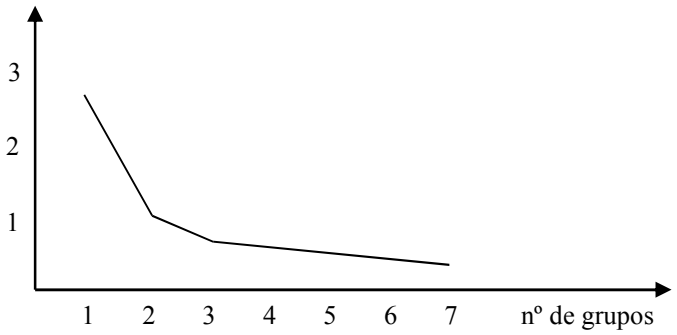


Figura 4 - Coeficientes de fusão (adaptada de Reis, 2001)

De acordo com Reis (2001) uma limitação deste método é a escolha do número de *clusters* quando a representação gráfica mostra apenas pequenos saltos. Para resolver este problema a autora refere que R. Mojena e D. Wishard desenvolveram um método no sentido de encontrar uma partição ideal.

Uma outra forma de avaliar o número ideal de *clusters*, apresentada por Maroco (2007), é a aplicação do critério do R-quadrado. O R-quadrado é uma medida de quão diferentes cada um dos grupos ou *clusters* são em cada passo do algoritmo e calcula-se através da razão entre a soma dos quadrados entre os grupos e a soma dos quadrados totais para cada uma das variáveis usadas na análise:

$$R-quadrado = \frac{SQC}{SQT} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{k} n_{ij}\left(\overline{x}_{ij} - \overline{x}_{i}\right)^2}{\sum_{i=1}^{p}\sum_{j=1}^{k}\sum_{l=1}^{n_i}\left(x_{ijl} - \overline{x}\right)^2} \tag{45}$$

onde *SQC* representa a soma dos quadrados dos *clusters*, *SQT* a soma dos quadrados total, *p* o número de variáveis, *k* o número de grupos, n_{ij} o tamanho do grupo *j* na variável *i*, $\overline{x}_{ij}$ a média da variável *i* no grupo *j*, $\overline{x}_{i}$ a média da variável *i* e $\overline{x}$ a média da amostra global.

Desta forma, esta medida é a percentagem da variabilidade total que é retida em cada uma das soluções dos *clusters*. Com um só *cluster* a variabilidade entre *clusters* é zero e no caso de existirem tantos *clusters* quantos objetos esta variabilidade é 100%. Interessa encontrar o número mínimo de *clusters* que retenha uma percentagem significativa de variabilidade (*e.g.* superior a 80%) (Maroco, 2007).

2.2.5. Interpretação dos *Clusters*

Definidos os *clusters*, poderá ter interesse saber qual ou quais deles apresentam diferenças significativas relativamente a algumas variáveis.

Um procedimento usual seria utilizar a análise da variância simples (ANOVA) que possibilita a comparação entre parâmetros de mais do que duas populações. Partindo da dispersão total presente num conjunto de dados, a análise da variância permite identificar os *clusters* que deram origem a essa dispersão e avaliar a contribuição de cada um deles (Guimarães e Cabral, 1997).

A aplicação desta técnica estatística envolve, segundo Pestana e Gageiro (2008), os seguintes pressupostos:

- As observações são independentes entre si;
- As observações dentro de cada grupo têm distribuição normal;
- As variâncias de cada grupo são iguais entre si, ou seja, há homocedasticidade.

A ANOVA é relativamente robusta face à perda de igualdade de variâncias (desde que as dimensões das amostras provenientes dos k grupos sejam aproximadamente iguais) e de normalidade (é pouco afetada por desvios moderados da normal, quer em assimetria quer em achatamento, sobretudo se n for grande). Caso o desvio da normalidade seja severo, deve usar-se uma ANOVA não paramétrica (Athayde, 2010).

A alternativa não paramétrica à ANOVA mais conhecida e usada é o teste de Kruskal--Wallis (Miller Jr., 1997). Este testa a hipótese nula (H0) de que $k \geq 2$ amostras independentes provêm da mesma população ou de populações idênticas, tendo em consideração os seus valores médios (Siegel, 1975).

As hipóteses a testar podem então ser apresentadas do seguinte modo:

H0: $\mu_1 = \mu_2 = \cdots = \mu_k$

H1: $\mu_i \neq \mu_j$, para alguns grupos i e j com $i \neq j$

Dadas k amostras aleatórias, sejam:

- $X_{i1}, X_{i2}, ..., X_{in_i}$ as observações da i-ésima amostra de tamanho n_i, $i \in \{1, 2, ..., k\}$;
- $n = n_1 + ... + n_k$ o número total de observações da amostra combinada;
- $R\left(X_{ij}\right)$ o *rank* atribuído à observação X_{ij}, $i \in \{1, ..., k\}$ e $j \in \{1, ..., n_i\}$;
- $R_i = \sum_{i=1}^{n_i} R\left(X_{ij}\right)$ a soma dos *ranks* atribuídos às observações da amostra i (Conover, 1980).

O teste de Kruskal-Wallis determina se estas somas são tão díspares que não seja provável que se refiram a amostras extraídas da mesma população (Siegel e Castellan, 1988).
A estatística de teste é dada por:

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{n(n+1)^2}{4}\right) \tag{46}$$

onde,

$$S^2 = \frac{1}{n-1}\left(\sum_{\substack{i\in\{1,\ldots,k\}\\ j\in\{1,\ldots,n_i\}}} R\left(X_{ij}\right)^2 - \frac{n(n+1)^2}{4}\right) \tag{47}$$

Se não houver empates nos *ranks*, tem-se:

$$\sum_{\substack{i\in\{1,\ldots,k\}\\ j\in\{1,\ldots,n_i\}}} R\left(X_{ij}\right)^2 = \sum_{l=1}^{n} l^2 = \frac{n(n+1)(2n+1)}{6} \tag{48}$$

então,

$$S^2 = \frac{1}{n-1}\left(\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}\right) = \frac{n(n+1)}{n-1}\left(\frac{2n-2}{24}\right) = \frac{n(n+1)}{12} \tag{49}$$

E a expressão de T fica assim simplificada para:

$$T = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1) \tag{50}$$

Quando $k=3$, todas as amostras têm dimensão inferior ou igual a 5 $(n_i \le 5)$ e não houver empates, os quantis de *T* são obtidos a partir da tabela A.8 que se encontra em Conover (1980). Nas outras situações, os quantis de *T* podem ser aproximados pelos da distribuição qui-quadrado com $k-1$ graus de liberdade. Ao nível de significância α rejeita-se H0 se *T* exceder o quantil $1-\alpha$ dado pela lei $\chi^2_{(k-1)}$ (Conover, 1980).

Quando H0 é rejeitada, é possível averiguar quais os pares de populações que têm médias significativamente distintas. Isto será feito à custa da comparação da soma dos *ranks* das respetivas amostras.

Assim, de acordo com Conover (1980), pode-se concluir que as populações *i* e *j* têm médias significativamente distintas se:

$$\left| \frac{R_i}{n_i} - \frac{R_j}{n_j} \right| > t_{1-\alpha/2} \sqrt{S^2 \frac{n-1-T}{n-k} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \tag{51}$$

em que:

- n_i é a dimensão da amostra retirada da população *i*;
- n_j é a dimensão da amostra retirada da população *j*;
- n é a dimensão da amostra combinada;
- T é a estatística de teste de Kruskal-Wallis;
- S^2 encontra-se descrita em (47);
- R_i é a soma dos *ranks* da amostra *i*;
- R_j é a soma dos *ranks* da amostra *j*;
- $t_{1-\alpha/2}$ é o quantil de ordem $1-\alpha/2$ da lei *t-student* com $n-k$ graus de liberdade, sendo α o mesmo nível de significância usado no teste de Kruskal-Wallis.

2.3. Medidas de correlação e seus testes de significância

Em muitas pesquisas surge a necessidade de verificar se dois conjuntos de dados estão relacionados e qual o grau desse relacionamento. Os coeficientes de correlação indicam o grau de associação entre dois conjuntos de dados e os testes de significância para cada coeficiente determinam, a um certo nível de probabilidade, se existe associação na população da qual se extraiu a amostra em estudo (Siegel e Castellan, 1988).

A medida usual de correlação, no caso paramétrico, é o coeficiente de correlação de Pearson entre duas variáveis X_1 e X_2, o qual mede a intensidade e a direção da associação de tipo linear entre essas variáveis e é dado por (Maroco, 2007):

$$r = \frac{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)\left(X_{2i} - \bar{X}_2\right)}{\sqrt{\sum_{i=1}^{n}\left(X_{1i} - \bar{X}_1\right)^2 \sum_{i=1}^{n}\left(X_{2i} - \bar{X}_2\right)^2}} \tag{52}$$

A aplicação do teste de significância para este coeficiente exige que os dados provenham de uma população normal bivariada (Siegel e Castellan, 1988).

O coeficiente de correlação de Spearman, r_s, é uma medida de associação não paramétrica entre duas variáveis (Maroco, 2007).

Dados *n* indivíduos ordenados em *ranks* segundo as variáveis X_1 e X_2, pode-se determinar o coeficiente de correlação de Spearman entre essas variáveis substituindo na fórmula do coeficiente de correlação de Pearson os valores das observações de X_1 e X_2 pelos respetivos *ranks* [1]. Obtém-se assim:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n^3 - n}, \tag{53}$$

onde $D_i = R_{1i} - R_{2i}$.

Em caso de empates nos *ranks*, atribui-se a cada indivíduo a média dos *ranks* que lhe seriam atribuídos caso não tivesse ocorrido empate.

[1] Esta dedução pode ser encontrada em Siegel (1975).

Segundo Siegel e Castellan (1988), se a proporção de empates não é grande o seu efeito sobre r_s é insignificante, contudo se esta proporção for grande deve-se incorporar um fator de correção para os empates de cada variável X_i $\left(T_{X_i}\right)$ ao cálculo de r_s :

$$T_{X_i} = \sum_{j=1}^{g_i} t_j^3 - t_j \tag{54}$$

onde g_i é o número de grupos de observações empatadas para a variável X_i e t_j é o número de observações empatadas em cada grupo de empates da variável X_i, obtendo-se assim, de acordo este autor:

$$r_s = \frac{\left(n^3 - n\right) - 6\sum_{i=1}^{n} d_i^2 - \frac{T_{X_1} + T_{X_2}}{2}}{\sqrt{\left(n^3 - n\right)^2 - \left(T_{X_1} + T_{X_2}\right)\left(n^3 - n\right) + T_{X_1} T_{X_2}}} \tag{55}$$

Se os indivíduos de onde foram retirados os valores observados das variáveis X_1 e X_2 para o cálculo de r_s são extraídos aleatoriamente de uma população, pode testar-se a hipótese nula de que as duas variáveis em estudo não sejam associadas na população e que o valor de r_s seja diferente de zero só por acaso.

Em Siegel e Castellan (1988) encontram-se tabelados os valores críticos de r_s obtidos por este método.

Para grandes amostras, valores de *n* superiores a aproximadamente 20 a 25, a significância de um valor obtido de r_s , sob H0, pode ser testada pela estatística:

$$z = r_s \sqrt{n-1} \tag{56}$$

Para *n* grande, o valor definido pela equação (56) tem distribuição aproximadamente normal com média 0 e desvio-padrão 1. A probabilidade associada, sob H0, a qualquer valor tão extremo quanto um valor observado de r_s pode ser obtida calculando-se o valor *z* associado àquele valor usando a equação (56) e, em seguida, determinando-se a significância do valor *z* com auxílio da tabela A que se encontra em Siegel e Castellan (1988).

CAPÍTULO 3: ESTUDO EXPERIMENTAL

No início deste capítulo será apresentada uma descrição da parcela e da base de dados em estudo. De seguida, é apresentado todo o tratamento estatístico efetuado, com aplicação das referidas técnicas estatísticas. Os *softwares* utilizados no tratamento dos dados foram o *IBM-SPSS-Statistics-19* e o *R*.

3.1. Descrição da parcela em estudo

A parcela do ensaio (C5) situa-se na Estação Vitivinícola Amândio Galhano no concelho dos Arcos de Valdevez (41º 48' de latitude Norte e 8º26'de longitude Oeste) (Mota, 2005). A parcela encontra-se a uma altitude média de 76 m, com ligeiro declive de 5 % e de exposição dominante S-SSO (Maciel, 2005).

Figura 5 - Localização da parcela C5 da EVAG (Fonte: *GoogleEarth*, 2011)

Segundo Armada (1990), o solo pertence ao grupo dos Antrossolos de Surriba Dístricos Normais, revelando-se espesso, com elevados riscos de erosão, de permeabilidade moderadamente lenta e de drenagem externa e interna regular, com teor baixo em coloides minerais e baixo a médio teor em matéria orgânica, e elevada capacidade de armazenamento de água útil (217 mm); é ácido, de teor baixo em azoto, teor muito baixo

em fósforo e potássio, baixo teor em bases de troca e muito fortemente lixiviados. A textura dominante franco-arenosa (Armada, 1990).

A vinha está instalada segundo um alinhamento no sentido Norte-Sul, e conduzida em cordão simples ascendente bilateral a 1m de altura do solo, com compasso de plantação de $2,5 \times 2$ m, o que corresponde a uma densidade de plantação de 2000 videiras/ha. Trata-se de uma vinha em condições de cultura biológica com enrelvamento permanente de gramíneas.

Para este ensaio escolheu-se a casta tinta Vinhão, casta de qualidade, tratando-se a única casta regional tintureira sendo por isso cultivada em toda a região; por outro lado, é uma casta que vem revelando sensibilidade à falta de água (e ao escaldão) nas atuais condições de cultura, isto é, em vinhas estremes e não consociadas, como tradicionalmente com culturas de regadio (ex. milho). Produz mostos ricos em açúcares, dando vinhos de cor intensa, vermelho granada, de aroma vinoso, encorpado e ligeiramente adstringente. Casta vigorosa e regular na produção tem boa afinidade com a maioria dos porta-enxertos usados na região. É uma casta de ciclo curto, sendo tardia no abrolhamento, recuperando depois na floração e na maturação, onde se situa numa posição intermédia entre as demais castas da região (Mota e Garrido, 2001).

O porta-enxerto em que está implantada esta vinha é o 1103P – *Rupestris X Berlandieri – 1103 Paulsen*. Caracteriza-se por ser bastante vigoroso e responde bem à enxertia (Duarte e Eiras-Dias, 1989). Segundo Pinho (1993), tem boa adaptação em terras argilo-calcários de subsolo fresco, resistindo bem a solos secos, ao calcário ativo e tem um sistema radicular profundo. Em condições de muito bom vigor é uma planta de frutificação regular, com desenvolvimento precoce assegurando uma maturação regular a boa.

3.2. Caracterização da amostra e variáveis

Neste trabalho, a base de dados em estudo tem cinco grupos de variáveis. Um primeiro relativo a variáveis do solo, o segundo relativo à planta (videiras), o terceiro ao mosto, um quarto grupo relativo às uvas e um quinto grupo relativo ao produto final (vinho).

O primeiro grupo é referente a um conjunto de variáveis relativas às características do solo da parcela em análise. Na parcela em estudo foram georreferenciados 45 pontos, com espaçamento regular, resultando numa malha de 15 linhas por 3 colunas, como se pode verificar na Figura 6 e no Anexo 1.

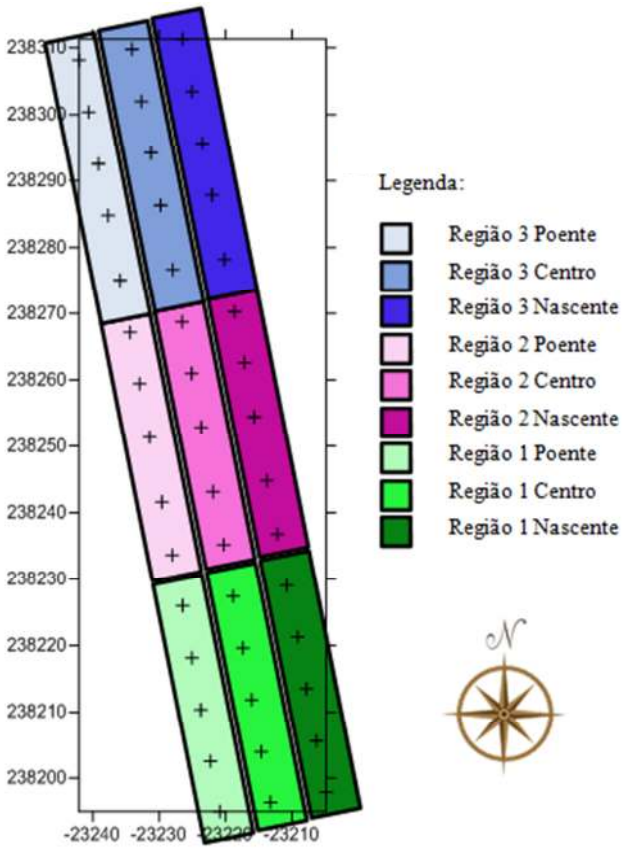


Figura 6 - Parcela com localização dos 45 pontos georreferenciados

As amostras de solo foram recolhidas em cada um dos 45 pontos georreferenciados à profundidade de 15-20 cm e para cada uma foram avaliadas dezasseis variáveis, a saber: pH em extrato aquoso (pH); matéria orgânica, em % (MO); densidade aparente, em gramas de solo seco por volume de solo não perturbado (DA); fração fina, em % (FF); fração grosseira, em % (FG); fósforo assimilável, em µg/g (P_2O_5); potássio assimilável, em µg/g

(K_2O); cálcio assimilável, em µg/g (Ca); magnésio assimilável, em µg/g (Mg); azoto total, em % (AzT); níquel, em µg/g (Ni); crómio, em µg/g (Cr); cádmio, em µg/g (Cd); nitratos, em µg/g (N); boro, em µg/g (B) e capacidade de troca catiónica, em m.e./100g (CTC).

O segundo grupo de variáveis é relativo à produtividade e vigor das videiras. A produção e vigor das videiras foram avaliados em consonância com os 45 pontos georreferenciados. Entre as componentes da produção das videiras foram avaliadas três variáveis: número de cachos por videira (Ncachos); peso médio do cacho por videira, em kg (Pcacho_kg) e peso de uvas por videira, em kg (Uvas_kg_vid). O vigor das videiras foi definido através de três variáveis: número de varas por videira (Nvaras); peso das varas por videira, em kg (Pvaras_kg) e peso médio da vara por videira, em gramas (Pvara_g).

O terceiro grupo de seis variáveis diz respeito às características físico-químicas do mosto. Em relação ao mosto, foram recolhidas nove amostras compósitas de bagos de acordo com as nove localizações esquematizadas na Figura 6. Para a caracterização do mosto foram efetuadas as seguintes análises de acordo com o Regulamento (CEE) nº 2676/90, de 17 de setembro de 1990: pH do mosto (pH_mosto); acidez total, em g/dm^3 (Acidez_T); ácido málico, em g/dm^3 (Acido_malico); ácido tartárico, em g/dm^3 (Acido_tartarico); açúcares, em g/dm^3 (Açucares) e teor de álcool provável, em % vv (TAP).

O quarto grupo de variáveis engloba os compostos voláteis existentes nas uvas. A análise da composição aromática e cromática foi também realizada em nove amostras compósitas de bagos de acordo com as nove localizações esquematizadas na Figura 6. Estas amostras de uvas deram origem a sumos, dos quais foram extraídos os compostos do aroma na forma livre e na forma de glicoconjugados, obtendo-se extratos que foram analisados por GC-MS (*Gas-chromatography-mass spectrometry*). A identificação dos compostos voláteis foi efetuada com recurso ao programa MS WorkStation versão 6.6 (Varian) comparando os espectros de massas e os índices de retenção com os de compostos de referência puros. Todos os compostos foram quantificados como equivalentes de 4--nonanol. Todas as análises foram efetuadas em triplicado encontrando-se, no Anexo 2, os valores médios das concentrações destas análises para todos os compostos assim como o agrupamento destes por famílias químicas.

Foram identificados e quantificados um total de vinte e dois compostos do aroma na forma livre que foram agrupados por 5 famílias. Estas famílias de compostos do aroma na forma

livre foram as variáveis em estudo: família dos cinco compostos em C_6 (FL1), família dos sete álcoois (FL2), família dos quatro álcoois monoterpénicos (FL3), família dos quatro fenóis voláteis (FL4) e família dos dois compostos carbonilados (FL5). Temos ainda um total de quarenta e três agliconas odoríferas nos extratos da fração glicosilada agrupadas em 7 famílias que correspondem às variáveis: família dos três compostos em C_6 (FG1), família dos sete álcoois (FG2), família dos quatro álcoois monoterpénicos (FG3), família dos sete óxidos e dióis monoterpénicos (FG4), família dos dez norisoprenóides em C_{13} (FG5), família dos onze fenóis voláteis (FG6) e família de apenas um composto carbonilado (FG7).

Finalmente o quinto grupo de variáveis é relativo às características organoléticas do vinho, obtidas por um painel de sete provadores. Foi efetuada a prova de cada um dos nove vinhos correspondentes às nove regiões assinaladas na Figura 6. Cada provador emitiu apreciações pessoais e independentes ao nível visual, olfativo e gustativo de cada vinho. Ao nível visual foram avaliadas a limpidez (limpidez_v) e a cor (cor_v), ao nível olfativo avaliou-se a limpidez (limpidez_o), a intensidade (intensidade_o) e a qualidade (qualidade_o) e ao nível gustativo, a limpidez (limpidez_g), a intensidade (intensidade_g), a persistência (persistência_g) e a qualidade (qualidade_g). Também foi pedida uma apreciação global do vinho (Avaliacao_Global). Por fim, a variável Nota_Final representa a soma de todas as classificações anteriores, tendo-se trabalhado apenas com esta variável por ser uma ponderação de todos os fatores avaliados em relação ao vinho.
Foi utilizada a ficha de prova descritiva da Câmara de Provas da CVRVV. A ficha de prova utilizada encontra-se no Anexo 3.

3.3. Análise e discussão dos resultados

Em primeiro lugar começou-se por fazer a análise dos valores em falta na base de dados. As variáveis que apresentaram valores em falta e as respetivas percentagens encontram-se assinaladas na Tabela 1.

	Variáveis	Número de valores em falta	Ponto de amostragem	%
SOLO	DA	2	A42 e A45	4,44
	FF e FG	1	A14	2,22
VIDEIRA	Ncachos Peso_kg_vid Pcacho_kg nvaras pvaras_kg pvara_g	4	A13, A28, A29 e A30	8,89

Tabela 1 - Valores em falta das variáveis em estudo

Relativamente aos valores em falta, a análise estatística dos resultados foi efetuada, em paralelo, de três formas distintas:

- substituindo estes valores pelo método do vizinho mais próximo;
- substituindo os valores em falta pelo método de interpolação por *krigagem*;
- retirando os indivíduos que apresentavam valores em falta.

Dado que os resultados foram semelhantes nas três análises efetuadas optou-se por realizar o estudo retirando os indivíduos que apresentavam valores em falta pois, para além da percentagem de valores em falta ser baixa, nos métodos do vizinho mais próximo e da interpolação por *krigagem* os valores em falta são substituídos por valores estimados, que não são os valores reais, pelo que poderão inserir maior variabilidade nos resultados.

De seguida, fez-se uma análise exploratória dos dados com o intuito de perceber o comportamento de cada uma das variáveis e até mesmo detetar a possível existência de *outliers*.

Na Tabela 2 apresenta-se um resumo com algumas das principais características amostrais relativas a cada uma das variáveis analisadas.

		Média	Mediana	Desvio Padrão	Mínimo	Máximo	Percentis	
							25	75
SOLO	pH	5,76	5,73	0,36	5,14	6,60	5,45	6,04
	MO	3,48	2,73	1,84	1,85	10,11	2,19	4,13
	DA	1,14	1,16	0,11	0,90	1,32	1,07	1,23
	FF	68,38	68,33	5,64	54,33	83,70	64,39	71,67
	FG	31,62	31,67	5,64	16,30	45,67	28,33	35,61
	P_2O_5	41,50	37,11	17,61	10,95	93,80	26,24	53,61
	K_2O	86,37	82,19	25,03	46,07	149,03	65,80	107,11
	Ca	482,63	450,00	168,32	244,50	835,50	350,25	624,00
	Mg	72,77	73,50	12,04	52,50	108,00	63,75	81,00
	AzT	0,15	0,12	0,07	0,08	0,43	0,10	0,17
	Ni	3,76	3,68	1,37	0,96	6,00	2,64	4,96
	Cr	0,43	0,43	0,21	0,02	0,94	0,25	0,62
	Cd	0,11	0,10	0,03	0,07	0,22	0,10	0,13
	N	1,33	1,36	0,26	0,94	1,99	1,12	1,49
	B	0,42	0,40	0,10	0,29	0,79	0,35	0,45
	CTC	11,34	10,92	3,26	6,60	18,25	8,61	14,02
MOSTO	pH_mosto	3,32	3,33	0,08	3,21	3,44	3,24	3,36
	Acidez_T	5,94	6,00	0,55	5,08	6,96	5,75	6,15
	Acido_malico	2,09	1,90	0,52	1,60	3,40	1,80	2,10
	Acido_tartarico	4,51	4,50	0,28	4,10	5,10	4,30	4,60
	Açucares	216,39	217,00	3,61	207,80	221,60	215,90	218,20
	TAP	12,86	12,89	0,21	12,35	13,17	12,83	12,96
VIDEIRA	Ncachos	38,07	34,00	13,58	12,00	68,00	29,00	47,00
	Uvas_kg_vid	5,12	4,75	2,15	1,35	9,65	3,45	7,20
	Pcacho_kg	0,13	0,14	0,03	0,08	0,18	0,11	0,15
	Nvaras	31,88	31,00	7,17	22,00	52,00	26,50	35,00
	Pvaras_kg	2,47	2,40	0,92	1,00	4,80	1,83	2,90
	Pvara_g	76,80	77,14	21,21	35,71	154,84	61,39	88,68
UVAS	FL1	101,21	100,55	20,83	69,15	135,86	88,82	115,75
	FL2	263,42	236,17	58,18	195,98	373,46	219,12	297,03
	FL3	22,17	6,38	33,97	3,24	113,02	4,35	17,79
	FL4	7,70	8,33	2,73	3,32	11,26	5,68	10,06
	FL5	14,24	7,88	16,10	4,71	58,26	5,84	12,12
	FG1	32,55	31,18	9,39	19,68	52,35	28,86	33,65
	FG2	155,25	165,14	34,99	81,24	191,58	140,09	181,18
	FG3	3,69	3,51	1,17	1,95	5,48	2,78	4,82
	FG4	10,46	11,20	2,53	5,29	14,17	9,96	11,78
	FG5	36,41	40,77	10,25	15,53	51,27	29,70	41,75
	FG6	47,27	51,34	15,40	19,70	73,42	48,18	52,91
	FG7	0,98	1,13	0,32	0,45	1,37	0,63	1,20
VINHO	Nota_Final	64,44	65,00	9,06	48,00	76,00	59,00	71,00

Tabela 2 - Algumas das principais características amostrais das variáveis em análise

Para complementar esta análise elaboraram-se os diagramas de extremos e quartis relativos às variáveis do solo.

O facto da escala dos dados variar torna estes gráficos de difícil interpretação. Uma solução natural é proceder a um reescalonamento dos dados, por exemplo usando como unidade a escala intrínseca de cada amostra. O mais usual é puxar a localização para a

origem (centrar os dados), e dividi-los pela sua escala (o que se chama reduzir os dados) por forma a terem escala unitária. O resultado é a amostra "estandardizada" (Pestana e Velosa, 2008).

Obtém-se assim os diagramas de extremos e quartis de mais fácil leitura para as dezasseis variáveis do solo em estudo, representados na Figura 7.

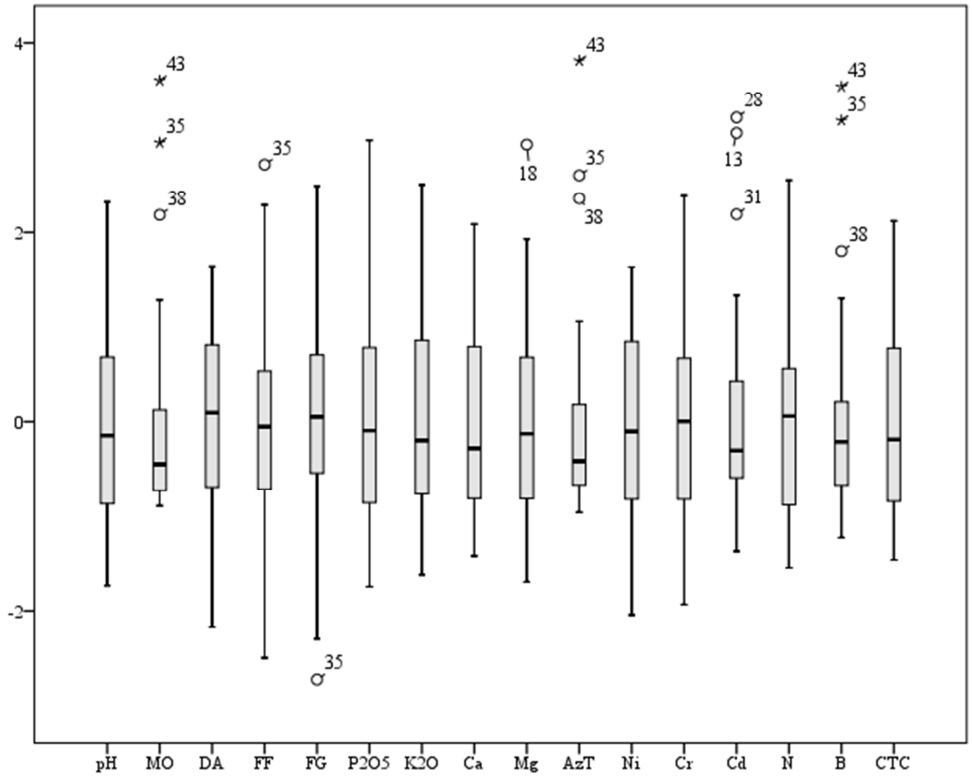


Figura 7 - Diagrama de extremos e quartis das variáveis do solo estandardizadas

Graficamente consegue perceber-se que algumas características do solo apresentam maior variabilidade que outras e que relativamente às variáveis MO, FF, FG, Mg, AzT, Cd e B verifica-se a presença de *outliers*. Pode-se mesmo verificar a existência de *outliers* severos relativamente às variáveis MO, AzT e B.

Os *outliers* severos aparecem nos pontos de amostragem A35 e A43. Efetuou-se então a análise em estudo sem estes indivíduos e os resultados foram similares pelo que se optou por trabalhar com estes valores pois a sua eliminação implicaria perda de informação.

Analisou-se de seguida a normalidade dos dados para cada variável em estudo através do teste de Shapiro-Wilk. Os resultados obtidos da aplicação deste teste encontram-se na Tabela 3.

		Estatística de teste	Graus de liberdade	Valor prova
SOLO	pH	0,973	42	0,401
	MO	0,738	42	**< 0,001** *
	DA	0,966	42	0,232
	FF	0,974	42	0,446
	FG	0,974	42	0,446
	P_2O_5	0,965	42	0,227
	K_2O	0,968	42	0,273
	Ca	0,939	42	**0,026** *
	Mg	0,968	42	0,289
	AzT	0,744	42	**< 0,001** *
	Ni	0,963	42	0,182
	Cr	0,975	42	0,473
	Cd	0,845	42	**< 0,001** *
	N	0,953	42	0,081
	B	0,818	42	**< 0,001** *
	CTC	0,946	42	**0,048** *
MOSTO	pH_mosto	0,885	45	**< 0,001** *
	Acidez_T	0,914	45	**0,003** *
	Acido_malico	0,737	45	**< 0,001** *
	Acido_tartarico	0,904	45	**0,001** *
	Açucares	0,816	45	**< 0,001** *
	TAP	0,814	45	**< 0,001** *
VIDEIRA	Ncachos	0,958	41	0,138
	Uvas_kg_vid	0,930	41	**0,014** *
	Pcacho_kg	0,966	41	0,248
	Nvaras	0,914	41	**0,004** *
	Pvaras_kg	0,921	41	**0,008** *
	Pvara_g	0,931	41	**0,015** *
UVAS	FL1	0,941	45	**0,024** *
	FL2	0,872	45	**< 0,001** *
	FL3	0,566	45	**< 0,001** *
	FL4	0,891	45	**0,001** *
	FL5	0,547	45	**< 0,001** *
	FG1	0,886	45	**< 0,001** *
	FG2	0,839	45	**< 0,001** *
	FG3	0,910	45	**0,002** *
	FG4	0,902	45	**0,001** *
	FG5	0,900	45	**0,001** *
	FG6	0,816	45	**< 0,001** *
	FG7	0,827	45	**< 0,001** *
VINHO	Nota_Final	0,911	45	**0,002** *

* significativo a 5%

Tabela 3 - Resultados do teste de normalidade de Shapiro-Wilk

Consultando os valores prova obtidos na aplicação deste teste, conclui-se que há evidência estatística suficiente para rejeitar a normalidade dos dados no que concerne às variáveis: MO, Ca, AzT, Cd, B, CTC, pH_mosto, Acidez_T, Acido_malico, Acido_tartarico, Açucares, TAP, Uvas_kg_vid, Nvaras, Pvaras_kg, Pvara_g, FL1, FL2, FL3, FL4, FL5, FG1, FG2, FG3, FG4, FG5, FG6, FG7 e Nota_Final.

Com o objetivo de verificar se existe um pequeno número de variáveis que seja responsável por explicar uma proporção elevada da variação total associada ao conjunto original de dados do solo, e assim reduzir a dimensionalidade do problema relativamente a estas variáveis, levou-se a cabo uma Análise de Componentes Principais (ACP) tendo-se para isso verificado se estas variáveis estão correlacionadas.

Neste sentido, começou-se por construir a matriz de correlações dos dados. A escolha da matriz de correlações para derivação das componentes principais (CPs), em lugar da matriz de covariâncias, tem a ver com o facto de existirem diferentes unidades de medida para as variáveis. Esta matriz está apresentada na Tabela 4.

	pH	MO	DA	FF	FG	P_2O_5	K_2O	Ca	Mg	AzT	Ni	Cr	Cd	N	B	CTC
pH	1,00															
MO	-0,30	1,00														
DA	-0,11	-0,04	1,00													
FF	-0,44	0,44	0,03	1,00												
FG	0,44	-0,44	-0,03	**-1,00**	1,00											
P_2O_5	0,47	-0,40	0,15	-0,27	0,27	1,00										
K_2O	0,07	0,14	-0,01	0,23	-0,23	0,18	1,00									
Ca	**0,77**	0,02	-0,10	-0,19	0,19	0,43	0,20	1,00								
Mg	0,42	-0,22	0,13	-0,10	0,10	**0,53**	0,22	**0,73**	1,00							
AzT	-0,19	**0,99**	-0,05	0,39	-0,39	-0,36	0,15	0,12	-0,17	1,00						
Ni	-0,25	**0,56**	-0,23	0,42	-0,42	-0,30	0,09	-0,18	-0,34	**0,55**	1,00					
Cr	0,19	-0,28	-0,02	-0,42	0,42	0,02	-0,18	0,04	0,01	-0,27	**-0,54**	1,00				
Cd	0,26	-0,11	0,23	-0,06	0,06	0,41	0,24	0,26	0,31	-0,09	-0,11	0,02	1,00			
N	0,22	0,24	0,05	-0,01	0,01	0,36	0,08	**0,53**	0,38	0,28	-0,17	0,07	0,26	1,00		
B	-0,28	**0,99**	-0,02	0,47	-0,47	-0,36	0,13	0,04	-0,19	**0,98**	**0,54**	-0,28	-0,08	0,26	1,00	
CTC	**0,74**	0,00	-0,07	-0,16	0,16	0,46	0,28	**0,99**	**0,78**	0,10	-0,20	0,02	0,29	**0,52**	0,02	1,00

Tabela 4 - Matriz de correlações das variáveis do solo

A leitura da Tabela 4 permite verificar que existe uma forte correlação positiva entre a variável pH com as variáveis Ca e CTC, entre a variável MO e as variáveis AzT e B (quase perfeitas), entre a variável Ca e as variáveis Mg e CTC (esta última quase perfeita), entre as variáveis Mg e CTC e entre as variáveis AzT e B (quase perfeita). Também estão correlacionadas de forma moderadamente positiva as variáveis P_2O_5 e Mg, a variável N com as variáveis Ca e CTC e a variável Ni com as variáveis MO, AzT e B.

Verifica-se ainda uma correlação negativa perfeita entre as variáveis FF e FG e uma correlação moderadamente negativa entre as variáveis Ni e Cr.

A análise da matriz de correlações revela a existência de variáveis perfeita e excessivamente correlacionadas. Relativamente às variáveis FF e FG, que apresentam uma correlação perfeita negativa, uma delas (FG) foi eliminada da análise porque ambas dão informação sobre a granulometria do terreno e os seus valores são complementares.

A adequação desta técnica de Estatística Multivariada ao conjunto de dados em estudo foi avaliada através do teste de esfericidade de Bartlett. Este testa a hipótese da matriz de correlações ser uma matriz identidade e o seu determinante ser igual a um, logo, de as variáveis não estarem correlacionadas entre si.

A estatística de teste é dada por $-\left[n-1-\left(2p+5\right)\div 6\right]\times \ln\left|\boldsymbol{R}\right|$, onde n é a dimensão da amostra (= 42), p é o número de variáveis em análise (= 15) e $\boldsymbol{R}$ é a matriz de correlações empírica. Esta estatística tem uma distribuição assimptótica de χ^2 com $\left[p\times\left(p-1\right)\div 2\right]$ graus de liberdade (Reis, 2001).

Estes cálculos foram realizados com recurso ao *software R*, tendo-se obtido:

Determinante da matriz de correlações empírica:

```
> R=cor(dadossolo,use='complete.obs')
> det(R)
[1] 2.080077e-28
```

Estatística de teste:

```
> chi_square= -(41-(2*15+5)/6)*log(det(R),base=exp(1))
> chi_square
[1] 2241.523
```

Graus de liberdade:

```
> df= (15*14)/2
> df
[1] 105
```

Valor prova:

```
> pv= 1-pchisq(chi_square,df)
> pv
[1] 0
```

Consultando o valor prova associado a este teste, conclui-se que há evidência estatística suficiente para rejeitar a hipótese nula, ou seja, conclui-se que existe correlação significativa entre as variáveis do solo, pelo que faz sentido levar a cabo uma ACP.
O determinante da matriz de correlações é próximo de zero o que também evidencia que as variáveis em estudo estão suficientemente correlacionadas.

Procedeu-se então à derivação das CPs começando por calcular a proporção de variância explicada pelas novas variáveis. Os resultados obtidos encontram-se na Tabela 5.

Componente	Valores Próprios Iniciais			Soma dos quadrados dos *loadings* após a extração			Soma dos quadrados dos *loadings* após a rotação		
	Total	% de Variância	% acumulada	Total	% de Variância	% acumulada	Total	% de Variância	% acumulada
1	4,764	31,757	31,757	4,764	31,757	31,757	4,113	27,417	27,417
2	3,580	23,870	55,627	3,580	23,870	55,627	3,591	23,942	51,359
3	1,489	9,929	65,556	1,489	9,929	65,556	1,887	12,582	63,942
4	1,283	8,551	74,107	1,283	8,551	74,107	1,525	10,166	74,107
5	0,829	5,527	79,634						
6	0,814	5,426	85,060						
7	0,678	4,522	89,583						
8	0,531	3,538	93,120						
9	0,374	2,491	95,612						
10	0,336	2,242	97,854						
11	0,245	1,634	99,488						
12	0,062	0,412	99,900						
13	0,014	0,093	99,993						
14	0,001	0,007	100						
15	6,41E-15	4,27E-14	100						

Tabela 5 - Variância Total Explicada

Analisando a linha relativa à Proporção Acumulada, pelo critério da variância total explicada, devem reter-se as quatro primeiras CPs, as quais explicam aproximadamente 74,11% da variabilidade dos dados. Esta conclusão é ainda válida utilizando o critério de Kaiser aplicado à matriz de correlações pois, de acordo com a Tabela 9, os valores próprios são superiores a 1 até à quarta CP.

Analisando o *screeplot* da Figura 8, pode-se observar um cotovelo com articulação de "3" a "5" e, de acordo como os dois critérios anteriores que foram consensuais, optou-se por reter as quatro primeiras CPs.

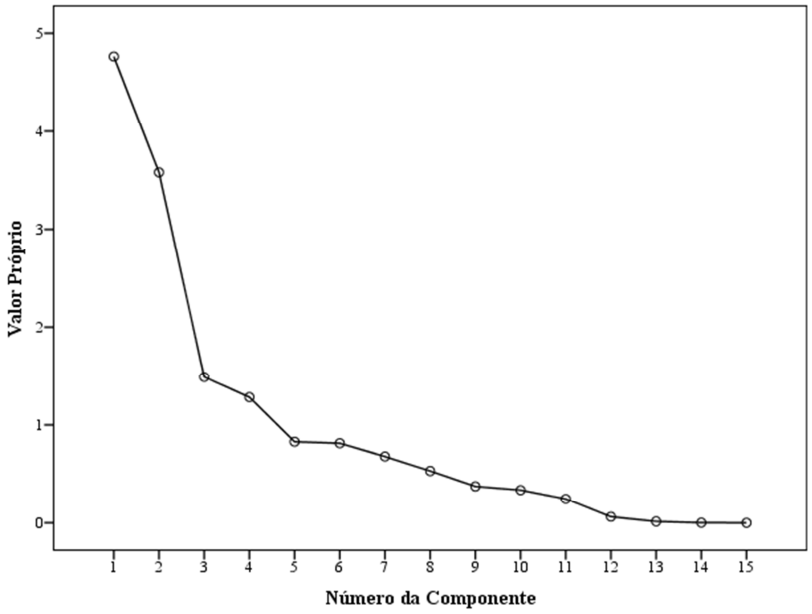


Figura 8 - *Scree Plot*

Na Tabela 6 encontram-se os *loadings* para as quatro primeiras CPs, ou seja, os coeficientes das combinações lineares de cada uma das CPs em função das variáveis originais. Para derivação das CPs efetuou-se a rotação Varimax dado que a interpretação destas componentes ficou mais de acordo com a opinião dos especialistas da área.

	Componente			
	1	2	3	4
pH	**0,796**	-0,244	-0,167	-0,188
MO	-0,068	**0,965**	0,192	-0,054
DA	-0,139	0,000	-0,057	**0,819**
FF	-0,255	0,390	**0,591**	0,123
P_2O_5	**0,583**	-0,398	0,086	0,365
K_2O	0,287	0,024	**0,584**	0,157
Ca	**0,967**	0,105	-0,032	-0,047
Mg	**0,762**	-0,169	0,082	0,278
AzT	0,032	**0,965**	0,171	-0,084
Ni	-0,223	0,440	**0,589**	-0,357
Cr	0,059	-0,135	**-0,795**	0,058
Cd	0,355	-0,111	0,173	**0,558**
N	**0,567**	0,407	-0,197	0,338
B	-0,048	**0,961**	0,193	-0,014
CTC	**0,974**	0,077	0,021	0,000

Tabela 6 - Matriz dos *loadings* das quatro primeiras CPs com rotação Varimax

Da análise desta tabela conclui-se que:

- A primeira CP tem *loadings* positivos elevados nas variáveis pH, P_2O_5, Ca, Mg, N e CTC;
- A segunda CP tem *loadings* positivos elevados nas variáveis MO, AzT e B;
- A terceira CP tem *loadings* positivos elevados nas variáveis FF, K_2O e Ni, e *loading* negativo elevado na variável Cr;
- A quarta CP tem *loadings* positivos elevados nas variáveis DA e Cd.

Verifica-se ainda, pela análise da Tabela 7, que as variáveis originais que mais informação perdem quando se transformam as quinze variáveis originais em quatro componentes são K_2O e Cd. As restantes variáveis originais apresentam valores elevados de variância explicada pelas CPs.

	Comunalidades
pH	0,755
MO	0,976
DA	0,693
FF	0,581
P_2O_5	0,639
K_2O	**0,448**
Ca	0,949
Mg	0,693
AzT	0,968
Ni	0,718
Cr	0,657
Cd	**0,480**
N	0,640
B	0,964
CTC	0,955

Tabela 7 - Comunalidades das variáveis originais (do solo)

Para tentar encontrar agrupamentos nos dados e reduzir a dimensionalidade do problema para dois ou três é útil fazer os diagramas de dispersão dos *scores* das primeiras duas (ou três) CPs. Assim, determinaram-se os *scores* das quatro primeiras CPs retidas, que se encontram na Tabela 8.

	CP1	CP2	CP3	CP4
A1	0,67	-0,18	**-1,97**	0,53
A2	0,26	-0,25	-0,10	-0,18
A3	-0,28	-0,40	**-1,08**	**1,08**
A4	0,98	-0,54	0,10	0,26
A5	-0,36	-0,21	-0,78	0,54
A6	**-1,01**	0,25	**-1,47**	**1,23**
A7	0,70	-0,54	-0,74	-0,57
A8	0,15	-0,30	**-1,23**	0,79
A9	**-1,12**	-0,80	0,05	0,52
A10	0,89	-0,62	-0,91	-0,34
A11	**1,52**	**-1,01**	0,89	0,22
A12	0,86	0,77	-0,36	0,39
A13	0,58	-0,48	-0,15	0,95
A14	*	*	*	*
A15	**1,10**	-0,19	-0,40	0,95
A16	**1,10**	-0,09	**-1,54**	0,53
A17	-0,99	0,37	**-1,59**	0,16
A18	**1,41**	0,24	0,38	**1,24**
A19	**-1,23**	-0,60	-0,96	-0,76
A20	0,45	-0,37	-0,64	**-1,21**
A21	**1,73**	-0,23	0,61	0,28
A22	**-1,44**	-0,93	-0,03	-0,49
A23	**-1,27**	-0,31	-0,93	0,64
A24	**1,96**	-0,01	**1,46**	**-1,50**
A25	-0,99	**-1,06**	0,54	-0,02
A26	-0,68	-0,71	0,32	**-1,92**
A27	0,03	-0,63	**-1,16**	**-1,22**
A28	-0,08	-0,88	**2,01**	**2,04**
A29	-0,12	**-1,25**	**1,26**	-0,39
A30	0,33	-0,41	-0,23	-0,95
A31	0,05	0,08	**1,91**	**2,66**
A32	**-1,23**	0,07	0,52	-0,23
A33	**1,28**	-0,39	0,94	-0,79
A34	-0,61	1,44	0,54	0,85
A35	**-1,39**	**2,60**	**1,38**	0,90
A36	**-1,64**	0,87	0,67	0,05
A37	-0,32	-0,29	0,48	0,10
A38	0,62	**1,80**	0,54	**-1,31**
A39	-0,73	-0,67	**1,84**	**-1,79**
A40	-0,64	0,45	0,42	-0,59
A41	**-1,18**	0,82	-0,07	-0,93
A42	**	**	**	**
A43	**1,32**	**4,01**	-0,68	-0,71
A44	-0,69	0,57	0,15	**-1,03**
A45	**	**	**	**

* A14 apresenta valores em falta para a variável FF

** A42 e A45 apresentam valores em falta para a variável DA

Tabela 8 - Matriz dos *scores* das quatro primeiras CPs

Para a primeira CP, os pontos de amostragem de solo A11, A15, A16, A18, A21, A24, A33 e A43 são os que apresentam *scores* mais elevados. Estando esta CP fortemente correlacionada de forma positiva com as variáveis pH, P_2O_5, Ca, Mg, N e CTC, então é nestes pontos de amostragem que estas variáveis apresentam maiores concentrações. Para esta componente os *scores* mais baixos observam-se nos pontos de amostragem A6, A9, A19, A22, A23, A32, A35, A36 e A41, pelo que é nestes pontos que se encontram as menores concentrações destas variáveis.

A segunda CP tem *loadings* positivos elevados para as variáveis MO, AzT e B. De acordo com a Tabela 8, tem-se que estas variáveis apresentam maiores concentrações nos pontos de amostragem A35, A38 e A43 (*scores* mais elevados) e concentrações menores nos pontos A11, A25 e A29 (*scores* mais baixos).

A terceira CP apresenta *loadings* positivos elevados para as variáveis FF, K_2O e Ni e *loading* negativo elevado para a variável Cr. Por observação da Tabela 8 tem-se que os pontos de amostragem A24, A28, A29, A31, A35 e A39 (*scores* mais elevados) apresentam as maiores concentrações de FF, K_2O e Ni e as menores concentrações de Cr, e que nos pontos A1, A3, A6, A8, A16, A17 e A27 (*scores* mais baixos) se encontram as menores concentrações de FF, K_2O e Ni e as maiores concentrações de Cr.

Relativamente à quarta CP, que apresenta *loadings* positivos elevados para as variáveis DA e Cd, é nos pontos de amostragem de solo A3, A6, A18, A28 e A31 (*scores* mais elevados) que se verificam as maiores concentrações destas variáveis, e as menores concentrações encontram-se nos pontos A20, A24, A26, A27, A38, A39 e A44 (*scores* mais baixos).

A Figura 9 corresponde a uma representação aproximada da nuvem de pontos originais, fazendo corresponder os *scores* das duas primeiras CPs que explicam aproximadamente 51,36% da variabilidade total dos dados.

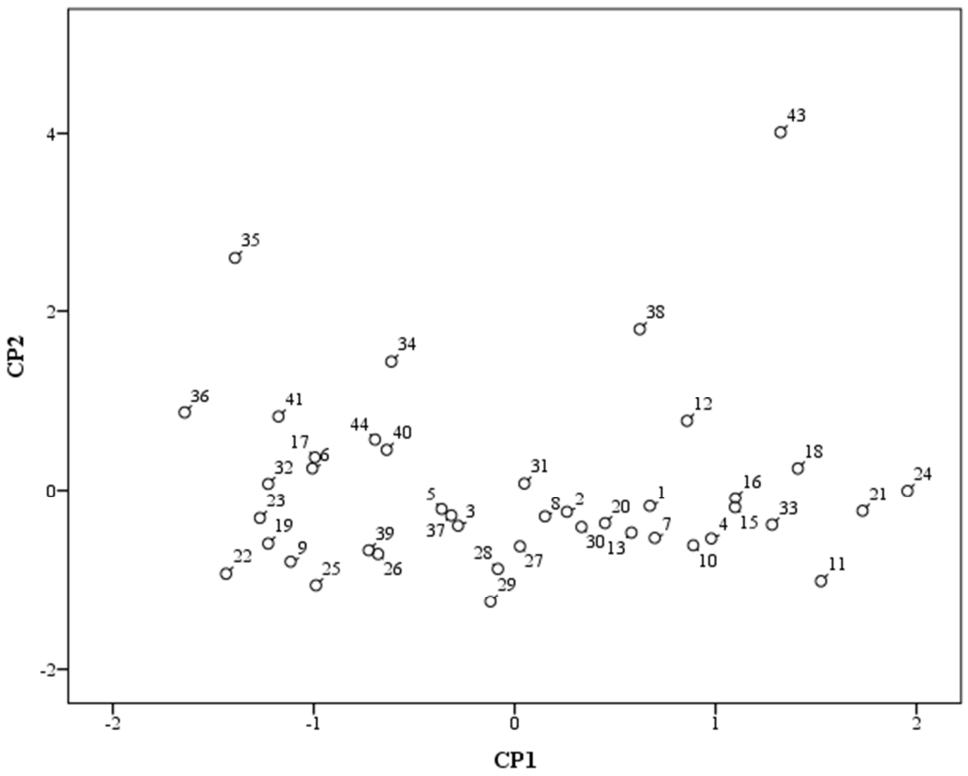


Figura 9 - Diagrama de dispersão das duas primeiras CPs

Por observação direta deste gráfico não é visível uma divisão clara dos dados em grupos.

De seguida, apresenta-se o *biplot* manual da ACP relativo às duas primeiras CPs, construído com auxílio do *software* estatístico *R* e com base no código para construção de *biplots* apresentado por Ferreira (2010).

Biplot manual[2]:

```
> N<-nrow(cps)
> xmin <- min(cps[,1])
> xmax <- max(cps[,1])
> ymin <- min(cps[,2])
> ymax <- max(cps[,2])
> plot(c(xmin,xmax),c(ymin,ymax),xlab="1ª Componente Principal (27,4%)",
ylab="2ª Componente Principal (23,9%)",type="n")
> text(cps[,1],cps[,2],1:N,cex=0.7)
```

[2] No código apresentado "cps" representa uma *dataframe* com os *scores* das quatro primeiras CPs e "Var" um vetor com os nomes das variáveis do solo.

```
> abline(v=0,lty=2); abline(h=0, lty=2)
> x1.min = min(0,min(loadings[,1]))
> x1.max = max(0,max(loadings[,1]))
> y1.min = min(0,min(loadings[,2]))
> y1.max = max(0,max(loadings[,2]))
> x1.scale = max(abs(xmax),abs(xmin))/max(abs(x1.max),abs(x1.min))*0.75
> y1.scale = max(abs(ymax),abs(ymin))/max(abs(y1.max),abs(y1.min))*0.75
> arrows(rep(0,100),rep(0,100),loadings[,1]*x1.scale,
loadings[,2]*y1.scale, col=2,lwd=2,length=0.15,angle=20)
> text(loadings[,1]*x1.scale, loadings[,2]*y1.scale,
labels=abbreviate(Var), cex=0.9)
```

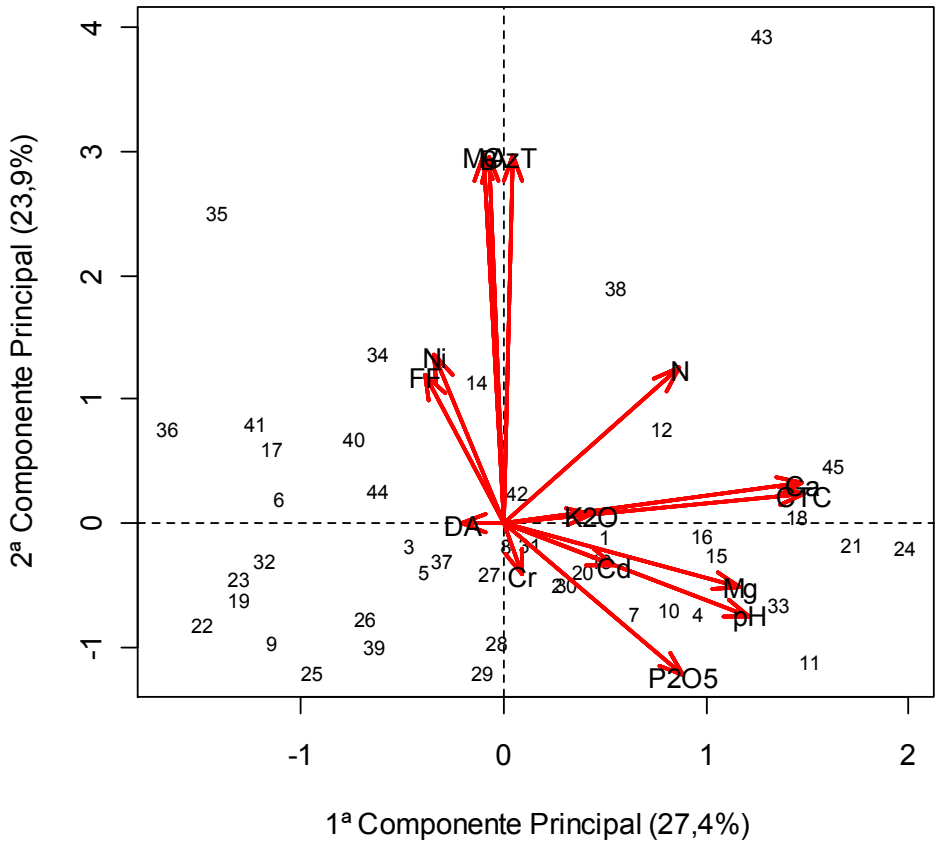


Figura 10 - *Biplot* das duas primeiras CPs

Da análise do gráfico da Figura 10 pode-se perceber uma correlação forte entre as variáveis AzT, B e MO, uma vez que as setas associadas aos respetivos marcadores apontam no mesmo sentido. O mesmo se pode concluir relativamente às variáveis Ca, CTC e K_2O, às variáveis Ni e FF e às variáveis Cd e pH.

A relação entre os marcadores de variáveis e os eixos associados à CP1 (horizontal) e à CP2 (vertical) permite ainda tirar algumas conclusões. Assim, o facto das setas que servem de marcadores correspondentes às variáveis Ca e CTC serem quase horizontais reflete a existência de uma correlação forte dessas variáveis com a CP1. Podemos ainda concluir que a variável K_2O se apresenta correlacionada positivamente com esta CP e a variável DA apresenta uma correlação negativa com CP1. Da mesma forma se conclui que as variáveis MO, AzT e B estão fortemente correlacionadas com a CP2 pelo facto dos marcadores dessas variáveis serem aproximadamente verticais. A variável Cr apresenta-se correlacionada negativamente com a segunda CP.

Da leitura deste gráfico podemos ainda observar que a variância das variáveis AzT, B, MO, Ca e CTC é maior que as restantes. Esta conclusão advém do facto dos respetivos marcadores terem maior comprimento.

A representação bidimensional das duas primeiras CPs é apenas uma aproximação da representação dos dados originais, pelo que é necessário ter prudência na leitura deste tipo de gráficos. Neste caso a proporção de variabilidade dos dados originais explicada pelas duas primeiras CPs é de aproximadamente 51,36%, pelo que as conclusões retiradas por observação gráfica podem não corresponder totalmente à realidade.

Procedeu-se de seguida a uma Análise de *Clusters* (AC) com o objetivo de identificar, na parcela, zonas com características semelhantes ao nível das variáveis do solo.

Dado que as primeiras quatro CPs explicam mais de 70% da variabilidade total dos dados originais então, segundo Chatfield e Collins (1995), os *scores* destas primeiras CPs podem ser usados em análises posteriores e, de acordo com Jolliffe (2002), a Análise de *Clusters* é uma das técnicas multivariadas onde mais frequentemente se realiza uma redução da dimensionalidade preliminar, referindo ainda que em muitos estudos se opta por utilizar as primeiras *m* CPs para tentar encontrar agrupamentos nos dados.

Assim, neste trabalho utilzaram-se os *scores* das quatro primeiras CPs para efetuar a AC.

Dado que se pretende um agrupamento de indivíduos utilizou-se uma medida de dissemelhança. Para os vários métodos aplicados efetuou-se a análise com diferentes distâncias e a alteração do tipo de distância não provocou grandes alterações nos resultados. Optou-se pelo quadrado da distância euclidiana visto ser a mais usual e

adequada aos dados, pois todas as variáveis (CPs) são de natureza contínua, têm a mesma variância (= 1) e não estão correlacionadas.

Foram aplicados vários métodos hierárquicos aglomerativos e os resultados obtidos na composição dos *clusters* foram semelhantes para os métodos da ligação completa, da ligação média e de Ward, pelo que é possível concluir que os resultados obtidos são fidedignos. Utilizou-se o método de Ward para identificar os grupos por ser aquele que tem em consideração a menor perda de informação resultante da aglutinação.

A tabela que mostra como o agrupamento se foi fazendo encontra-se no Anexo 4. Em cada etapa agrupam-se dois indivíduos, como se está a trabalhar com 42 indivíduos (pontos de amostragem de solo) realizaram-se 41 etapas de fusão.
Da análise da tabela de agrupamento (Anexo 4) pode-se verificar que na primeira etapa se agruparam os indivíduos 7 e 10 pois são os menos distantes, como se pode constatar pelas colunas "*clusters* que se combina" e "coeficientes". A primeira etapa realiza-se sempre com casos individuais, característica dos métodos hierárquicos aglomerativos, e a partir desta etapa estes indivíduos constituem um grupo e são indivisíveis nas etapas seguintes.
As colunas seguintes indicam qual a etapa anterior em que cada um daqueles indivíduos se associou num *cluster* e qual a próxima etapa em que o *cluster* acabado de se formar voltará a associar-se a outro indivíduo.

A Figura 11 ilustra o dendograma obtido de acordo com as novas variáveis independentes obtidas na ACP, o qual também mostra as várias fases do processo de agrupamento.

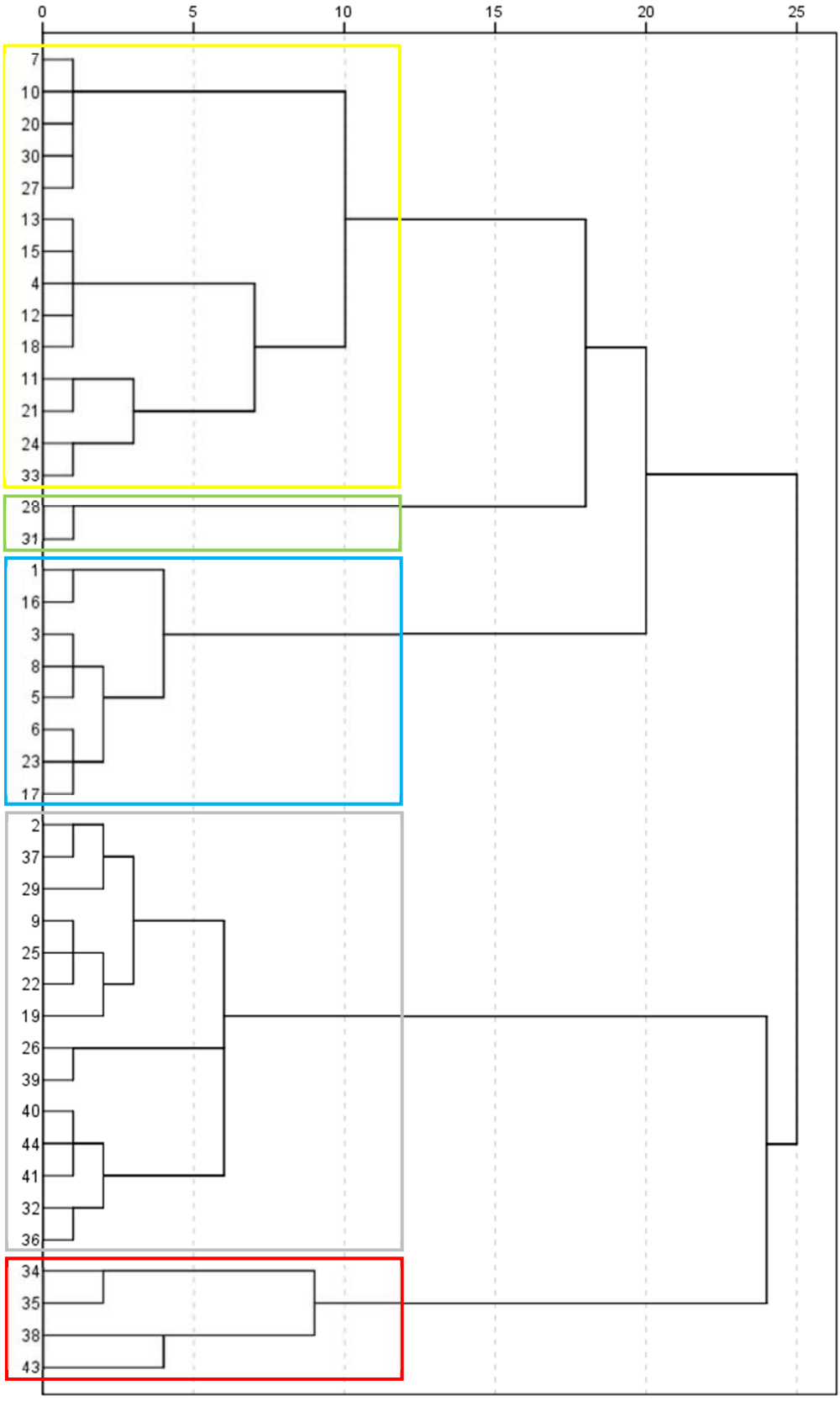


Figura 11 - Dendograma segundo o método de Ward

O problema que se coloca é por onde cortar o dendograma de modo a obter-se o número de grupos ideal, o que é sempre uma decisão difícil.

A visualização do dendograma (método empírico) sugere uma divisão em 5 *clusters*, fazendo um corte na distância aproximadamente 12.

Outro método de avaliar o número de *clusters* ideal baseia-se no valor da medida de proximidade usada para juntar os *clusters*, que se encontra na coluna "coeficientes" da tabela de agrupamento do Anexo 4. O gráfico da Figura 12, onde estão representados para cada etapa os respetivos coeficientes de aglomeração, permite visualizar as diferenças das distâncias entre duas etapas consecutivas e verifica-se que o primeiro grande aumento destas distâncias ocorre da etapa 37 para a 38. Seguindo a junção de indivíduos assinalada na coluna "*Cluster* que se combina" da tabela de agrupamento do Anexo 4 até à etapa 37 obtêm-se 5 *clusters*.

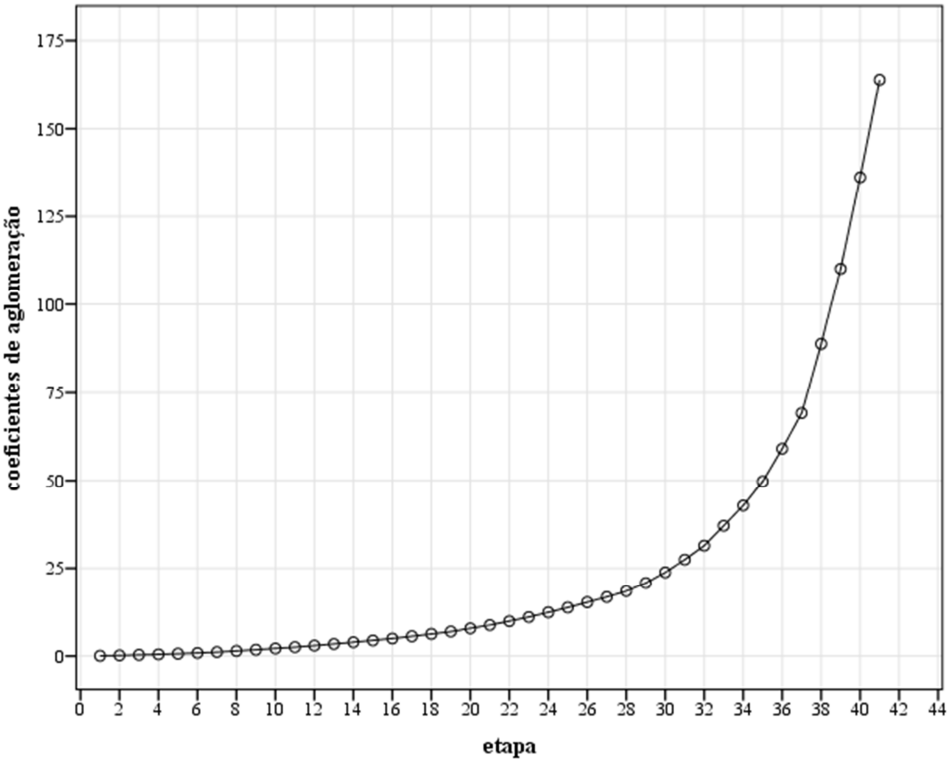


Figura 12 - Representação dos coeficientes de aglomeração para cada etapa

Outro critério é o do R-quadrado que, de acordo com Maroco (2007), indica que se deve encontrar o número mínimo de *clusters* que retenha uma percentagem significativa de variabilidade (*e.g.* superior a 80%). Por observação da Tabela 9, pode-se afirmar que uma solução de dez *clusters* seria aceitável, retendo-se assim aproximadamente 81% da variabilidade total.

Número de *Clusters*	R-quadrado
1	0
2	0,16950232
3	0,3285359
4	0,45818323
5	0,57818819
6	0,63988689
7	0,69655363
8	0,73886794
9	0,77392376
10	0,80834567

Tabela 9 - Resultados do critério do R-quadrado

Na escolha do número de *clusters,* reter poucos grupos pode levar a que estes sejam demasiado heterogéneos e demasiados grupos tornam a sua interpretação difícil, pelo que se entendeu que o método anterior está a sugerir um elevado número de *clusters* e se optou pela utilização do método empírico por ser o mais frequentemente usado na prática e ir ao encontro do sugerido pelo método gráfico do número de *clusters* a reter apresentado na Figura 12.

Tem-se então a divisão da parcela em 5 *clusters* que se encontram assinalados na Figura 13.

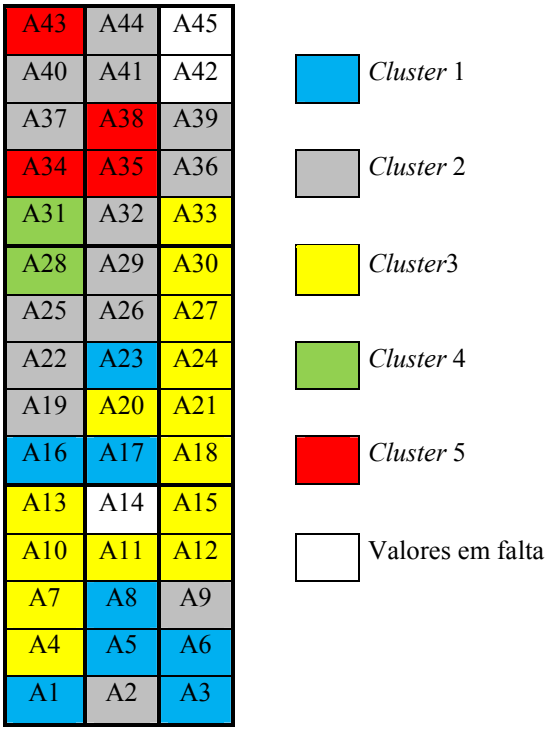


Figura 13 - Distribuição dos 5 grupos na parcela em estudo

No *cluster* 1 tem-se 8 pontos de amostragem: A1, A3, A5, A6, A8, A16, A17 e A23 que se localizam na região 1 e nas regiões 2 poente e centro da parcela.
Esta subdivisão sugere ainda um 2º *cluster* constituído por 14 pontos de amostragem, a saber: A2, A9, A19, A22, A25, A26, A29, A32, A36, A37, A39, A40, A41 e A44, os quais se encontram por quase todas as regiões da parcela.
No 3º *cluster* tem-se 14 pontos de amostragem: A4, A7, A10, A11, A12, A13, A15, A18, A20, A21, A24, A27, A30 e A33, que se localizam na região 1, nas regiões 2 centro e nascente e na região 3 nascente.
O *cluster* 4 engloba 2 pontos de amostragem: A28 e A31, que se localizam na fronteira das regiões 2 e 3 poentes.
Finalmente o 5º *cluster* é constituído por 4 pontos de amostragem: A34, A35, A38 e A43, que se encontram nas regiões 3 centro e poente da parcela.

Conhecidos os *clusters* pretende-se saber se existem diferenças significativas entre eles, tendo em consideração as variáveis do solo e as variáveis que traduzem a qualidade do vinho.
Para testar a existência destas diferenças utiliza-se a ANOVA quando os pressupostos de normalidade e homocedasticidade dos grupos são garantidos.
O pressuposto da normalidade foi testado com o teste de Shapiro-Wilk. Os resultados obtidos da aplicação deste teste encontram-se nas Tabelas A5.1 a A5.5 do Anexo 5.
Em relação às variáveis do solo temos que este pressuposto não é verificado em alguns grupos das variáveis pH, MO, Mg, AzT, Ni, Cd e N.
Relativamente às variáveis do mosto verifica-se que em todas elas há grupos que não cumprem o pressuposto da normalidade.
No que concerne às variáveis da videira o pressuposto da normalidade apenas não é verificado para o grupo 2 da variável Uvas_kg_vid e para o grupo 5 da variável Pvara_g.
Em relação às variáveis dos compostos voláteis do sumo das uvas verifica-se que em todas elas há grupos que não cumprem este pressuposto.
Finalmente, para a variável Nota_Final do vinho também se verifica que há grupos que não cumprem o pressuposto da normalidade.
Para testar a homocedasticidade utilizou-se o teste de Levene por ser robusto no caso de perda da normalidade. Os resultados obtidos da aplicação deste teste encontram-se no Anexo 6.

Conclui-se que não existe igualdade de variâncias entre os cinco grupos no caso das variáveis MO, Ca, B, CTC, CP2, CP4, pH_mosto, Acido_malico, Acido_tartarico, FL2, FL4, FG1 e FG2.

A ANOVA é relativamente robusta face à perda de igualdade de variâncias desde que as dimensões das amostras provenientes dos 5 grupos sejam aproximadamente iguais, e face à perda de normalidade se o número de casos for grande, o que não se verifica neste caso, tem-se inclusive um quarto grupo com apenas dois elementos. Optou-se então pela utilização do teste não paramétrico de Kruskal-Wallis, visto que os grupos encontrados têm dimensão pequena e desigual.

Tal como referido anteriormente, a qualidade do vinho foi avaliada pelas características físico-químicas do mosto, pela análise da composição aromática das uvas, por uma classificação atribuída ao vinho por um painel de provadores e por parâmetros da produtividade e do vigor da videira. Os resultados da aplicação do teste de Kruskal-Wallis, para avaliar se existem diferenças significativas entre os 5 grupos relativamente às variáveis do solo e às que traduzem a qualidade do vinho, encontram-se nas Tabelas 10 e 11. Nestas tabelas encontram-se também registados os valores médios de cada um dos grupos para cada variável.

		Valores médios nos *Clusters*					Estatística de teste	Graus de liberdade	Valor Prova
		1	2	3	4	5			
SOLO	pH	5,730	5,510	6,129	5,445	5,510	25,208	4	**< 0,001***
	MO	2,645	3,205	2,717	3,065	8,088	11,068	4	**0,026***
	DA	1,209	1,099	1,126	1,287	1,128	8,428	4	0,077
	FF	64,456	69,593	66,340	73,934	74,016	12,137	4	**0,016***
	P2O5	46,522	33,694	50,356	69,562	22,551	17,703	4	**0,001***
	K2O	69,763	81,883	92,384	115,691	102,494	8,954	4	0,062
	Ca	426,563	335,464	639,964	426,750	503,250	25,089	4	**< 0,001***
	Mg	69,750	65,250	82,179	76,500	66,375	14,611	4	**0,006***
	AzT	0,112	0,130	0,123	0,123	0,331	10,998	4	**0,027***
	Ni	2,295	4,406	3,254	4,540	5,210	19,686	4	**0,001***
	Cr	0,664	0,353	0,459	0,128	0,248	18,001	4	**0,001***
	Cd	0,115	0,095	0,127	0,205	0,105	15,528	4	**0,004***
	N	1,453	1,123	1,379	1,515	1,590	15,638	4	**0,004***
	B	0,380	0,397	0,379	0,399	0,676	11,085	4	**0,026***
	CTC	10,125	8,509	14,425	10,738	11,663	24,561	4	**< 0,001***
	CP1	-0,249	-0,838	0,987	-0,018	-0,015	23,362	4	**< 0,001***
	CP2	-0,110	-0,270	-0,314	-0,401	2,463	12,379	4	**0,015***
	CP3	-1,324	0,365	-0,015	1,959	0,447	22,508	4	**< 0,001***
	CP4	0,688	-0,546	-0,164	2,352	-0,067	15,942	4	**0,003***

* significativo a 5%

Tabela 10 - Resultados do teste de Kruskal-Wallis para as variáveis do solo

Consultando os valores prova obtidos na aplicação deste teste, conclui-se que há evidência estatística suficiente nos dados para rejeitar a igualdade de valores médios nos 5 *clusters* em todas as variáveis do solo, exceto para as variáveis DA e K_2O.

		Valores médios nos *Clusters*					Estatística de teste	Graus de liberdade	Valor Prova
		1	2	3	4	5			
MOSTO	pH_mosto	3,332	3,300	3,319	3,306	3,356	1,771	4	0,778
	Acidez_T	5,670	6,127	5,699	6,676	6,418	9,878	4	**0,043***
	Acido_malico	1,750	2,179	1,929	2,650	2,900	15,115	4	**0,004***
	Acido_tartarico	4,588	4,550	4,507	4,600	4,200	5,813	4	0,214
	Açucares	215,438	216,129	216,450	217,050	216,450	1,744	4	0,783
	TAP	12,803	12,842	12,859	12,895	12,860	1,744	4	0,783
VIDEIRAS	Ncachos	44,250	33,385	41,000	29,000	35,000	3,554	4	0,470
	Uvas_kg_vid	4,963	4,635	5,767	5,050	5,175	2,333	4	0,675
	Pcacho_kg	0,111	0,136	0,140	0,174	0,145	9,275	4	0,055
	Nvaras	34,875	29,077	32,583	51,000	29,750	6,489	4	0,165
	Pvaras_kg	2,538	2,188	2,700	3,950	2,238	2,743	4	0,602
	Pvara_g	70,570	74,654	82,148	77,451	75,187	,541	4	0,969
UVAS	FL1	118,256	101,382	95,513	104,880	84,850	8,642	4	0,071
	FL2	227,601	268,426	252,554	296,290	354,260	10,834	4	**0,028***
	FL3	42,441	18,127	16,982	12,085	5,365	9,325	4	0,053
	FL4	6,603	7,241	8,083	7,015	10,985	8,841	4	0,065
	FL5	20,149	12,884	11,341	9,150	13,235	3,550	4	0,470
	FG1	36,695	30,620	34,707	32,145	25,250	6,254	4	0,181
	FG2	140,455	162,778	150,623	174,715	171,570	2,651	4	0,618
	FG3	3,639	3,428	4,389	3,285	3,145	5,317	4	0,256
	FG4	9,415	10,649	11,016	10,065	10,760	3,124	4	0,537
	FG5	31,951	36,276	39,999	36,630	30,940	4,500	4	0,343
	FG6	36,724	49,372	46,637	61,740	60,800	6,473	4	0,167
	FG7	0,716	1,041	0,986	1,170	1,220	5,596	4	0,231
VINHO	Nota_Final	60,6250	63,4286	69,2143	58,0000	60,5000	8,180	4	0,085

* significativo a 5%

Tabela 11 - Resultados do teste de Kruskal-Wallis para as variáveis relativas à qualidade do vinho

Verificou-se ainda que os valores médios apresentam diferenças significativas entre *clusters* em relação às variáveis Acidez_total, Acido_malico e FL2 relativas à qualidade do vinho.

Interessa então identificar qual ou quais os *clusters* que diferem entre si. Neste sentido aplicou-se o teste não paramétrico de comparações múltiplas (*All Pairwise*) associado ao teste de Kruskal-Wallis no SPSS, o qual é equivalente à aplicação da desigualdade (51). Os resultados obtidos encontram-se nas Tabelas 12 e 13.

Variável	*Cluster i*	*Cluster j*	Valor Prova
CP1	1	2	0,173
		3	**0,007***
		4	0,728
		5	0,702
	2	3	**< 0,001***
		4	0,245
		5	0,139
	3	4	0,221
		5	0,088
	4	5	0,962
CP2	1	2	0,238
		3	0,233
		4	0,463
		5	**0,035***
	2	3	0,988
		4	0,939
		5	**0,001***
	3	4	0,945
		5	**0,001***
	4	5	**0,030***
CP3	1	2	**< 0,001***
		3	**0,003***
		4	**< 0,001***
		5	**0,002***
	2	3	0,317
		4	0,096
		5	0,754
	3	4	**0,030***
		5	0,327
	4	5	0,212
CP4	1	2	**0,001***
		3	**0,027***
		4	0,321
		5	0,157
	2	3	0,242
		4	**0,004***
		5	0,327
	3	4	**0,020***
		5	0,841
	4	5	0,057

* significativo a 5%

Tabela 12 - Teste de comparações múltiplas para as CPs do solo

Para as CPs do solo verifica-se, ao nível de significância de 5%, que:

- na CP1, o *cluster* 3 difere dos *clusters* 1 e 2 apresentando um valor médio superior;
- na CP2, o *cluster* 5 difere dos restantes apresentando um valor médio superior;
- na CP3, o *cluster* 1 difere dos restantes apresentando um valor médio inferior e o *cluster* 4 difere do *cluster* 3 apresentando um valor médio superior;
- na CP4, o *cluster* 2 difere dos *clusters* 1 e 4 apresentando valor médio inferior, o mesmo acontecendo ao *cluster* 3.

Os resultados da aplicação do teste de comparações múltiplas para as variáveis originais do solo encontram-se no Anexo 7. As variáveis que traduzem a qualidade do vinho que apresentaram diferenças significativas nos valores médios entre *clusters* encontram-se assinaladas na Tabela 13.

Variável	*Cluster i*	*Cluster j*	Valor Prova
Acidez_T	1	2	0,065
		3	0,694
		4	**0,020***
		5	0,073
	2	3	0,088
		4	0,179
		5	0,624
	3	4	**0,028***
		5	0,104
	4	5	0,394
Acido_malico	1	2	**0,016***
		3	0,056
		4	**0,042***
		5	**< 0,001***
	2	3	0,570
		4	0,471
		5	**0,037***
	3	4	0,315
		5	**0,014***
	4	5	0,461
FL2	1	2	0,128
		3	0,172
		4	0,173
		5	**0,001***
	2	3	0,852
		4	0,595
		5	**0,021***
	3	4	0,533
		5	**0,015***
	4	5	0,297

* significativo a 5%

Tabela 13 - Teste de comparações múltiplas

Relativamente à variável Acidez_T existem diferenças significativas entre o *cluster* 4 e os *clusters* 1 e 3, apresentando o *cluster* 4 valor médio superior.

Para a variável Acido_malico verifica-se que, ao nível de significância de 5%, o *cluster* 1 difere dos *clusters* 2, 4 e 5 apresentando um valor médio inferior e o *cluster* 5 difere dos *clusters* 1, 2 e 3 apresentando um valor médio superior.

Em relação à família de álcoois do aroma das uvas na forma livre (FL2) existem diferenças significativas entre o *cluster* 5 e os *clusters* 1, 2 e 3, apresentando o 5º *cluster* valor médio superior.

Com o objetivo de verificar quais as variáveis do solo que influenciam a qualidade do vinho determinaram-se as correlações entre as novas variáveis do solo e as que determinam a qualidade do vinho.

Para testar se a associação linear, indicada pelo coeficiente de correlação, é estatisticamente significativa, não se pôde utilizar o teste de independência de Pearson dado que não se consegue garantir a binormalidade dos dados. Desta forma, foi usado na análise o teste baseado no coeficiente de correlação não paramétrico de Spearman. Os resultados obtidos encontram-se na Tabela 14.

			CP2	CP3	CP4
MOSTO	Acidez_T	Coeficiente de correlação		**0,349***	
		Valor Prova		0,023	
	Acido_malico	Coeficiente de correlação	**0,480****	**0,408****	
		Valor Prova	0,001	0,007	
	Acido_tartarico	Coeficiente de correlação	**-0,405****		
		Valor Prova	0,008		
VIDEIRA	Ncachos	Coeficiente de correlação		**-0,367***	
		Valor Prova		0,024	
UVAS	FL1	Coeficiente de correlação		**-0,314***	**0,317***
		Valor Prova		0,043	0,041
	FL2	Coeficiente de correlação	**0,455****	**0,331***	
		Valor Prova	0,002	0,032	
	FL3	Coeficiente de correlação		**-0,344***	**0,397****
		Valor Prova		0,026	0,009
	FL4	Coeficiente de correlação	**0,530****		
		Valor Prova	0		
	FG1	Coeficiente de correlação	**-0,425****	**-0,378***	
		Valor Prova	0,005	0,014	
	FG4	Coeficiente de correlação			**-0,458****
		Valor Prova			0,002
	FG5	Coeficiente de correlação			**-0,372***
		Valor Prova			0,015
VINHO	Nota_Final	Coeficiente de correlação			**-0,402****
		Valor Prova			0,008

* Significativo a 5%
** Significativo a 1%

Tabela 14 - Correlações significativas entre as CPs do solo e as variáveis relacionadas com a qualidade do vinho

Com base nos resultados obtidos, e considerando a regra de decisão $\alpha = 5\%$, podemos rejeitar a hipótese nula e, consequentemente, concluir que a correlação é significativamente diferente de zero, bem como que a mesma pode ser inferida para a população da qual a amostra foi extraída, para as variáveis assinaladas na Tabela 14.

A CP2 do solo (que apresenta *loadings* positivos elevados para variáveis MO, AzT e B) apresenta uma correlação estatisticamente significativa positiva com as variáveis Acido_malico, FL2 (família de álcoois do aroma das uvas na forma livre) e FL4 (família de fenóis voláteis do aroma das uvas na forma livre). Portanto, a elevadas concentrações destes constituintes do solo corresponderão elevadas concentrações de ácido málico e dos compostos das famílias FL2 e FL4 das uvas. O *cluster* 5 é o que apresenta valores médios maiores nestas variáveis e o *cluster* 1 é o que apresenta valores médios menores.
Em relação às correlações obtidas entre a CP2 do solo e acidez do mosto, existem já estudos que referem que altos valores de MO no solo induzem alta acidez nos mostos (Delas *et al.*, 1992; Araújo, 2004).

A CP2 do solo também se correlaciona de forma significativa mas negativa com as variáveis Acido_tartarico do mosto e FG1 (família de compostos em C_6 do aroma das uvas na forma glicosilada), ou seja, a concentrações elevadas destes constituintes no solo corresponderão baixas concentrações de ácido tartárico no mosto e dos compostos da família FG1 nas uvas, o que se verifica no *cluster* 5.

A CP3 do solo, para a qual têm *loadings* elevados positivos as variáveis FF, K_2O e Ni, está correlacionada de forma significativa positiva com a acidez total e o ácido málico do mosto e com a FL2 (família de álcoois do aroma das uvas na forma livre). Portanto a baixas concentrações destes constituintes no solo corresponderão baixas concentrações de acidez total e de ácido málico no mosto e de FL2 nas uvas, como podemos verificar no *cluster* 1. A variável Cr tem *loading* pesado negativo nesta CP, assim, a altas concentrações de Cr no solo corresponderão baixas concentrações de acidez total, ácido málico e FL2, o que também se verifica no *cluster* 1.

A CP3 do solo também se correlaciona de forma significativa mas negativa com as variáveis Ncachos da videira e com as famílias FL1 (família de compostos em C_6 do aroma das uvas na forma glicosilada), FL3 (família de álcoois monoterpénicos do aroma das uvas

na forma livre) e FG1 (família de compostos em C_6 do aroma das uvas na forma glicosilada), ou seja, a baixas concentrações de FF, K_2O e Ni no solo corresponderá um elevado número de cachos por videira e altas concentrações dos compostos das famílias FL1, FL3 e FG1 nas uvas, o que se verifica no *cluster* 1. A elevadas concentrações de Cr corresponderão muitos cachos por videira e altas concentrações de FL1, FL3 e FG1 nas uvas, o que também se verifica no *cluster* 1.

A CP4 do solo (que apresenta *loadings* positivos elevados nas variáveis DA e Cd) correlaciona-se de forma significativa positiva com as variáveis FL1 (família de compostos em C_6 do aroma na forma livre) e FL3 (família de álcoois monoterpénicos do aroma das uvas na forma livre). Portanto a baixas concentrações destes constituintes do solo corresponderão baixas concentrações dos compostos das famílias FL1 e FL3. Na CP4, os *clusters* 2 e 3 diferem significativamente dos *clusters* 1 e 4, apresentando valores médios inferiores.

Finalmente, a CP4 do solo também se correlaciona de forma significativa mas negativa com as variáveis FG4 (família de óxidos e dióis monoterpénicos do aroma das uvas na forma glicosilada), FG5 (família de norisoprenóides em C_{13} do aroma das uvas na forma glicosilada) e Nota_Final, ou seja, a valores elevados destes constituintes do solo corresponderão baixas concentrações dos compostos das famílias FG4 e FG5 e baixa nota final ao vinho, o que se verifica essencialmente no *cluster* 4.

CAPÍTULO 4: CONCLUSÃO E TRABALHOS FUTUROS

Neste trabalho pretendeu-se identificar quais os parâmetros físico-químicos do solo que influenciaram o desenvolvimento das videiras e a qualidade das uvas e dos vinhos, numa parcela da EVAG, no concelho dos Arcos de Valdevez. Os dados analisados reportam-se apenas ao ano 2010.

As parcelas de vinha, mesmo plantadas com a mesma casta, não são homogéneas e esta variabilidade espacial tem origem diversa. Uma boa gestão espacial das diversas variáveis relacionadas com a produtividade dentro de uma vinha, irá permitir melhorar o rendimento económico da atividade vitícola, reduzindo os custos de produção e o impacto ambiental.

Para a concretização deste objetivo foram usadas diferentes técnicas de Estatística Multivariada que foram bastante úteis para reduzir a dimensionalidade do vasto conjunto de informação e detetar a existência de zonas homogéneas na parcela em estudo que se distinguem ao nível das características do solo.

A Análise de Componentes Principais permitiu uma redução da dimensionalidade do problema tendo em consideração as variáveis do solo analisadas. A aplicação desta técnica resultou em quatro componentes principais as quais explicam aproximadamente 74,11% da variabilidade total dos dados originais.

A Análise de *Clusters* permitiu um agrupamento dos pontos de amostragem em cinco grupos homogéneos relativamente às características do solo. Com a aplicação do teste não paramétrico de Kruskal-Wallis verificou-se que, à exceção das variáveis DA e K_2O, todas as restantes relativas ao solo apresentam diferenças significativas entre *clusters*. Estas diferenças foram ainda detetadas para as variáveis acidez total e ácido málico do mosto e para a família de compostos FL2 das uvas. O teste de comparações múltiplas não paramétrico permitiu ainda avaliar em que *clusters* é que a variação destes parâmetros foi significativa.

Por fim, a qualidade do vinho e o equilíbrio vegetativo da videira foram avaliados através das suas correlações com os parâmetros do solo e identificaram-se as zonas da parcela que apresentam défice/excesso nas concentrações dessas variáveis.

Este estudo providenciou um conjunto de indicadores que poderão ajudar no processo de tomada de decisão para a racionalização de utilização dos fatores de produção, como os nutrientes e os produtos fitofármacos, contribuindo para a minimização dos efeitos secundários da atividade. Desta forma, o trabalho efetuado poderá orientar a prática de uma agricultura sustentável.

Como trabalhos futuros sugere-se que se efetue o mesmo estudo em vários anos para assim se poder tirar conclusões fidedignas e até mesmo fazê-lo em diferentes parcelas ou regiões. Em termos de análise estatística seria interessante explorar outras técnicas nomeadamente a modelação, ou seja, a construção e validação de modelos matemáticos que permitam descrever a relação existente entre as variáveis solo-vinha-vinho.

BIBLIOGRAFIA

Araújo, I. (2004). Características aromáticas e cromáticas das castas Amaral e Vinhão. Tese de Mestrado em Viticultura e Enologia. Universidade do Porto e Universidade Técnica de Lisboa, Porto.

Armada, N. (1990). Caracterização dos Solos da "Estação Vitivinícola Amândio Galhano" e sua Relação com a Vinha. Relatório de Estágio da Licenciatura em Engenharia Agrícola. UTAD, Vila Real.

Athayde, M. E. (2010). Apontamentos da disciplina Modelos Estatísticos. Departamento de Matemática, Universidade do Minho.

Azevedo, C. (2010). Apontamentos da disciplina Estatística Multivariada. Departamento de Matemática, Universidade do Minho.

Branco, A. J. (2004). Uma Introdução à Análise de Clusters. Sociedade Portuguesa de Estatística.

Cadima, J. (2010). Apontamentos de Estatística Multivariada, Departamento de Matemática – Instituto Superior de Agronomia – Universidade Técnica de Lisboa.

Chatfield,C.; Collins, A.J. (1995). *Introduction to Multivariate Analysis*. Chapman & Hall.

Clímaco P., Castro R. (1991). Adaptación de variedades y portainjertos en viticultura. Vitivinicultura. **3**: 51-54.

Clímaco, P. (1997). Influência da cultivar e da maturação da uva e na produtividade da videira (*Vitis vinifera L.*), Tese de Doutoramento. Instituto Superior de Agronomia, Universidade Técnica de Lisboa.

Conover, W. J. (1980). Pratical Nonparametric Statistics. 2ª edição. John Wiley & Sons.

Delas J., Molot C., Soyer J.P. (1992). Influence du porte-greffe et de la fertilization azotée sur la composition des baies de Merlot. IV Simposio Internazionale di Fisiologia della Vite, Torino, Itália. pp. 397-400.

Duarte, M. T., Eiras-Dias, E. (1989). Catálogo de Porta-Enxertos mais utilizados em Portugal. IVV/CNPPA/EVN/MAPA, Tipografia Guerra, Viseu.

Ferreira, E. (2010). Métodos Biplot aplicados a dados de Biologia Molecular. Tese de Mestrado em Matemática e Aplicações. Universidade de Aveiro.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. Biometrika, Great Britain. Volume 58, **3**: 453-467.

Galet, P. (1993). Précis de Viticulture. Imprimerie Déhan. Montpellier.

Galindo, M. P. (1985). Contribuciones a la Representación Simultánea de Datos Multidimensionales. Tese de Doutoramento. Universidade de Salamanca.

González, F. A. (1999). *Algunas aportaciones al Análisis de Datos, utilizando técnicas de representación Multivariante*. Tese de Doutoramento. Faculdade de Ciências, Departamento de Matemática, Universidade de Cádiz.

Guimarães, R. C., Cabral, J. A. S. (1999). *Estatística*. Mc Graw Hill.

Johnson, R. A., Wichern, D. W. (1982). Applied Multivariate Statistical Analysis. Prentice Hall.

Jolliffe, I. T. (2002). Principal Component Analysis. 2ª Edição. Springer.

Layon, D. M., Cass, A., Hansen, D. (2004). The effect of soil properties on vine performance. Technical Report, n.º 34. CSIRO Land and Water.

Leme, P. C., Malheiro, P. (1998). The behaviour of the grapevine "Azal Branco" In fields conditions. Congresso Mundial da Vinha e do Vinho, 78ª Assembleia Geral O.I.V., Lisboa, Portugal, 22-27 junho. pp. 76-82.

Maciel, A. (2005). A Pertinência dos Estudos de Microclimatologia para a Prevenção dos Riscos Climáticos num Vinhedo do "Entre Douro e Minho". Estação Vitivinícola Amândio Galhano. Tese de Mestrado em Gestão dos Riscos Naturais. FLUP, Porto.

Maroco, J. (2007). Análise Estatística com utilização do SPSS. 3ª Edição. Edições Sílabo.

Miller Jr., R. G. (1997). Beyond Anova – Basics of applied statistics. Chapman & Hall.

Mota, M. T. (2005). Potencialidades e condicionalismos da condução LYS. *CV*. Loureiro. Região dos Vinhos Verdes. Tese de Doutoramento em Engenharia Agronómica. Instituto Superior de Agronomia, Universidade Técnica de Lisboa.

Mota, T., Garrido, J. (2001). Implantação da Vinha. Castas, Porta-enxertos, Sistemas de condução e Plantação. Manual técnico, CVRVV. Tip. Artes Gráficas Bacelar & Irmãos, Arcos de Valdevez.

Peña, D. (2002). Analisis de datos multivariantes. McGrawHill.

Pestana, D. D., Velosa, S. F. (2008). Introdução à Probabilidade e à Estatística. Volume 1, 3ª Edição. Fundação Calouste Gulbenkian.

Pestana, M. H., Gageiro, J. N. (2008). Análise de dados para Ciências Sociais - A complementaridade do SPSS. 5ª Edição. Edições Sílabo.

Pinho, J. O. (1993). Compêndio de Ampelologia I. Edição Figueirinhas, Porto/Lisboa.

Reis,E. (2001). Estatística Multivariada Aplicada. 2ª Edição. Edições Sílabo.

Siegel, S. (1975). Estatística não-paramétrica (para as ciências do comportamento). Mc Graw Hill.

Siegel, S., Castellan Jr., N. J. (1988). Nonparametric Statistics for the Behavioral Sciences. 2ª Edição. Mc Graw Hill.

Silva, T. C. F. (2007). Técnicas de Análise Estatística Multivariada. Tese de Mestrado em Matemática (Área de Especialização Ensino). Universidade do Minho.

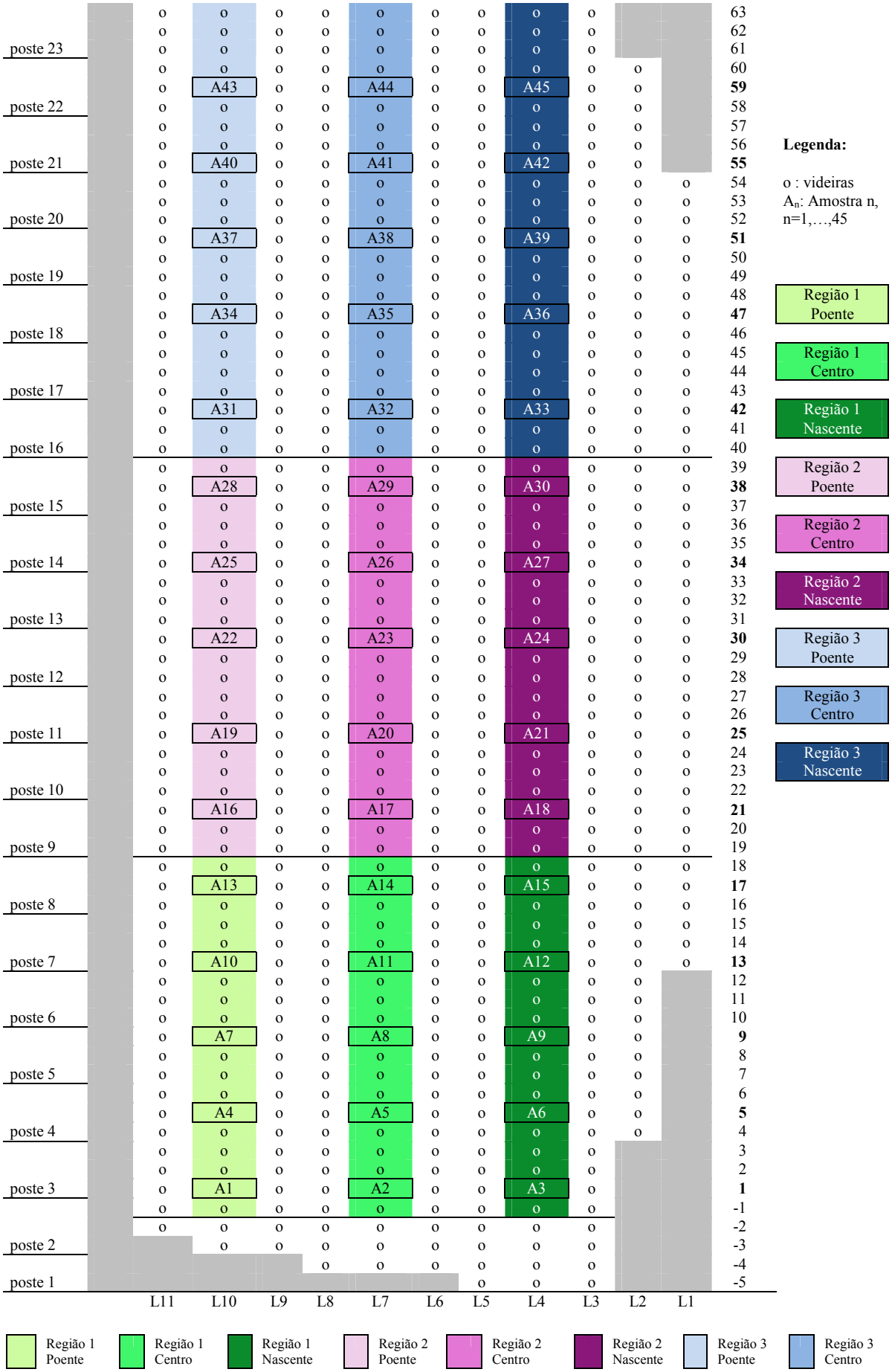
Tomasi, D., Calò, A., Biscaro, S., Panero, L., Di Stefano, R. (1998). Influence des caracteristiques du sol dans la composition polyphenolique et antnthocyaniques des raisins de Cabernet Sauvignon. Congresso Mundial da Vinha e do Vinho, 78ª Assembleia Geral O.I.V., Lisboa, Portugal, 22-27 junho. pp. 190-198.

Vairinhos, V. M.; Galindo M. P. (2004). Biplots PMD - Data Mining Centrada em Biplots. Apresentação de um Protótipo. XI Jornadas de Classificação e Análise de Dados, Associação Portuguesa de Classificação e Análise de Dados, Lisboa.

Villardón, J. L. V., Analisis de Componentes Principales, Departamento de Estadística, Universidad de Salamanca. Acedido em 14 de maio de 2011, em: http://biplot.usal.es/ALUMNOS/BIOLOGIA/OPTATIVAS/ANALMUL/ACP.pdf

ANEXOS

ANEXO 1 - Parcela C5



ANEXO 2 - Concentrações dos compostos voláteis do aroma das uvas

	B1			B2			B3		
	N	**C**	**P**	**N**	**C**	**P**	**N**	**C**	**P**
Compostos em C_6									
(*E*)-2-hexenal	68,8	36,9	30,0	24,3	45,3	29,7	42,8	23,1	46,1
1-hexanol	44,8	62,7	44,6	38,3	51,9	52,5	34,6	38,6	42,6
(*Z*)-3-hexen-1-ol	2,7	3,9	2,3	3,3	3,6	4,2	3,0	3,9	2,5
(*E*)-2-hexeno-1-ol	19,3	11,9	11,6	9,6	21,4	22,1	12,6	3,3	8,3
(*Z*)-2-hexeno-1-ol	0,3	0,2	0,3	0,2	0,2	0,8	0,4	0,2	1,1
TOTAL	135,9	115,7	88,8	75,7	122,4	109,2	93,4	69,1	100,6
Álcoois									
3-metil-3-buteno-1-ol	5,6	9,0	6,1	7,1	4,8	7,2	11,3	15,0	16,3
(*Z*)-2-penten-1-ol	4,6	7,1	6,5	6,1	6,4	5,2	3,3	4,8	2,6
3-metil-2-buteno-1-ol	4,4	6,9	7,2	7,8	6,7	11,1	8,3	7,9	14,9
1-octeno-3-ol	2,2	3,4	1,7	1,0	1,8	0,7	1,0	1,2	1,3
álcool benzílico	68,8	98,7	100,9	115,2	82,5	86,4	81,1	155,5	192,8
2-feniletanol	105,2	146,2	111,8	158,2	107,1	107,0	119,6	145,7	143,7
2-fenoxietanol	5,2	4,1	1,9	1,7	2,1	1,5	2,6	5,0	1,8
TOTAL	196,0	275,4	236,2	297,0	211,4	219,1	227,2	335,1	373,5
Álcoois monoterpénicos									
linalol	2,3	2,3	1,9	0,4	1,2	0,8	1,0	0,8	2,2
4-terpineol	31,5	108,0	0,8	1,0	6,3	14,6	1,1	1,5	1,4
nerol	1,1	0,5	2,2	0,8	0,4	1,2	0,5	0,4	1,6
geraniol	1,0	2,2	1,1	1,1	1,1	1,1	1,3	1,7	1,2
TOTAL	35,9	113,0	6,0	3,2	9,0	17,8	3,9	4,4	6,4
Fenóis voláteis									
salicilato de metilo	1,0	0,5	1,5	1,9	1,3	1,0	1,0	3,8	5,4
vanilina	3,5	7,4	7,0	4,6	2,8	1,0	3,8	4,7	3,3
acetovanilona	0,5	1,0	0,9	0,8	0,4	0,7	0,7	1,2	0,8
zingerona	0,8	0,4	0,6	1,0	0,3	0,6	0,3	1,6	1,2
TOTAL	5,7	9,3	10,1	8,3	4,8	3,3	5,9	11,3	10,7
Compostos carbonilados									
benzaldeído	3,6	56,3	9,9	3,8	6,2	4,7	3,0	13,8	7,3
feniletanal	1,1	1,9	2,2	2,0	1,7	2,7	2,5	1,8	3,6
TOTAL	4,7	58,3	12,1	5,8	7,9	7,4	5,5	15,5	10,9

Tabela A2.1 - Concentrações médias (µg/l) dos compostos voláteis da fração livre do aroma das uvas da casta Vinhão em função do 4-nonanol

	B1			B2			B3		
	N	**C**	**P**	**N**	**C**	**P**	**N**	**C**	**P**
Compostos em C_6									
1-hexanol	25,8	16,9	34,5	15,3	31,5	24,0	23,1	14,5	20,4
(*Z*)-3-hexen-1-ol	0,6	0,5	0,8	0,6	0,7	0,7	0,9	0,8	0,6
(*E*)-2-hexeno-1-ol	7,3	11,4	17,1	5,7	9,1	8,9	7,2	4,3	9,8
TOTAL	33,6	28,9	52,4	21,6	41,4	33,5	31,2	19,7	30,8
Álcoois									
3-metil-3-buteno-1-ol	1,4	1,1	2,7	1,5	3,0	2,2	2,3	1,2	1,7
3-metil-2-buteno-1-ol	2,8	2,1	3,9	1,8	5,1	3,8	4,3	3,3	3,6
1-octeno-3-ol	0,3	0,1	0,3	0,3	0,3	0,3	0,4	0,3	0,2
1-octanol	0,6	0,6	0,8	1,0	1,1	0,7	0,7	0,6	0,5
1-feniletanol	0,8	0,7	1,8	1,7	2,4	1,7	1,8	1,2	1,7
álcool benzílico	51,5	40,4	95,5	72,0	100,4	90,3	82,7	96,9	112,7
2-feniletanol	58,4	36,2	76,1	61,7	79,3	66,2	86,8	55,3	63,8
TOTAL	115,8	81,2	181,2	140,1	191,6	165,1	179,1	158,8	184,3
Álcoois monoterpénicos									
linalol	0,3	0,1	0,4	0,8	0,3	0,3	0,5	0,6	0,3
α-terpineol	0,2	0,3	0,3	0,3	0,2	0,2	0,3	0,4	0,4
nerol	0,4	0,4	0,4	1,0	0,9	0,4	0,9	0,5	0,7
geraniol	1,1	2,4	3,8	3,3	3,7	2,9	0,8	2,0	1,4
TOTAL	2,0	3,2	4,8	5,5	5,1	3,8	2,6	3,5	2,8
Óxidos e dióis monoterpénicos									
óxido furânico de linalol, *trans*-	0,5	0,5	0,3	0,6	0,8	0,6	0,6	1,0	0,6
óxido furânico de linalol, *cis*-	2,3	1,5	3,1	3,5	4,1	2,1	3,3	3,6	2,4
óxido pirânico de linalol, *trans*-	0,3	0,1	0,5	0,7	0,8	0,5	0,7	0,8	0,7
óxido pirânico de linalol, *cis*-	0,3	0,5	0,8	0,8	1,1	0,5	0,7	0,8	0,5
(*E*)-8-hidroxilinalol	1,3	1,3	2,0	2,3	2,6	2,2	1,8	1,8	2,5
(*Z*)-8-hidroxilinalol	2,3	1,1	3,4	3,3	3,5	3,6	3,8	2,5	2,6
ácido gerânico	0,7	0,3	1,3	1,3	1,3	0,5	0,9	0,9	0,8
TOTAL	7,6	5,3	11,2	12,6	14,2	10,0	11,8	11,3	10,2
Norisopprenóides em C_{13}									
3,4-dihidro-3-oxo-actinidol I	1,0	0,6	2,0	1,7	0,7	1,4	1,6	1,2	4,5
3,4-dihidro-3-oxo-actinidol II	2,0	1,4	1,8	2,7	2,3	1,0	1,2	0,7	3,3
3,4-dihidro-3-oxo-actinidol III	0,7	0,5	1,3	1,6	1,3	0,9	1,8	0,9	0,8
3-hidroxi-β-damascona	3,0	2,1	4,0	4,7	4,6	12,1	5,1	3,9	3,1
3-oxo-α-ionol	8,5	3,3	11,0	12,3	12,0	10,3	15,5	8,2	6,1
3-hidroxi-7,8-dihidro-β-ionol	2,2	1,2	3,6	3,3	3,3	1,6	4,3	2,4	2,3
4-oxo-7,8-dihidro-β-ionol	2,2	0,7	2,1	2,2	2,1	1,7	2,5	1,5	1,8
3-oxo-7,8-dihidro-α-ionol	3,7	1,9	4,9	4,9	4,7	3,6	6,0	4,3	4,0
3-hidroxi-7,8-dehidro-β-ionol	1,3	1,0	1,5	2,9	1,7	1,2	2,2	1,4	1,2
vomifoliol	4,9	2,8	8,4	9,6	9,0	7,3	11,1	5,2	5,1
TOTAL	29,6	15,5	40,8	45,8	41,7	41,1	51,3	29,7	32,2
Fenóis voláteis									
salicilato de metilo	2,8	1,7	5,9	6,2	6,9	7,9	8,6	10,6	15,9
guaiacol	0,2	0,2	0,4	0,7	0,4	0,4	0,5	0,5	0,6
4-vinilguaiacol	3,5	2,8	7,3	4,5	6,3	5,0	3,9	4,0	8,6
4-vinilfenol	0,8	1,0	2,7	2,5	2,3	2,1	2,0	1,7	4,5
vanilato de metilo	0,6	0,4	1,0	1,1	0,9	0,6	1,3	1,0	1,0
acetovanilona	1,8	1,0	2,4	1,5	2,0	1,8	3,7	2,4	3,9
3,4-dimetoxifenol	0,4	0,3	0,6	0,4	0,4	0,7	0,8	0,6	0,6
zingerona	2,8	1,3	3,6	5,3	4,1	4,2	4,0	2,0	5,3
álcool 3,4,5-trimetoxibenzílico	1,4	0,7	1,3	3,4	2,4	1,5	2,7	1,3	2,2
2,5-dihidroxibenzoato de metilo	3,8	4,6	12,9	14,4	12,1	14,6	12,8	12,4	20,8
3,4,5-trimetoxifenol	6,3	5,5	14,6	13,9	13,6	11,2	11,3	11,9	10,0
TOTAL	24,4	19,7	52,9	54,0	51,3	50,1	51,4	48,2	73,4
Compostos carbonilados									
benzaldeído	0,5	0,5	1,2	1,1	0,6	1,3	1,1	1,4	1,1
TOTAL	0,5	0,5	1,2	1,1	0,6	1,3	1,1	1,4	1,1

Tabela A2.2 - Concentrações médias (µg/l) dos compostos voláteis da fração glicosilada do aroma das uvas da casta Vinhão em função do 4-nonanol

ANEXO 3 - Ficha de prova descritiva

Provador _____ N.º ____ Amostra ____

Categoria ____

Data __/__/____

		Excelente	Muito Bom	Bom	Aceitável	Insuficiente	Observações
Exame Visual	Limpidez	☐ 5	☐ 4	☐ 3	☐ 2	☐ 1	
	Cor	☐ 10	☐ 8	☐ 6	☐ 4	☐ 2	
Exame Olfactivo	Limpidez	☐ 6	☐ 5	☐ 4	☐ 3	☐ 2	
	Intensidade	☐ 8	☐ 7	☐ 6	☐ 4	☐ 2	
	Qualidade	☐ 16	☐ 14	☐ 12	☐ 10	☐ 8	
Exame Gustativo	Limpidez	☐ 6	☐ 5	☐ 4	☐ 3	☐ 2	
	Intensidade	☐ 8	☐ 7	☐ 6	☐ 4	☐ 2	
	Persistência	☐ 8	☐ 7	☐ 6	☐ 5	☐ 4	
	Qualidade	☐ 22	☐ 19	☐ 16	☐ 13	☐ 10	
Apreciação Global		☐ 11	☐ 10	☐ 9	☐ 8	☐ 7	**Total** / Rúbrica
Sub-total							

Ficha de prova baseada na ficha de prova do O.I.V. / U.I.OE

ANEXO 4 - Tabela de agrupamento da Análise de *Clusters*

Etapa	*Cluster* que se combina		Coeficientes	Etapa em que o *cluster* aparece pela primeira vez		Próxima Etapa
	Cluster 1	*Cluster* 2		*Cluster* 1	*Cluster* 2	
1	7	10	0,061	0	0	20
2	20	30	0,183	0	0	14
3	40	44	0,322	0	0	9
4	3	8	0,473	0	0	7
5	1	16	0,658	0	0	32
6	13	15	0,861	0	0	13
7	3	5	1,128	4	0	28
8	9	25	1,437	0	0	16
9	40	41	1,767	3	0	25
10	11	21	2,134	0	0	29
11	2	37	2,503	0	0	26
12	32	36	2,953	0	0	25
13	4	13	3,411	0	6	22
14	20	27	3,902	2	0	20
15	6	23	4,409	0	0	17
16	9	22	4,949	8	0	27
17	6	17	5,579	15	0	28
18	28	31	6,242	0	0	38
19	24	33	6,929	0	0	29
20	7	20	7,836	1	14	37
21	12	18	8,766	0	0	22
22	4	12	9,880	13	21	35
23	26	39	11,045	0	0	33
24	34	35	12,385	0	0	36
25	32	40	13,763	12	9	33
26	2	29	15,259	11	0	30
27	9	19	16,761	16	0	30
28	3	6	18,446	7	17	32
29	11	24	20,795	10	19	35
30	2	9	23,822	26	27	34
31	38	43	27,442	0	0	36
32	1	3	31,431	5	28	39
33	26	32	37,077	23	25	34
34	2	26	42,826	30	33	40
35	4	11	49,765	22	29	37
36	34	38	59,059	24	31	40
37	4	7	69,177	35	20	38
38	4	28	88,858	37	18	39
39	1	4	110,120	32	38	41
40	2	34	136,202	34	36	41
41	1	2	164,000	39	40	0

Tabela A4.1 - Tabela de agrupamento da AC

ANEXO 5 - Testes de normalidade nos 5 grupos

	Clusters	Shapiro-Wilk		
		Estatística de teste	g.l.	Valor prova
pH	1	0,782	8	**0,018***
	2	0,974	14	0,924
	3	0,961	14	0,739
	5	0,800	4	0,103
MO	1	0,819	8	**0,046***
	2	0,861	14	**0,032***
	3	0,888	14	0,076
	5	0,989	4	0,951
DA	1	0,909	8	0,345
	2	0,908	14	0,147
	3	0,924	14	0,249
	5	0,890	4	0,384
FF	1	0,831	8	0,061
	2	0,954	14	0,621
	3	0,897	14	0,103
	5	0,835	4	0,182
P_2O_5	1	0,913	8	0,373
	2	0,949	14	0,542
	3	0,914	14	0,178
	5	0,851	4	0,230
K_2O	1	0,913	8	0,374
	2	0,915	14	0,186
	3	0,970	14	0,872
	5	0,906	4	0,462
Ca	1	0,877	8	0,177
	2	0,953	14	0,614
	3	0,982	14	0,985
	5	0,955	4	0,745
Mg	1	0,799	8	**0,028***
	2	0,936	14	0,364
	3	0,944	14	0,467
	5	0,870	4	0,296
AzT	1	0,756	8	**0,010***
	2	0,874	14	**0,048***
	3	0,898	14	0,107
	5	0,977	4	0,887
Ni	1	0,960	8	0,814
	2	0,939	14	0,409
	3	0,973	14	0,913
	5	0,683	4	**0,007***
Cr	1	0,953	8	0,743
	2	0,958	14	0,697
	3	0,946	14	0,502
	5	0,915	4	0,507
Cd	1	0,831	8	0,061
	2	0,947	14	0,522
	3	0,846	14	**0,020***
	5	0,963	4	0,796

N	1	0,936	8	0,574
	2	0,826	14	**0,011***
	3	0,920	14	0,223
	5	0,944	4	0,681
B	1	0,941	8	0,620
	2	0,944	14	0,475
	3	0,938	14	0,395
	5	0,903	4	0,446
CTC	1	0,917	8	0,407
	2	0,953	14	0,616
	3	0,976	14	0,949
	5	0,952	4	0,727
CP1	1	0,939	8	0,600
	2	0,962	14	0,757
	3	0,989	14	0,999
	5	0,964	4	0,804
CP2	1	0,870	8	0,152
	2	0,922	14	0,237
	3	0,907	14	0,143
	5	0,926	4	0,569
CP3	1	0,974	8	0,928
	2	0,936	14	0,375
	3	0,965	14	0,797
	5	0,931	4	0,599
CP4	1	0,939	8	0,605
	2	0,943	14	0,454
	3	0,939	14	0,407
	5	0,848	4	0,218

* significativo a 5%

Tabela A5.1 - Testes de Normalidade dos dados do solo por grupos

	Clusters	Shapiro-Wilk		
		Estatistica de teste	g.l.	Valor prova
pH_mosto	1	0,867	8	0,141
	2	0,897	14	0,102
	3	0,802	14	**0,005***
	5	0,729	4	**0,024***
Acidez_T	1	0,910	8	0,352
	2	0,919	14	0,213
	3	0,762	14	**0,002***
	5	0,729	4	**0,024***
Acido_malico	1	0,897	8	0,274
	2	0,794	14	**0,004***
	3	0,884	14	0,067
	5	0,729	4	0,024
Acido_tartarico	1	0,822	8	0,049
	2	0,876	14	0,051
	3	0,770	14	**0,002***
	5	0,729	4	**0,024***
Açucares	1	0,851	8	0,098
	2	0,844	14	**0,019***
	3	0,704	14	**< 0,001***
	5	0,729	4	**0,024***
TAP	1	0,851	8	0,097
	2	0,840	14	**0,016***
	3	0,706	14	**< 0,001***
	5	0,729	4	**0,024***

* significativo a 5%

Tabela A5.2 - Testes de Normalidade dos dados do mosto por grupos

	Clusters	Shapiro-Wilk		
		Estatistica de teste	g.l.	Valor prova
Ncachos	1	0,919	8	0,421
	2	0,954	13	0,659
	3	0,969	12	0,905
	5	0,899	4	0,428
Uvas_kg_vid	1	0,854	8	0,105
	2	0,828	13	**0,015***
	3	0,974	12	0,947
	5	0,876	4	0,321
Pcacho_kg	1	0,956	8	0,776
	2	0,890	13	0,098
	3	0,949	12	0,616
	5	0,862	4	0,268
Nvaras	1	0,938	8	0,591
	2	0,966	13	0,843
	3	0,898	12	0,148
	5	0,989	4	0,952
Pvaras_kg	1	0,932	8	0,532
	2	0,918	13	0,238
	3	0,926	12	0,340
	5	0,922	4	0,547
Pvara_g	1	0,988	8	0,991
	2	0,984	13	0,994
	3	0,892	12	0,127
	5	0,760	4	**0,048***

* significativo a 5%

Tabela A5.3 - Testes de Normalidade dos dados da videira por grupos

	Clusters	Shapiro-Wilk		
		Estatistica de teste	g.l.	Valor prova
FL1	1	0,915	8	0,387
	2	0,924	14	0,249
	3	0,803	14	**0,005***
	5	0,729	4	**0,024***
FL2	1	0,839	8	0,074
	2	0,812	14	**0,007***
	3	0,844	14	**0,018***
	5	0,729	4	**0,024***
FL3	1	0,752	8	**0,009***
	2	0,532	14	**< 0,001***
	3	0,525	14	**< 0,001***
	5	0,729	4	**0,024***
FL4	1	0,871	8	0,154
	2	0,830	14	**0,012***
	3	0,853	14	**0,024***
	5	0,729	4	**0,024***
FL5	1	0,644	8	**0,001***
	2	0,541	14	**< 0,001***
	3	0,471	14	**< 0,001***
	5	0,729	4	**0,024***
FG1	1	0,868	8	0,144
	2	0,870	14	**0,043***
	3	0,815	14	**0,008***
	5	0,729	4	**0,024***
FG2	1	0,854	8	0,105
	2	0,789	14	**0,004***
	3	0,880	14	0,058
	5	0,729	4	**0,024***
FG3	1	0,882	8	0,197
	2	0,921	14	0,230
	3	0,758	14	**0,002***
	5	0,729	4	**0,024***
FG4	1	0,896	8	0,268
	2	0,909	14	0,155
	3	0,826	14	**0,011***
	5	0,729	4	**0,024***
FG5	1	0,795	8	**0,025***
	2	0,915	14	0,185
	3	0,809	14	**0,006***
	5	0,729	4	**0,024***
FG6	1	0,751	8	**0,008***
	2	0,801	14	**0,005***
	3	0,589	14	**< 0,001***
	5	0,729	4	**0,024***
FG7	1	0,751	8	**0,008***
	2	0,835	14	**0,014***
	3	0,683	14	**< 0,001***
	5	0,729	4	**0,024***

* significativo a 5%

Tabela A5.4 - Testes de Normalidade dos dados das uvas por grupos

	Clusters	Shapiro-Wilk		
		Estatistica de teste	g.l.	Valor prova
Nota_Final	1	0,867	8	0,142
	2	0,951	14	0,582
	3	0,789	14	**0,004***
	5	0,729	4	**0,024***

* significativo a 5%

Tabela A5.5 - Testes de Normalidade dos dados do vinho por grupos

ANEXO 6 - Teste de homogeneidade das variâncias (Teste de Levene)

		Estatística de teste	g.l. 1	g.l. 2	Valor prova
SOLO	pH	1,385	4	37	0,258
	MO	3,947	4	37	**0,009***
	DA	0,867	4	37	0,493
	FF	0,735	4	37	0,574
	P_2O_5	1,137	4	37	0,354
	K_2O	0,518	4	37	0,723
	Ca	6,286	4	37	**0,001***
	Mg	1,824	4	37	0,145
	AzT	2,497	4	37	0,059
	Ni	0,885	4	37	0,482
	Cr	1,712	4	37	0,168
	Cd	1,245	4	37	0,309
	N	1,733	4	37	0,163
	B	4,189	4	37	**0,007***
	CTC	4,617	4	37	**0,004***
	CP1	2,512	4	37	0,058
	CP2	3,054	4	37	**0,029***
	CP3	0,795	4	37	0,536
	CP4	2,720	4	37	**0,044***
MOSTO	pH_mosto	4,151	4	37	**0,007***
	Acidez_T	0,709	4	37	0,591
	Acido_malico	8,384	4	37	**< 0,001***
	Acido_tartarico	2,753	4	37	**0,042***
	Açucares	0,869	4	37	0,492
	TAP	0,904	4	37	0,472
VIDEIRAS	Ncachos	1,954	3	33	0,140
	Uvas_kg_vid	0,121	3	33	0,947
	Pcacho_kg	2,568	3	33	0,071
	Peso_ton_ha	0,121	3	33	0,947
	Nvaras	1,165	3	33	0,338
	Pvaras_kg	2,464	3	33	0,080
	Pvara_g	1,790	3	33	0,168
UVAS	FL1	1,101	4	37	0,371
	FL2	8,197	4	37	**< 0,001***
	FL3	2,013	4	37	0,113
	FL4	9,095	4	37	**< 0,001***
	FL5	2,469	4	37	0,062
	FG1	3,688	4	37	**0,013***
	FG2	3,019	4	37	**0,030***
	FG3	2,207	4	37	0,087
	FG4	2,484	4	37	0,060
	FG5	1,780	4	37	0,154
	FG6	,847	4	37	0,505
	FG7	1,011	4	37	0,414
VINHO	Nota_Final	0,600	4	37	0,665

* significativo a 5%

Tabela A6.1 - Testes de homogeneidade das variâncias (teste de Levene) para

ANEXO 7 - Teste de comparações múltiplas para as variáveis originais do solo

Variável	*Cluster i*	*Cluster j*	Valor Prova
pH	1	2	0,251
		3	**0,006***
		4	0,377
		5	0,439
	2	3	**<0,001***
		4	0,802
		5	0,951
	3	4	**0,011***
		5	**0,003***
	4	5	0,796
P_2O_5	1	2	0,083
		3	0,783
		4	0,180
		5	**0,006***
	2	3	**0,018***
		4	**0,016***
		5	0,104
	3	4	0,215
		5	**0,001***
	4	5	**0,001***
AzT	1	2	0,743
		3	0,581
		4	0,643
		5	**0,002***
	2	3	0,793
		4	0,770
		5	**0,002***
	3	4	0,872
		5	**0,004***
	4	5	0,082
Cd	1	2	0,090
		3	0,314
		4	0,051
		5	0,758
	2	3	**0,002***
		4	**0,002***
		5	0,321
	3	4	0,146
		5	0,262
	4	5	**0,045***
CTC	1	2	0,211
		3	**0,004***
		4	0,671
		5	0,538
	2	3	**<0,001***
		4	0,239
		5	0,100
	3	4	0,209
		5	0,109
	4	5	0,962
MO	1	2	0,689
		3	0,984
		4	0,643
		5	**0,003***
	2	3	0,622
		4	0,802
		5	**0,004***
	3	4	0,619
		5	**0,001***
	4	5	0,099
Ca	1	2	0,135
		3	**0,006***
		4	0,782
		5	0,659
	2	3	**<0,001***
		4	0,243
		5	0,100
	3	4	0,189
		5	0,097
	4	5	0,953
Ni	1	2	**0,001***
		3	0,182
		4	**0,034***
		5	**0,001***
	2	3	**0,016***
		4	0,823
		5	0,284
	3	4	0,153
		5	**0,007***
	4	5	0,613
N	1	2	**0,005***
		3	0,822
		4	0,752
		5	0,566
	2	3	**0,002***
		4	**0,048***
		5	**0,005***
	3	4	0,644
		5	0,426
	4	5	0,906
FF	1	2	**0,034***
		3	0,280
		4	**0,016***
		5	**0,006***
	2	3	0,224
		4	0,201
		5	0,190
	3	4	0,059
		5	**0,034***
	4	5	0,796
Mg	1	2	0,384
		3	**0,032***
		4	0,360
		5	0,617
	2	3	**<0,001***
		4	0,142
		5	0,888
	3	4	0,764
		5	**0,027***
	4	5	0,234
Cr	1	2	**0,002***
		3	**0,043***
		4	**0,002***
		5	**0,002***
	2	3	0,190
		4	0,153
		5	0,332
	3	4	**0,037***
		5	0,065
	4	5	0,540
B	1	2	0,701
		3	0,950
		4	0,652
		5	**0,004***
	2	3	0,600
		4	0,805
		5	**0,004***
	3	4	0,611
		5	**0,001***
	4	5	0,099

*significativo a 5%

Tabela A7.1 - Teste de comparações múltiplas para as variáveis originais do solo