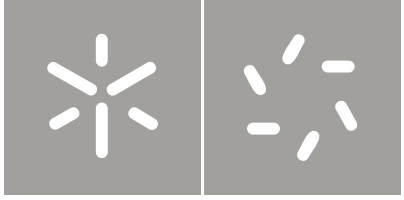
Universidade do Minho
Escola de Ciências

Daniela Botelho Catalão

Metodologia Estatística para a Avaliação de um Recurso Natural (Minho e Galiza)

Março de 2014

Metodologia Estatística para a Avaliação de um Recurso Natural (Minho e Galiza)
Daniela Botelho Catalão
UMinho | 2014



Universidade do Minho
Escola de Ciências

Daniela Botelho Catalão

Metodologia Estatística para a Avaliação de um Recurso Natural (Minho e Galiza)

Dissertação de Mestrado
Estatística

Trabalho efetuado sob a orientação de
Professora Doutora Arminda Manuela Gonçalves
Professora Doutora Susana Faria

Março de 2014

DECLARAÇÃO

Nome: Daniela Botelho Catalão

Correio electrónico: danielacatalao@hotmail.com

Tel./Tlm.: 966358389

Número do Bilhete de Identidade: 13102496

Título da dissertação:

Metodologia Estatística para a Avaliação de um Recurso Natural (Minho e Galiza)

Ano de conclusão: 2014

Orientador(es): Doutora Arminda Manuela Gonçalves e Doutora Susana Faria

Designação do Mestrado: Mestrado em Estatística

Área de Especialização: Estatística

Escola: Ciências

Departamento: Matemática e Aplicações

Guimarães, ___/___/_____

Assinatura: _____

Agradecimentos

Os meus agradecimentos:

- *às professoras Doutora Arminda Manuela Gonçalves e Doutora Susana Faria pela orientação, disponibilidade, paciência, pelo acompanhamento e esclarecimentos ao longo de toda a realização deste trabalho;*
- *à minha amiga Helena Piairo pelos bons conselhos em todos os momentos de indecisão, apoio e pela amizade, simpatia e tempo despendido para me ajudar;*
- *aos meus colegas, por toda a cooperação, dedicação, troca de ideias mas acima de tudo pela amizade;*
- *por último, um profundo agradecimento para as pessoas que são mais importantes para mim, a minha família, por todos os sacrifícios, carinho, paciência, força e por terem sempre acreditado em mim.*

A todos, o meu sincero agradecimento.

Resumo

A monitorização e a gestão de projetos ambientais, em particular, a avaliação de recursos naturais, envolvem um conjunto de problemas e desafios relacionados com a natural complexidade dos fenómenos envolvidos e com a escassez da informação disponível. As metodologias estatísticas têm sido largamente aplicadas para a resolução destes problemas.

Perante um mercado tão globalizado e cada vez mais competitivo, é cada vez maior a preocupação do viticultor em fazer uma gestão da informação detalhada sobre os solos aptos à viticultura, uma vez que ao conhecer melhor a área produtiva permite aprimorar a qualidade e a quantidade do vinho. O principal fator que influencia a qualidade do vinho é o nível de maturação da uva e, o bom desenvolvimento deste processo de maturação está intimamente relacionado com as castas e com as condições ambientais, tais como o solo e o clima.

De modo a proteger e a melhorar este património cultural e económico, que é a vinha e a uva, é fundamental aperfeiçoar a rentabilidade dos sistemas produtivos e minimizar a degradação do recurso solo, para garantir a sustentabilidade dos sistemas de produção vitivinícola.

Neste estudo pretendeu-se desenvolver uma metodologia geral, recorrendo a metodologias da área da Estatística Multivariada e da Inferência Estatística (em particular, testes de hipóteses), com o objetivo de avaliar e interpretar a variabilidade de um alargado conjunto de informação sobre as características físico-químicas do solo que influenciam o desenvolvimento das videiras, a qualidade do mosto e das uvas e, por consequência, a qualidade dos vinhos. Com este conhecimento pretende-se otimizar a produção no que diz respeito à qualidade e quantidade de vinho.

Este estudo incidiu sobre dois talhões distintos, um deles situado no Minho (Portugal) e o outro localizado na Galiza (Espanha). O primeiro talhão encontra-se sobre um solo residual granítico e compreende a casta tinta Vinhão, enquanto o talhão da Galiza encontra-se sobre um terreno que é um aluvião, sendo constituído pela casta branca Alvarinho.

A análise dos resultados obtidos a partir das duas bases de dados que incorporavam um conjunto de dados recolhidos em campo e em laboratório, relativos ao ano 2011, pôs em evidência a influência do solo sobre o rendimento e qualidade das uvas e do vinho. Foi possível identificar algumas das características do solo que estão correlacionadas com o desenvolvimento das videiras e com a qualidade das uvas e respetivos vinhos, nestas duas vinhas em estudo.

Abstract

The monitoring and management of environmental projects, in particular the assessment of natural resources involve a set of issues and challenges related to the natural complexity of the phenomena involved and the scarcity of information available. The statistical methods have been widely applied to solve these problems.

Facing such a globalized and increasingly competitive market, the concern of the grape grower is focus in doing the best management of the suitable soils for viticulture, which is, together with the clime conditions, one of the most important factors that influences the quality of the wine, as well as the level of ripeness of the grapes and the good development of this maturation process.

Aiming to protect and enhance this cultural and economic heritage, which is the vineyard and the grape is crucial to improve the profitability of production systems and minimize the degradation of soil resources, to ensure the sustainability of wine production methods.

In this study we pretend to develop a methodology, using Multivariate Statistical Analysis and Statistical Inference, in order to evaluate and interpret the variability of a broad set of information on the physical-chemical soil properties that influences the growth of vines, the quality of wine and grapes and, consequently, the quality of wines. With this knowledge we pretend to optimize the wine production, as regards quality and quantity of wine.

This study refers two different parcels of land, one located in Minho (Portugal) and the other in Galiza (Spain). The first plot is shown on a soil residual granitic and understands the grape variety Vinhão, while the plot of Galicia is on land that is an alluvial, being constituted by white grape Alvarinho.

The analysis of the results from those two different databases that incorporated a set of data collected in field and laboratory, for the year 2011, highlighted the influence of the soil on the yield and quality of grapes and wine. It was possible to identify the physical and chemical properties of the soil that influence the vineyard grouth, the quality of the grapes and therefore their respective wines.

Lista de Variáveis em Estudo

Variáveis do Solo

AzT: Azoto Total

B: Boro

Ca: Cálcio assimilável

Cd: Cádmio

Cr: Crómio

CTC: Capacidade de Troca Catiónica

DA: Densidade Aparente

FF: Fração Fina

FG: Fração Grosseira

$K2O$: Potássio Assimilável

Mg: Magnésio Assimilável

MO: Matéria Orgânica

N: Nitratos

Ni: Níquel

pH: pH

$P2O5$: Fósforo Assimilável

Variáveis da Videira

$Ncachos$: Número de cachos por videira

$Pcacho$: Peso médio do cacho por videira

$Pbruto$: Peso bruto por videira

$Uvas\_kg\_vid$: Peso de cachos por videira

Variáveis do Mosto

$pH.mosto$: pH do mosto

$Ac.tart$: Ácido tartárico

$Ac.mal$: Ácido málico

$Ac.TOT$: Acidez Total

TAP: Teor de Álcool Provável

$FL1M$: Família de compostos em C_6 do aroma do mosto na fração livre

$FL2M$: Família de álcoois do aroma do mosto na fração livre

$FL3M$: Família de álcoois monoterpénicos do aroma do mosto na fração livre

$FL4M$: Família de fenóis voláteis do aroma do mosto na fração livre

$FL5M$: Família de compostos carbonilados do aroma do mosto na fração livre

$FG1M$: Família de compostos em C_6 do aroma do mosto na fração glicosilada

$FG2M$: Família de álcoois do aroma do mosto na fração glicosilada

$FG3M$: Família de álcoois monoterpénicos do aroma do mosto na fração glicosilada

$FG4M$: Família de óxidos e dióis monoterpénicos do aroma do mosto na fração glicosilada

$FG5M$: Família de norisoprenóides em C_{13} do aroma do mosto na fração glicosilada

$FG6M$: Família de fenóis voláteis do aroma do mosto na fração glicosilada

$FG7M$: Família de compostos carbonilados do aroma do mosto na fração glicosilada

$Amarelo\_M$: Percentagem de cor amarela

$Vermelho\_M$: Percentagem de cor vermelha

$Azul\_M$: Percentagem de cor azul

$Sumo\_T\_M$: Sumo turvo

$Sumo\_L\_M$: Sumo límpido

$Rend\_sumo\_M$: Percentagem do rendimento em sumo

Variáveis das Uvas

$FL1U$: Família de compostos em C_6 do aroma das uvas na fração livre

$FL2U$: Família de álcoois do aroma das uvas na fração livre

$FL3U$: Família de álcoois monoterpénicos do aroma das uvas na fração livre

$FL4U$: Família de fenóis voláteis do aroma das uvas na fração livre

$FL5U$: Família de compostos carbonilados do aroma das uvas na fração livre

$FL6U$: Família de óxidos e dióis monoterpénicos do aroma das uvas na fração livre

$FL7U$: Família de ácidos gordos voláteis do aroma das uvas na fração livre

$FG1U$: Família de compostos em C_6 do aroma do mosto na fração glicosilada

$FG2U$: Família de álcoois do aroma do mosto na fração glicosilada

$FG3U$: Família de álcoois monoterpénicos do aroma do mosto na fração glicosilada

$FG4U$: Família de óxidos e dióis monoterpénicos do aroma do mosto na fração glicosilada

$FG5U$: Família de norisoprenóides em C_{13} do aroma do mosto na fração glicosilada

$FG6U$: Família de fenóis voláteis do aroma do mosto na fração glicosilada

$FG7U$: Família de compostos carbonilados do aroma do mosto na fração glicosilada

$Amarelo\_U$: Percentagem de cor amarela

$Vermelho\_U$: Percentagem de cor vermelha

$Azul\_U$: Percentagem de cor azul

$Massa$: Massa

$Sumo\_T\_U$: Sumo turvo

$Sumo\_L\_U$: Sumo límpido

$Rend\_sumo\_U$: Percentagem do rendimento em sumo

Variáveis do vinho

$Nota.Final$: Avaliação final atribuída ao vinho

Abreviaturas

AC: Análise de *Clusters*

ACP: Análise de Componentes Principais

AF: Análise Fatorial

AFE: Análise Fatorial Exploratória

$BSR\ 1$: Bodega Santiago Ruiz 1

$BSR\ 2$: Bodega Santiago Ruiz 2

$BSR\ 3$: Bodega Santiago Ruiz 3

$BSR\ 4$: Bodega Santiago Ruiz 4

$BSR\ 5$: Bodega Santiago Ruiz 5

$BSR\ 6$: Bodega Santiago Ruiz 6

$BSR\ 7$: Bodega Santiago Ruiz 7

$BSR\ 8$: Bodega Santiago Ruiz 8

$BSR\ 9$: Bodega Santiago Ruiz 9

$BSR\ 10$: Bodega Santiago Ruiz 10

$BSR\ 11$: Bodega Santiago Ruiz 11

$BSR\ 12$: Bodega Santiago Ruiz 12

$BSR\ 1-2$: Bodega Santiago Ruiz $1-2$

$BSR\ 3-4$: Bodega Santiago Ruiz $3-4$

$BSR\ 5-6$: Bodega Santiago Ruiz $5-6$

$BSR\ 7-8$: Bodega Santiago Ruiz $7-8$

$BSR\ 9-10$: Bodega Santiago Ruiz $9-10$

$BSR\ 11-12$: Bodega Santiago Ruiz $11-12$

$B1N$: Bloco 1 Nascente

$B1C$: Bloco 1 Centro

$B1P$: Bloco 1 Poente

$B2N$: Bloco 2 Nascente

$B2C$: Bloco 2 Centro

$B2P$: Bloco 2 Poente

$B3N$: Bloco 3 Nascente

$B3C$: Bloco 3 Centro

$B3P$: Bloco 3 Poente

Cl: Cloro

CP: Componente Principal

Cu: Cobre

CVRVV: Comissão de Viticultura da Região dos Vinhos Verdes

C_6: 6 Átomos de Carbono

C_{13}: 13 Átomos de Carbono

ENE-WSW: Este-Nordeste-Oeste-Sudoeste

ET: Estatística de Teste

EVAG: Estação Vitivinícola Amândio Galhano

Fe: Ferro

K: Potássio

KMO: Kaiser-Meyer-Olkin

LSD: Least Significant Difference

Mb: Molibdeno

MMQ: Método dos Mínimos Quadrados

Mn: Manganês

NA: *Not Available*

P: Fósforo

S: Enxofre

S-SSO: Sul-Sul-Sudoeste

SPSS: *Statistical Package for Social Sciences*

SQE: Soma de Quadrados dos Resíduos

SQR: Soma de Quadrados da Regressão

SQT: Soma de Quadrados Total

Zn: Zinco

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Enquadramento - Influência do Solo na Qualidade do Vinho

A competição do mercado a nível mundial, no que diz respeito aos preços e qualidade das uvas e produtos derivados, criou uma pressão económica que tem motivado os viticultores a procurar novos meios para melhorar a produção (Pacheco, 1999).

O cultivo da videira é uma das atividades agrícolas que proporciona grande rentabilidade aos viticultores. Porém, para que isso se concretize, o viticultor deve produzir uvas de boa qualidade e com boa produtividade para se tornar mais competitivo (Siqueira *et al.*, 2008).

A produção de vinhos de qualidade resulta da interação de um conjunto de fatores, tais como o solo, o clima e as atividades humanas; portanto, é fundamental avaliar estes fatores para aumentar e melhorar a rentabilidade dos sistemas produtivos e minimizar a degradação do recurso do solo e, assim, garantir a sustentabilidade dos sistemas de produção vitivinícola.

Nos dias de hoje há um grande empenho para conseguir encontrar os parâmetros do solo que apresentam efetivamente maior influência na vinha e consequentemente no vinho, ou seja, que permitam um aumento da qualidade do produto final (Flores, 2011).

O solo é a camada superficial da crosta terrestre que serve de suporte e fornece nutrição às plantas nele existente. Toda a vida à superfície da terra depende desta camada. É no solo que vivem as videiras e que dele produzem as uvas e, consequentemente, dá origem ao vinho; a qualidade do vinho irá depender das características físicas (cor, textura, estrutura, infiltração, trocas gasosas, dureza do solo, porosidade, permeabilidade,...), químicas (poder de absorção, pH, composição química) e biológicas do solo, pois são estas que irão regular a sua fertilidade (Afonso, 2009).

A máxima qualidade e quantidade de produção depende principalmente do estado híbrido

do solo e da nutrição mineral das plantas, havendo uma nutrição adequada se o nível de água disponível no solo for suficiente para assegurar a absorção dos nutrientes. Se por um lado aumenta a absorção de elementos nutritivos principais como o azoto e o potássio, por outro reduz a absorção do cálcio e sobretudo o magnésio, podendo provocar graves desequilíbrios nutricionais e doenças fisiológicas (Zaballa *et al*., 1997).

No entanto, para uma vinha ser de qualidade, os solos querem-se preferencialmente pobres ou pouco férteis, uma vez que a videira tem baixas necessidades nutritivas. A vinha essencialmente requer, uma moderada ou baixa quantidade de azoto, potássio e fósforo e quantidades residuais de magnésio, manganês, ferro, zinco, cobre e boro, não necessitando também de muita água. Se a videira obtiver excesso de nutrientes (em particular de azoto ou potássio), e rega em demasia, irá produzir uma vegetação excessiva cujos frutos serão constituídos por bagos gordos mal expostos e mais sujeitos ao ataque de fungos e bactérias; o vinho produzido será de fraca ou apenas aceitável qualidade (Pacheco, 1999).

De acordo com Araújo (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, no entanto, apenas são absorvidos pela vinha em reduzidas quantidades. A sua carência provoca doenças e causam fitotoxicidade quando absorvidos em excesso.

Contudo, é muito importante e necessário conhecer as castas e as suas capacidades de absorção de certos elementos minerais. De facto, estudos referenciados em Pacheco (1999) comprovam que as mesmas concentrações de alguns nutrientes em diferentes tipos de castas e diferentes tipos de solo provocam diferentes resultados na videira e, consequentemente, nas uvas.

Segundo Somers (1977), o potássio desempenha uma função importante na qualidade organolética e de conservação de um vinho através da sua influência sobre o pH do mesmo. A carência de potássio diminui ligeiramente o teor de açúcares nos bagos sendo este, contudo, independente do nível de nutrição fora da zona de carência; a boa produção de uma parcela só pode ser atingida em condições de nutrição potássica elevada. A rega de uma vinha leva a uma maior absorção de potássio e, consequentemente, ao aumento do pH do vinho implicando uma menor proporção de pigmentos corados, uma vez que a concentração destes aumenta com a diminuição do tamanho dos bagos (Mota, 2005). Estudos referenciados em Pacheco (1999) indicam que para uma determinada casta o aumento da concentração de potássio leva a uma maior produção, e para as mesmas condições de solo na mesma região, mas diferente casta, não se obtiveram resultados significativos com as mesmas concentrações de potássio.

As respostas de produção da videira a diferentes concentrações de azoto diferem de acordo com a casta e o clima. Para além disso, um teor de humidade do solo inadequado pode limitar a

absorção do azoto e, consequentemente, conduzir a efeitos negativos da *performance* dos vinhos (Pacheco, 1999). Estudos referenciados por Pacheco (1999) revelam-se contraditórios no que diz respeito à influência do azoto nos ácidos do mostos e, consequentemente, nos vinhos.

A videira não requer grandes quantidades de fósforo, mas constata-se que é um nutriente essencial uma vez que a sua carência provoca a redução do crescimento das raízes, das sementes e dos bagos. Estudos referenciados em Pacheco (1999) indicam que níveis crescentes de fósforo correspondem a uma maior produção numa determinada casta e com outras castas não existem diferenças significativas com as variações de concentração de fósforo.

Sabe-se também que o boro é um micronutriente essencial para o crescimento da videira e quando presente em doses incorretas causa prejuízos na viticultura (Melo *et al.*, 2008).

No que diz respeito à relação entre a nutrição mineral e as características qualitativas do mosto é complexa devido à dificuldade em se definir com exatidão, o conceito de qualidade do mosto, mas o interesse por parte dos investigadores é cada vez maior (Pacheco, 1999). Segundo Failla *et al.* (1996) o papel da nutrição mineral na acidez do mosto é complexa, pelo motivo de os nutrientes poderem ter um efeito direto ou indireto. Estudos referenciados em Pacheco (1999) afirmam que maiores concentrações de azoto levam a aumento do pH, do teor de ácido málico e do ácido tartárico do mosto; indicam também que existe uma relação positiva entre o potássio e o teor de álcool de ácido málico dos bagos. Elevado ácido málico no mosto pode influenciar as capacidades gustativas do vinho.

Relativamente a algumas características físicas do solo, sabe-se que a existência de seixos/calhaus na vinha são muito importantes no solo, uma vez que facilitam a sua drenagem e o crescimento da raiz e permitem um melhor arejamento e aquecimento do solo; à superfície impedem parte da evaporação, retendo alguma humidade no solo essencial à planta; evitam ainda a erosão, absorvem calor e conseguem transmiti-lo em profundidade, o que se torna vantajoso em climas frios (Borges, 2008).

Como se pode constatar, a ligação da geologia à viticultura é evidente; surge da aplicação da cartografia geológica e de solos, climatologia, hidrologia e medição de parâmetros pontuais e globais do solo. A geologia permite 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 (Layon *et al.*, 2004).

O conhecimento adequado das características do solo e das exigências da vinha é fundamental para se conseguir atingir níveis de produção economicamente rentáveis e de qualidade (Jordão, 2007).

1.2 Objetivos do Estudo

Neste estudo pretende-se desenvolver uma metodologia geral, recorrendo a técnicas da área da Estatística Multivariada e da Inferência Estatística (em particular, testes de hipóteses), com o objetivo de avaliar e interpretar a variabilidade de um alargado conjunto de informação sobre as características físico-químicas do solo que influenciam o desenvolvimento das videiras, a qualidade do mosto e das uvas e, por consequência, a qualidade dos vinhos. Pretende-se com este conhecimento, e num mercado cada vez mais competitivo, otimizar qualitativamente e quantitativamente a produção.

Em particular, pretende-se a:

- realização de uma análise descritiva para descrever o comportamento das variáveis e detetar possíveis *outliers*;
- identificação, nas parcelas em estudo, de zonas com semelhantes características físicas e químicas do solo;
- identificação das variáveis do solo que influenciam a qualidade do vinho e quais as zonas das parcelas que apresentam grandes/pequenas concentrações dessas variáveis.

Este estudo recaiu sobre duas zonas distintas: uma zona em terrenos situados na Estação Vitivinícola Amândio Galhano (EVAG), no concelho dos Arcos de Valdevez pertencente à Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV) e a outra zona em Bodega Santiago Ruiz, situado na Galiza em Espanha. Note-se que as vinha em estudo do Minho compreendem apenas a casta tinta Vinhão, enquanto a vinha da Galiza a casta branca Alvarinho. Na quinta Campos Lima, situada no concelho de Arcos de Valdevez, o solo é um solo residual granítico, enquanto na Bodega Santiago Ruiz em *Tomiño* (Galiza) o solo é um aluvião, constituído na sua maioria por acumulações de seixos, frequentemente cobertos por camadas arenoargilosas e matéria orgânica (Azevedo e Oliveira, 2010a; Azevedo e Oliveira, 2010b).

Os dados presentes neste trabalho são referentes ao ano de 2011. Disponibiliza-se ainda a informação que na quinta do Minho, a recolha das uvas para o estudo da produtividade, mostos e prova foi efectuada a $16/09/2011$, e para a quinta da Galiza realizada no dia $23/08/2011$. Quanto à recolha das amostras do solo para análise, esta foi realizada em Abril de 2010, em ambas as quintas, tendo sido posteriormente analisadas no Laboratório de Análise de Solos da Escola Superior Agrária do Instituto Politécnico de Viana do Castelo.

As várias metodologias de análise estatística foram realizadas utilizando o *software* R, versão $2.15.2$, com o auxílio de várias *packages*, e o *software* SPSS (*Statistical Package for Social Sciences*).

1.3 Estrutura do Trabalho

O presente trabalho está organizado em 5 capítulos e respetivos subcapítulos.

Inicialmente é apresentada uma introdução onde se faz um breve enquadramento do tema em estudo, explicando a influência do solo na qualidade do vinho, bem como são apresentados os principais objetivos deste trabalho.

No 2^{o} capítulo descrevem-se, de forma sucinta, as metodologias estatísticas utilizadas no âmbito deste trabalho, fazendo-se uma breve exposição das bases teóricas para a análise Estatística Multivariada e para a Inferência Estatística.

No 3^{o} capítulo apresenta-se a caracterização das parcelas de cultivo e respetivas castas em estudo, a caracterização das variáveis em estudo e a realização de uma análise descritiva univariada para analisar/avaliar os seus comportamentos.

No 4^{o} capítulo apresentam-se e discutem-se os resultados obtidos pela aplicação da metodologia geral desenvolvida, na área da Estatística Multivariada e da Inferência Estatística, para cada uma das parcelas em Estudo (Minho e Galiza), com o objectivo de se analisar a influência de algumas características do solo na produtividade da videira, na qualidade das uvas, do mosto e do vinho.

Finalmente, no 5^{o} capítulo apresentam-se as principais conclusões decorrentes do trabalho desenvolvido, indicando alguns comentários acerca dos resultados obtidos, das dificuldades encontradas ao longo deste trabalho e de algumas linhas de investigação para trabalho futuro.

Capítulo 2

Enquadramento Teórico

2.1 Associação e Correlação

Ao iniciarem-se alguns estudos estatísticos, surge naturalmente a pergunta se as variáveis aleatórias em análise se encontram associadas/correlacionadas, e portanto, para responder a esta questão é importante determinar o sentido e intensidade desta relação, quando existe.

Seja X e Y duas variáveis aleatórias. Uma medida que estuda a relação entre as variáveis é a covariância, σ_{XY}, definida pelo valor esperado dos produtos dos desvios em relação à média

$$\sigma_{XY} = cov(X, Y) = E[(X - E(X))(Y - E(Y))]. \tag{2.1}$$

Dada uma amostra bivariada de n valores (x_i, y_i), $i = 1, \ldots, n$, provenientes dessas variáveis, a covariância pode ser estimada pela seguinte expressão

$$c_{xy} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}, \tag{2.2}$$

onde $\bar{x} = \frac{\sum_i^n x_i}{n}$ e $\bar{y} = \frac{\sum_i^n y_i}{n}$ são as médias amostrais de X e Y, respetivamente.

Apesar da covariância ser uma estatística adequada para analisar a relação linear entre duas variáveis, ela não é adequada para medir a intensidade da relação entre variáveis, dado que é influenciada pelas unidades de medida de cada variável. É útil uma vez que o seu sinal indica o tipo de relação linear presente: positiva, negativa ou ausência de correlação. Para evitar a influência da ordem de grandeza das unidades de medida de cada variável, divide-se a covariância pelo produto dos desvios padrão das duas variáveis em estudo, σ_X e σ_Y, dando origem ao coeficiente de correlação populacional de *Pearson*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \times \sigma_Y}. \tag{2.3}$$

Dada uma amostra bivariada de n valores (x_i, y_i), o coeficiente de correlação amostral de *Pearson* é dado por

$$r_{xy} = \frac{c_{xy}}{s_x \times s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2.4}$$

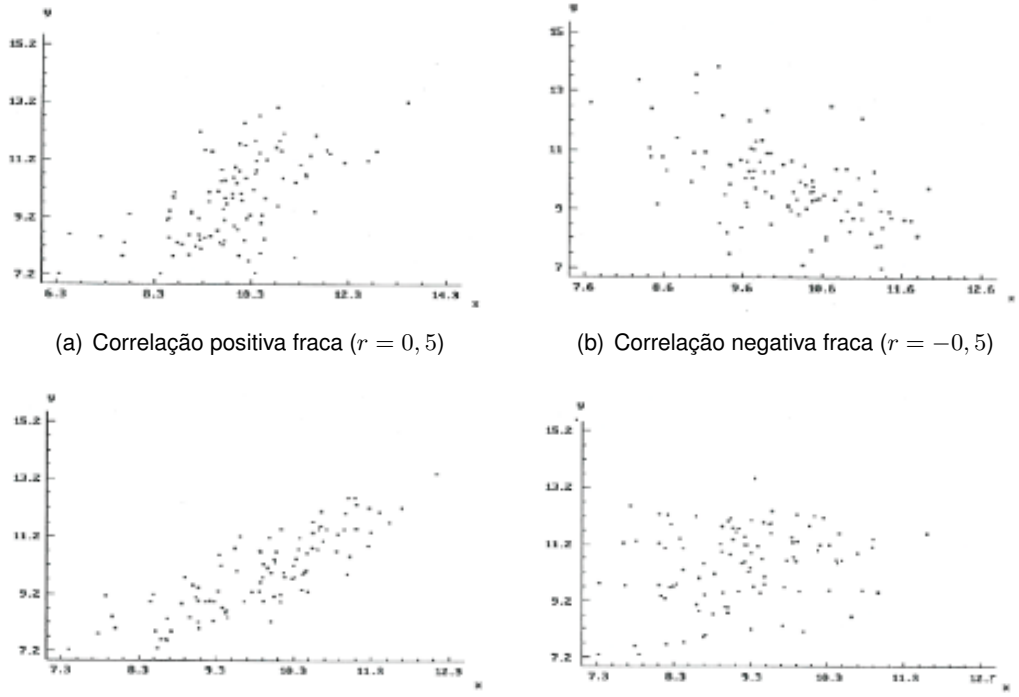
onde $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}$ e $s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})}{n-1}$ são estimativas para as variâncias populacionais das variáveis em estudo, X e Y, sendo $\bar{x}$ e $\bar{y}$ as suas estimativas para as médias populacionais (Athayde, 2011; Milton, 2003).

Note-se que este coeficiente apenas toma valores no intervalo $[-1, 1]$ conseguindo-se a partir dele classificar a relação linear de acordo com a Tabela 2.1.

Tabela 2.1: Classificação dos valores do coeficiente de correlação de *Pearson*.

valor da correlação	Classificação
$r_{xy} = -1$	relação linear negativa perfeita
$-1 < r_{xy} < 0$	relação linear negativa
$r_{xy} = 0$	ausência de relação linear
$0 < r_{xy} < 1$	relação linear positiva
$r_{xy} = 1$	relação linear positiva perfeita

A Figura 2.1 apresenta exemplos de diagramas de dispersão, associados aos coeficientes de correlação $r = 0,5$, $r = -0,5$, $r = 0,85$ e $r = 0$.



(a) Correlação positiva fraca ($r = 0,5$)

(b) Correlação negativa fraca ($r = -0,5$)

(c) Correlação positiva forte ($r = 0,85$)

(d) Ausência de correlação ($r = 0$)

Figura 2.1: Diagramas de dispersão com indicação do coeficiente de correlação associado (Murteira *et al.*, 2002).

Nos casos em que a relação entre duas variáveis não é linear, uma delas não é contínua, as observações não são selecionadas aleatoriamente ou ambas as variáveis não seguem uma distribuição Normal, outras alternativas de coeficientes são aplicadas, como por exemplo, o coeficiente de correlação de *Spearman*.

O coeficiente de correlação de *Spearman*, R_s, é uma medida de associação não paramétrica, que apenas exige que ambas as variáveis em estudo sejam medidas pelo menos em escala ordinal de modo que as observações possam ser ordenadas por *ranks*, consoante o valor assumido em cada variável.

Seja (x_i, y_i), $i = 1, ..., n$, pares de observações e denote-se por $R(x_i)$ e $R(y_i)$ os *ranks* das observações x_i e y_i, respetivamente. Ordenando os n indivíduos por *ranks* segundo as duas variáveis, a correlação seria perfeita se $R(x_i) = R(y_i)$ para todos os $i's$. Assim, para o cálculo do coeficiente de correlação de *Spearman* é essencial determinarem-se as diferenças entre estes *ranks*

$$D_i = R(x_i) - R(y_i), \tag{2.5}$$

com o objetivo de indicar a disparidade entre os dois conjuntos de *ranks*. Quanto maior for a magnitude de D_i, menos perfeita é a associação entre as duas variáveis (Siegel, 1975).

Como afirma Higgins (2004), o coeficiente de correlação de *Spearman* é obtido aplicando a fórmula do coeficiente de correlação de *Pearson* não ao par de observações (x_i, y_i), mas sim ao par dos *ranks* $(R(x_i), R(y_i))$, com $i = 1, ..., n$, obtendo-se assim

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n^3 - n}. \tag{2.6}$$

Acontece por vezes existirem observações que apresentam o mesmo *rank* na mesma variável. Nestas situações atribui-se a cada observação a média dos *ranks* que lhes seriam atribuídos caso não tivesse ocorrido empate. No caso da proporção de empates ser pequena, continua a usar-se a fórmula (2.6), uma vez que o seu efeito sobre r_s é insignificante, caso contrário é necessário incorporar um fator de correção para os empates no cálculo de r_s (Siegel e Castellan, 1988)

$$r_s = \frac{(n^3 - n) - 6 \sum_{i=1}^{n} d_i^2 - \frac{T_{x_i} + T_{y_i}}{2}}{\sqrt{(n^3 - n)^2 - (T_{x_i} + T_{y_i})(n^3 - n) + T_{x_i} T_{y_i}}}, \tag{2.7}$$

onde

$$T_{x_i} = \sum_{j=1}^{g_i} t_j^3 - t_j \qquad e \qquad T_{y_i} = \sum_{j=1}^{p_i} l_j^3 - l_j, \tag{2.8}$$

representando g_i e p_i o número de grupos de observações empatadas na variável x_i e y_i, respetivamente, e t_j e l_j são o número de observações empatadas em cada grupo de empates da

variável x_i e y_i, respetivamente.

Após o cálculo destes coeficientes, é importante determinar a "significância" da relação observada, ou seja, determinar, com um certo grau de confiança, se existe esta associação na população da qual se recolheu a amostra e que serviu para o cálculo do coeficiente.

No caso paramétrico, para querer comprovar a significância de um valor observado de correlação r_{xy}, deve satisfazer-se a exigência relativa à normalidade dos dados. Caso as duas variáveis não provenham de uma população Normal bivariada, deve-se utilizar um coeficiente não paramétrico, como por exemplo, o coeficiente de correlação de Spearman, cujos testes de hipóteses não exigem nenhuma suposição sobre a distribuição das amostras, mas apenas que a escala de medida seja pelo menos ordinal (Siegel, 1975).

Neste secção apenas se irá abordar o teste de significância não paramétrico sobre o coeficiente de correlação de Spearman, R_s, uma vez que é este o usado neste trabalho.

Dada uma amostra cujas observações foram retiradas aleatoriamente de uma população, pretende-se testar (Higgins, 2004):

H_0: Não existe associação entre as variáveis X e Y na população ($R_s = 0$)

vs

H_1: Existe associação entre as variáveis X e Y na população ($R_s \neq 0$).

Sob H_0, a estatística de teste é dada por

$$Z = \frac{r_s}{\sqrt{var(r_s)}} = r_s\sqrt{n-1}. \tag{2.9}$$

A distribuição de Z é aproximada à distribuição Normal padrão com média nula e desvio-padrão 1, desde que as amostras sejam grandes (Higgins, 2004), ou seja, desde que o número de observações de cada amostra seja pelo menos igual a 20, como refere Siegel e Castellan (1988). Ao nível de significância α rejeita-se H_0 se $|Z| \geq z_{\frac{\alpha}{2}}$, onde $z_{\frac{\alpha}{2}}$ representa o quantil $\frac{\alpha}{2}$ da distribuição $N(0,1)$.

Para amostras mais reduzidas, a aproximação à distribuição Normal padrão para a distribuição amostral Z não é suficiente e, como tal, devem ser utilizadas tabelas especiais[2] que contêm valores críticos, c, os quais devem ser comparados com o valor de r_s, permitindo assim tornar o teste mais exato. Rejeita-se H_0 se $|r_s| \geq c$, sendo o valor de prova dado por $2P(r_s \geq c) = 2P(r_s \leq -c)$ (Higgins, 2004).

[2] Esta tabela pode-se encontrar no livro de Siegel e Castellan (1988).

2.2 Análise Estatística Multivariada

Muitos dos problemas envolvidos na Estatística são constituídos por dados de natureza multivariada, ou seja, os dados correspondem a observações não de uma, mas de um determinado número de características num conjunto de indivíduos.

O objetivo das técnicas de redução de dimensionalidade é, precisamente, reduzir a dimensionalidade dos dados sem perder, contudo, informação. A técnica de análise que permite reduzir tal dimensionalidade é a Análise Fatorial (AF), sendo a Análise de Componentes Principais (ACP) muitas vezes utilizada para o mesmo fim.

Estas análises multivariadas muitas vezes também têm como objetivo tentar agrupar/classificar indivíduos consoante as características que lhes são atribuídas (valores das variáveis). Para tal efeito é usual realizar-se uma Análise de *Clusters* (AC).

2.2.1 Análise de Componentes Principais (ACP)

A ACP tem como objetivo principal transformar um conjunto original de p variáveis que se encontram correlacionadas num novo conjunto de p variáveis não correlacionadas, as chamadas componentes principais (CPs), que resultam de combinações lineares do conjunto inicial. Uma análise de componentes principais envolve um elevado número de variáveis originais num problema com um número reduzido de variáveis, as CPs, uma vez que, se as primeiras componentes principais explicarem a maior parte da variabilidade total dos dados, pode-se reter apenas estas componentes, reduzindo assim a dimensionalidade dos dados que, posteriormente, irão ser usados em outros estudos. Tais CPs são calculadas por ordem decrescente de importância, ou seja, a primeira CP é a que explica a maior quantidade de variância dos dados; a segunda CP é a componente que de seguida explica a maior quantidade de variabilidade dos dados e, assim, sucessivamente. Note-se que não se está a desperdiçar nenhuma variável original uma vez que todas as CPs são combinações lineares de todas as variáveis originais; tal resultado leva a que não faça sentido usar variáveis originais categóricas (Reis, 2001).

Do ponto de vista geométrico, o que acontece numa ACP é uma rotação ortogonal dos eixos no espaço p dimensional (p variáveis originais), ou seja, as CPs representam um novo sistema de coordenadas obtido pela rotação dos sistemas de eixos originais. Os novos eixos fornecem as direções da máxima variabilidade.

Para uma melhor compreensão, considere-se uma amostra de duas variáveis de n observações, representadas na Figura 2.2. Obtêm-se, por meio da rotação dos eixos originais X_1 e X_2 , um novo sistema de coordenadas, em que Y_1 representa o principal eixo e Y_2 o eixo secundário

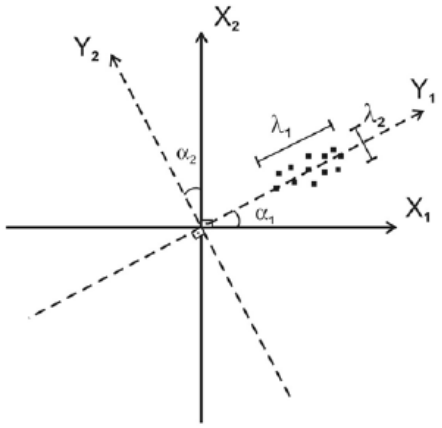


Figura 2.2: Sistema de coordenadas antes e após a rotação, (X_1, X_2) e (Y_1, Y_2) respetivamente, ângulos de rotação e valores próprios (Caten, 2008).

(componentes principais), sendo α_1 o ângulo formado entre o eixo original X_1 e Y_1 e α_2 o ângulo formado entre o eixo original X_2 e Y_2. Os valores próprios da matriz de variâncias/covariâncias dos dados, λ_1 e λ_2, representam a variabilidade contida em cada um dos novos eixos (Johnson e Wichern, 2007).

2.2.1.1 Propriedades da matriz de variâncias/covariâncias

Quando as variáveis em estudo são todas do tipo quantitativo, o conjunto de dados multivariados são usualmente apresentados numa matriz, denominada por $\mathbf{X}$ de dimensão $n \times p$,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \tag{2.10}$$

onde n representa o número de indivíduos, p é o número de variáveis medidas em cada indivíduo e x_{ij} é o valor da j-ésima variável no indivíduo i. Pode considerar-se esta matriz como uma realização de um vetor p-variado $\mathbf{X}$, onde $\mathbf{X}^{\mathrm{T}} = [X_1 X_2 \dots X_p]$, com:

- um vetor valor médio dado por $\boldsymbol{\mu}^{\mathrm{T}} = [\mu_1\ \mu_2\ \cdots\ \mu_p]$, onde $\mu_i = \mathbb{E}(X_i)$,
- um vetor variância definido por $\boldsymbol{\sigma^2}^{\mathrm{T}} = \left[\sigma_1^2\ \sigma_2^2\ \cdots\ \sigma_p^2\right]$, onde $\sigma_i^2 = Var[X_i] = \mathbb{E}[(X_i - \mathbb{E}(X_i))^2]$,
- uma matriz de variância/covariância $\boldsymbol{\Sigma}$,

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{bmatrix} \tag{2.11}$$

onde cada elemento $\sigma_{ij},\ i,j = 1,\dots,p$ e $i \neq j$, representa a covariância entre o indivíduo i e j, sendo o seu valor dado por $\sigma_{ij} = \mathbb{E}[(x_i - \mathbb{E}(x_i))(x_j - \mathbb{E}(x_j))]$. Note-se que quando $i = j$, $\sigma_{ii} = \sigma_i^2$ representam as variâncias das variáveis em estudo, estando estas localizadas na diagonal principal da matriz $\mathbf{\Sigma}$. Esta matriz também pode ser expressa na forma matricial por

$$\mathbf{\Sigma} = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})^T(\mathbf{X} - \boldsymbol{\mu})] = \mathbb{E}[\mathbf{X}^T\mathbf{X}] - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\mu}. \tag{2.12}$$

Note-se que o facto de esta matriz ser simétrica e definida positiva, ou seja, $\mathbf{\Sigma} = \mathbf{\Sigma}^T$ e $\forall \mathbf{a} \in \mathbb{R}^p, \mathbf{a}^T\mathbf{\Sigma}\mathbf{a} > 0$ e $\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} = 0 \Leftrightarrow \mathbf{a} = 0$, garante a existência de p valores próprios distintos, reais e positivos, associados a p vetores próprios ortogonais (Jobson, 1992). Tais propriedades desta matriz são essenciais para a aplicação da ACP, uma vez que é a partir desta matriz que se derivam as componentes principais (CPs).

A ACP depende criticamente das escalas utilizadas para medir as variáveis, uma vez que os vetores próprios não são invariantes para mudanças de escala. Em alguns casos, as variâncias das variáveis originais são muito heterogéneas (muitas vezes pelo facto de estarem a serem medidas em escalas diferentes), e como tal as que têm maior variância vão ter maior influência na determinação das primeiras componentes; por outro lado, quando as escalas de medidas são muito diferentes, as que possuem maiores valores terão mais peso na análise. Para contornar tal situação, deve-se utilizar como estrutura de variabilidade a matriz de correlações em vez da matriz de variâncias/covariâncias, isto é, deve-se derivar as CPs a partir da matriz de variâncias/covariâncias obtida a partir da matriz dos dados estandardizados, que não é mais do que a matriz de correlações dos dados (Reis *et al*., 1997).

Antes de se realizar uma ACP é necessário testar a validade da aplicação deste tipo de análise. Quando calculada a matriz de correlações, pode acontecer existir um número elevado de variáveis que não estejam correlacionadas e, como tal, deve-se por em causa a significância desta análise. Vários testes podem ser utilizados para este fim, como por exemplo o teste de Esfericidade de *Bartlett* e o teste de *Kaiser-Meyer-Olkin* (KMO):

- Teste de Esfericidade de *Bartlett*: testa se não existem correlações significativas entre as variáveis na população, ou seja, a aplicação de uma técnica multivariada pressupõe que se rejeite a hipótese nula que afirma que a matriz de correlações da população, $\mathbf{R}$, é uma matriz identidade, $\mathbf{I}$. Temos, então, como hipóteses H_0: $\mathbf{R} = \mathbf{I}$ *vs* H_1: $\mathbf{R} \neq \mathbf{I}$.

A estatística de teste (ET) definida por Bartlett, para testar a hipótese anterior, é dada por

$$ET = -\left[(n-1) - \frac{2p+5}{6}\right] ln(|\mathbf{R}|), \tag{2.13}$$

onde n é a dimensão da amostra, $|\mathbf{R}|$ é o determinante da matriz de correlações amostrais dos dados e p é o número de variáveis em estudo. Sob H_0 e n grande, esta estatística tem distribuição aproximada a uma Qui-quadrado, χ^2, com $\frac{p(p-1)}{2}$ graus de liberdade.

Assim, rejeita-se a hipótese nula se o valor observado da ET for maior ou igual ao valor crítico da distribuição de Qui-quadrado para o nível de significância escolhido para o teste (Reis, 2001).

Este teste é pouco utilizado uma vez que é muito sensível à dimensão da amostra (para grandes amostras muitas vezes rejeita-se a hipótese nula, mesmo quando as correlações são muito reduzidas) e, além disso, este teste exige que as variáveis provenham de uma distribuição Normal multivariada, sendo muito sensível à violação deste pressuposto, pelo que se torna preferível usar a estatística de Kaiser-Meyer-Olkin (KMO).

- Estatística de Kaiser-Meyer-Olkin (KMO): medida que varia entre 0 e 1 que serve para comparar as correlações simples com as correlações parciais observadas entre as variáveis.

  A estatística KMO é definida por

$$KMO = \frac{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} r_{x_i x_j}^2}{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} r_{x_i x_j}^2 + \sum_{i=1}^{p-1}\sum_{j=i+1}^{p} r_{x_i x_j|x_k}^2}, \tag{2.14}$$

  onde $r_{x_i x_j}$ é o coeficiente de correlação observado entre as variáveis x_i e x_j e o termo $r_{x_i x_j|x_k}^2 = \frac{(r_{x_i x_j} - r_{x_i x_k} r_{x_j x_k})}{\sqrt{(1-r_{x_i x_j}^2)(1-r_{x_j x_k}^2)}}$ representa o quadrado da correlação parcial entre as variáveis x_i e x_j após eliminada a influência das variáveis x_k, ($k \neq i \neq j = 1, \ldots, p$). Esta correlação parcial é a correlação que existe entre duas variáveis depois de se ter eliminado a influência de outras variáveis, que também se apresentam correlacionadas com estas duas (Norman e Streiner, 2000).

  Friel (2003) sugere a escala apresentada na Tabela 2.2 para interpretar o valor da estatística KMO.

Tabela 2.2: Escala para a interpretação da estatística KMO dada por Friel (2003).

KMO	Classificação
] 0.9 - 1]	Excelente
] 0.8 - 0.9]	Boa
] 0.7 - 0.8]	Médio
] 0.6 - 0.7]	Razoável
] 0.5 - 0.6]	Mau mas ainda aceitável
$\leq$ 0.5	Inaceitável

2.2.1.2 Derivação das componentes principais e suas propriedades

Numa ACP pretende-se encontrar um novo conjunto de variáveis (CPs), $Y_1, Y_2, \ldots, Y_p$, não correlacionadas e cujas variâncias decresçam da primeira até à última variável. Cada Y_j, $j = 1, \ldots, p$, é obtida a partir de uma combinação linear das variáveis originais $X_1, X_2, \ldots, X_p$, isto é

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p = \mathbf{a}_j^{\mathrm{T}}\mathbf{X} \tag{2.15}$$

onde $\mathbf{a}_j^{\mathrm{T}} = [a_{1j}, \ldots, a_{pj}]$ é um vetor de constantes tal que, para $j = 1, \ldots, p$,

$$\mathbf{a}_j^{\mathrm{T}}\mathbf{a}_j = \sum_{k=1}^{p} a_{kj}^2 = 1. \tag{2.16}$$

Esta restrição fixa, assim, a escala para as novas variáveis. Note-se que este vetor de constantes corresponde ao vetor próprio normalizado da matriz de covariância (correlação) associado à variável X_j.

A primeira componente principal, Y_1, é determinada pela escolha de $\mathbf{a}_1$ de modo a que Y_1 tenha a maior variância possível. Como $Var(Y_1) = Var(\mathbf{a}_1^{\mathrm{T}}\mathbf{X}) = \mathbf{a}_1^{\mathrm{T}}\mathbf{\Sigma}\mathbf{a}_1$, onde $\mathbf{\Sigma}$ é a matriz de variâncias/covariâncias do vetor p-variado $\mathbf{X}$, esta vai estar sujeita à restrição referida anteriormente.

O método usual para maximizar uma função de várias variáveis, sujeita a uma ou várias restrições, é o método dos multiplicadores de *Lagrange* cujas soluções são as raízes do sistema homogéneo $(\mathbf{\Sigma} - \lambda\mathbf{I})\mathbf{a}_1 = 0$. Note-se que $(\mathbf{\Sigma} - \lambda\mathbf{I})$ deve ser uma matriz singular de modo a que haja mais soluções para além da solução nula, ou seja, $|\mathbf{\Sigma} - \lambda\mathbf{I}| = 0$, o que significa que λ é um valor próprio de $\mathbf{\Sigma}$. Como a matriz $\mathbf{\Sigma}$ apresenta p valores próprios distintos, escolhe-se o que apresenta maior valor, uma vez que o objetivo é maximizar a variância da componente que é dada agora por

$$Var(Y_1) = Var(\mathbf{a}_1^{\mathrm{T}}\mathbf{X}) = \mathbf{a}_1^{\mathrm{T}}\mathbf{\Sigma}\mathbf{a}_1 = \mathbf{a}_1^{T}\lambda\mathbf{I}\mathbf{a}_1 = \lambda. \tag{2.17}$$

Este valor próprio associado à primeira CP será designado por λ_1 e $\mathbf{a}_1$ o vetor próprio associado de norma 1.

A segunda componente principal, $Y_2 = \mathbf{a}_2^T\mathbf{X}$, é determinada pela escolha do vetor $\mathbf{a}_2$ de modo a que Y_2 tenha a segunda maior variância de entre as CPs e que seja não correlacionada com Y_1. É obtida como uma extensão do argumento anterior, mas para além da restrição (2.16), tem ainda que obedecer à condição

$$cov(Y_2, Y_1) = Cov(\mathbf{a}_2^T\mathbf{X}, \mathbf{a}_1^T\mathbf{X}) = \mathbb{E}[\mathbf{a}_2^T(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\mathbf{a}_1] = \mathbf{a}_2^T\Sigma\mathbf{a}_1 = 0. \tag{2.18}$$

Ora, uma vez que $\mathbf{\Sigma a}_1 = \lambda \mathbf{a}_1$, a condição (2.18) é equivalente a $\mathbf{a}_2^T \mathbf{a}_1 = 0$, ou seja, $\mathbf{a}_1$ e $\mathbf{a}_2$ devem ser ortogonais.

De forma a maximizar $Var(Y_2) = \mathbf{a}_2^T \mathbf{\Sigma a}_2$ sujeita às duas restrições definidas anteriormente, utiliza-se de novo o método dos multiplicadores de *Lagrange* que permite resolver o sistema homogéneo $(\mathbf{\Sigma} - \lambda \mathbf{I})\mathbf{a}_2 = 0$ e cuja solução é o segundo maior valor próprio, λ_2, e $\mathbf{a}_2$ o correspondente vetor próprio de norma 1.

Por um processo semelhante são obtidas as restantes componentes principais, sendo os valores da j-ésima CP definidos pelo vetor próprio de norma 1 correspondente ao j-ésimo maior valor próprio da matriz $\mathbf{\Sigma}$ (Chatfield e Collins, 1980).

Denotando-se por $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p]$ a matriz $p \times p$ dos vetores próprios de $\mathbf{\Sigma}$ e por $\mathbf{Y}$ o vetor de ordem $p \times 1$ constituído pelas CPs, $\mathbf{Y}^T = [Y_1, \ldots, Y_p]$, tem-se

$$\mathbf{Y} = \mathbf{A}^T \mathbf{X}, \tag{2.19}$$

(Chatfield e Collins, 1980), onde:

- $\mathbf{\Lambda}$ representa a matriz de variâncias/covariâncias de $\mathbf{Y}$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ... & \lambda_p \end{bmatrix}. \tag{2.20}$$

  Note-se que $\mathbf{\Lambda}$ é uma matriz diagonal uma vez que as CPs não estão correlacionadas e, além disso, percebe-se que as variâncias de cada CP são dadas pelos valores próprios da matriz $\mathbf{\Sigma}$;

- $Var(\mathbf{Y}) = \mathbf{A}^T \mathbf{\Sigma A}$, pelo que $\mathbf{\Lambda} = \mathbf{A}^T \mathbf{\Sigma A}$. Esta relação é muito importante uma vez que faz o elo de ligação entre matriz de variâncias/covariâncias de $\mathbf{X}$ e as componentes principais. Desta expressão pode-se deduzir que $\mathbf{\Sigma} = \mathbf{A\Lambda A}^T$, uma vez que $\mathbf{A}^T$ é uma matriz ortogonal tal que $\mathbf{A}^T\mathbf{A} = I$;

- $\sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} \lambda_j = tr(\mathbf{\Lambda}) = tr(\mathbf{A}^T\mathbf{\Sigma A}) = tr(\mathbf{\Sigma AA}^T) = tr(\mathbf{\Sigma}) = \sum_{j=1}^{p} Var(X_j)$, ou seja, a soma das variâncias das p componentes principais é igual à soma das variâncias das p variáveis originais;

- a CP_i, $i = 1, \ldots, p$, explica uma proporção $\pi_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$ da variabilidade total dos dados, concluindo assim que as primeiras $m < p$ CPs explicarão $\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100\%$ da variabilidade total do conjunto dos dados;

- a covariância entre a j-ésima CP e a i-ésima variável original é dada por $Cov(Y_j, X_i) = Cov(Y_j, \sum_{j=1}^{p} a_{ij} Y_j) = a_{ij} Var(Y_j) = a_{ij}\lambda_j$, onde a_{ij} é o coeficiente de X_i na combinação linear que define a j-ésima CP, Y_j e λ_j é a variância da j-ésima CP (Reis, 2001).

Numa ACP é usual determinarem-se as componentes principais a partir de um conjunto de variáveis previamente estandardizadas, de maneira a que as diferentes escalas com que foram medidas as variáveis originais não influenciem os resultados finais; isto significa que se derivam as componentes não a partir da matriz de variâncias/covariâncias $\mathbf{\Sigma}$, mas através da matriz de correlação $\mathbf{P}$ dos dados. Todo o processo descrito anteriormente, para o cálculo das CPs, é idêntico à exceção dos vetores próprios associados agora à matriz $\mathbf{P}$ (Chatfield e Collins, 1980). Neste caso:

- a componente principal j explica $\frac{\lambda_j}{p} \times 100\%$ da variabilidade total uma vez que $tr(\mathbf{P}) = p$ (Chatfield e Collins, 1980);
- a correlação entre a i-ésima variável X_i e a j-ésima CP é dada por $Cor(Y_j, X_i) = \frac{Cov(Y_j, X_i)}{\sqrt{Var(Y_j)Var(X_i)}} = \frac{a_{ij}\lambda_j}{\sqrt{\lambda_j \sigma_i^2}} = a_{ij}\frac{\sqrt{\lambda_j}}{\sigma_i}$, onde σ_i é o desvio padrão da variável X_i (Cadima, 2010).

Resumindo, a ACP encontra p novas variáveis, designadas por componentes principais, que são combinações lineares das p variáveis originais, sendo os coeficientes destas combinações os vetores próprios da matriz de variâncias/covariâncias (correlações) e os respetivos valores próprios são interpretados como as suas variâncias; cada CP explica π_i, $i = 1, \ldots, p$, da variabilidade total dos dados.

2.2.1.3 Critérios de seleção de CPs

Quando o objetivo da ACP passa por se obter uma redução da dimensionalidade dos dados, é necessário fazer-se uma ponderação na escolha do número de componentes a reter, tendo em consideração a proporção de variância total explicada por estas.

Existem vários critérios para determinar o número de componentes principais a reter.

Uma ferramenta visual importante para auxiliar a escolha do número de componentes a ser retido é o *scree plot*, também conhecido por "gráfico do cotovelo". O *scree plot* é um gráfico onde estão representados os valores próprios em função do índice de cada uma das respetivas componentes principais, $\hat{\lambda}\ vs\ i$. A partir da sua análise visual, devem-se selecionar as componentes até que a linha que as une comece a ter um declive reduzido (Brown *et al.*, 2012).

No exemplo apresentado na Figura 2.3, selecionar as duas primeiras, ou talvez as três primeiras componentes, seriam suficientes para resumir a variação amostral total, uma vez que se

observa a formação de um cotovelo na posição $i = 3$, o que pode indicar que as componentes acima de 2 possuem aproximadamente a mesma magnitude e são relativamente pequenas.

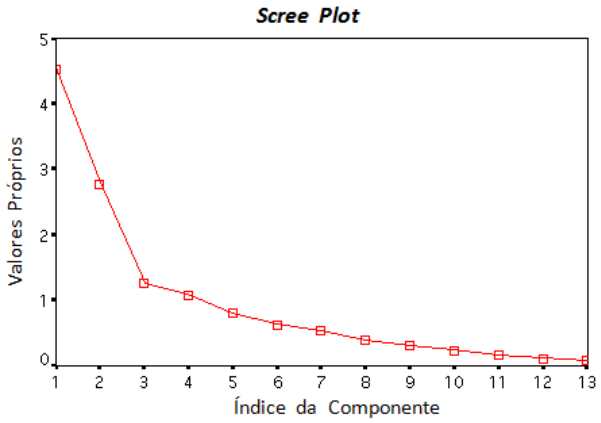


Figura 2.3: Representação de um *Scree Plot*.

Um outro critério usado para a escolha do número de CPs passa por reter as primeiras Cps que expliquem entre 70% e 90% da variabilidade total dos dados, como sugere Jollife (2002).

O Critério de *Kaiser*, também muito utilizado, prende-se com o excluir as CPs cujos valores próprios são inferiores à média dos valores próprios. Na situação em que a análise seja feita a partir da matriz de correlações, deve-se reter as CPs cujos valores próprios são maiores que 1 (Reis, 2001).

Existe ainda um critério mais formal que apenas pode ser aplicado quando as componentes principais derivam de uma matriz de variâncias/covariâncias amostral e consiste em reter apenas as CPs cuja variância é significativamente diferente de zero. *Bartlett* testa a hipótese de que os últimos $p - k$ valores próprios de $\mathbf{\Sigma}$ são iguais. Se esta hipótese não for rejeitada, apenas se irá reter as k primeiras componentes (Reis *et al.*, 1997).

Não existe um critério que seja considerado o melhor, até porque esta tomada de decisão muitas vezes é feita de acordo com a opinião dos peritos no assunto que está a ser discutido. No entanto, existem diversas opiniões que sugerem os critérios a utilizar, de acordo com os dados existentes, e algumas dicas para a obtenção de CPs fiáveis: segundo Hakstian *et al.* (1982), quando o número de variáveis em análise é relativamente reduzido, $p \leq 30$, ou o número de observações elevado, $n > 250$, o critério de *Kaiser* e o *scree plot* geram soluções credíveis no que diz respeito ao verdadeiro número de componentes principais a reter; segundo Reis *et al.* (1997), para se obterem CPs fiáveis deve-se ter um quociente mínimo de $\frac{n}{p} = 2$, o que vai contra a opinião de outros autores, que afirmam que no mínimo esta razão deveria ser 5 e sempre com $n > 100$ (Reis *et al.*, 1997).

2.2.1.4 Observações das CPs - *scores*

A equação 2.19 estabelece uma relação entre o vetor aleatório observado $\mathbf{X}$ e as componentes principais $\mathbf{Y}$. Em geral, $\mathbf{Y}$ tem valor médio não nulo. O que é frequente fazer-se é adicionar um vetor apropriado de constantes, de modo a que todas as componentes principais tenham média nula. Sendo $\bar{\mathbf{X}}^{\mathrm{T}} = [\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p]$ uma estimativa para $\boldsymbol{\mu}$, onde $\bar{x}_j$, $j = 1, \ldots, p$, é o valor observado da média amostral da variável X_j, a transformação usual que se faz é

$$\mathbf{Y} = \mathbf{A}^T(\mathbf{X} - \bar{\mathbf{X}}), \tag{2.21}$$

que consiste numa translação seguida de uma rotação ortogonal. Usando a equação (2.21), para a observação x_i do i-ésimo indivíduo, tem-se

$$y_i = \mathbf{A}^T(x_i - \bar{x}_i), \tag{2.22}$$

onde y_i, $i = 1, \ldots, n$, é designado por *score* do i-ésimo indivíduo (Chatfield e Collins, 1980).

2.2.2 Análise Factorial (AF)

A Análise Fatorial (AF) apresenta alguns objetivos semelhantes à Análise de Componentes Principais. A ideia de uma AF passa por determinar novas variáveis chamadas de fatores ou variáveis latentes ou fatores comuns, em número menor relativamente ao conjunto de variáveis originais, de modo a descrever e ter uma melhor compreensão sobre um conjunto de dados mas sem perda significativa de informação contida nesse conjunto. Mas, enquanto a ACP produz uma transformação ortogonal das variáveis que não dependem de nenhum modelo subjacente, a AF é baseada num modelo estatístico adequado, existindo maior interesse em explicar a estrutura de covariância das variáveis do que em explicar as variâncias. Qualquer variação que não é explicada pelos fatores comuns pode ser descrita a partir dos erros residuais, ou também designados por fatores específicos (Chatfield e Collins, 1980).

De um modo geral, inicialmente deve-se perceber as correlações existentes entre as variáveis, utilizando para isso um coeficiente de correlação/associação. A partir da matriz de correlações poderão ser identificados subgrupos de variáveis que estão altamente correlacionadas entre si dentro de cada subgrupo, mas pouco correlacionadas com variáveis de outros subgrupos. A AF permitirá concluir se é possível explicar este padrão de correlações a partir de um menor número de variáveis; este terá o papel de uma análise exploratória que reduzirá a dimensionalidade do problema. Já a análise confirmatória deverá ser utilizada para testar uma hipótese inicial de que os dados poderão ser reduzidos a uma determinada dimensão e qual a distribuição das variáveis

segundo essa dimensão. No entanto, esta divisão de análises nem sempre é clara quando aplicada uma AF (Reis, 2001).

Segundo Chatfield e Collins (1980), apesar de haver diferenças entre a AF e a ACP, muitas vezes estas são confundidas por quem as utiliza.

2.2.2.1 Propriedades do modelo

A Análise Fatorial Exploratória (AFE) assenta num modelo de regressão múltipla cujas variáveis respostas são as variáveis observadas e as variáveis explicativas são os fatores (Everitt e Hothorn, 2011). Contudo, a estimação dos coeficientes de regressão, denominados por *loadings*, não é tão direta como numa regressão.

Suponha que se tem observações de p variáveis aleatórias, X_1, X_2, ..., X_p, com vetor valor médio $\boldsymbol{\mu}$ e matriz de variâncias/covariâncias $\boldsymbol{\Sigma}$. Com o interesse de explicar a estrutura de covariância das variáveis, sem perda de generalidade assume-se que $\boldsymbol{\mu} = 0$. Também é assumido que a característica de $\boldsymbol{\Sigma}$ é máxima $(= p)$, ou seja, que a matriz é invertível.

O modelo da AF assume que existem $m < p$ fatores comuns, f_1, f_2, ..., f_m, subjacentes às variáveis observáveis, e que cada variável observada X_j, com $j = 1, \ldots, p$ é uma função linear deste conjunto de fatores mais um resíduo, isto é

$$X_j = \lambda_{j1} f_1 + \cdots + \lambda_{jm} f_m + e_j, \quad j = 1, \ldots, p, \tag{2.23}$$

onde o coeficiente λ_{jk}, $k = 1, \ldots, m$, que é designado por *loading* da j-ésima variável no k-ésimo fator mede a contribuição do fator j na variável k e e_j representa o resíduo associado à j-ésima variável, também designado por fator específico.

Em notação matricial tem-se $\mathbf{X} = \boldsymbol{\Lambda}\mathbf{f} + \mathbf{e}$ onde $\mathbf{f}^T = [f_1, f_2, \ldots, f_m]$ é o vetor constituído pelos fatores comuns, $\mathbf{e}^T = [e_1, e_2, \ldots, e_p]$ são os fatores específicos e $\boldsymbol{\Lambda}$

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \ldots & \lambda_{pm} \end{bmatrix}, \tag{2.24}$$

é a matriz de dimensão $p \times m$ dos *loadings*, que não deve ser confundida com a matriz diagonal dos valores próprios apresentada com o mesmo símbolo na ACP (Chatfield e Collins, 1980).

Como em todos os modelos, no modelo anterior é necessário assumir-se alguns prossupostos exigidos, tais como (Jobson, 1992):

- os fatores comuns, (f_j com $j = 1, \ldots, m$), são independentes (ortogonais) entre si, sendo que esta hipótese do modelo pode ser relaxada uma vez que se pode efetuar uma rotação não ortogonal (oblíqua) da matriz dos fatores; têm a mesma distribuição, com valor médio nulo, $E[f_j] = 0$, e variância unitária, $Var[f_j] = 1$, onde $j = 1, \ldots, p$;
- os fatores específicos (e_j, com $j = 1, \ldots, p$) são independentes entre si, $cov(e_i, e_j) = 0$ com $i \neq j$, e independentes dos fatores comuns, $cov(f, e) = 0$; encontram-se igualmente distribuídos com valor médio nulo, $E[e_j] = 0$ e variância $Var[e_j] = \psi_j$.

De acordo com as hipóteses referidas e pela equação (2.23), a variância da variável X_j, para $j = 1, \ldots, p$, pode ser escrita como

$$Var(X_j) = Var(\lambda_{j1} f_1 + \cdots + \lambda_{jm} f_m + e_j) = \lambda_{j1}^2 + \lambda_{j2}^2 \ldots \lambda_{jm}^2 + Var(e_j) = \sum_{k=1}^{m} \lambda_{jk}^2 + \psi_j. \quad (2.25)$$

Esta variância, como se pode observar, divide-se em duas partes aditivas: $h_j^2 = \sum_{k=1}^{m} \lambda_{jk}^2$, designada por comunalidade da j-ésima variável, que é a variância explicada pelos fatores comuns; ψ_j, denominada por variância específica, que é a variância única desta variável que não é partilhada com as restantes variáveis.

Ainda pela equação (2.23), pode-se concluir que

$$Cov(X_i, X_j) = \sum_{k=1}^{m} \lambda_{ik} \lambda_{jk}, \quad (2.26)$$

onde $i, j = 1, \ldots, p$, podendo, assim, a matriz de variâncias/covariâncias de $\mathbf{X}$, $\mathbf{\Sigma}$, ser escrita como

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\Psi}, \quad (2.27)$$

onde $\mathbf{\Psi}$ é uma matriz diagonal cujas entradas representam as variâncias específicas, ψ_1, ψ_2, ..., ψ_m.

Se a decomposição desta matriz de covariância se verificar, então o modelo com m fatores está correto (Chatfield e Collins, 1980).

2.2.2.2 Estimação dos parâmetros do modelo

Antes de se iniciar uma AF, tal como acontece na ACP, é necessário testar a validade desta análise. Para este efeito é também comum utilizar-se o teste de Esfericidade de *Bartlett* e o teste *KMO* descritos na Secção *2.2.1.1*.

O problema numa AF consiste em determinar os parâmetros do modelo que está subjacente nesta análise (pm *loadings* dos fatores comuns e p variâncias dos fatores específicos) de

modo a que se verifique a equação (2.27) para um conjunto de m fatores latentes. Acontece que quando encontrada uma solução para a matriz dos *loadings*, esta não é única; utilizando a equação (2.27), pode-se verificar que efetuando uma qualquer rotação ortogonal dos fatores no espaço m-dimensional, obtém-se um conjunto novo de fatores que satisfazem a mesma a equação (2.27), ou seja, suponha-se que $\mathbf{T}$ é uma matriz ortogonal de ordem $m \times m$. Então,

$$(\mathbf{\Lambda T})(\mathbf{\Lambda T})^{\mathbf{T}} = \mathbf{\Lambda T T}^T\mathbf{\Lambda}^T = \mathbf{\Lambda\Lambda}^T, \tag{2.28}$$

o que mostra que apesar da matriz dos *loadings* $\mathbf{\Lambda}$ e $\mathbf{\Lambda T}$ serem diferentes, ambas constroem a mesma matriz de variâncias/covariâncias de $\mathbf{X}$ (Chatfield e Collins, 1980). Esta falta de unicidade vai permitir fazer várias rotações apropriadas, providenciando uma solução para a incapacidade de interpretação dos fatores comuns. De um ponto de vista geométrico, esta rotação passa por transladar os eixos fatoriais no espaço fatorial sem alterar a orientação dos vetores que representam as variáveis originais.

Os parâmetros do modelo associados ao modelo de uma AF quase sempre são desconhecidos e, como tal, é necessário estimá-los a partir dos dados observados. Para esta estimação, o comum é a utilização da matriz de correlações amostral $\mathbf{R}$, em vez da matriz de variâncias/covariância amostral $\mathbf{S}$. Sendo assim, as variáveis $X_1, \ldots, X_p$ implícitas na equação 2.23 deverão ser estandardizadas tendo, consequentemente, valor médio nulo e variância unitária, pelo que (Chatfield e Collins, 1980)

$$1 = \sum_{k=1}^{m} \lambda_{jk}^2 + \psi_j, \tag{2.29}$$

onde ψ_j é a variância específica da variável X_j e λ_{jk} é o *loading* da j-ésima variável no k-ésimo fator, para $j = 1, \ldots, p$.

Atualmente, existe um conjunto variado de métodos iterativos, que com recurso a computadores, permitem a estimação dos parâmetros do modelo. Entre eles, encontra-se o (Bartholomew e Knott, 1999):

- Método do fator principal: tem por base o cálculo de estimativas de comunalidades, tendo como base o cálculo dos valores e vetores próprios da matriz de variâncias/covariâncias estimadas $\mathbf{S}$ (ou da matriz de correlações estimadas $\mathbf{R}$) como acontece na ACP. A diferença é que o cálculo é efetuado sobre a matriz de variâncias/covariâncias reduzida, $\mathbf{S}^*$, definida por $\mathbf{S}^* = \mathbf{S} - \mathbf{\Psi}$, onde $\mathbf{\Psi}$ representa uma matriz diagonal cujas entradas são as variâncias específicas, $\psi_1, \psi_2, \ldots, \psi_p$;
- Método de máxima verosimilhança: passa por definir uma distância F, $F = \log(|\mathbf{\Lambda\Lambda}^{\mathrm{T}} + \mathbf{\Psi}|) + \mathrm{tr}\,(\mathbf{S}|\mathbf{\Lambda\Lambda}^{\mathrm{T}} + \mathbf{\Psi}|^{-1}) - \log(|\mathbf{S}|) - p$, que toma valor nulo se $\mathbf{S} = \mathbf{\Lambda\Lambda}^{\mathrm{T}} + \mathbf{\Psi}$ e valores superiores a 0, caso contrário. Minimizando a função F vão ser encontradas as estimativas para os

loadings dos fatores comuns e para as variâncias específicas. A estimação dos parâmetros através deste método requer a multinormalidade dos dados;

- Método das componentes principais: utiliza o método de extração das Componentes Principais para a extração dos fatores.

Nestes dois últimos métodos há a possibilidade de se obterem estimativas de comunalidades não admissíveis, como por exemplo, excederem a variância da variável original correspondente, o que faz com que haja uma estimativa com sinal negativo para a variância, o que é um absurdo.

2.2.2.3 Critérios de seleção de fatores

A etapa mais importante na estimação do modelo da AF é a escolha do número de fatores m a reter. Esta decisão é geralmente crítica, uma vez que uma solução com k fatores produz *loadings* muito diferentes de uma solução com $k+1$ fatores (ao contrário da ACP que se mantêm os mesmos *loadings* independentemente do número de CP´s que se escolham). Além disso, se o número de fatores for elevado a interpretação dos *loadings* vai-se tornar bastante difícil uma vez que vai existir uma maior fragmentação da informação; por outro lado se m for muito pequeno, importantes fatores comuns serão omitidos e ter-se-ão demasiados *loadings* elevados (Jobson, 1992). Os critérios de escolha apresentados para a ACP também são válidos para a AF.

2.2.2.4 Rotação de fatores

Nem sempre a solução fatorial encontrada para um modelo de AF é de fácil interpretação, pelo que atribuir um significado empírico a cada um dos fatores extraídos torna-se difícil. Não tendo esta solução o poder da unicidade, pode facilmente aplicar-se uma rotação de fatores que é equivalente a considerar um modelo com *loadings* $\mathbf{\Lambda}^* = \mathbf{\Lambda T}$, equação (2.28), não havendo alterações nas propriedades subjacentes, ou seja, não há alteração na estrutura geral da solução, mas apenas como a solução é descrita.

Segundo Reis (2001), nesta análise existem dois tipos de rotação: a rotação ortogonal (métodos restritos à condição da preservação da ortogonalidade dos fatores) e a rotação oblíqua (métodos que permitem obter fatores correlacionados). Entre as rotações ortogonais, encontram-se entre os métodos mais usuais:

- Método *Varimax*: este método foi proposto por Kaiser (1958); pretende que para cada fator existam apenas alguns *loadings* com valores elevados e todos os outros com valores muito próximos de zero, ou seja, vai produzir fatores com correlações elevadas com apenas um

pequeno número de variáveis. Este método tem como objetivos encontrar uma matriz ortogonal $\mathbf{T}$ tal que $\mathbf{\Lambda}^* = \mathbf{\Lambda T}$ e maximizar a variação entre os pesos de cada fator sob a restrição de que as comunalidades não se alteram. Tal procedimento é conseguido a partir de um processo iterativo de maximização de uma função quadrática destes pesos (Reis, 2001);

- Método *Quartimax*: este método foi proposto por Carroll (1953); tem como objetivo tornar os pesos de cada variável elevados para um número reduzido de fatores e próximos de zero para todos os outros fatores, ou seja, força a que uma dada variável fique fortemente correlacionada com apenas um fator (Reis, 2001).

Uma das vantagens da rotação ortogonal reside no facto dos *loadings* representarem as correlações entre os fatores e as variáveis originais, o que não acontece quando a rotação é oblíqua, devido à correlação entre os fatores (Chatfield e Collins, 1980).

De entre as rotações oblíquas, as mais usadas são a *Oblimin*, proposta por Jennrich e Sampson (1966), e *Promax*, originada por Hendrickson e White (1964). Estes métodos fazem com que se perca o pressuposto de independência entre os fatores, mas permitindo que estes façam rotações com qualquer amplitude de maneira a simplificarem o agrupamento das variáveis e a interpretação dos fatores (Reis, 2001).

A Figura 2.4 ilustra em simultâneo uma rotação ortogonal de dois fatores (após a rotação, os eixos são mantidos com uma amplitude de 90^o graus) e uma rotação oblíqua destes mesmos dois fatores (após a rotação os eixos não mantêm uma amplitude de 90^o graus.)

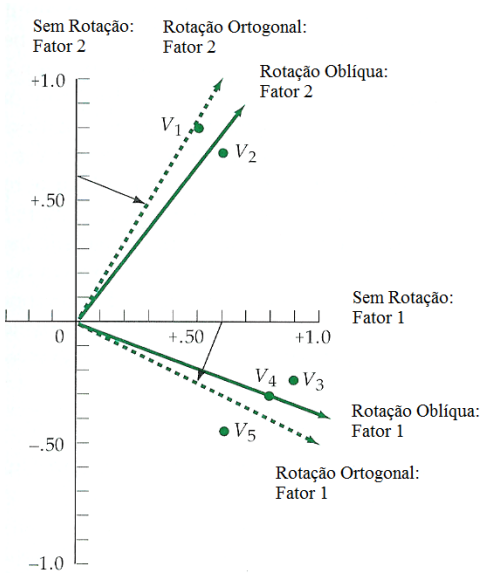


Figura 2.4: Rotação ortogonal e oblíqua de dois fatores (Adaptada de Hair *et al.* (1995)).

2.2.2.5 Observações das CPs - *scores*

Para análises posteriores à AF, por vezes torna-se necessário o cálculo dos *scores* dos fatores para cada indivíduo; este cálculo não é tão simples como no caso de uma ACP. Note-se que a equação que define o modelo da AF, equação (2.23), não é invertível; as variáveis originais encontram-se expressas em termos dos fatores, mas para se calcularem os *scores* deve-se ter essa relação na direção oposta.

Os *scores* são estimativas dos valores dos fatores aleatórios não observados $f_j, j = 1, \ldots, m$. Esta estimação torna-se complicada pelo facto das quantidades não observadas, f_j, mais as quantidades e_j serem em número mais elevado do que o número de x_j observado (Johnson e Wichern, 2007).

Para contornar tais problemas, existem vários métodos para o cálculo dos *scores*. Entre eles encontram-se:

- Método de *Thompson* (ou método de regressão): no caso das variáveis iniciais e os fatores possuírem distribuição Normal, a distribuição condicional de f dado x tem distribuição $\mathbf{N}(\mathbf{\Lambda}^T\mathbf{\Sigma}^{-1}\mathbf{X}, \mathbf{I} - \mathbf{\Lambda}^T\mathbf{\Sigma}^{-1}\mathbf{\Lambda})$, pelo que é usada a média amostral desta distribuição para se calcular $\widehat{\boldsymbol{f}}$, $\widehat{\boldsymbol{f}} = \hat{\mathbf{\Lambda}}^T\mathbf{S}^{-1}\mathbf{X}$, onde as médias amostrais de cada variável do vetor $\mathbf{X}$ foram já subtraídas, apresentando assim médias nulas;

- Método de *Bartlett* (ou método dos mínimos quadrados ponderados): uma vez obtidas as estimativas para $\mathbf{\Lambda}$ e $\mathbf{\Psi}$ as estimativas para os *scores* dos fatores, $\widehat{\boldsymbol{f}}$, são determinadas minimizando a função $(\mathbf{X} - \widehat{\mathbf{\Lambda}}\widehat{\boldsymbol{f}})\widehat{\mathbf{\Psi}}^{-1}(\mathbf{X} - \widehat{\mathbf{\Lambda}}\widehat{\boldsymbol{f}})$ e cuja solução ocorre para $\widehat{\boldsymbol{f}} = \left(\widehat{\mathbf{\Lambda}}^{\mathrm{T}}\widehat{\mathbf{\Psi}}^{-1}\widehat{\mathbf{\Lambda}}\right)^{-1}\widehat{\mathbf{\Lambda}}^{\mathrm{T}}\widehat{\mathbf{\Psi}}^{-1}\mathbf{X}$ (Johnson e Wichern, 2007).

2.2.3 Análise de *Clusters* (AC)

Desde muito cedo que o ser humano aprende a classificar e a agrupar os objetos que o rodeiam e a associar os resultados dessa classificação a palavras da sua linguagem (Reis, 2001). Constituir grupos é uma característica da atividade humana e um suporte essencial do método de aprendizagem e, em geral, do próprio método científico (Branco, 2004).

Tanto o desenvolvimento de técnicas de agrupamento bem como a sua aplicação têm aparecido em muitos campos de diferentes estudos: engenharia, zoologia, medicina, psicologia e *marketing* são apenas alguns dos campos de aplicação destas técnicas (Jobson, 1992). Mas a maior contribuição para estas aplicações foi dada por Sokal e Sneath em 1963 com o seu livro *Principles of Numerical Taxonomy*. A partir deste livro o número de publicações sobre este assunto aumentou consideravelmente, pelo motivo de ter havido um grande desenvolvimento de computadores com

elevado poder de cálculo e a importância da classificação como método científico (Reis, 2001).

A Análise de *Clusters* (AC) é uma técnica que envolve um conjunto de procedimentos de Estatística Multivariada e que pode ser utilizada para classificar objetos observando apenas as semelhanças e dissemelhanças entre eles e não definindo, à priori, nenhum critério de inclusão em qualquer agrupamento (os grupos são sugeridos pelos dados e não previamente definidos); a AC tenta organizar uma coleção de objetos em grupos relativamente homogéneos e heterogéneos entre si, ou seja, dado um conjunto de n indivíduos, estes são agrupados de acordo com a informação que existe sobre eles, dada na forma de p variáveis, de modo a que indivíduos dentro do mesmo grupo sejam mais semelhantes entre si do que indivíduos pertencentes a agrupamentos diferentes (Reis, 2001).

Tal como a Análise de Componentes Principais, a Análise de *Clusters* pode ser vista como uma técnica de redução dos dados; em vez de se reduzir o número de variáveis ou colunas necessários para caracterizar $\mathbf{X}_{n\times p}$, como na ACP, a Análise de *Clusters* reduz o número de objetos distintos, ou linhas de $\mathbf{X}_{n\times p}$, através da criação de grupos de objetos chamados *clusters* (Jobson, 1992).

O processo de formação de *clusters* tem por base ideias de semelhança e de dissemelhança: dois indivíduos pertencem a agrupamentos diferentes se não são semelhantes, ou seja, se são dissemelhantes, e pertencem a agrupamentos iguais se são semelhantes. Como tal é essencial definir-se uma escala quantitativa para quantificar esta proximidade e afastamento entre os objetos (Johnson e Wichern, 2007). A dissemelhança mede o grau de diferença/afastamento entre dois objetos, enquanto a semelhança mede o grau de parecença ou proximidade. Assim, é primordial a construção de uma matriz de proximidade cujos elementos indiquem o grau de proximidade ou afastamento entre os indivíduos.

Acontece que uma dificuldade inicial desta análise é o facto de existir uma grande quantidade de coeficientes de (dis)semelhança, tornando-se, assim, por vezes difícil a escolha desta medida de proximidade. Contudo é usual utilizarem-se coeficientes de dissemelhança, muitos deles baseados em distâncias, para agrupar indivíduos e medidas de correlação ou associação (semelhanças) para se agruparem variáveis. Na escolha de algumas destas medidas pesa a natureza dos dados: se são quantitativos ou qualitativos (nominais e ordinais).

Ao iniciar-se uma Análise de *Clusters* é necessário ter atenção as variáveis que vão caracterizar cada indivíduo, principalmente quando estão definidas em diferentes unidades de medida, uma vez que aplicada uma AC sem uma estandardização prévia, as variáveis com maiores valores e maior dispersão vão intervir com pesos diferentes na determinação das dissemelhanças, seja qual for a medida de (dis)semelhança escolhida. Contudo, esta estandardização pode não ser aconselhável em alguns casos, uma vez que este processo muitas vezes reduz as diferenças entre os indivíduos, anulando os agrupamentos naturais que possam existir nos dados (Reis, 2001). Em

alternativa a este processo, existem outras formas de se tratar estes dados, como por exemplo, atribuir pesos diferentes às variáveis de forma a homogeneizar a sua contribuição na construção dos índices de semelhança (Branco, 2004).

Resumindo, na AC estão envolvidas várias etapas essenciais (Reis, 2001; Branco, 2004):

- seleção de uma amostra de indivíduos e definição de um conjunto de variáveis que dão informação acerca dos indivíduos, necessária para os agrupar. Estes dados são representados através de uma matriz, $\mathbf{X} = [x_{ij}]$, $i = 1, \ldots, n$ e $j = 1, \ldots, p$ onde x_{ij} representa o valor da variável j observada no indivíduo i;
- definição de uma medida de dissemelhança/distância (semelhança). Os valores resultantes desta medida são representados por uma matriz designada por matriz de dissemelhanças (semelhanças), $\mathbf{D} = [d_{ij}]$ ($\mathbf{S} = [s_{ij}]$), $i, j \in 1, \ldots, n$, onde d_{ij} (s_{ij}) representa o valor da dissemelhança (semelhança) entre os objetos i e j. Esta matriz é quadrada, simétrica e tem a diagonal nula no caso de uma matriz $\mathbf{D}$ e diagonal formada por 1's no caso de uma matriz $\mathbf{S}$;
- escolha de um método de agrupamento, ou seja, definir um algoritmo de partição/classificação;
- validação dos resultados obtidos.

2.2.3.1 Medidas de dissemelhança

Como referido anteriormente, a dissemelhança reflete o grau de diferença ou afastamento entre os indivíduos; define-se dissemelhança entre dois indivíduos i e j, com $i, j = 1, \ldots, n$, de uma dada amostra, como a função d_{ij} cujos valores verificam as seguintes propriedades:

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

Caso se verifique também a desigualdade triangular, $d_{ij} \leq d_{ik} + d_{kj}$, $\forall i, j, k$, a dissemelhança diz-se uma semidistância; acrescentando a propriedade $d_{ij=0}$ se e só se $i = j$, a dissemelhança passa a designar-se como distância.

Muitas das dissemelhanças não satisfazem a propriedade triangular, mas outras satisfazem uma mais forte, a propriedade ultramétrica, $d_{ij} \leq max\ \{d_{ik}, d_{kj}\}$, $\forall i, j, k$.

Note-se que para muitas situações práticas é suficiente que se satisfaçam as propriedades 1, 2 e 3.

No caso de se estar perante variáveis quantitativas, a medida de dissemelhança mais utilizada é a distância euclidiana; esta distância entre os objetos i e j é dada por

$$d_{ij} = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{1/2}, \tag{2.30}$$

ou na forma vetorial

$$d_{ij} = \left[(\mathbf{X}_i - \mathbf{X}_j)^T (\mathbf{X}_i - \mathbf{X}_j) \right]^{1/2}, \tag{2.31}$$

onde $\mathbf{X}_i$ e $\mathbf{X}_j$ são vetores das linhas da matriz dos dados $\mathbf{X}$, ou seja, são os vetores das observações relativas aos indivíduos i e j, respetivamente.

Esta distância nem sempre é aconselhável quando se tem variáveis com diferentes unidades de medida, variâncias muito diferentes ou mesmo se são correlacionadas, pois estas variáveis vão contribuir com pesos diferentes na determinação das distâncias. Para contornar tal situação definem-se novas distâncias derivadas da dissemelhança (2.31), introduzindo nesta uma matriz de pesos, $\mathbf{A}$

$$d_{ij} = \left[(\mathbf{X_i} - \mathbf{X_j})^T \mathbf{A} (\mathbf{X_i} - \mathbf{X_j}) \right]^{1/2}. \tag{2.32}$$

Obtém-se, assim, a:

- Distância euclidiana se $\mathbf{A} = \mathbf{I}$ ($\mathbf{I}$ é a matriz identidade);
- Distância euclidiana média se $\mathbf{A} = \frac{1}{p}\mathbf{I}$ (p é o número de variáveis em estudo);
- Distância euclidiana estandardizada se $\mathbf{A} = \mathbf{D}^{-1} = (diag(s_1^2, s_2^2, \ldots, s_p^2))^{-1}$ ($\mathbf{D}$ é a matriz diagonal das variâncias das colunas de $\mathbf{X}$);
- Distância de *Mahalanobis* se $\mathbf{A} = \mathbf{S}^{-1}$ ($\mathbf{S}$ é a matriz de covariância empírica das p variáveis em estudo). Esta distância reduz a dependência das unidades de medição e a influência da correlação, o que pode mascarar ainda mais os resultados de uma AC (Branco, 2004).

2.2.3.2 Métodos hierárquicos de Análise de *Clusters*

Após escolhida a dissemelhança pretendida e construída a matriz das distâncias entre os indivíduos em estudo, surge a questão de como usar esta matriz para a formação dos *clusters*.

Em AC, as técnicas podem-se classificar em dois grandes grupos: técnicas hierárquicas e não hierárquicas, sendo a mais abordada e a mais comum a classificação hierárquica. Como o próprio nome indica, este método permite formar *clusters* de forma hierárquica; sequencialmente, vai havendo uma série de partições/fusões gerando uma ligação entre um único grupo que contém todos os indivíduos em estudo e n grupos singulares (a cada grupo pertence um único indivíduo). Para se aplicar os métodos hierárquicos, recorre-se geralmente a dois tipos de algoritmos (Chatfield e Collins, 1980):

- Aglomerativos (ou ascendentes) - inicia-se com n grupos/*clusters* formados por apenas um indivíduo e vão-se realizando sucessivas fusões destes *clusters*, aglomerações, até se formar um único grupo que contenha todos os indivíduos;
- Divisivos (ou descendentes) - inicia-se com um único *cluster*, onde reúne todos os indivíduos e vão-se realizando sucessivas separações destes indivíduos, criando grupos mais pequenos, até se formarem *clusters* com apenas um indivíduo.

Estas classificações destes dois processos podem ser representadas a partir de um diagrama bidimensional denominado por dendrograma (ou árvore hierárquica) no qual estão ilustradas as fusões ou divisões elaboradas em cada etapa do processo (Figura 2.5).

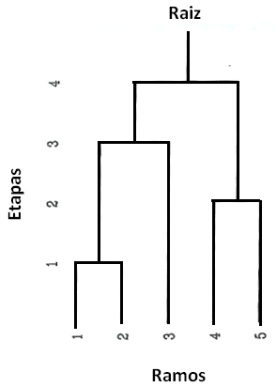


Figura 2.5: Dendrograma ou árvore hierárquica.

Cada nó do dendrograma representa um *cluster*, enquanto que o comprimento do ramo que lhe está associado (altura) indica a medida de proximidade usada para unir/separar os dois *clusters*. Note-se que se estes ramos apresentarem alturas pequenas (distâncias pequenas entre os grupos) indicam que a agregação é feita entre *clusters* razoavelmente homogéneos, ou seja, semelhantes (Branco, 2004).

De facto, o dendrograma inicia-se nos nós terminais e termina na raiz se o processo for aglomerativo. Caso se recorra a um processo divisivo, o dendrograma será elaborado em sentido contrário.

Neste trabalho, apenas se irá falar no método hierárquico aglomerativo.

2.2.3.3 Método hierárquico aglomerativo

O método hierárquico aglomerativo inicia-se com a procura, na matriz das distâncias $\mathbf{D}$, do par de indivíduos mais próximos, que corresponde ao valor do elemento mais pequeno que a matriz contém. Na Figura 2.5 olhando de baixo para cima, esses dois indivíduos correspondem aos *clusters* 1 e 2 formando-se, assim, um único *cluster* com estes dois indivíduos. Note-se que a partir do momento em que um *cluster* se forme, este é indivisível.

Numa segunda etapa, para se proceder à seleção do seguinte par de indivíduos ou *clusters* mais próximos a agrupar, deve-se atualizar a matriz das distâncias de modo a que apresentem as distâncias entre o grupo recém-formado e os restantes *clusters* singulares. Observando a Figura 2.5, os *clusters* com distância mais próxima são o 4 e 5, formando em conjunto um novo grupo.

Este processo repete-se até que todos os indivíduos fiquem num mesmo grupo (Jobson, 1992).

Repare-se que na terceira etapa representada na Figura 2.5, há a fusão entre um *cluster* singular, 3, e um *cluster* constituído por mais de um indivíduo, 1 e 2. Nestes casos é necessário a aplicação de alguns critérios que permitam quantificar a distância entre um indivíduo e um *cluster* que contenha vários indivíduos, ou entre dois grupos não singulares.

Existem vários métodos de agregação em AC e cada um deles dá origem, em princípio, a agrupamentos diferentes; não se consegue dizer qual o melhor, pelo que se deve utilizar vários e, no fim, comparar os resultados. No caso destes serem idênticos, pode-se concluir que se obtiveram resultados bastante fiáveis.

2.2.3.4 Métodos de aglomeração

Segundo Reis (2001), os métodos hierárquicos aglomerativos mais utilizados na AC são:

- Método de ligação simples ou método do vizinho mais próximo - a distância entre dois grupos, I e J, é dada pela maior das semelhanças que existe entre dois elementos de cada grupo, ou seja, é dada pela menor das distâncias que existe entre dois quaisquer elementos de cada grupo (dada pelos vizinhos mais próximos) (Reis, 2001):

$$d_{IJ} = min\left\{d_{ij} : i \in I,\ j \in J\right\}. \tag{2.33}$$

  Na Figura 2.6 encontra-se ilustrado este método.

  Neste método, uma só ligação é suficiente para se juntar dois grupos o que pode abarcar alguns problemas no sentido em que, dois grupos podem ser bastante diferentes mas basta haver uma única ligação muito próxima (a distância entre dois indivíduos (um de cada grupo)

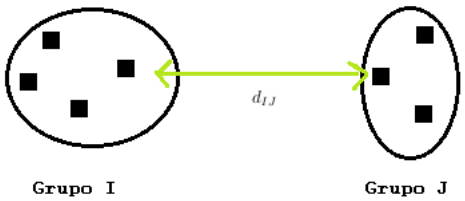


Figura 2.6: Método do vizinho mais próximo.

ser muito pequena) para que estes dois grupos fiquem agregados.

Neste processo, cada vez que se adiciona um indivíduo a um *cluster*, as distâncias deste novo *cluster* aos restantes já formados são menores ou não se alteram, fazendo assim com que os novos grupos aglomerados se tornem cada vez maiores, deixando os indivíduos isolados resistentes na sua posição, o que mostra a capacidade deste método em detetar *outliers* (Branco, 2004).

O método da ligação simples apresenta uma propriedade única: se houver duas distâncias iguais e menores que as restantes, pode-se escolher qualquer uma delas para a agregação dos respetivos grupos que o resultado final não se irá alterar; este também é o que envolve procedimentos matemáticos mais simples e o que tem maiores vantagens computacionais: é um método cuja aplicação é muito mais rápida que todos os outros e, além disso, este pode mesmo ser implementado quando existem vários milhares de indivíduos para serem comparados (Chatfield e Collins, 1980).

- Método de ligação completa ou método do vizinho mais afastado - este processo de agrupamento é inverso ao descrito anteriormente uma vez que, a distância entre dois grupos, I e J, é dada pela menor das semelhanças que existe entre dois elementos de cada grupo, ou seja, é dada pela maior das distâncias que existe entre dois quaisquer elementos de cada grupo (dada pelos vizinhos mais afastados) (Reis, 2001):

$$d_{IJ} = max\left\{d_{ij} : i \in I,\ j \in J\right\}. \tag{2.34}$$

Na Figura 2.7 encontra-se ilustrado este método.

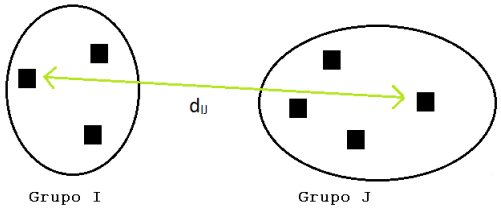


Figura 2.7: Método de vizinho mais afastado.

Ao contrário do método anterior, ao acrescentar um indivíduo a um grupo, a distância do novo grupo aos restantes aumenta ou não se altera, havendo assim uma tendência para que grupos grandes não cresçam mais (Branco, 2004).

- Método da ligação média - define a distância entre dois grupo, I e J, como a média de todas as distâncias existentes entre cada dois indivíduos dos respetivos grupos, (Reis, 2001):

$$d_{IJ} = \frac{\sum_{i=1}^{n_I} \sum_{j=1}^{n_J} d_{ij}}{n_I n_J}, \tag{2.35}$$

onde n_I e n_J representam o número de indivíduos do grupo I e J, respetivamente.

Na Figura 2.8 encontra-se ilustrado este método.

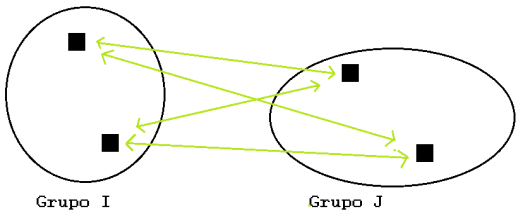


Figura 2.8: Distâncias possíveis entre dois grupos, I e J.

O processo utilizado por este método parece intermédio aos utilizados anteriormente, uma vez que não usa a distância do vizinho mais próximo nem a do vizinho mais afastado, mas sim um valor médio de todas as distâncias existentes, tendo assim a vantagem de evitar valores extremos e tomar em consideração toda a informação dos grupos (Reis, 2001).

- Método do centróide - a dissemelhança entre dois grupos é definida como a distância entre os seus centróides, isto é, é definida como a diferença entre as suas médias, para todas as variáveis

$$d_{IJ} = d\left(\bar{\mathbf{X}}_I, \bar{\mathbf{X}}_J\right), \tag{2.36}$$

onde $\bar{\mathbf{X}}_I = \frac{\sum_{i \in I} \mathbf{X}_i}{n_I}$ e $\bar{\mathbf{X}}_J = \frac{\sum_{i \in J} \mathbf{X}_i}{n_J}$, com $\mathbf{X}_i$ o vetor das p observações do indivíduo i, são os centróides dos respetivos grupos.

Em cada passo, os grupos a aglutinar são os que têm os centróides mais próximos.

A interpretação dos resultados a partir deste método torna-se difícil devido ao facto de as distâncias de fusão entre dois grupos poderem aumentar ou diminuir de etapa para etapa. Qualquer distância entre *clusters* pode ser usada neste processo mas, para contornar tal dificuldade e obter um maior sucesso em termos de facilidade de aplicação, o quadrado da distância euclidiana é a medida mais aconselhada (Branco, 2004).

- Método de Ward (1963) ou método de variância mínima - relativamente aos que já foram descritos, este método é o único que é aplicado diretamente aos indivíduos de cada grupo

sem ser necessário construir uma matriz de dissemelhanças.

Este processo tem como objetivo, de etapa a etapa, minimizar a soma dos quadrados dos erros (soma dos quadrados dos desvios das observações individuais relativamente às médias dos grupos em que são classificadas), isto é, em cada etapa, de todos os *clusters* existentes, são retidos os que apresentarem menor soma de quadrados dos erros, maximizando, assim, a homogeneidade no interior dos grupos.

Inicialmente, cada um dos n indivíduos representam um *cluster*, pelo que a soma dos quadrados dos erros é nulo. De seguida, em cada passo deste processo, fazem-se todas as combinações possíveis de pares de *clusters*, calculando o incremento da soma dos quadrados dos erros resultantes da reunião dos *clusters* de cada par

$$SSW_{I\cup J} - (SSW_I + SSW_J)\,, \tag{2.37}$$

onde $SSW_I = \sum_{i\in I}\sum_{j=1}^{p}\left(x_{ijI} - \bar{x}_{jI}\right)^2$ é a soma dos quadrados dentro do grupo A, x_{ijI} é a observação do indivíduo i do grupo I na variável j e $\bar{x}_{jI}$ é a média da variável j no grupo I. Do mesmo modo se obtém SSW_J e $SSW_{I\cup J}$ (Branco, 2004) .

O par de *clusters* escolhidos para serem aglomerados são os que apresentam menor incremento, ou seja, menor perda de informação resultante do agrupamento.

Segundo Reis (2001), este método é desenvolvido em 4 fases:

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. otimizar a variância mínima dentro dos grupos.

Segundo Chatfield e Collins (1980) os métodos de ligação simples e completa baseiam-se apenas numa única distância para representar a proximidade de dois grupos (a menor ou maior das distâncias) e, portanto, são muito influenciados pelas observações extremas. No processo de ligação simples, um único *outlier* situado entre dois aglomerados pode resultar num eventual agrupamento dos dois grupos. No caso de um processo de ligação completa, pequenas alterações na localização de pontos particulares ou erros, pode ter um impacto substancial sobre a solução hierárquica. A ligação média, o método do centróide e *Ward* são geralmente preferíveis devido à sua insensibilidade em relação a extremos ou *outliers*. Dependendo do tipo de aglomerados esperados esta propriedade também pode ser uma desvantagem.

2.2.3.5 Validação dos resultados obtidos

As opiniões de alguns autores são muitas vezes divergentes no que diz respeito à escolha do melhor método, pois todos eles apresentam vantagens e desvantagens (Chatfield e Collins, 1980). Como tal, o melhor será utilizar vários critérios e para cada um destes experimentar várias (dis)semelhanças e comparar os resultados obtidos, como sugere Reis (2001).

Contudo, Dubes e Jain (1988) referem um critério que permite validar o método utilizado: usar o coeficiente de correlação cofenético para comparar a matriz de proximidades associada aos dados (matriz de dissemelhança, $\mathbf{D}$) e a matriz cofenética associada ao dendrograma (matriz cujos elementos correspondem às distâncias ultramétricas entre cada dois indivíduos no momento em que se juntam pela primeira vez para formar *clusters*), de modo a medir o grau de deformação provocado pela construção do dendrograma. O coeficiente de correlação cofenético é o conhecido coeficiente de correlação de Pearson entre os pares de dissemelhança entre estas duas matrizes. Quanto maior for este coeficiente maior é a concordância entre os dados e o dendrograma, o que significa que ambas as estruturas revelam a mesma informação.

Após escolhido e validado o método a utilizar, a estrutura hierárquica proveniente deste procedimento costuma ser representada pelo designado dendrograma, já referido anteriormente. Visualizando a formação dos *clusters*, torna-se necessário decidir quantos grupos reter.

Um método simples e informal que sugere quantos *clusters* se deve reter é a análise gráfica, onde se representa o número de *clusters* contra o índice de fusão que não é mais do que o valor numérico (distância ou semelhança) para o qual vários objetos se unem para formar um grupo. A partição ótima poderá ser considerada quando a divisão de um novo grupo não introduz alterações significativas no coeficiente de fusão, ou seja, quando o declive da reta que une a distância entre dois *clusters* passa a ser relativamente pequeno. Geralmente, a zona de cotovelo do gráfico, dá indicação do número de *clusters* a reter (Reis, 2001).

2.3 Testes Estatísticos

Quando se está no contexto de Inferência Estatística paramétrica, as técnicas clássicas utilizadas que incidem sobre parâmetros (valor esperado, variância, proporção, ...) partem, geralmente, de pressupostos a que a população ou populações, a partir das quais as observações são retiradas, têm de obedecer. Tais pressupostos são, por exemplo, assumir que os dados seguem uma distribuição Normal, as distribuições das populações têm de ter a mesma variância, entre outros.

O problema surge quando estes pressupostos não são verificados. Nestes casos recorrem-se a técnicas não paramétricas. Na maioria destas técnicas apenas existem suposições básicas, como por exemplo, as observações têm de ser independentes e tem de haver continuidade da distribuição subjacente aos dados.

Os testes não paramétricos têm a vantagem de poderem ser utilizados em observações categóricas; muitos destes testes dão maior importância aos *ranks*, sinais, *scores* atribuídos às observações do que propriamente aos dados recolhidos. Nestes casos vai haver algum desperdício de informação mas, por outro lado, pode-se moderar os pesos das observações consideradas perturbadoras (*outliers*, ...) na tomada de decisão, em vez de as ignorar.

Apesar de tudo, os testes paramétricos são os "mais poderosos" para rejeitar a hipótese nula quando ela é falsa e, como tal, devem ser sempre os escolhidos quando verificados todos os pressupostos exigidos.

2.3.1 Testes à validade de pressupostos exigidos para a aplicação de alguns testes paramétricos

2.3.1.1 Teste à Normalidade dos dados

Uma condição necessária para se aplicarem alguns testes de hipóteses paramétricos é a normalidade dos dados. Para testar a normalidade de uma variável aleatória é utilizado, frequentemente, um teste designado por teste de *Shapiro-Wilk*. A estatística de teste é dada por

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.38}$$

onde x_i são os valores da variável X ordenados por ordem crescente, $\bar{x}$ é a média amostral de X e a_i são constantes que se encontram tabelados[7] (Shapiro e Francia, 1972). Pequenos valores de W indicam que a variável não possui distribuição Normal, encontrando-se os valores críticos para W tabelados[8]. Apenas é apropriado para amostras pequenas ($n < 30$) (Hair *et al.*, 1995).

[7] Esta tabela pode-se encontrar no livro de Pearson e Hartley (1972) (Tabela 15).

[8] Esta tabela pode-se encontrar no livro de Pearson e Hartley (1972) (Tabela 16).

2.3.1.2 Teste à Homogeneidade de variâncias

Uma outra condição que é necessária ser verificada para se efetuarem alguns testes paramétricos é a homogeneidade das variâncias das k amostras em estudo. O teste de *Levene* é um dos testes mais potentes utilizados para esta finalidade, sendo robusto a desvios de normalidade. As hipóteses a testar são

H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

vs

H_1: $\exists i, j : \sigma_i^2 \neq \sigma_j^2$ $(i \neq j; i, j = 1, \ldots, k)$,

onde σ_i^2, com $i = 1, \ldots, k$ são as variâncias populacionais das variáveis X_i.

A estatística de teste é dada por

$$W = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} n_i (\bar{z}_i - \bar{z})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{z}_{ij} - \bar{z}_i)^2}, \tag{2.39}$$

onde:

- n_i, $i = 1, \ldots, k$, é a dimensão de cada uma das k amostras;
- $N = n_1 + n_2 + \cdots + n_k$ é a dimensão da amostra global;
- $z_{ij} = |x_{ij} - \bar{x}_i|$, com $i = 1, \ldots, k$ e $j = 1, \ldots, n_i$ em que x_{ij} é a observação j da amostra i e $\bar{x}_i$ é a média da amostra i;
- $\bar{z}_i$ é a média de z_{ij} na amostra i e $\bar{z}$ é a média de z_{ij} na amostra global.

Note-se que, caso se suspeite que X não provém de uma distribuição Normal, então $z_{ij} = |x_{ij} - \bar{x}_i|$, onde $\bar{x}_i$ é a mediana da amostra i. Esta fórmula de cálculo com recurso à mediana é robusta e potente para desvios de normalidade.

Sob a hipótese nula tem-se $W \sim F(k-1, N-k)$, ou seja, W provém de uma distribuição de probabilidade contínua designada por *Fisher-Snedecor* (F) com $(k-1)$ e $(N-k)$ graus de liberdade; a um nível de significância α rejeita-se H_0 se $W \geq f_{1-\alpha;(k-1,N-k)}$ (Levene, 1960).

2.3.2 Teste de Localização

O teste de *Mann - Whitney* é um teste não paramétrico que é utilizado para verificar se duas amostras provêm da mesma população. Uma generalização para k amostras é proporcionada pelo teste de *Kruskal-Wallis*, ou teste H (Spiegel, 1993). A utilização deste teste apenas requer

que os dados provenham de amostras independentes de populações com distribuições inerentes contínuas e medidas no mínimo numa escala ordinal (Siegel, 1975).

Formalmente, as hipóteses subjacentes a este teste podem-se escrever como

H_0: $F(X_1) = F(X_2) = \cdots = F(X_k)$ (a distribuição dos valores da variável são idênticas nas k populações)

vs

H_1: $\exists\, i, j$: $F(X_i) \neq F(X_j)$, com $i \neq j$; $i, j = 1 \ldots, k$ (existe pelo menos uma população onde a distribuição da variável é diferente de uma das distribuições das outras populações).

As hipóteses do teste de *Kruskal-Wallis* são porém, frequentemente, escritas como (Siegel e Castellan, 1988)

H_0: $\eta_1 = \cdots = \eta_k$ (as medianas das populações são iguais)

vs

H_1: $\exists\, i, j$: $\eta_i \neq \eta_j$, com $i \neq j$; $i, j = 1 \ldots, k$ (existe pelo menos um par de medianas significativamente diferentes).

Seja k o número de amostras em estudo de dimensão $n_1, \ldots, n_k$ e seja N o número total de observações de todas as amostras, $N = \sum_{j=1}^{k} n_j$. A realização deste teste baseia-se na ordenação por ordem crescente das N observações atribuindo a cada uma delas o respetivo *rank* (à menor observação atribui-se a ordem 1, ao seguinte a ordem 2, $\ldots$, e à maior a ordem N). De seguida determina-se a soma dos *ranks* em cada uma das amostras (R_k) (Spiegel, 1993).

Obtidas estas somas, $R_1, \ldots, R_k$, pode-se obter o valor da estatística de teste (Branco, 2004)

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1), \tag{2.40}$$

cuja distribuição se aproxima de uma Qui-quadrado com $k-1$ graus de liberdade, desde que o número de observações de cada amostra seja pelo menos igual a 5.

Para amostras mais reduzidas ($n_j \leq 5,\ j = 1,\ \ldots,\ k$), a aproximação Qui-quadrado para a distribuição amostral de H não é suficiente e, como tal, devem ser utilizadas tabelas especiais[3] que contêm valores críticos, os quais devem ser comparados com o valor de H, permitindo assim tornar o teste mais exato (Dagnelie, 1973).

No caso de haver g grupos de observações empatadas, com τ_i observações no i-ésimo grupo

[3] Esta tabela pode-se encontrar no livro de Spiegel (1993).

de empates, a expressão acima é multiplicada por um fator de correção

$$\frac{1}{1-\frac{\sum_{i=1}^{g}(\tau_i-1)\tau_i(\tau_i+1)}{(N-1)N(N+1)}}, \tag{2.41}$$

obtendo-se

$$H^* = \frac{\frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_j^2}{n_j} - 3(N+1)}{1-\frac{\sum_{i=1}^{g}(\tau_i^3-\tau_i)}{N^3-N}}. \tag{2.42}$$

Ao nível de significância α rejeita-se H_0 se H exceder o quantil $1-\alpha$ da distribuição χ^2_{k-1} (Conover, 1980).

2.3.2.1 Teste de Comparações Múltiplas

Quando utilizado o teste não paramétrico de *Kruskal-Wallis* e a hipótese nula é rejeitada, existe a necessidade de, à posteriori, se efetuarem comparações múltiplas para perceber qual ou quais das amostras diferem. O facto de o número de comparações (testes) a realizar poder ser elevado, $\binom{k}{2}$, e o facto de não se conhecer com exatidão o nível de significância simultâneo devido à não independência entre os vários testes, surge o aparecimento de testes de hipóteses realizados simultaneamente, dois a dois, para encontrar as possíveis diferenças entre as k distribuições populacionais (Reis, 2001; Reis *et al.*, 1997).

Existem vários testes de comparação múltipla, sendo um dos mais utilizados o teste de LSD (Least Significant Difference) de *Fisher*; procedendo à comparação múltipla das médias das ordens das respetivas amostras, este teste vai identificar qual ou quais dos grupos as distribuições são estatisticamente diferentes (Siegel e Castellan, 1988). As hipóteses a testar, para todas as combinações possíveis de grupos 2 a 2 são

H_0: $F(X_i) = F(X_j)$, $i \neq j$, $i, j = 1, \ldots, k$ $(\eta_i = \eta_j)$

vs

H_1: $F(X_i) \neq F(X_j)$, $i \neq j$, $i, j = 1, \ldots, k$ $(\eta_i \neq \eta_j)$.

Segundo Conover (1980), a estatística de teste é dada por

$$T_r = \frac{\bar{R}_i - \bar{R}_j}{\sqrt{S_r^2\left(\frac{N-1-H}{N-k}\right)\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}, \tag{2.43}$$

onde:

- $\bar{R}_i$ e $\bar{R}_j$ são as médias das ordens das amostras i e j, respetivamente;

- S_r^2 é a variância de todas as ordens;
- H é a estatística de teste de *Kruskal-Wallis*;
- n_i e n_j são as dimensões das amostras retiradas das populações i e j, respetivamente;
- N é a dimensão da amostra combinada.

Sob a hipótese nula, rejeita-se H_0 se $T_r \geq t_{\frac{1-\alpha}{2}}$, onde $t_{\frac{1-\alpha}{2}}$ é o quantil de ordem $\frac{1-\alpha}{2}$ da distribuição *t-student* com $(N-k)$ graus de liberdade, sendo α o mesmo nível de significância usado no teste de *Kruskal-Wallis*.

Capítulo 3

Caracterização e Análise Exploratória dos Dados

3.1 Caracterização das Parcelas e Castas em Estudo

A parcela em estudo situada no Minho (Portugal), sobre a qual recaiu parte deste trabalho, situa-se na Estação Vitivinícola Amândio Galhano (EVAG) no concelho dos Arcos de Valdevez (Mota, 2005), tendo de coordenadas $41°48'46.70''$ de latitude e $8°24'45.61''$ de longitude. Esta parcela tem uma área de $4116\ m^2$ e encontra-se a uma altitude média de $76\ m$ com um ligeiro declive de 5% e de exposição dominante a Sul-Sul-Sudoeste (S-SSO) (Maciel, 2005), Figura 3.1.

Figura 3.1: Imagem aérea da parcela em estudo da EVAG (Fonte: GoogleEarth, 2011).

Em literatura datada do ano de 1990, o solo onde se situa esta parcela foi caracterizado como espesso com elevado risco de erosão, de permeabilidade moderadamente lenta e de drenagem externa e interna regular, com teor baixo em coloides minerais, com baixo a médio teor em matéria orgânica e com elevada capacidade de armazenamento de água útil. Foi considerado um solo ácido, de teor baixo em azoto e muito baixo em fósforo e potássio; com baixo teor em bases de troca e muito fortemente lixiviados. A textura dominante é franco-arenosa (Armanda, 1990).

A casta vinhão presente na parcela em estudo (Figura 3.2) foi plantada no ano de 1999 com um compasso de plantação de $2\ m \times 2,5\ m$. É uma casta de qualidade sendo a única casta regional tintureira; é sensível à falta de água, produz mostos ricos em açúcares e o seu vinho tem uma cor intensa, vermelho granada, de aroma vinoso e encorpado. É uma casta vigorosa e regular na produção; é de ciclo curto, sendo tardia no abrolhamento, recuperando depois na floração e na maturação (Mota e Garrido, 2001).

Figura 3.2: Exemplo do cacho e da folha da casta tinta Vinhão (Fonte: $http://www.winesofportugal.info/$).

A segunda parcela em estudo neste trabalho situa-se na Bodega Santiago Ruiz, em *Tomiño* na Região da Galiza (Espanha), tendo de coordenadas $8°43'09.7''$ de longitude e $41°59'41.0''$ de latitude. Este terreno apresenta uma área igual a $3268\ m^2$ e encontra-se a uma altitude de $18\ m$ (Araújo, 2010), Figura 3.3.

Figura 3.3: Imagem aérea da Bodega Santiago Ruiz com indicação da parcela da vinha estudada (Fonte: GoogleEarth, 2011).

A Bodega Santiago Ruiz encontra-se próxima de um curso de água, o rio Minho, correspondendo a uma forte depressão tectónica que se alonga na direção Este-Nordeste - Oeste-Sudoeste (ENE - WSW). É uma zona que apresenta uma baixa altitude e baixo relevo, não havendo grandes variações de cota; desenvolve-se sobre sedimentos recentes, que constituem um terraço fluvial.

A quinta que contém a parcela em estudo encontra-se numa zona de acumulação de sedimentos recentes, formando um antigo terraço do Rio Minho que é constituído na maior parte por

seixos geralmente cobertos por camadas arenoargilosas e matéria orgânica; é um terraço principalmente de tipologia conglomerática, com calhaus rolados de quartzito e quartzo (Azevedo e Oliveira, 2010a).

A vinha em estudo foi plantada no ano 2001 com um compasso de plantação de $2\ m \times 2,75\ m$ (Araújo, 2010). Nesta está representada a casta Alvarinho, Figura 3.4, que é considerada uma das mais notáveis castas de uvas branca da Região Demarcada dos Vinhos Verdes. É uma casta branca de grande qualidade, medianamente vigorosa, de baixa produção e elevada rusticidade; origina vinhos com aroma acentuado e são considerados muito saborosos com elevada graduação[1].

O seu cacho é de tamanho pequeno, sendo o seu bago de tamanho médio mas não uniforme; apresenta uma cor verde amarelada, ficando com um tom rosado quando demasiado posto ao sol[2].

Esta casta requer terrenos secos para potencializar a qualidade do vinho; é uma casta precoce no abrolhamento e na maturação. Produz mostos muito ricos em açúcares e apresenta um razoável teor em ácidos orgânicos[3].

Figura 3.4: Exemplo do cacho e da folha da casta branca Alvarinho (Fonte: $http://www.winesofportugal.info/$).

3.2 Caracterização das Variáveis em Estudo

Na parcela em estudo do Minho foram georreferenciados 45 pontos com espaçamento regular, $A1, A2, \ldots, A45$, em 9 regiões (parcelas), cada uma contendo 5 destes pontos, Figura 3.5.

Nestes pontos foram recolhidas as amostras do solo a uma profundidade de $15 - 20\ cm$, com o objetivo de se analisarem algumas variáveis referentes a características químicas do solo (Silva, 2011). Nestas localizações também se avaliaram as variáveis referentes à produtividade das videiras.

Depois de realizada uma divisão deste terreno em 9 parcelas - Bloco 1 Nascente ($B1N$), Bloco 1 Centro ($B1C$), Bloco 1 Poente ($B1P$), Bloco 2 Nascente ($B2N$), Bloco 2 Centro ($B2C$), Bloco

[1] Informação retirada de http://www.valedominhodigital.pt
[2] Informação retida de http://www.infovini.com/
[3] Informação retirada de http://www.alvarinho.pt

2 Poente ($B2P$), Bloco 3 Nascente ($B3N$), Bloco 3 Centro ($B3C$), Bloco 3 Poente ($B3P$) - foram recolhidas 9 amostras de mosto a partir das quais foram avaliadas algumas variáveis relativas às características físico-químicas do mosto; também foi feita uma análise do rendimento em sumo e da composição cromática e aromática do mosto. Por último, foram analisados os vinhos correspondentes às nove localizações referidas anteriormente.

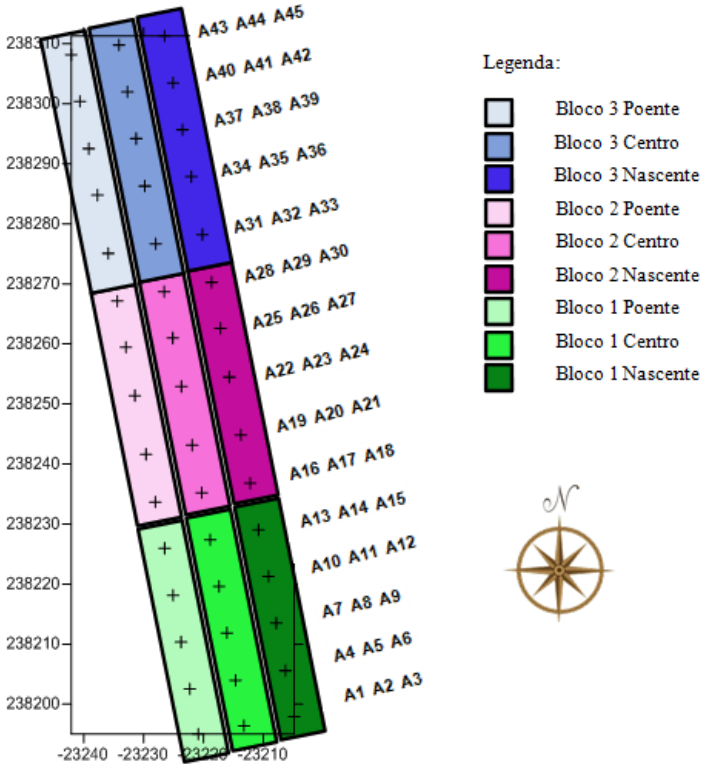


Figura 3.5: Parcela do Minho com a localização dos 45 pontos georreferenciados bem como a sua divisão em 9 parcelas (Adaptada de Silva (2011)).

Segundo os responsáveis pela recolha da amostra, estas variáveis não foram medidas em 45 localizações/videiras, mas apenas em 9, fundamentalmente, devido à indisponibilidade financeira para efectuar ensaios laboratoriais a mais de 9 amostras mas também, porque o aumento da quantidade de uvas recolhidas tem como consequência a diminuição da quantidade de vinho produzido e o rendimento financeiro obtido. No entanto, é de referir, que apesar de existir 45 pontos de amostragem no solo, tal número não seria alcançável para o mosto. O mosto analisado resulta do somatório de vários pontos de recolha de uvas e, portanto, será sempre inferior aos pontos de amostragem obtidos no solo.

Na parcela em estudo na Galiza também foram georreferenciados 45 pontos, denominados por $B1$, $B2$, $\ldots$, $B45$, como mostra a Figura 3.6, de onde foram recolhidas as amostras do solo e analisadas algumas variáveis relativas às características químicas do solo.

Relativamente às variáveis alusivas à produção das videiras, estas foram avaliadas de acordo com 12 localizações/videiras - Bodega Santiago Ruiz 1 (BSR 1), Bodega Santiago Ruiz 2 (BSR

2), Bodega Santiago Ruiz 3 (BSR 3), Bodega Santiago Ruiz 4 (BSR 4), Bodega Santiago Ruiz 5 (BSR 5), Bodega Santiago Ruiz 6 (BSR 6), Bodega Santiago Ruiz 7 (BSR 7), Bodega Santiago Ruiz 8 (BSR 8), Bodega Santiago Ruiz 9 (BSR 9), Bodega Santiago Ruiz 10 (BSR 10), Bodega Santiago Ruiz 11 (BSR 11) e Bodega Santiago Ruiz 12 (BSR 12) - representadas por um círculo na Figura 3.6.

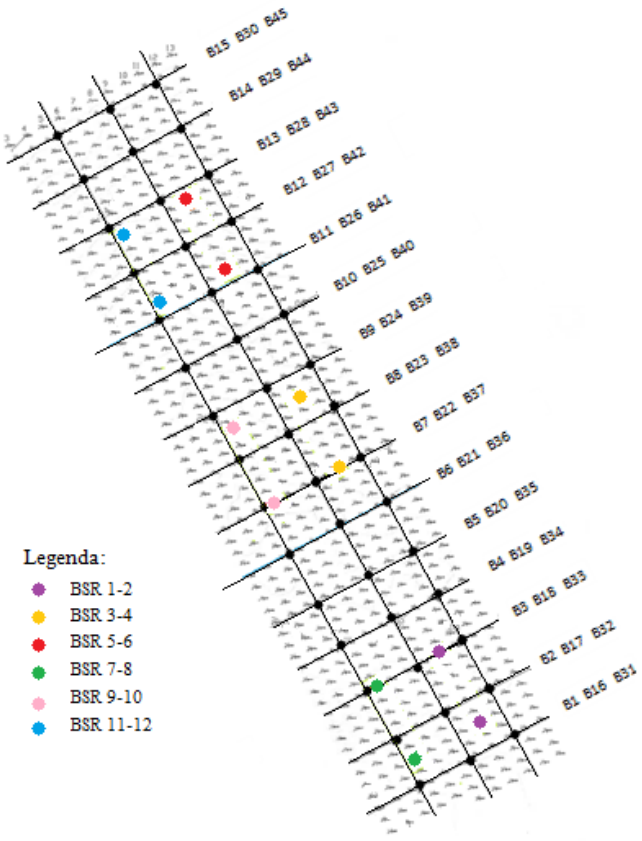


Figura 3.6: Parcela da Galiza com a localização dos 45 pontos georreferenciados bem como a sua divisão em 6 parcelas.

Segundo os responsáveis pela recolha da amostra, estas variáveis não foram medidas em 45 localizações/videiras, mas apenas em 6 devido à indisponibilidade financeira para realizar ensaios laboratoriais a mais de 6 amostras mas também, porque o aumento da quantidade de uvas recolhidas tem como consequência a diminuição da quantidade de vinho produzido e o rendimento financeiro obtido. No entanto, é de referir, que apesar de existir 45 pontos de amostragem no solo, tal número não seria alcançável para o mosto. O mosto analisado resulta do somatório de vários pontos de recolha de uvas e, portanto será sempre inferior aos pontos de amostragem obtidos no solo.

A localização dos 12 pontos teve como objectivo definir uma malha uniforme que abarcasse toda a parcela.

As variáveis analisadas referentes às uvas e ao mosto foram avaliadas apenas a partir de 6 amostras de uvas provenientes de 6 localizações - Bodega Santiago Ruiz $1-2$ (BSR $1-2$), Bodega

Santiago Ruiz $2-3$ ($BSR\,3-4$), Bodega Santiago Ruiz $5-6$ ($BSR\,5-6$), Bodega Santiago Ruiz $7-8$ ($BSR\,7-8$), Bodega Santiago Ruiz $9-10$ ($BSR\,9-10$) e Bodega Santiago Ruiz $11-12$ ($BSR\,11-12$); cada uma destas localizações corresponde a duas das videiras sinalizadas com um círculo. Na Figura 3.6 cada par destas videiras encontram-se representadas com a mesma cor.

Segundo os responsáveis pela recolha da amostra, estas variáveis não foram medidas em 12 localizações, mas sim em 6, uma vez que para ter material suficiente para analisar as uvas foi necessário juntar as uvas de duas videiras e, assim, se conseguir analisar o pretendido.

Por fim foram analisados os 6 vinhos correspondentes a estas 6 localizações.

Nas Tabelas 3.1, 3.2, 3.3, 3.4 e 3.5 encontra-se uma descrição de todas as variáveis em estudo na quinta da Galiza e na quinta do Minho, apresentando-se as unidades de medida de cada uma delas, bem como as suas designações pelas quais vão ser referenciadas ao longo do trabalho.

Tabela 3.1: Variáveis do solo em estudo nas parcelas do Minho e da Galiza.

	Variáveis		Descrição	Existência	
	Designação	Unidades		Minho	Galiza
SOLO	DA	(g/vol)	Densidade Aparente	×	×
	MO	$(\%)$	Matéria Orgânica	×	×
	pH		pH em Estrato Aquoso	×	×
	FF	$(\%)$	Fração Fina ($\%FF + \%FG = 100\%$)	×	×
	FG	$(\%)$	Fração Grosseira ($\%FF + \%FG = 100\%$)	×	×
	$P2O5$	(ug/g)	Fósforo Assimilável	×	×
	$K2O$	(ug/g)	Potássio Assimilável	×	×
	Ca	(ug/g)	Cálcio Assimilável	×	×
	Mg	(ug/g)	Magnésio Assimilável	×	×
	AzT	$(\%)$	Azoto Total	×	×
	Ni	(ug/g)	Níquel	×	×
	Cr	(ug/g)	Crómio	×	×
	Cd	(ug/g)	Cádmio	×	×
	N	(ug/g)	Nitratos	×	×
	B	(ug/g)	Boro	×	×
	CTC	$(m.e./100g)$	Capacidade de Troca Catiónica	×	×

Na Tabela 3.1 apresentam-se as 16 variáveis alusivas às características do solo que foram avaliadas em ambas as quintas. Note-se que a soma das percentagens das variáveis Fração Fina (FF) e Fração Grosseira (FG) para cada observação dá 100%, ou seja, $\%FF + \%FG = 100\%$, o que será uma informação importante para a interpretação de alguns resultados obtidos.

Tabela 3.2: Variáveis referentes à produtividade da videira em estudo nas parcelas do Minho e da Galiza.

	Variáveis		Descrição	Existência	
	Designação	Unidades		Minho	Galiza
VIDEIRA	$Ncachos$		Número de cachos por videira	×	×
	$Uvas\_kg\_vid$	(kg/vid)	Peso de cachos por videira	×	×
	$Pbruto$	(kg)	Peso bruto por videira	×	×
	$Pcacho$	(kg)	Peso médio de cacho por videira ($Pbruto \div Ncachos$)	×	×

Existe um outro conjunto de quatro variáveis referentes à produtividade das videiras que se podem observar na Tabela 3.2. Existe a relação $Pcacho = Pbruto \div Ncachos$, podendo-se assim, caso algumas análises o sugiram, proceder à remoção de algumas destas variáveis uma vez que umas dão informações totais sobre as outras. Também as variáveis aleatórias $Pbruto$ e

$Uvas\_kg\_vid$ são variáveis que apresentam valores muito próximos pelo que também há a possibilidade de apenas considerar-se uma delas e, assim, facilmente se reduzir o número de variáveis a estudar.

Tabela 3.3: Variáveis referentes à qualidade do mosto em estudo nas parcelas do Minho e da Galiza.

	Variáveis		Descrição	Existência	
	Designação	Unidades		Minho	Galiza
MOSTO					
	$pH.mosto$		pH do mosto	×	
	$Ac.tart$	(g/dm^3)	Ácido Tartárico	×	
	$Ac.mal$	(g/dm^3)	Acido Málico	×	
	$Ac.TOT$	$(gc.tar./dm^3)$	Acidez Total	×	×
	TAP	$(\%vv)$	Teor de Álcool Provável		×
	Aroma				
	$FL1M$	$(\mu g/l)$	Família de compostos em $C6$ do aroma do mosto na fração livre	×	
	$FL2M$	$(\mu g/l)$	Família de álcoois do aroma do mosto na fração livre	×	
	$FL3M$	$(\mu g/l)$	Família de álcoois monoterpénicos do aroma do mosto na fração livre	×	
	$FL4M$	$(\mu g/l)$	Família de fenóis voláteis do aroma do mosto na fração livre	×	
	$FL5M$	$(\mu g/l)$	Família de compostos Carbonilados do aroma do mosto na fração livre	×	
	$FG1M$	$(\mu g/l)$	Família de compostos em $C6$ do aroma do mosto na fração glicosilada	×	
	$FG2M$	$(\mu g/l)$	Família de álcoois do aroma do mosto na fração glicosilada	×	
	$FG3M$	$(\mu g/l)$	Família de álcoois monoterpénicos do aroma do mosto na fração glicosilada	×	
	$FG4M$	$(\mu g/l)$	Família de óxidos e dióis monoterpénicos do aroma do mosto na fração glicosilada	×	
	$FG5M$	$(\mu g/l)$	Família de norisoprenóides em C13 do aroma do mosto na fração glicosilada	×	
	$FG6M$	$(\mu g/l)$	Família de fenóis voláteis do aroma do mosto na fração glicosilada	×	
	$FG7M$	$(\mu g/l)$	Família de compostos carbonilados do aroma do mosto na fração glicosilada	×	
	Cor				
	$Amarelo\_M$	$(\%)$	Percentagem de cor amarelo	×	
	$Vermelho\_M$	$(\%)$	Percentagem de com vermelho	×	
	$Azul\_M$	$(\%)$	Percentagem de cor azul $(\%Amarelo\_M + \%Vermelho\_M + \%Azul\_M = 100\%)$	×	
	Sumo				
	$Sumo\_T\_M$	(ml)	Sumo turvo	×	
	$Sumo\_L\_M$	(ml)	Sumo límpido	×	
	$Rend\_sumo\_M$	$(\%)$	Percentagem do rendimento em sumo $(Sumo\_L\_M \div Sumo\_T\_M \times 100)$	×	

Dentro de um terceiro grupo de variáveis designado por Mosto, Tabela 3.3, existe um conjunto de 5 variáveis que diz respeito às características físico-químicas do mosto[4]; apenas as variáveis $Ac.TOT$ e TAP foram analisadas na quinta da Galiza. Dentro deste conjunto também se encontram 3 variáveis relativas à cor do mosto, bem como 3 variáveis relativas ao sumo analisado, mas apenas a variável correspondente ao rendimento em sumo será utilizada por ser a que apresenta maior interesse. Note-se que é possível estabelecerem-se as relações

$$\%Amarelo\_M + \%Vermelho\_M + \%Azul\_M = 100\% \tag{3.1}$$

$$Sumo\_L\_M \div Sumo\_T\_M \times 100 = Rend\_sumo\_M. \tag{3.2}$$

A partir do mosto obtido da vinha do Minho, foi também realizada uma análise ao seu aroma[5]: do mosto foram extraídos os compostos do aroma na forma livre e glicoconjugada, obtendo-se extratos que de seguida foram analisados e identificados os compostos voláteis, tendo sido agrupados por famílias químicas, como é ilustrado no Anexo I. Foram identificados e quantificados um

[4] Este dados foram obtidos a partir de uma análise efetuada de acordo com o Regulamento (CEE) nº 2676/90, de 17 de setembro de 1990 (Silva, 2010).

[5] Todos os procedimentos executados para a obtenção dos dados relativos aos aromas, sumos e cor encontram-se descritos nos relatórios disponibilizados no site http://www.sinergeo.pt/

total de 16 compostos do aroma na forma livre que foram agrupados em 5 famílias. Tem-se ainda um total de 42 agliconas odoríferas nos extratos da fração glicosilada agrupadas em 7 famílias.

Tabela 3.4: Variáveis referentes à qualidade das uvas em estudo na parcela da Galiza.

	Variáveis		Descrição	Existência	
	Designação	Unidades		Minho	Galiza
UVAS	**Aroma**				
	$FL1U$	$(\mu g/l)$	Família de compostos em $C6$ do aroma das uvas na fração livre		×
	$FL2U$	$(\mu g/l)$	Família de álcoois do aroma das uvas na fração livre		×
	$FL3U$	$(\mu g/l)$	Família de álcoois monoterpénicos do aroma das uvas na fração livre		×
	$FL4U$	$(\mu g/l)$	Família de fenóis voláteis do aroma das uvas na fração livre		×
	$FL5U$	$(\mu g/l)$	Família de compostos carbonilados do aroma das uvas na fração livre		×
	$FL6U$	$(\mu g/l)$	Família de óxidos e dióis monoterpénicos do aroma das uvas na fração livre		×
	$FL7U$	$(\mu g/l)$	Família de ácidos gordos voláteis do aroma das uvas na fração livre		×
	$FG1U$	$(\mu g/l)$	Família de compostos em C6 do aroma das uvas na fração glicosilada		×
	$FG2U$	$(\mu g/l)$	Família de álcoois do aroma das uvas na fração glicosilada		×
	$FG3U$	$(\mu g/l)$	Família de álcoois monoterpénicos do aroma das uvas na fração glicosilada		×
	$FG4U$	$(\mu g/l)$	Família de óxidos e dióis monoterpénicos do aroma das uvas na fração glicosilada		×
	$FG5U$	$(\mu g/l)$	Família de norisoprenóides em C13 do aroma das uvas na fração glicosilada		×
	$FG6U$	$(\mu g/l)$	Família de fenóis voláteis do aroma das uvas na fração glicosilada		×
	$FG7U$	$(\mu g/l)$	Família de compostos carbonilados do aroma das uvas na fração glicosilada		×
	Cor				
	$Amarelo\_U$	$(\%)$	Percentagem de cor amarelo		×
	$Vermelho\_U$	$(\%)$	Percentagem de com vermelho		×
	$Azul\_U$	$(\%)$	Percentagem de cor azul		×
			$(\%Amarelo\_U + \%Vermelho\_U + \%Azul\_U = 100\%)$		
	Sumo				
	$Massa$	(g)	Massa		×
	$Sumo\_T\_U$	(ml)	Sumo turvo		×
	$Sumo\_L\_U$	(ml)	Sumo límpido		×
	$Rend\_sumo\_U$	$(\%)$	Percentagem do rendimento em sumo $(Sumo\_L\_U \div Massa \times 100)$		×

Toda esta análise da composição aromática, cromática e rendimento em sumo também foi realizada às uvas (em vez do mosto) na quinta da Galiza, Tabela 3.4. Foram identificados e quantificados um total de 29 compostos do aroma na forma livre que foram agrupados em 7 famílias, Anexo II. Tem-se ainda um total de 51 agliconas odoríferas nos extratos da fração glicosilada agrupadas em 7 famílias. Note-se que relativamente às variáveis do sumo obtido a partir das uvas, tem-se a seguinte relação

$$Sumo\_L\_M \div Massa \times 100 = Rend\_sumo\_U. \tag{3.3}$$

Tabela 3.5: Variável referente à qualidade do vinho em estudo nas parcelas do Minho e da Galiza.

	Variáveis		Descrição	Existência	
	Designação	Unidades		Minho	Galiza
VINHO	$Nota.Final$		Nota atribuída por um painel de provadores de vinho	×	×

Finalmente, o último grupo constituído apenas por uma variável (do Vinho), Tabela 3.5, refere-se a uma classificação atribuída aos vinhos resultante do preenchimento de uma ficha de prova descritiva da Câmara de Provas da CVRVV que se encontra no Anexo III. Cada um dos sete provadores de vinho após dar o seu parecer, atribuindo uma classificação a nível visual, olfativo e gustativo de cada vinho, também fez uma apreciação global do vinho. No final, foi atribuída uma nota final que 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.

3.2.1 Análise Exploratória dos Dados

Numa primeira fase desta análise exploratória dos dados, começou-se por verificar a existência de valores em falta (designados usualmente por *Not Available* (*NA*)) nas bases de dados referentes ao Minho e à Galiza. A Tabela 3.6 apresenta um resumo da localização dos *NA*, bem como a respetiva percentagem, num total de 45 observações.

Tabela 3.6: Valores em falta das variáveis em estudo referentes ao Minho e Galiza, sua localização e percentagem.

Localização		Variáveis	Número de valores em falta	Ponto de amostragem	%
	S	DA	2	A44 A45	4,44
Minho	O	FF	1	A14	2,22
	L	FG	1	A14	2,22
Galiza	O	DA	21	B8-B15 B25-B30 B39-B45	46,67

Dada a grande percentagem de valores em falta na variável Densidade Aparente (DA) do solo na Galiza, aproximadamente $46,67\%$, foi decidido, em concordância com o perito na área em estudo, remover tal variável. Note-se que estes 21 dados perdidos, deve-se ao facto de esta área do solo ser constituída por seixos de média e grande dimensão que impossibilitou que o trado fosse introduzido para recolher material para análise. No que diz respeito aos restantes *NA*, foram seguidas várias alternativas para a resolução da situação, isto é, formas de estimar os valores em falta, mas apenas duas pareceram as mais adequadas:

- substituir estes valores pelo método do vizinho mais próximo;
- substituir os valores pelo valor médio das observações destas variáveis.

De acordo com a opinião do perito na área em estudo, optou-se pelo primeiro método: a partir do cálculo da distância Euclidiana encontrou-se o ponto georreferenciado, que estava mais próximo do que continha o *NA* e partilharam o respetivo valor da variável.

Numa segunda fase desta análise, após se proceder à substituição dos *NA* e à remoção da variável DA, fez-se uma análise preliminar dos dados com o objetivo de maximizar a obtenção de informações sobre os dados e, assim, descobrir tendências, detetar alguns padrões, *outliers*, etc., que poderão guiar os procedimentos a realizar posteriormente.

As Tabelas 3.7 e 3.8 apresentam um resumo de algumas das principais características amostrais relativas a cada uma das variáveis analisadas nas quintas do Mnho e da Galiza, respetivamente.

Tabela 3.7: Algumas das principais características amostrais das variáveis em análise no Minho.

	Variáveis	nºobs.	Mínimo	Máximo	Mediana	Média	Desvio padrão	Percentis 25%	Percentis 75%
SOLO	DA		0,90	1,32	1,16	1,14	0,11	1,08	1,23
	MO		1,85	10,11	2,73	3,48	1,84	2,23	4,12
	pH		5,14	6,60	5,73	5,76	0,36	5,45	6,01
	FF		54,33	83,70	68,33	68,38	5,64	64,42	71,54
	FG		16,30	45,67	31,67	31,62	5,64	28,46	35,58
	$P2O5$		10,95	93,80	37,11	41,50	17,61	26,47	51,91
	$K2O$		46,07	149,00	82,19	86,37	25,03	67,33	106,28
	Ca		244,50	835,50	450,00	482,63	168,32	354,00	616,50
	Mg	45	52,50	108,00	73,50	72,77	12,04	64,50	81,00
	AzT		0,08	0,43	0,12	0,15	0,07	0,01	0,17
	Ni		0,96	6,00	3,68	3,76	1,37	2,64	4,92
	Cr		0,02	0,94	0,43	0,43	0,21	0,26	0,61
	Cd		0,07	0,22	0,10	0,11	0,03	0,10	0,13
	N		0,94	1,99	1,36	1,34	0,26	1,13	1,48
	B		0,29	0,79	0,40	0,42	0,10	0,35	0,45
	CTC		6,60	18,25	10,92	11,34	3,26	8,61	13,88
VIDEIRA	$Ncachos$		19,00	71,00	44,00	45,27	12,33	39,00	53,00
	$Uvas\_kg\_vid$		2,25	17,05	7,75	8,24	3,29	5,75	9,05
	$Pcacho$	45	0,12	0,43	0,17	0,18	0,05	0,15	0,20
	$Pbruto$		3,00	17,80	8,50	8,99	3,29	6,50	9,80
MOSTO	$pH.mosto$		3,05	3,30	3,14	3,16	0,08	3,12	3,19
	$Ac.tart$		5,55	7,26	6,77	6,66	0,52	6,71	6,91
	$Ac.mal$		1,49	2,74	2,10	2,09	0,43	1,81	2,43
	$Ac.TOT$		5,74	7,92	6,94	6,90	0,65	6,64	7,23
	TAP								
	Aroma								
	$FL1M$		684,90	908,60	741,50	758,00	68,51	720,20	788,00
	$FL2M$		74,17	277,54	120,00	138,91	64,91	103,71	179,12
	$FL3M$		28,67	96,77	43,70	54,54	22,41	39,12	67,94
	$FL4M$		0,22	4,83	2,31	2,42	1,29	2,04	2,91
	$FL5M$	9	5,39	42,98	19,14	22,24	11,12	15,44	25,18
	$FG1M$		24,43	83,07	66,53	64,18	16,87	62,38	74,18
	$FG2M$		83,77	276,11	221,80	211,74	61,21	200,80	250,32
	$FG3M$		5,52	21,96	11,27	12,97	5,62	8,96	17,64
	$FG4M$		8,80	34,31	15,53	19,10	8,63	13,76	26,77
	$FG5M$		14,26	110,33	35,69	42,02	27,87	29,82	36,78
	$FG6M$		15,59	116,34	53,36	54,68	28,60	37,09	63,15
	$FG7M$		2,31	5,63	2,96	3,54	1,15	2,69	4,11
	Cor								
	$Amarelo\_M$		36,27	41,94	39,62	39,37	1,66	38,44	40,20
	$Vermelho\_M$		43,88	57,31	52,75	55,26	5,21	47,03	55,26
	$Azul\_M$		4,25	16,41	8,54	9,34	4,33	5,95	12,19
	Sumo								
	$Sumo\_T\_M$		40,00	60,00	50,00	50,00	5,87	47,00	54,00
	$Sumo\_L\_M$		35,00	52,00	44,00	44,11	5,51	41,00	48,00
	$Rend\_sumo\_M$		82,00	95,70	88,00	88,21	3,91	86,70	89,80
VINHO	$Nota.Final$	9	59,00	65,14	62,71	62,40	2,33	60,14	64,57

Tabela 3.8: Algumas das principais características amostrais das variáveis em análise na Galiza.

	Variáveis	nºobs.	Mínimo	Máximo	Mediana	Média	Desvio padrão	Percentis 25%	Percentis 75%
SOLO	MO		2,26	22,37	7,18	8,35	5,02	4,46	11,87
	pH		4,04	5,86	5,06	5,05	0,31	4,93	5,24
	FF		48,65	95,50	72,27	71,45	13,59	59,27	82,74
	FG		4,50	51,35	27,73	28,55	13,59	17,26	40,73
	$P2O5$		0,00	64,83	5,20	8,47	12,02	1,83	9,11
	$K2O$		29,41	151,90	55,34	61,75	25,68	45,19	68,66
	Ca		232,50	916,50	327,00	365,30	123,81	292,50	400,50
	Mg	45	49,50	204,00	67,50	75,20	24,76	61,50	82,50
	AzT		0,08	0,67	0,26	0,30	0,17	0,18	0,45
	Ni		2,40	7,16	5,48	5,03	1,50	3,52	6,16
	Cr		0,10	0,80	0,37	0,42	0,19	0,29	0,56
	Cd		0,07	0,15	0,08	0,09	0,02	0,07	0,10
	N		1,43	5,41	2,36	2,47	0,86	1,80	2,90
	B		0,35	1,15	0,58	0,64	0,21	0,48	0,79
	CTC		5,81	22,97	8,16	9,14	3,00	7,45	10,21
VIDEIRA	$Ncachos$		29,00	116,00	80,00	78,25	24,58	71,25	92,75
	$Uvas\_kg\_vid$		2,95	19,15	10,55	10,54	4,16	9,83	12,18
	$Pcacho$	12	0,12	0,19	0,14	0,14	0,02	0,13	0,15
	$Pbruto$		3,70	19,90	11,30	11,29	4,16	10,57	12,93
MOSTO	$Ac.TOT$	6	9,07	11,50	10,32	10,29	0,84	9,85	10,73
	TAP		11,30	12,90	11,90	12,00	0,59	11,62	12,32
UVAS	**Aroma**								
	$FL1U$		98,30	178,00	158,90	149,70	32,28	131,80	175,90
	$FL2U$		53,71	86,23	58,38	64,16	12,78	55,90	69,46
	$FL3U$		10,44	13,36	12,21	12,05	1,21	11,15	13,03
	$FL4U$		4,15	6,40	4,70	4,86	0,79	4,50	4,78
	$FL5U$	6	25,85	49,72	30,63	33,49	9,08	27,06	36,34
	$FL6U$		5,88	19,17	11,77	11,95	4,73	9,08	14,12
	$FL7U$		1,91	18,44	4,37	6,16	6,17	2,86	5,23
	$FG1U$		4,21	6,99	4,54	5,16	1,21	4,30	5,96
	$FG2U$		31,64	70,50	49,66	50,63	12,86	46,18	55,41
	$FG3U$		19,73	39,17	22,30	24,68	7,40	20,18	24,38
	$FG4U$		68,86	101,53	80,59	81,14	11,97	72,22	84,53
	$FG5U$		23,86	42,33	30,43	31,41	6,14	29,00	32,29
	$FG6U$		12,64	34,12	22,89	23,27	7,91	18,38	28,40
	$FG7U$		0,58	1,09	0,71	0,80	0,21	0,66	0,95
	Cor								
	$Amarelo\_U$		52,49	55,12	53,82	53,85	0,99	53,29	54,51
	$Vermelho\_U$		28,32	29,54	29,08	29,04	0,45	28,84	29,36
	$Azul\_U$		16,04	18,25	16,99	17,11	0,82	16,63	17,66
	Sumo								
	$Massa\_U$		561,10	601,20	576,80	577,00	13,97	568,30	579,90
	$Sumo\_T\_U$		370,00	430,00	390,00	398,30	22,29	390,00	412,50
	$Sumo\_L\_U$		330,00	415,00	380,00	377,50	29,96	365,00	395,00
	$Rend\_sumo\_U$		58,80	72,30	66,35	65,45	5,24	61,33	68,53
VINHO	$Nota.Final$	6	56,29	66,00	60,45	61,20	3,43	59,99	63,23

Fazendo uma análise gráfica comparativa entre as variáveis dos solos das duas regiões (Minho e Galiza) a partir da construção de *boxplots*, Figura 3.7, parece existir uma maior concentração de MO, AzT, Ni, N e B no solo da Galiza, havendo uma grande variabilidade de concentrações de MO e AzT neste mesmo solo.

Relativamente às variáveis pH, $P2O5$, $K2O$, Ca, Cd e CTC têm uma maior concentração no solo do Minho, notando-se uma elevada variabilidade nas variáveis $P2O5$ e Ca neste solo comparativamente ao solo da Galiza. Em relação às variáveis FF, FG, Mg e Cr os gráficos não sugerem grandes diferenças de concentrações, percebendo-se apenas que há uma maior variabilidade de FF e FG no solo da Galiza.

A partir da observação da Figura 3.7 pode-se ainda verificar a existência de observações

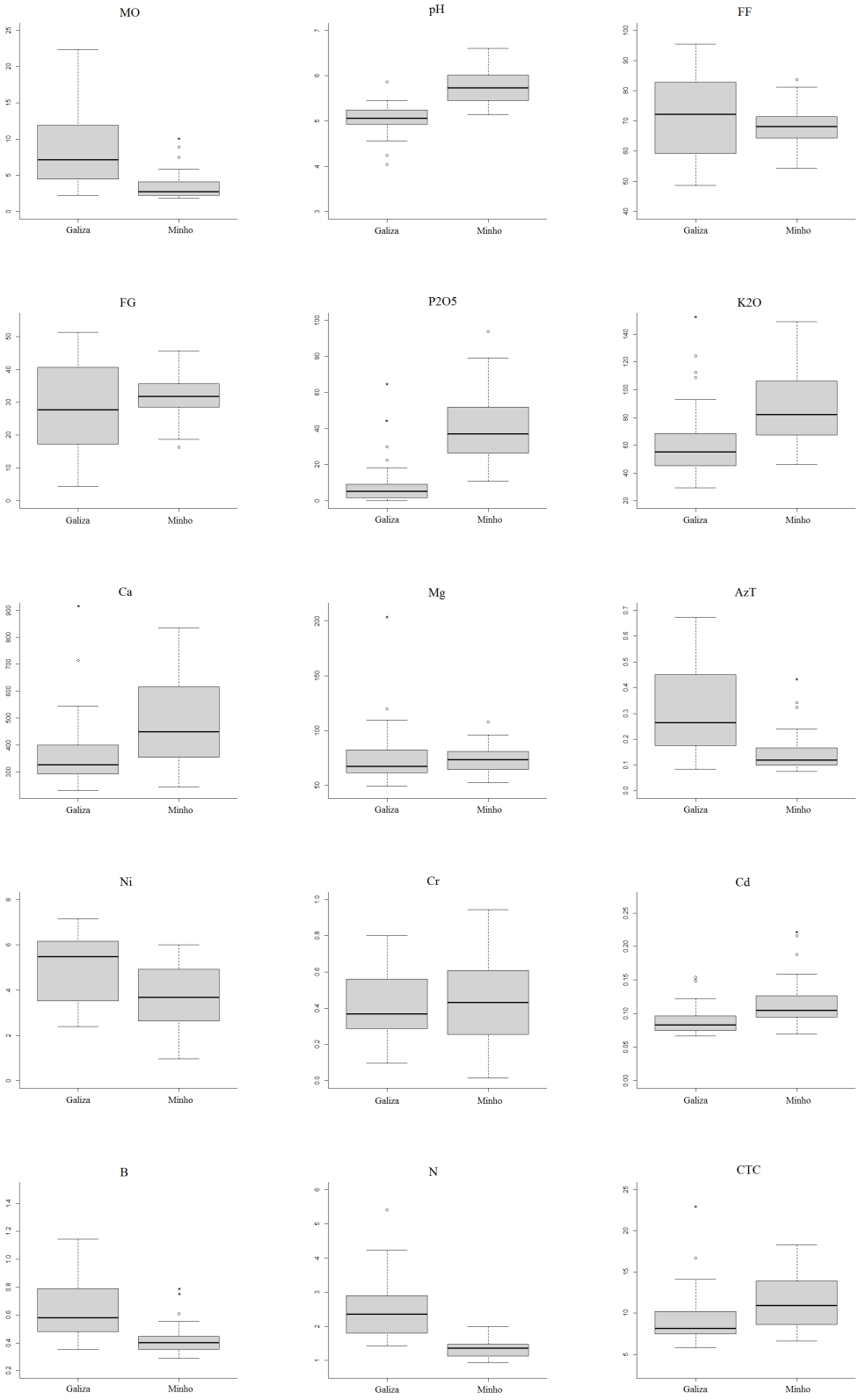


Figura 3.7: *Boxplots* referentes às variáveis do solo da Galiza e do Minho.

outliers, ou seja, observações que apresentam valores muito discrepantes comparativamente aos das restantes observações. Estes valores atípicos podem ter a capacidade de alterar os resultados finais e, como tal, muitas vezes é necessário dar-lhes uma atenção especial.

Na Tabela 3.9 apresenta-se o número de *outliers* moderados e severos das variáveis relativas ao solo do Minho e da Galiza.

Tabela 3.9: *Outliers* severos e moderados presentes nas variáveis do solo do Minho e da Galiza.

Localização		Variáveis	Número de *outliers*	Ponto de amostragem dos *outliers* moderados	Ponto de amostragem dos *outliers* severos
Minho	SOLO	B	3	A38	A43 A35
		MO	3	A35 A38	A43
		AzT	3	A35 A38	A43
		Cd	3	A13 A31	A28
		FF e FG	1	A35	
		Mg	1	A18	
		$P2O5$	1	A11	
Galiza	SOLO	$P2O5$	4	B41 B29	B42 B8
		$K2O$	4	B42 B22 B29	B8
		pH	3	B43 B37 B29	
		Cd	3	B42 B30 B29	
		Ca	2	B41	B42
		Mg	2	B41	B42
		CTC	2	B41	B42
		N	1	B29	

Os *outliers* severos aparecem nos pontos de amostragem $A43$, $A35$, $A28$, $B42$, $B8$ e $B42$. Repare-se que existem pontos de amostragem que apresentam *outliers* em diferentes variáveis, tais como $A43$, $A35$, $B41$, $B42$,..., o que pode indicar que estas partes do solo apresentam características diferentes (discordantes) relativamente a diferentes variáveis e, como tal, devem ser sinalizados no estudo para se poder estudar a sua particularidade.

No entanto, quando se realizou uma Análise Fatorial, esta foi efetuada com e sem os *outliers* severos, e os resultados foram similares, pelo que se optou por trabalhar com estes valores, pois a sua eliminação implicaria perda de informação uma vez que não foram erros de registo mas dados observados.

Capítulo 4

Apresentação dos Resultados e Discussão

Neste capítulo serão apresentados e discutidos os resultados obtidos pela aplicação da metodologia geral desenvolvida neste trabalho, na área da Estatística Multivariada e da Inferência Estatística (aplicação de testes de hipóteses não paramétricos), com o intuito de se analisar a influência de algumas características do solo na qualidade do vinho. Pretende-se, também, caracterizar as parcelas dos solos em estudo de acordo com as várias concentrações dos seus constituintes que influenciam a produtividade da videira e a qualidade das uvas, do mosto e do vinho.

Para atingir este fim, foram realizados dois estudos pela aplicação de Análises de *Clusters* ao conjunto dos dados do solo: um estudo a partir das variáveis originais do solo e o outro a partir dos fatores retidos pela aplicação de uma Análise Fatorial a essas mesmas variáveis. O objetivo deste último estudo prende-se com o objetivo de reduzir a dimensionalidade do problema.

Esta análise foi efetuada nas duas parcelas em estudo, no Minho e na Galiza e, no final, os resultados de ambas foram analisados e comparados, retirando-se as principais conclusões.

Note-se que, inicialmente, também foi objetivo deste estudo realizar Análises de Regressão Linear Múltipla de modo a estudar como é que as variáveis do solo (variáveis explicativas, independentes ou regressoras) influenciam cada uma das variáveis referentes à videira, à uva/mosto ou ao vinho (variáveis resposta ou dependentes). Devido ao número muito reduzido de observações que a maioria das variáveis resposta apresentaram, apenas foi possível realizar-se este tipo de análise com duas variáveis referentes à produtividade da videira: peso bruto de cachos por videira e peso médio de cacho por videira. Obtendo-se resultados estatisticamente pouco significativos (permitindo escassas interpretações e conclusões), decidiu-se apenas analisar a relação de apenas uma variável do solo (explicativa) com as variáveis resposta, isto é, aplicar Análises de Regressão Linear Simples.

A interpretação e a discussão dos resultados provenientes das Análises de Regressão Linear Simples foi efetuada com a ajuda de peritos na área de estudo, concluindo-se que os resulta-

dos provenientes dos dados da Galiza não fariam qualquer sentido tendo como possível causa o facto deste estudo estar enviesado, uma vez que a análise referida foi realizada com um total de 12 observações. Quanto aos resultados provenientes da análise dos dados da quinta do Minho, verificaram-se algumas relações lineares significativas, indo ao encontro do conhecimento que os peritos possuíam, bem como aos resultados provenientes de estudos que foram realizados nesta área.

Mas, contudo, esta análise foi criticada por um outro perito na área da geologia, argumentando não se poder estabelecer uma relação direta entre as concentrações químicas do solo e a produtividade da videira, uma vez que existem vários constituintes no solo, como por exemplo a água disponível, que não se integraram como variáveis neste estudo, e que sabe-se que influenciam a capacidade de absorção da planta destes nutrientes. Desta forma, as concentrações das variáveis que vão influenciar a produtividade da videira não são as que existem no solo mas sim as que a planta absorve. Assim, os resultados desta análise apenas foram importantes para colocar questões e discutir possíveis relações que determinadas variáveis do solo poderiam ter com as variáveis peso bruto de cachos por videira e peso médio de cacho por videira tendo em atenção o conhecimento destes processos. Por estas razões decidiu-se, que estas análises iriam ser apresentadas de forma sucinta em anexo (Anexo V).

4.1 Quinta Campos de Lima: Resultados e Discussão

Pelo facto de as variáveis em análise não terem sido medidas o mesmo número de vezes, e por isso impossibilitou a análise de algumas relações/associações entre elas e a aplicação de alguns testes estatísticos, este estudo iniciou-se com o objetivo de reduzir o número de observações em algumas variáveis de modo a que todas elas apresentassem o mesmo número de elementos para, posteriormente, se conseguir aplicar metodologias estatístiocas mais adequadas. Calculando-se as médias amostrais das variáveis do solo e da videira de cada conjunto de observações em cada uma das 9 parcelas ($B1N$, $B1C$, $B1P$, $B2N$, $B2C$, $B2P$, $B3N$, $B3C$ e $B3P$), de um conjunto de 45 observações passou-se a ter 9 observações (número de elementos igual à de todas as outras variáveis relativas à qualidade do vinho).

Haveria uma outra alternativa que passava por aplicar o método do vizinho mais próximo ou o método de krigagem para estimar novas observações e, assim, aumentar o número de observações de 9 para um total de 45. Contudo a utilização deste último método para resolver este problema de igualar as dimensões das amostras não pareceu o mais indicado, uma vez que seriam poucas as observações para estimar um número tão elevado de novas observações.

Resolvido este problema, procedeu-se de seguida a uma Análise de *Clusters* (AC) realizada a

partir das variáveis do solo estandardizadas, com o objetivo de identificar, na parcela em estudo, zonas com características semelhantes ao nível dos constituintes do solo, para de seguida se realizarem comparações entre estas zonas relativamente às variáveis referentes à produtividade da videira, da qualidade do mosto/uvas e da qualidade do vinho.

A escolha dos métodos e distâncias a serem utilizadas nesta análise é sempre complexa, uma vez que cada opção vai representar uma perspetiva diferente de formação dos *clusters* e, consequentemente, os resultados obtidos poderão não ser idênticos. Neste trabalho, por serem os que pareceram mais adequados, foram utilizados os métodos hierárquicos com o algoritmo aglomerativo.

Não sendo considerado nenhum método hierárquico aglomerativo como o melhor, é comum utilizarem-se diversos métodos e, de seguida, fazer-se a comparação dos resultados, podendo concluir que se obtiveram resultados com elevado grau de confiança se estes forem semelhantes. Seguindo tal procedimento, utilizando a distância euclidiana por ser a mais usual e apresentar resultados semelhantes relativamente a outras distâncias aplicadas a este caso, foram operados 5 métodos (método de ligação completa, de ligação média, de ligação simples, método de *Ward* e do centróide) e no final selecionou-se a solução que pareceu a mais adequada e que obteve uma correlação cofenética mais elevada.

Os resultados obtidos pelos métodos da ligação completa, da ligação média e de *Ward* mostraram os *clusters* com melhores definições e, para além disso, apresentaram *clusters* praticamente semelhantes, o que levou a concluir que há uma clara estrutura de grupos subjacentes aos dados. O resultado dos métodos da ligação simples e do centróide distanciaram-se muito dos outros.

Com o objetivo de minimizar a perda de informação em cada passo do processo aglomerativo e perante a elevada correlação cofenética que apresentou, um valor aproximadamente de $0,74$, utilizou-se o método de *Ward* para identificar os grupos.

Após a formação dos *clusters* a partir do método escolhido, torna-se necessário decidir quantos grupos reter. Percebe-se que reter poucos grupos pode levar a que estes sejam muito heterogéneos e um elevado número pode levar a uma difícil interpretabilidade dos mesmos.

A visualização do dendrograma, que é um método empírico frequentemente utilizado para sugerir quantos grupos se devem considerar, e a opinião dos peritos nas áreas em estudo pesaram nesta tomada de decisão. Para além disso, esta decisão também se baseou na visualização da representação gráfica do valor da medida de proximidade entre dois grupos, juntos em cada etapa.

Observando a Figura 4.1, o gráfico representado indica que o primeiro grande aumento significativo na distância entre duas etapas adjacentes foi registado entre as etapas 6 e 7, pelo que será razoável parar com a junção de grupos após a etapa 6, obtendo-se, assim, 3 *clusters*.

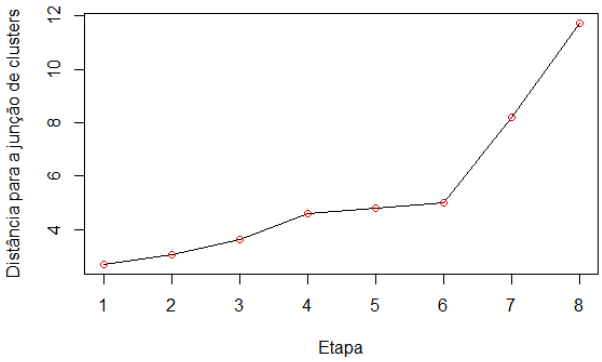


Figura 4.1: Critério para obter o número de *clusters*.

A Figura 4.2 ilustra o dendrograma obtido, assinalando-se os 3 *clusters* a considerar, de acordo com os constituintes do solo e, a Figura 4.3 apresenta o mapeamento dos grupos obtidos.

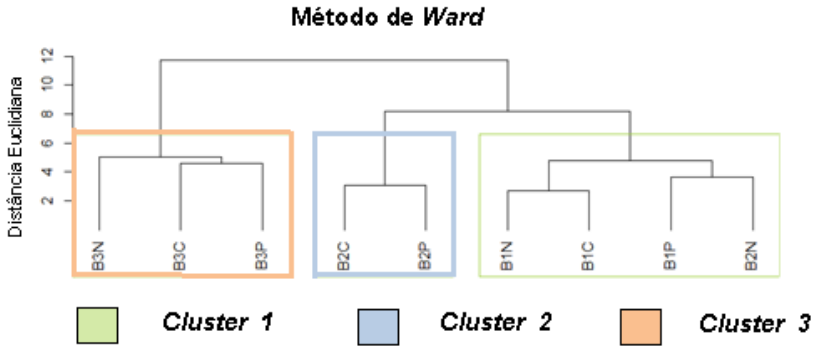


Figura 4.2: Dendrograma resultante do método de *Ward* assinalado com o número de *clusters* a considerar.

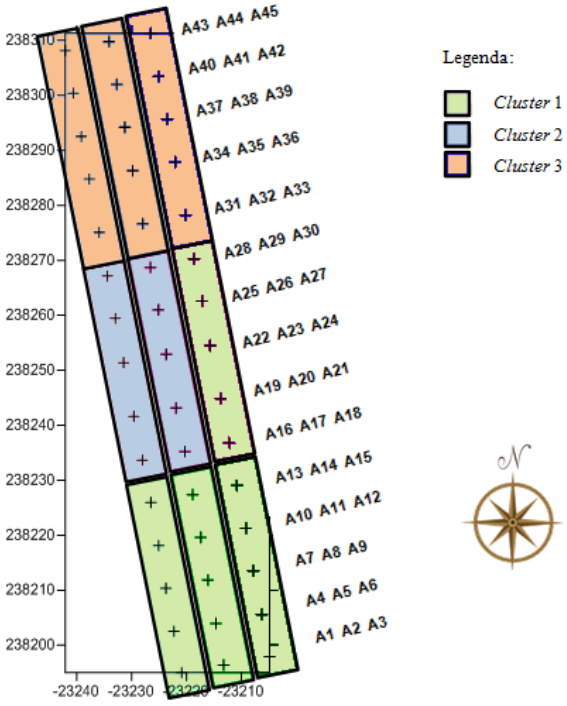


Figura 4.3: Distribuição dos 3 grupos na parcela em estudo de acordo com os resultados obtidos a partir do método *Ward*.

Fazendo uma análise ao dendrograma, observa-se que o 1^{o} *cluster* é constituído pelas parcelas $B1P$, $B1C$, $B1N$ e $B2N$, o 2^{o} *cluster* engloba a zona $B2P$ e $B2C$ e ao 3^{o} *cluster* pertencem os terrenos $B3P$, $B3C$ e $B3N$.

A parcela escolhida na quinta do Minho é caracterizada por um solo residual granítico, com um ligeiro aumento da cota de Sul para Norte. A drenagem das águas superficiais ocorre de Norte para Sul e de Oeste para Este. Empiricamente, esta parcela é dividida em 3 blocos (Bloco 1, Bloco 2 e Bloco 3). O Bloco 1 corresponde aos lotes $B1P$, $B1C$ e $B1N$; o bloco 2 corresponde a $B2P$, a $B2C$ e a $B2N$ e o Bloco 3 corresponde a $B3P$, a $B3C$ e a $B3N$. Face ao exposto, consideram-se que os *clusters* 1, 2 e 3 estão separados pela diferença de cota que se verifica no terreno. O *cluster* 3 está a uma cota superior ao *cluster* 2 e este está a uma cota superior ao *cluster* 1.

A Tabela 4.1 apresenta um resumo das principais características amostrais de cada um dos constituintes do solo em cada um dos *clusters*. Analisando os resultados obtidos, observa-se que:

- no *cluster* 1, as variáveis DA, pH, FG, $P2O5$, Ca, Mg, Cr, Cd, N e CTC apresentam o valor mediano amostral mais elevado, sendo que as variáveis FF e Ni são as que possuem o valor mediano mais baixo;
- no *cluster* 2, as variáveis MO, $K2O$, Ca, Mg, AzT, N, B e CTC têm o valor mediano mais baixo;
- no *cluster* 3, as variáveis MO, FF, $K2O$, AzT, Ni e B mostram o valor mediano mais elevado, sendo que a variável DA, pH, FG, $P2O5$, Cr e Cd são as que apresentam o valor mediano mais baixo.

Claramente se percebe que no *cluster* 1 a maioria das variáveis não apresentam o valor mediano mais baixo, comparativamente ao *cluster* 2 e 3; muitas delas apresentam o maior valor mediano. No *cluster* 2 não existem variáveis cujo valor mediano seja o mais elevado, mas metade das variáveis têm o menor valor mediano. Quanto ao *cluster* 3, existe uma mesma percentagem de variáveis que apresentam o valor mediano mais elevado e o valor mediano mais baixo.

Observando os resultados obtidos, também se pode verificar que:

- no *cluster* 1, as variáveis Ca, Mg, Ni, Cr e CTC mostram o maior desvio padrão e, apenas a variável DA e N possuem o mais baixo;
- no *cluster* 2, apenas a variável Cd tem o maior desvio padrão mas, a grande maioria apresenta o desvio padrão mais baixo, nomeadamente, MO, pH, FF, FG, $P2O5$, $K2O$, Ca, Mg, AzT, B e CTC;
- no *cluster* 3, as variáveis DA, MO, pH, FF, FG, $P2O5$, $K2O$, AzT, N e B apresentam o desvio padrão mais elevado e a variável Ni, Cr e Cd têm o menor desvio padrão.

Tabela 4.1: Caracterização dos 3 *clusters* retidos de acordo com os elementos do solo.

	DA	MO	pH	FF	FG	$P2O5$	$K2O$	Ca	Mg	AzT	Ni	Cr	Cd	N	B	CTC
***CLUSTER* 1**																
Mínimo	1,121	2,208	5,710	64,831	32,612	43,771	88,158	474,900	73,500	0,098	2,000	0,400	0,093	1,366	0,338	11,246
Máximo	1,176	3,068	6,150	67,388	35,169	56,590	100,151	656,700	84,000	0,135	4,104	0,710	0,138	1,458	0,397	14,774
Mediana	1.135	2,942	5,882	65,716	34,284	46,837	90,671	541,350	81,300	0,128	2,376	0,526	0,120	1,385	0,396	12,692
Média	1,142	2,790	5,906	65,913	34,087	48,509	92,413	553,575	80,025	0,122	2,714	0,541	0,118	1,399	0,382	12,851
Desvio padrão	0,026	0,393	0,199	1,072	1,072	5,663	5,537	81,751	4,539	0,017	0,947	0,130	0,019	0,041	0,029	1,564
1^oQuartil	1,122	2,744	5,764	65,407	33,779	45,174	88,549	496,050	79,125	0,118	2,204	0,474	0,108	1,378	0,381	11,796
3^oQuartil	1,155	2,989	6,024	66,221	34,593	50,172	94,535	598,875	82,200	0,132	2,886	0,593	0,129	1,405	0,397	13,747
***CLUSTER* 2**																
Mínimo	1,100	2,152	5,780	65,150	33,676	35,539	69,696	378,900	63,600	0,092	3,424	0,342	0,102	1,112	0,333	9,157
Máximo	1,162	2,697	5,806	66,324	34,850	43,083	74,749	409,800	63,600	0,111	3,904	0,438	0,131	1,254	0,368	9,652
Mediana	1,131	2,425	5,793	65,737	34,263	39,311	72,222	394,350	63,600	0,102	3,664	0,390	0,116	1,183	0,351	9,404
Média	1,131	2,425	5,793	65,737	34,263	39,311	72,222	394,350	63,600	0,102	3,664	0,390	0,116	1,183	0,351	9,404
Desvio padrão	0,044	0,386	0,018	0,830	0,830	5,335	3,573	21,850	0,000	0,014	0,339	0,068	0,021	0,100	0,025	0,350
1^oQuartil	1,116	2,288	5,787	65,444	33,969	37,425	70,959	386,625	63,600	0,097	3,544	0,366	0,109	1,147	0,342	9,280
3^oQuartil	1,147	2,561	5,800	66,031	34,556	41,197	73,485	402,075	63,600	0,107	3,784	0,414	0,124	1,218	0,360	9,528
***CLUSTER* 3**																
Mínimo	1,104	3,942	5,362	69,960	25,101	19,573	80,164	372,300	64,200	0,167	5,064	0,278	0,101	1,242	0,453	9,205
Máximo	1,215	5,884	5,772	74,899	30,040	45,101	92,022	507,300	72,000	0,234	5,384	0,362	0,127	1,538	0,548	11,697
Mediana	1,130	5,602	5,516	73,850	26,150	36,177	91,043	461,100	71,400	0,230	5,192	0,298	0,102	1,274	0,539	10,978
Média	1,150	5,143	5,550	72,903	27,097	33,617	87,743	446,900	69,200	0,210	5,213	0,313	0,110	1,351	0,513	10,627
Desvio padrão	0,058	1,049	0,207	2,602	2,602	12,955	6,582	68,611	4,341	0,038	0,161	0,044	0,015	0,162	0,052	1,283
1^oQuartil	1,117	4,772	5,439	71,905	25,625	27,875	85,604	416,700	67,800	0,199	5,128	0,288	0,101	1,258	0,496	10,091
3^oQuartil	1,173	5,743	5,644	74,375	28,095	40,639	91.533	484,200	71,700	0,232	5,288	0,330	0,114	1,406	0,543	11,337

Conhecido o número de *clusters* pretende-se perceber se existem diferenças significativas entre eles, relativamente às concentrações dos constituintes do solo e às variáveis que traduzem a produtividade da videira e da qualidade do vinho. Para testar a possibilidade de existência destas diferenças recorreu-se ao teste não paramétrico de *Kruskal-Wallis*, uma vez que não é verificado o pressuposto da normalidade da distribuição dos grupos e o número de observações é bastante reduzido.

Os resultados significativos obtidos da aplicação deste teste, assumindo-se como hipótese nula que não existem diferenças significativas entre os três *clusters* relativamente às variáveis indicadas, encontram-se na Tabela 4.2. Também se pode encontrar nesta tabela os valores medianos amostrais das variáveis, em cada um dos *clusters*.

Tabela 4.2: Resultados significativos do teste de *Kruskal-Wallis* para as variáveis em estudo.

		Medianas			Estatística	Graus de	Valor-prova
	Variáveis	*Cluster* 1	*Cluster* 2	*Cluster* 3	de teste	liberdade	
SOLO	MO	2,942	2,425	5,602	6,300	2	**0,043****
	FF	65,716	65,737	73,850	5,400	2	0,067*
	FG	34,284	34,263	26,150	5,400	2	0,067*
	$P2O5$	46,837	39,311	36,177	4,878	2	0,087*
	Mg	81,300	63,600	71,400	7,059	2	**0,029****
	AzT	0,128	0,102	0,230	6,300	2	**0,043****
	Ni	2,376	3,664	5,192	5,800	2	0,058*
	Cr	0,526	0,390	0,298	5,611	2	0,060*
	B	0.396	0,351	0,539	6,353	2	**0,042****
	CTC	12.692	9,404	10,978	5,500	2	0,064*
MOSTO	$Ac.mal$	1,745	1,975	2,630	4,544	2	0,083*
	$FL1M$	707,035	768,430	788,050	6,111	2	**0,047****

**significativo a 5% *significativo a 10%

Pela aplicação deste teste, conclui-se que há evidência estatística, a um nível de significância de $0,05$, para rejeitar a igualdade de valores medianos nos 3 *clusters* nas variáveis do solo MO, Mg, AzT, B e na variável aromática do mosto $FL1M$; a um nível de significância mais elevado, $0,1$, passa a haver evidência estatística para rejeitar também a igualdade de valores medianos nas variáveis do solo FF, FG, $P2O5$, Ni, Cr, CTC e no $Ac.mal$ do mosto. Quanto às variáveis do solo DA, $K2O$, Cd, N e pH apresentam concentrações idênticas em todos os *clusters* considerados.

Verificada a existência de diferenças entre os *clusters* interessa saber quais os *clusters* que diferem entre si. Para este efeito aplicou-se o teste não paramétrico de comparações múltiplas LSD. Os resultados obtidos encontram-se na Tabela 4.3.

Feita uma análise aos resultados obtidos, a um nível de significância de 0,05, há evidência estatística para afirmar que, relativamente à(s) variável(eis):

Tabela 4.3: Resultados do teste LSD para as variáveis em estudo.

Variável	*cluster*	*cluster*	Valor-prova
MO	1	2	0,125
		3	**0,015****
	2	3	**0,004****
Mg	1	2	**0,001****
		3	**0,005****
	2	3	**0,044****
AzT	1	2	0,125
		3	**0,015****
	2	3	**0,004****
B	1	2	0,119
		3	**0,014****
	2	3	**0,004****
$FL1M$	1	2	**0,024****
		3	**0,006****
	2	3	0,574
FF	1	2	1,000
		3	**0,017****
	2	3	**0,034****
FG	1	2	1,000
		3	**0,017****
	2	3	**0,034****
$P2O5$	1	2	**0,048****
		3	**0,041****
	2	3	0,859
Cr	1	2	0,116
		3	**0,010****
	2	3	0,219
CTC	1	2	**0,014****
		3	**0,050****
	2	3	0,261
Ni	1	2	**0,001****
		3	**0,005****
	2	3	**0,044****
$Ac.mal$	1	2	0,513
		3	**0,032****
	2	3	0,146

**significativo a 5%

- MO, AzT, B, FF e FG o *cluster* 3 difere dos *clusters* 1 e 2 apresentando valores medianos superiores para MO, AzT, B e FF, e valor mediano inferior para FG;
- Mg, os 3 *clusters* diferem uns dos outros, apresentando maior valor mediano o grupo 1 e menor valor o grupo 2;
- Ni, os 3 *clusters* também diferem todos, apresentando maior valor mediano o grupo 3 e menor

valor o grupo 1;

- FL1M, P2O5 e CTC, o *cluster* 1 difere dos *clusters* 2 e 3 apresentando um valor mediano inferior relativamente à variável FL1M e superior em relação a P2O5 e CTC;
- Ac.mal e Cr, o *cluster* 3 difere do *cluster* 1, apresentando maior valor mediano.

Tendo como objetivo perceber qual a influência de cada um dos constituintes do solo na qualidade do vinho, determinaram-se os valores das correlações entre estas variáveis. De seguida para testar se estas associações eram estatisticamente significativas, realizou-se o teste baseado no coeficiente de correlação não paramétrico de *Spearman*, evitando assim nesta análise o uso do teste de independência de *Pearson*, dado que as dimensões das amostras são demasiado pequenas e, consequentemente, não se consegue provar a multinormalidade dos dados.

Na Tabela 4.4 encontram-se os valores das correlações entre as variáveis, significativamente diferentes de zero, ou seja, em que a hipótese nula do teste foi rejeitada a um nível de significância de $0,05$. Analisando os resultados obtidos, observa-se que:

- a variável DA apresenta uma forte correlação negativa com um ácido encontrado no mosto, $Ac.tart$, concluindo-se, assim, que quanto maior for a densidade aparente no solo menor será a concentração de ácido tartárico. Não havendo diferenças significativas na concentração de DA nos diferentes *clusters*, a sua influência no $Ac.tart$ nos diferentes *clusters* é semelhante;
- a matéria orgânica (MO) e o azoto total (AzT) existentes no solo estão correlacionadas positivamente com $FL3M$ e negativamente com a nota final atribuída ao vinho, o que significa que aumentando a concentração de MO e de AzT irão aumentar os compostos das famílias $FL3M$ e piorar a pontuação atribuída ao vinho pelos provadores. É no *cluster* 3 que se verifica maior concentração destes constituintes do solo e portanto onde se obtém o vinho com pior classificação e mostos com maior quantidade de compostos de $FL3M$;
- o aumento do pH do solo e da percentagem de FG faz diminuir os compostos da família $FG6M$ e aumentar a nota final atribuída ao vinho. É no *cluster* 3 que se encontram as menores percentagens de FG (e maiores de FF), sendo por isso responsáveis pela existência nestas parcelas de terreno a maior concentração de compostos voláteis de aroma da família $FG3M$ e pela atribuição da pior pontuação ao vinho proveniente desta parcela;

Tabela 4.4: Correlações estatisticamente significativas, ao nível de significância de $0,05$, entre as variáveis originais do solo e as variáveis relacionadas com a videira, o mosto e a qualidade do vinho.

			DA	MO	pH	FF	FG	$P2O5$	AzT	Ni	Cr	Cd	N	B	CTC
VIDEIRA	$Ncachos$	Corr						0,800					0,700		
		Valor-prova						0,010					0,036		
	$Pbruto$	Corr						0,867					0,817		
		Valor-prova						0,002					0,007		
	$Pcacho$	Corr										0,800			
		Valor-prova										0,010			
MOSTO	$Ac.tart$	Corr	-0,800									0,817			
		Valor-prova	0,010									0,007			
	$Ac.TOT$	Corr										0,783			
		Valor-prova										0,013			
	$FL1M$	Corr								0,733	-0,733				
		Valor-prova								0,025	0,025				
	$FL3M$	Corr		0,700					0,667			-0,767		0,686	
		Valor-prova		0,036					0,049			0,016		0,041	
	$FL5M$	Corr									0,717				
		Valor-prova									0,030				
	$FG3M$	Corr										-0,867			
		Valor-prova										0,002			
	$FG4M$	Corr										-0,700			
		Valor-prova										0,036			
	$FG5M$	Corr										-0,767			
		Valor-prova										0,016			
	$FG6M$	Corr			-0,833	0,783	-0,783								
		Valor-prova			0,005	0,013	0,013								
	$FG7M$	Corr								-0,683	0,733				
		Valor-prova								0,042	0,025				
	$Vermelho\_M$	Corr								0,683	-0,783				
		Valor-prova								0,042	0,013				
	$Azul\_M$	Corr									0,750				
		Valor-prova									0,020				
VINHO	$Nota.Final$	Corr		-0,683	0,717	-0,750	0,750		-0,717						-0,703
		Valor-prova		0,042	0,030	0,020	0,020		0,030						0,035

- existe uma forte ligação entre as variáveis $P2O5$, N e com a produtividade das videiras; o aumento do fósforo e de nitratos leva a um aumento significativo do número de cachos por videira e do peso bruto de cachos por videira. É no *cluster* 1 que existem as maiores concentrações de $P2O5$, pelo que haverá neste terreno maior produtividade das videiras responsável por este elemento;

- Ni e Cr são dois constituintes que influenciam alguns compostos aromáticos e cor do mosto de forma contrária: enquanto que o aumento de níquel no solo faz aumentar $FL1M$, diminuir $FG7M$ e aumentar a percentagem de cor vermelha no mosto, Cr tem um efeito contrário nestas três variáveis. Comparando o *cluster* 3 com o *cluster* 1, é o *cluster* 3 que possui maiores concentrações de Ni e menores concentrações de Cr, pelo que será nesta parcela que se obtêm maiores concentrações de compostos da família $FL1M$, menores concentrações de compostos da família $FG7M$ e uma maior percentagem de cor vermelha no mosto;

- o cádmio (Cd) é um constituinte que está correlacionado positivamente com variáveis ligadas à acidez do mosto e ao peso médio de cacho por videira e negativamente correlacionado com alguns compostos aromáticos do mosto, ou seja, as concentrações elevadas de cádmio no solo corresponderão a altas concentrações de ácido tartárico e de ácido total no mosto bem como aumento de $Pcacho$ e corresponderão a baixas concentrações dos compostos da família $FL3M$, $FG3M$, $FG4M$ e $FG5M$ no mosto. Como observado na Tabela 4.2, não existem diferenças significativas relativamente à concentração deste elemento químico do solo nos três *clusters* considerados, concluindo-se assim que a influência destes elementos do solo nestas variáveis referentes à qualidade do vinho nos 3 *clusters* é idêntica;

- o boro (B) correlaciona-se positivamente com as variável $FL3M$, concluindo-se que para valores elevados deste constituinte do solo corresponderão altas concentrações dos compostos das famílias $FL3M$; o que se verifica no *cluster* 3;

- o aumento da capacidade de troca catiónica (CTC) leva a um aumento significativo na nota final ao vinho. Verificada a existência de maior concentração de CTC no *cluster* 1, conclui-se que os provadores de vinho atribuíram melhor nota ao vinho proveniente do *cluster* 1, ou seja, o *cluster* 1 produz melhor vinho que o *cluster* 2 e 3, na opinião dos provadores;

- as variáveis $K2O$, Ca e Mg não apresentam qualquer tipo de correlação com as variáveis responsáveis pela qualidade do vinho.

Agora, vai-se apresentar a Análise de *Clusters* realizada a partir dos fatores retidos resultantes de uma Análise Fatorial (AF).

Com o objetivo de transformar um problema que envolve um elevado número de variáveis originais e correlacionadas num problema com um número reduzido de variáveis, efetuou-se uma Análise Fatorial; com esta análise pretendeu-se verificar se existia um pequeno número de variáveis que fossem responsáveis por explicar grande parte da variabilidade total associada aos dados originais relativos ao solo e, portanto, reduzir a dimensionalidade do problema.

Para dar início à realização de uma AF, analisaram-se as relações existentes entre os diferentes constituintes do solo em estudo e, para tal, calcularam-se os valores das correlações de *Spearman* entre cada par de constituintes. Estas correlações encontram-se na Tabela 4.5. Como referido anteriormente, nestes cálculos participam todos os *outliers* existentes nos dados.

Tabela 4.5: Matriz de correlações das variáveis do solo do Minho.

	DA	MO	pH	FF	FG	$P2O5$	$K2O$	Ca	Mg	AzT	Ni	Cr	Cd	N	B	CTC
DA	1,00															
MO	-0,17	1,00														
pH	-0,04	-0,27	1,00													
FF	0,18	0,40	-0,42	1,00												
FG	-0,18	-0,40	0,42	***-1,00***	1,00											
$P2O5$	-0,03	-0,42	0,44	-0,26	0,26	1,00										
$K2O$	-0,19	0,14	0,09	0,18	-0,18	0,17	1,00									
Ca	-0,10	0,05	***0,76***	-0,12	0,12	0,36	0,19	1,00								
Mg	0,10	-0,20	0,42	-0,08	0,08	***0,50***	0,20	***0,72***	1,00							
AzT	-0,19	***0,99***	-0,16	0,35	-0,35	-0,38	0,15	0,15	-0,15	1,00						
Ni	-0,08	***0,53***	-0,23	0,46	-0,46	-0,31	0,03	-0,10	-0,28	***0,51***	1,00					
Cr	-0,04	-0,25	0,17	-0,43	0,43	0,01	-0,18	-0,01	0,01	-0,24	***-0.52***	1.00				
Cd	-0,21	-0,13	0,25	-0,05	0,05	0,42	0,24	0,25	0,29	-0,11	-0,10	-0,00	1,00			
N	-0,14	0,23	0,22	0,03	-0,03	0,34	0,06	***0,52***	0,38	0,26	-0,11	0,05	0,27	1,00		
B	-0,14	***0,98***	-0,26	0,44	-0,44	-0,38	0,13	0,06	-0,17	***0,98***	***0,52***	-0,26	-0,10	0,25	1,00	
CTC	-0,09	0,03	***0,74***	-0,11	0,11	0,39	0,26	***0,99***	***0,77***	0,13	-0,12	-0,02	0,27	***0,52***	0,05	1,00

Da Tabela 4.5 verifica-se que algumas variáveis do solo do Minho possuem uma correlação positiva quase perfeita ($r = 0,99/0,98$), como é o caso de:

- MO com AzT; MO com B; Ca com CTC; AzT com B.

Encontram-se também variáveis que estão fortemente correlacionadas positivamente, tais como:

- pH com Ca; pH com CTC; Mg com Ca; Mg com CTC.

No que se refere às variáveis FF e FG, estas têm uma correlação negativa perfeita; isto deve-se ao facto de ambas darem informações sobre a granulometria do terreno, sendo os seus valores complementares. Por este motivo a variável FG foi eliminada da análise.
Verifica-se ainda uma correlação moderadamente positiva entre as variáveis:

- Ni e MO; Ni e AzT; Ni e B; N e CTC; N e Ca; Mg e $P2O5$.

Existe também uma correlação moderadamente negativa entre as variáveis Ni e Cr.

Conclui-se que a análise da matriz de correlações revela a existência de variáveis que se encontram fortemente correlacionadas (num caso de uma forma mesmo perfeita).

Analisou-se a matriz de correlações das variáveis do solo e conclui-se que não é definida positiva. Usualmente, quando a matriz não possui esta propriedade é porque existem variáveis aleatórias demasiadamente correlacionadas ou é porque o número de variáveis aleatórias originais é elevado relativamente ao número de observações. Para contornar tal situação, realizou-se de novo a AF sem a variável Ca visto ser uma das variáveis que quando excluída desta análise torna a matriz de correlação definida positiva. Além disso, esta variável está fortemente correlacionada com a variável CTC, ($r = 0,99$) e sendo considerada menos importante no estudo, foi eliminada.

Uma AF só faz sentido se existir uma associação significativa entre as variáveis originais; o facto do determinante da matriz de correlações ser muito próximo de zero, $1,312 \times 10^{-7}$, é um indicador de que as variáveis revelam suficiente grau de associação entre si. Esta adequação desta técnica aos dados em estudo também foi avaliada a partir do teste de Esfericidade de *Bartlett* e da estatística de KMO, apresentados na Tabela 4.6.

Tabela 4.6: Resultados do teste de esfericidade de *Bartlett* e da estatística de KMO.

Estatística de KMO (*Kaiser-Meyer-Olkin*)		0,618
Teste de esfericidade de *Bartlett*	Estatística de Teste	610,092
	Graus de Liberdade	91
	Valor-prova	< 0,001

Apesar da estatística de KMO não ser muito elevada, conclui-se, mesmo assim, que há uma correlação razoável entre as variáveis; para além disso, como no teste de esfericidade de *Bartlett* se obteve um valor de prova próximo de 0, rejeitou-se a hipótese nula, isto é, há evidência estatística de que existe correlação significativa entre as variáveis e, como tal, é apropriada a aplicação da AF.

Confirmada a existência da correlação dos dados, passou-se à derivação dos fatores. Neste sentido, aplicou-se a AF usando o método de Componentes Principais, e a rotação *varimax* para facilitar a interpretabilidade dos fatores. A Tabela 4.7 apresenta os *loadings* - valores da correlação entre os elementos em estudo, os fatores extraídos e o respetivo valor da variância explicada por cada um deles; apresenta também os valores das comunalidades das variáveis em estudo.

Feita uma análise ao *scree plot*[4] e aos dados da Tabela 4.7, optou-se por reter os 4 primeiros fatores uma vez que estes permitem explicar aproximadamente 72% da variabilidade total dos dados; além disso, esta opção também foi tomada seguindo o critério de Kaiser aplicado à matriz de correlação dos dados, visto que os valores próprios são superiores a 1 até ao 4^{o} fator.

[4] O *screeplot* encontra-se no Anexo IV.

Tabela 4.7: *Loadings* dos 4 primeiros fatores com rotação *Varimax* e comunalidades das variáveis em estudo; proporção de variância explicada por cada fator.

	Fator				Comunalidades
	1	2	3	4	
DA	-0,268	0,121	0,307	-0,770	0,774
MO	0,961	-0,051	0,206	0,029	0,969
pH	-0,169	0,686	-0,297	0,109	0,599
FF	0,288	-0,101	0,756	-0,122	0,679
$P2O5$	-0,457	0,574	-0,003	0,321	0,641
$K2O$	0,036	0,196	0,352	0,566	**0,484**
Mg	-0,191	0,844	0,090	-0,003	0,757
AzT	0,968	0,034	0,165	0,042	0,967
Ni	0,462	-0,246	0,563	0,065	0,594
Cr	-0,093	0,020	-0,801	-0,109	0,662
Cd	-0,183	0,327	0,074	0,631	**0,544**
N	0,332	0,645	-0,110	0,114	**0,552**
B	0,946	-0,012	0,233	0,016	0,950
CTC	0,113	0,920	-0,028	0,116	0,874
Valor próprio	3,563	3,011	1,986	1,487	
Proporção de variância (%)	25,451	21,509	14,183	10,619	
Proporção acumulada (%)	25,451	46,960	61,143	71,763	

Um dos pontos fundamentais da AF é a interpretação dos fatores, obtida através da análise dos valores dos *loadings* das variáveis latentes. Assim, uma vez que havia uma melhor interpretabilidade quando escolhidos 4 fatores em vez de 5, decidiu-se que o mais adequado seria reter 4 fatores.

Analisando os valores dos *loadings* das variáveis latentes, ou seja, avaliando o peso que cada variável original tem em cada um dos fatores extraídos, verifica-se que:

- o 1^o fator está fortemente correlacionado positivamente com os constituintes MO, AzT, e B;
- o 2^o fator apresenta uma correlação positiva elevada com os constituintes pH, $P2O5$, Mg, N e CTC;
- o 3^o fator apresenta valores positivos de correlações mais elevados para as variáveis FF e Ni e um valor negativo mais elevado para Cr;
- o 4^o fator é influenciado positivamente por $K2O$ e Cd e negativamente por DA.

Observando a comunalidade associada a cada variável original, ou seja, a proporção da variância de cada variável explicada pelos fatores retidos, verifica-se que $K2O$, N e Cd são as que mais informação perdem quando utilizadas para a formação dos fatores; estes quatro fatores apenas conseguem explicar $48,4\%$, $55,2\%$ e $54,4\%$ da variância destas variáveis, respetivamente.

Observando os *scores* de cada fator extraídos através do método dos mínimos quadrados ponderados, Tabela 4.8, observa-se que os pontos de amostragem do solo $A14$, $A34$, $A35$, $A38$ e

$A43$ são os que apresentam maiores *scores* para o 1^o fator. Assim, é nestes pontos que as variáveis MO, Azt e B apresentam maiores concentrações uma vez que são estas as variáveis que estão fortemente correlacionadas positivamente com o 1^o fator; $A11$, $A25$ e $A28$ são os pontos que apresentam menores *scores* e, consequentemente, os que apresentam menores concentrações destes quatro constituintes do solo.

Relativamente ao 2^o fator, os *loadings* mais elevados encontram-se nas variáveis pH, $P2O5$, Mg, N e CTC pelo que as concentrações destes constituintes são mais altas nos pontos de amostragem $A11$, $A12$, $A15$, $A16$, $A18$, $A21$, $A24$, $A33$, $A43$ e $A45$, por apresentarem os *scores* mais elevados, e mais pequenas em $A19$, $A22$, $A26$, $A36$, $A39$, $A40$ e $A41$ (*scores* mais pequenos).

Tabela 4.8: *Scores* dos quatro primeiros fatores.

Terreno	Fator 1	Fator 2	Fator 3	Fator 4	Terreno	Fator 1	Fator 2	Fator 3	Fator 4
A1	-0,219	0,710	-2,030	0,166	A24	0,361	1,028	0,432	1,733
A2	-0,302	0,403	0,024	-0,325	A25	-1,129	-0,983	0,683	-0,404
A3	-0,367	-0,675	-1,621	1,930	A26	-0,530	-1,086	0,120	-0,186
A4	-0,584	0,887	-0,152	0,715	A27	-0,497	-0,220	-1,240	-0,790
A5	-0,429	0,106	-0,273	-1,023	A28	-1,097	-0,215	1,489	2,780
A6	-0,003	-0,644	-0,957	-0,691	A29	-0,995	-0,753	0,403	1,810
A7	-0,560	0,485	-0,776	0,000	A30	-0,214	-0,223	-0,747	1,032
A8	-0,393	0,451	-1,131	-0,331	A31	-0,452	0,519	2,118	1,551
A9	-0,974	-0,761	0,394	-0,813	A32	-0,165	-0,921	1,158	-1,480
A10	-0,720	0,998	-0,712	-1,023	A33	-0,500	1,371	1,041	-0,397
A11	-1,255	1,822	1,012	-0,550	A34	1,119	-0,153	1,073	-0,329
A12	0,634	1,049	-0,162	-0,521	A35	2,293	-0,878	1,972	-0,327
A13	-0,612	0,383	-0,472	1,271	A36	0,692	-1,427	1,136	-0,852
A14	1,043	-0,066	-1,380	0,005	A37	-0,402	-0,065	0,641	-0,441
A15	-0,289	1,163	-0,497	0,348	A38	2,050	0,028	0,071	0,733
A16	-0,157	1,028	-1,673	0,717	A39	-0,524	-1,166	1,594	0,163
A17	0,373	-0,928	-1,410	-0,683	A40	0,744	-1,230	-0,307	1,525
A18	-0,097	1,918	0,743	-0,363	A41	0,952	-1,669	-0,290	0,736
A19	-0,568	-1,339	-0,806	-0,817	A42	0,475	0,161	0,139	-1,384
A20	-0,349	0,132	-0,586	-0,867	A43	4,124	1,003	0,684	0,387
A21	-0,408	1,926	0,640	-0,296	A44	0,614	-0,842	0,310	-0,297
A22	-0,797	-1,781	-0,279	0,270	A45	0,752	1,273	1,149	-0,934
A23	-0,639	-0,820	-0,157	-1,745					

O 3^o fator encontra-se mais correlacionado positivamente com as variáveis FF e Ni e negativamente com Cr, pelo que se pode afirmar que são nos pontos $A11$, $A28$, $A31$, $A32$, $A33$, $A34$, $A35$, $A36$, $A39$ e $A45$ (os que apresentam maiores *scores*) que existem maior concentração de FF e Ni e menor concentração de Cr; para esta fator os *scores* mais baixos observam-se em $A1$, $A2$, $A8$, $A14$, $A16$, $A17$ e $A27$ e, como tal, são nestes pontos que se encontram as menores concentrações de FF e Ni e maiores concentrações de Cr.

Quanto ao 4^o fator, este apresenta *scores* mais elevados nos pontos $A3$, $A13$, $A24$, $A28$, $A29$, $A30$, $A31$ e $A40$ o que indica que são nestes pontos que as concentrações de $K2O$ e Cd são mais elevadas (*loadings* positivos) tendo a concentração mais pequena de DA (*loading* negativo). No

que diz respeito aos pontos $A5$, $A10$, $A23$, $A32$ e $A42$ (apresentam *scores* mais baixos) são os que apresentam menores concentrações de $K2O$ e Cd e maiores concentrações de DA.

Encontrados assim os 4 fatores extraídos a partir da AF e os respetivos *scores*, procedeu-se, de seguida, a uma Análise de *Clusters* com o objetivo de identificar na parcela em estudo zonas com características semelhantes e dissemelhantes ao nível das variáveis do solo. Segundo Jollife (2002), a Análise de *Clusters* é uma das técnicas multivariadas onde à priori se realiza uma redução de dimensionalidade com mais frequência. Note-se que, tal como aconteceu no estudo realizado com as variáveis originais, a partir do cálculo de médias amostrais, os 45 *scores* de cada fator foram reduzidos a 9, que representarão as diversas quantidades observadas de cada fator nas parcelas $B1N$, $B1C$, $B1P$, $B2N$, $B2C$, $B2P$, $B3N$, $B3C$ e $B3P$.

Na aplicação da Análise de *Clusters*, a utilização da distância euclidiana e a aplicação de vários métodos aglomerativos verificou-se que três deles, nomeadamente, a ligação completa, a ligação média e de *Ward* conduziram a composições de *clusters* semelhantes, concluindo-se que os dados revelaram uma estrutura de grupos com base nestes três métodos. Com o objectivo de minimizar a perda de informação em cada passo do processo aglomerativo e, perante a elevada correlação cofenética apresentada, um valor de aproximadamente de $0,85$, utilizou-se o método de *Ward* para a identificação dos grupos. O resultado dos métodos da ligação simples e do centróide distanciaram-se muito dos outros.

Após a formação dos *clusters* a partir do método escolhido, torna-se necessário decidir quantos grupos reter. Pela observação do gráfico da Figura 4.4, constata-se que ocorre um aumento significativo na distância de junção após a etapa 5, pelo que se consideraram um total de quatro grupos.

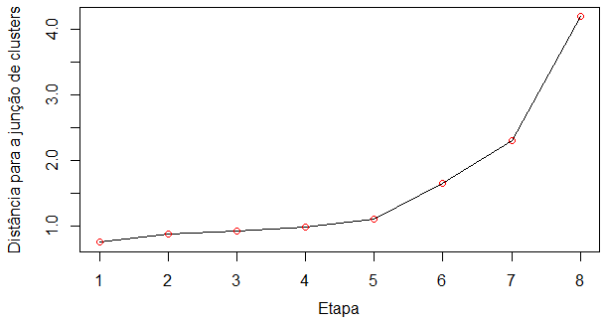


Figura 4.4: Critério para obter o número de *clusters*.

Representado o dendrograma obtido de acordo com as novas variáveis retidas na AF, Figura 4.5, este sugere a escolha dos mesmos 3 *clusters* obtidos no estudo anterior, apesar de também mostrar evidências que a parcela $B3N$ poderia formar um só grupo (o método de ligação média vincava bem o isolamento desta parcela, como mostra a Figura 4.6). A Figura 4.7 apresenta o mapeamento dos grupos obtidos resultantes da utilização destes dois métodos.

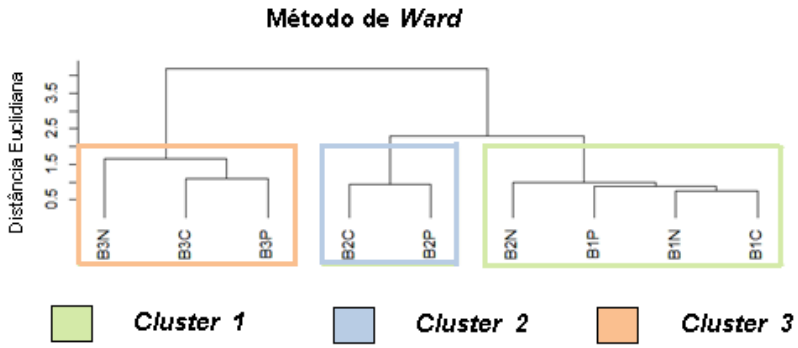


Figura 4.5: Dendrograma resultante do método de Ward assinalado com o número de *clusters* a considerar.

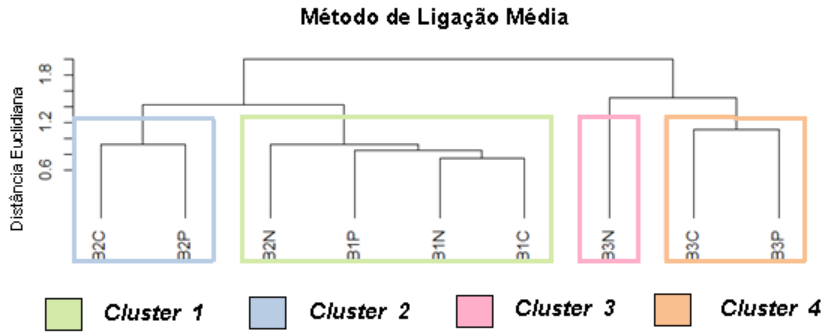


Figura 4.6: Dendrograma resultante do método de ligação média assinalado com o número de *clusters* a considerar.

Mas de acordo com a opinião dos peritos nas áreas em estudo e pelo facto de no estudo anterior os resultados terem sido idênticos, optou-se por reter os mesmos 3 *clusters*[7].

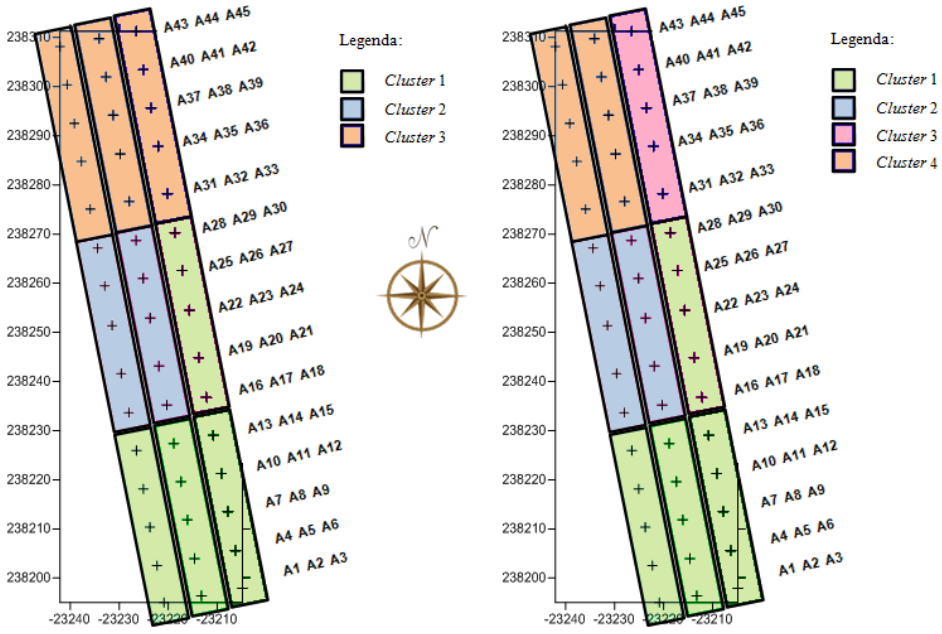


Figura 4.7: Distribuição dos 3 grupos na parcela em estudo de acordo com os resultados obtidos a partir do método de *Ward* e método de ligação média.

[7] Caso se retesse este 4^{o} grupo, este não podira ser considerado nos testes, quando se pretendesse verificar se há diferenças entre os grupos, devido a ser constiuído por uma única observação.

A Tabela 4.9 apresenta um resumo de algumas das principais características amostrais relativas a cada um dos fatores retidos, para cada um dos *clusters* resultantes da utilização do método de ligação *Ward*.

Analisando os resultados obtidos, observa-se que:

- no *cluster* 1, o Fator 2 e o Fator 4 apresentam o valor mediano mais elevado, sendo que o Fator 3 é o que tem o valor mediano mais baixo;
- no *cluster* 2, o Fator 1 e o Fator 2 têm o valor mediano mais baixo;
- no *cluster* 3, o Fator 1 e o Fator 3 mostram o valor mediano mais elevado, sendo que o Fator 4 é o que apresenta o valor mediano mais baixo.

Repare-se que no *cluster* 2 não existem fatores cujo valor mediano seja o mais elevado, mas metade das variáveis têm o menor valor mediano. Estas conclusões vão ao encontro das retiradas no estudo anterior.

Tabela 4.9: Caracterização dos 3 *clusters* retidos de acordo com os fatores resuntantes da AF.

	Fator 1	Fator 2	Fator 3	Fator 4
***CLUSTER* 1**				
Mínimo	-0,539	0,027	-0,828	-0,445
Máximo	-0,171	0,886	-0,035	0,263
Mediana	-0,233	0,618	-0,459	0,138
Média	-0,294	0,537	-0,445	0,024
Desvio Padrão	0,168	0,368	0,337	0,326
1^o Quartil	-0,335	0,414	-0,633	-0,074
3^o Quartil	-0,193	0,741	-0,271	0,235
***CLUSTER* 2**				
Mínimo	-0,749	-0,691	-0,326	-0,334
Máximo	-0,428	-0,658	-0,117	0,509
Mediana	-0,589	-0,675	-0,222	0,087
Média	-0,589	-0,675	-0,222	0,087
Desvio Padrão	0,227	0,023	0,148	0,596
1^o Quartil	-0,669	-0,683	-0,274	-0,124
3^o Quartil	-0,508	-0,666	-0,169	0,298
***CLUSTER* 3**				
Mínimo	0,179	-0,856	0,568	-0,681
Máximo	1,149	0,042	1,012	0,539
Mediana	1,026	0,015	0,644	-0,127
Média	0,785	-0,266	0,741	-0,090
Desvio Padrão	0,528	0,511	0,237	0,611
1^o Quartil	0,603	-0,421	0,606	-0,404
3^o Quartil	1,088	0,029	0,828	0,206

Observando os resultados obtidos, também se pode verificar que:

- no *cluster* 1, o Fator 3 apresenta o maior desvio padrão e os Fatores 1 e 4 possuem o mais baixo;
- no *cluster* 2, os Fatores 2 e 3 apresentam o desvio padrão mais baixo;
- no *cluster* 3, os Fatores 1, 2 e 4 possuem o desvio padrão maior.

Conhecido o número de *clusters*, recorreu-se ao teste não paramétrico de *Kruskal-Wallis* para verificar se existem diferenças significativas entre eles relativamente às concentrações dos constituintes do solo, agora representadas pelos fatores, e às variáveis que traduzem a produtividade da videira e da qualidade do vinho. Os resultados significativos a um nível de significância de $0,05$ e $0,1$ resultantes da aplicação deste teste, encontram-se na Tabela 4.10. Também se pode encontrar nestas tabelas os valores medianos de cada um dos grupos para cada variável.

Tabela 4.10: Resultados significativos do teste de *Kruskal-Wallis* para as variáveis em estudo.

	Medianas			Estatística	Graus de	*Valor-prova*
Variáveis	*Cluster* 1	*Cluster* 2	*Cluster* 3	de teste	liberdade	
Fator 1	-0,233	-0,589	1,026	6,300	2	**0,043****
Fator 2	0,618	-0,675	0,015	5,078	2	0,079*
Fator 3	-0,459	-0,222	0,644	5,800	2	0,055*
Ac.mal	1,745	1,975	2,630	4,544	2	0,083*
FL1M	707,035	768,430	788,050	6,111	2	**0,047****

*significativo a 10% **significativo a 5%

Da análise dos resultados da Tabela 4.10, conclui-se que há evidência estatística, a um nível de significância $0,05$, para rejeitar a igualdade de valores medianos nos 3 *clusters* no Fator 1 (MO, AzT e B) e na variável aromática do mosto $FL1M$; a um nível de significância mais elevado, $0,1$, passa a haver evidência estatística para rejeitar também a igualdade de valores medianos no Fator 2 (pH, $P2O5$, Mg, N e CTC), no Fator 3 (FF, Ni e Cr) e no $Ac.mal$ do mosto. O Fator 4 não revela quaisquer diferenças nos diversos *clusters*.

Analisando a Tabela 4.11, verifica-se que estes resultados estão de acordo com os encontrados no estudo anterior, à exceção do constituinte Mg que, correlacionado fortemente com o Fator 2, passou apenas a ser significativo a um nível de significância de $0,1$ (no estudo anterior era significativo para um nível de confiança de 5%). O Fator 4 (DA, $K2O$ e Cd) não apresenta diferenças significativas, não contrariando, assim, os resultados obtidos no estudo anterior, uma vez que estes elementos do solo também não diferiam significativamente no que diz respeito ao valor mediano entre os 3 *clusters*. As variáveis N e pH, estando correlacionadas com o Fator 2, passaram a ter significância estatística em relação à diferença das sua concentrações nos 3 *clusters* a um nível de significância de 10%.

Tabela 4.11: Comparação dos resultados do teste de *Kruskal-Wallis* entre os dois estudos.

Estudos	Estatisticamente significativo a 5%	Estatisticamente significativo a 10%	Estatísticamente não significativo
Variáveis originais	$MO\ AzT\ B\ Mg$ $FL1M$	$P2O5\ CTC$ $FF\ Ni\ Cr$ $Ac.mal$	$DA\ K2O\ Cd$ $N\ pH$
Scores	$MO\ AzT\ B$ (Fator 1) $FL1M$	$P2O5\ CTC$ **Mg pH N** (Fator 2) $FF\ Ni\ Cr$ (Fator 3) $Ac.mal$	$DA\ K2O\ Cd$ (Fator 4)

Verificada a existência de diferenças entre os *clusters*, interessa agora saber quais os *clusters* que diferem entre si. Para este efeito aplicou-se o teste não paramétrico de comparações múltiplas LSD. Os resultados obtidos encontram-se na Tabela 4.12.

Tabela 4.12: Resultados do teste LSD para as variáveis que apresentaram diferentes valores medianos em pelo menos um dos *clusters*.

Variável	*cluster*	*cluster*	*p-value*
Fator 1	1	2	0,125
		3	**0,015****
	2	3	**0,004****
Fator 2	1	2	**0,028****
		3	**0,049****
	2	3	0,529
Fator 3	1	2	0,337
		3	**0,008****
	2	3	**0,050****
FL1	1	2	**0,024****
		3	**0,006****
	2	3	0,574
Ac.mal	1	2	0,513
		3	**0,032****
	2	3	0,146

**significativo a 5%

A análise da Tabela 4.12 mostra, a um nível de significância de $0,05$, que relativamente:

- ao Fator 1, o *cluster* 3 difere dos *clusters* 1 e 2 apresentando um valor mediano superior;
- ao Fator 2, o *cluster* 1 difere dos *clusters* 2 e 3 apresentando um valor mediano superior;
- ao Fator 3, o *cluster* 3 difere dos *clusters* 1 e 2 apresentando um valor mediano superior;
- à variável FL1, o *cluster* 1 difere dos *clusters* 2 e 3 apresentando um valor mediano inferior;
- à variável Ac.mal, apenas existem diferenças entre o *cluster* 3 e *cluster* 1, apresentando maior valor mediano o *cluster* 3.

Observe-se que estes resultados são, na sua maioria, semelhantes aos discutidos no estudo anterior.

Para perceber qual a influência de cada um dos fatores na qualidade do vinho, determinaram-se os valores das correlações entre estas variáveis e testaram-se a sua significância, realizando-se para tal efeito o teste baseado no coeficiente de correlação não paramétrico de *Spearman*. Este teste foi aplicado devido às dimensões das amostras serem demasiado pequenas e, consequentemente, não se conseguir provar a multinormalidade dos dados.

Na Tabela 4.13 encontram-se os resultados das correlações entre as variáveis que são significativamente diferentes de zero, ou seja, em que a hipótese nula foi rejeitada a um nível de significância de $0,1$, havendo situações que foram rejeitadas apenas considerando como nível de significância $0,5$.

Com base nos resultados obtidos:

Tabela 4.13: Correlações estatisticamente significativas, ao nível de significância de $0,05$ e $0,1$, entre as variáveis originais do solo e as variáveis relacionadas com a videira, o mosto e a qualidade do vinho.

			Fator 1	Fator 2	Fator 3	Fator 4
VIDEIRA	$Pbruto$	Corr		0,600	-0,667	
		Valor-prova		0,088*	0,050**	
	$Pcacho$	Corr		0,650		0,617
		Valor-prova		0,058*		0,077*
MOSTO	$Ac.tart$	Corr				0,850
		Valor-prova				0,004**
	$Ac.TOT$	Corr				0,600
		Valor-prova				0,088*
	$FL1M$	Corr			0,750	
		Valor-prova			0,020**	
	$FL3M$	Corr	0,617			
		Valor-prova	0,077*			
	$FG3M$	Corr		-0,583		0,700
		Valor-prova		0,099*		0,036**
	$FG4M$	Corr		-0,633		
		Valor-prova		0,067*		
	$FG5M$	Corr		-0,633		
		Valor-prova		0,067*		
	$FG6M$	Corr	0,633			
		Valor-prova	0,067*			
	$FG7M$	Corr	-0,667		-0,867	
		Valor-prova	0,054*		0,002**	
	$Vermelho\_M$	Corr			0,600	
		Valor-prova			0,088*	
VINHO	$Nota.Final$	Corr	-0,783		-0,733	
		Valor-prova	0,013**		0,025**	

** significativo a 5% * significativo a 10%

- o Fator 1 (que apresenta *loadings* positivos muito elevados para as variáveis MO, AzT e B) apresenta uma correlação estatisticamente significativa e negativa com a variável $Nota.Final$ e, portanto, a elevadas concentrações destes constituintes do solo corresponderão a uma nota final baixa atribuída ao vinho. O *cluster* 3 é o que apresenta valores medianos maiores neste fator e, consequentemente, valores medianos maiores nas variáveis MO, AzT e B pelo que será nesta parcela que serão atribuídas as piores pontuações ao vinho;

- o Fator 2 não apresenta qualquer tipo de correlação estatisticamente significativa com as variáveis referentes à videira, à qualidade do mosto e do vinho, a um nível de significância de 0,05;

- o Fator 3 do solo, com *loadings* elevados positivos para FF e Ni, está correlacionado de forma significativa e positiva com $FL1M$ e, portanto, elevados valores destes constituintes do solo levam a elevadas concentrações de $FL1M$, o que acontece no *cluster* 3; a variável Cr tem *loading* pesado negativo neste Fator, pelo que baixas concentrações de Cr no solo corresponderão a altas concentrações de $FL1M$, o que também se verifica no *cluster* 3.

  O Fator 3 do solo também se correlaciona de forma significativa mas negativamente com as variáveis $Pbruto$, $FG7M$ e $Nota.Final$, ou seja, altos valores para as variáveis FF e Ni no solo corresponderão a um menor peso bruto por videira, uma menor concentração de $FG7M$ e a uma pior nota final atribuída ao vinho (o que se verifica no *cluster* 3) e, baixas concentrações de Cr levam a baixas concentrações destas variáveis o que também se verifica no *cluster* 3;

- o Fator 4 do solo apresenta *loadings* positivos para os constituintes $K2O$ e Cd e negativos para DA, está correlacionado de forma significativa e positiva com o $Ac.tart$ e $FG3M$; assim, grandes concentrações dos elementos $K2O$ e Cd levam a grandes concentrações deste ácido e dos compostos aromáticos da família $FG3M$ e, altas concentrações do elemento DA proporciona diminuição das concentrações de $Ac.tart$ e $FG3M$. Como observado na Tabela 4.11, não existem diferenças significativas relativamente à concentração do Fator 4 (e, consequentemente, deste elemento químico do solo) nos três *clusters* considerados, concluindo-se assim que a influência deste Fator (destes elementos do solo) nestas variáveis referentes à qualidade do vinho nos 3 *clusters* é idêntica.

Com base nos resultados obtidos, estes quatro fatores ainda se encontram correlacionados com outras variáveis do solo, mas apenas considerando-se a regra de decisão com um nível de significância de $0,10$:

- o Fator 1 do solo apresenta *loadings* positivos para os constituintes MO, AzT e B, está correlacionado de forma estatisticamente significativa e positiva com $FL3M$ e $FG6M$; assim,

grandes concentrações destes elementos do solo levam a grandes concentrações destas duas famílias, o que se verifica no *cluster* 3. Este fator também se encontra estatisticamente correlacionado negativamente com $FG7M$, o que indica que elevadas concentrações de MO, AzT e B corresponderão a baixas concentrações desta família de compostos aromáticos, o que ocorre também no *cluster* 3;

- o Fator 2 do solo (que apresenta *loadings* positivos para as variáveis pH, $P2O5$, Mg, N e CTC) correlaciona-se de forma significativa e positiva com as variáveis $Pbruto$ e $Pcacho$. Portanto, altas concentrações destes elementos do solo corresponderão a altos pesos brutos por videira e a altos pesos médios por cacho, o que se verifica no *cluster* 1. O Fator 2 do solo também se correlaciona de forma significativa mas negativa com as variáveis $FG3M$, $FG4M$ e $FG5M$, pelo que os valores elevados destes constituintes do solo corresponderão a baixas concentrações dos compostos das famílias $FG3M$, $FG4M$ e $FG5M$, verificando-se tal situação também no *cluster* 1;

- o Fator 3 do solo, relacionado positivamente com a variável FF, Ni e negativamente com a variável Cr, encontra-se correlacionado positivamente e de forma significativa com a variável $Vermelho\_M$: elevadas percentagens e concentrações de FF e Ni, respetivamente, vão proporcionar uma maior percentagem de cor vermelha no mosto e baixas concentrações de Cr, desencadeando também a existência de maior percentagem de cor vermelha no mosto; esta situação ocorre no *cluster* 3;

- o Fator 4 do solo (que apresenta *loadings* positivos para as variáveis $K2O$, Cd, e *loading* negativo para DA) encontra-se correlacionado positivamente e de forma significativa com a variável $Pcacho$ e Acidez Total: elevadas concentrações de $K2O$ e Cd vão levar a um maior peso médio de cacho por videira e a uma maior concentração de acidez total; baixas concentrações de DA vai desencadear também o aumento destas duas variáveis referentes à produtividade da videira e da acidez do mosto. Tais situações ocorrem no *cluster* 3 comparativamente aos restantes grupos.

4.2 Quinta Bodega Santiago Ruiz: Resultados e Discussão

À semelhança da análise efetuada aos dados da quinta do Minho, este estudo iniciou-se com o objetivo de reduzir o número de observações em algumas destas variáveis e, assim, obterem-se amostras com as mesmas dimensões e, posteriormente, se conseguir aplicar as metodologias estatísticas pretendidas. O número de observações correspondentes às variáveis do solo e às variáveis da videira sofreram uma redução para um total de 6 observações (igual dimensão das restantes variáveis em estudo), correspondendo cada uma delas a uma das 6 parcelas:

$BSR\,1-2$, $BSR\,3-4$, $BSR\,5-6$, $BSR\,7-8$, $BSR\,9-10$ e $BSR\,11-12$.

O processo de redução do número de observações das variáveis da videira (de 12 para 6 observações) baseou-se no cálculo das médias amostrais de cada par de observações ($BSR\ 1$ e $BSR\ 2$, ..., $BSR\ 11$ e $BSR\ 12$) correspondentes às duas videiras existentes em cada uma das 6 parcelas. Quanto à redução do número de observações das variáveis do solo (de 45 para 6 observações), esta foi efetuada também a partir do cálculo de médias amostrais dos pontos de amostragem do solo que se encontravam mais próximos de cada umas das 6 parcelas.

Resolvido este problema, procedeu-se a uma Análise de *Clusters*, realizada a partir das variáveis originais do solo estandardizadas e, paralelamente, a partir dos *scores* dos fatores obtidos através de uma Análise Fatorial, tendo sido os agrupamentos resultantes dos dois processos muito diferentes.

Após definido o número de *clusters* a reter, aplicou-se o teste de *Kruskal-Wallis*, tal como foi efetuado aos dados do Minho, e em ambos os processos não se obtiveram resultados estatisticamente significativos, o que indica que não haveria diferenças significativas nas variáveis em estudo nos diferentes agrupamentos. Igualmente, realizando o teste de correlação de *Spearman* verificou-se que a maioria das correlações entre as variáveis eram estatisticamente não significativas.

Face ao exposto destes resultados e constatando-se que estes poderiam estar muito enviesados devido ao reduzido número de observações existentes (6 observações correspondentes às parcelas $BSR\,1-2$, $BSR\,3-4$, $BSR\,5-6$, $BSR\,7-8$, $BSR\,9-10$ e $BSR\,11-12$) optou-se por se iniciar de novo o estudo, mas agora redimensionando o número de observações de todas as variáveis para um total de 12 (correspondentes às parcelas $BSR\ 1$ a $BSR\ 12$), com o objetivo de tornar o estudo o mais correto possível. Uma vez que as variáveis relativamente à produção das videiras foram medidas em 12 videiras (correspondentes às localizações $BSR\ 1$ a $BSR\ 12$, respetivamente), fazia todo o sentido duplicar cada uma das observações de cada variável relativas à qualidade do vinho (duplicar 6 observações), uma vez que cada observação foi obtida com base nas uvas provenientes de duas destas 12 videiras; com este procedimento está-se a atribuir as mesmas características, relativas à qualidade das uvas, do mosto e do vinho, para cada par de videiras.

No que diz respeito ao problema de como reduzir o número de observações correspondentes às variáveis do solo (transformar 45 observações em 12), vários métodos foram analisados, optando-se no final pelo método do vizinho mais próximo: para cada videira fez-se corresponder o ponto de amostragem mais próximo que foi utilizado para se analisarem as características do solo.

Feito o ajustamento do número de observações dos dados, realizou-se uma Análise de *Clusters* (AC) com o objetivo de identificar na parcela em estudo da Galiza, zonas com características semelhantes ao nível dos constituintes do solo.

Tal como anteriormente foi feito no caso da quinta do Minho, para esta análise foram utilizados os métodos hierárquicos com o algoritmo aglomerativo por parecerem os mais adequados. Diversos métodos foram utilizados (método de ligação completa, de ligação média, de ligação simples, método de *Ward* e do centróide) e no final os resultados foram comparados. Seguindo tal procedimento, utilizando a distância euclidiana, foram operados 5 métodos e, no final, selecionou-se a solução que pareceu mais adequada.

Uma vez que as agregações obtidas pelo método da ligação completa, da ligação média e de *Ward* não foram idênticas, apresentaram-se os resultados aos peritos nas áreas em estudo e, de acordo com a suas opiniões resultantes do conhecimento do terreno e das suas características, optou-se no final por escolher os *clusters* obtidos a partir do método *Ward*, apresentando um valor de correlação cofenética, aproximadamente, de $0,62$ (valor inferior aos obtidos a partir do restantes métodos). O resultado dos métodos da ligação simples e do centróide distanciaram-se muito dos outros.

Após a formação dos *clusters* a partir do método escolhido, torna-se necessário decidir quantos grupos reter.

A visualização do dendrograma, que é um método empírico frequentemente utilizado para sugerir quantos grupos se devem considerar, e a opinião dos peritos nas áreas em estudo pesou nesta tomada de decisão. Para além disso, esta decisão também se baseou no resultado obtido no valor da medida de proximidade entre dois grupos, juntos em cada etapa.

O número de grupos que se deveria considerar está indicado pelo gráfico da Figura 4.8, onde ocorre o primeiro aumento significativo na distância de junção após a etapa 8, pelo que se deveria considerar o número de grupos após a junção dos grupos nessa etapa, ou seja, 4 *clusters*.

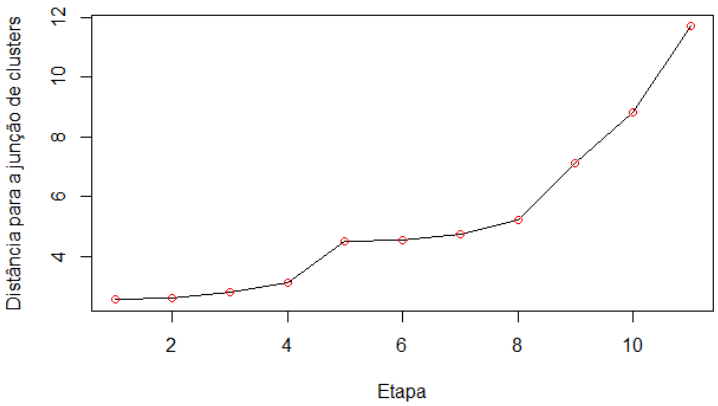


Figura 4.8: Critério para obter o número de *clusters*.

A Figura 4.9 ilustra o dendrograma obtido, assinalando-se os *clusters* considerados mais adequados neste terreno em estudo e a Figura 4.10 apresenta o mapeamento dos grupos obtidos.

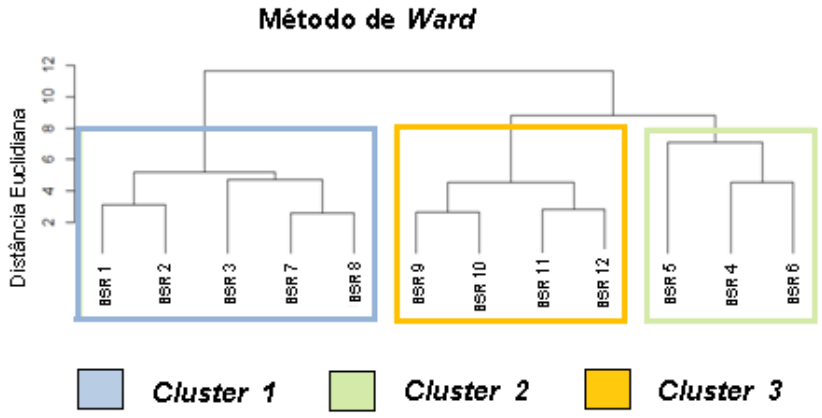


Figura 4.9: Dendrograma resultante do método de *Ward* assinalado com número de *clusters* a considerar.

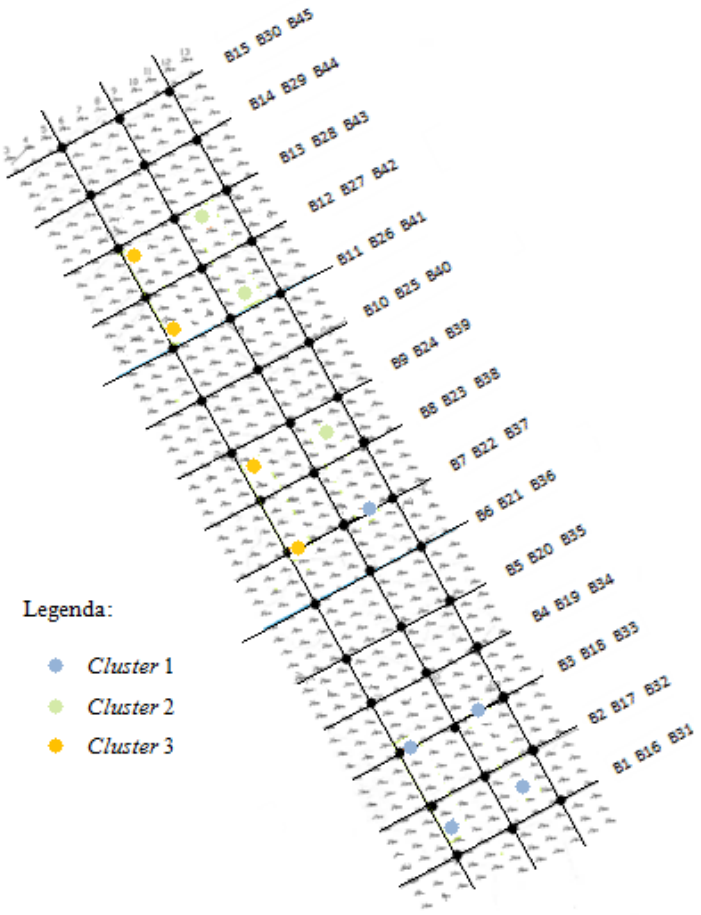


Figura 4.10: Distribuição dos 3 grupos na parcela em estudo de acordo com os resultados obtidos a partir do método de *Ward*.

Apesar dos critérios sugerirem a retenção de 4 *clusters* e o dendrograma mostrar que reter este número de *clusters* também seria adequado, apenas foram considerados 3 grupos.

Segundo o método *Ward*, o 4^{o} *cluster* a ser considerado continha apenas uma observação, $BSR\,5$; realizado um estudo a este elemento, verificou-se que este representa um *outlier* moderado nas variáveis $P2O5$, Ca, Mg e CTC. De acordo com os peritos nas áreas em estudo, não haveria necessidade de se criar um quarto *cluster* contendo esta única videira. Para além disso, caso se retivesse este 4^{o} grupo, não poderia ser considerado na aplicação de testes estatísticos, quando se pretendesse verificar se havia diferenças significativas entre os grupos, devido a este ser constituído por uma única observação. Face ao exposto, optou-se por não se fazer o isolamento desta videira dos restantes grupos, mesmo sendo sugerida tal divisão, também, pelos métodos da ligação completa e da ligação média.

Feita uma análise ao dendrograma, observa-se que o 1^{o} *cluster* é constituído pelas parcelas $BSR\,1$, $BSR\,2$, $BSR\,3$, $BSR\,7$ e $BSR\,8$, o 2^{o} *cluster* é formado pelas parcelas $BSR\,4$, $BSR\,5$ e $BSR\,6$ e o 3^{o} *cluster* formado por $BSR\,9$, $BSR\,10$, $BSR\,11$ e $BSR\,12$.

Sendo caracterizada por terraços fluviais do Rio Minho, a parcela total em estudo na Galiza evolui de granulometria baixa (silte e argila) para uma granulometria mais grosseira com seixos que podem ir até aos 10 cm de comprimento. O *cluster* 1 corresponde à zona da parcela constituída exclusivamente por argila e silte. A transição para granulometrias mais extensas (argila, areia, seixos) ocorre nos *clusters* 2 e 3. Sabe-se também que existe uma ligeira subida na cota de Sul para Norte, ou seja, a cota do *cluster* 1 é inferior à dos *clusters* 2 e 3.

A Tabela 4.14 apresenta um resumo de algumas das principais características amostrais relativas a cada um dos constituintes do solo, em cada um dos *clusters* encontrados. Analisando os resultados obtidos, observa-se que:

- no *cluster* 1, as variáveis MO, FG, $P2O5$, $K2O$, Ca, Mg, AzT, Ni, N, B e CTC têm o valor mediano mais baixo e a variável FF é a única que possui o valor mediano mais alto;
- no *cluster* 2, as variáveis MO, pH, $K2O$, Ca, Mg, AzT, Cr, Cd, N, B e CTC têm o valor mediano mais elevado, não havendo variáveis cujo valor mediano seja o mais baixo;
- no *cluster* 3, as variáveis FG, $P2O5$ e Ni mostram o valor mediano mais elevado, sendo que a variável pH, FF, Cr e Cd são as que apresentam o valor mediano mais baixo.

Verifica-se, claramente, que no *cluster* 1 a maior parte das variáveis apresenta o valor mediano mais baixo, comparativamente ao *cluster* 2 e ao *cluster* 3. Contrariamente, no *cluster* 2, a maioria das variáveis apresenta o maior valor mediano. Quanto ao *cluster* 3 existe uma mesma percentagem de variáveis que mostram o valor mediano mais elevado e o valor mediano mais baixo.

Tabela 4.14: Caracterização dos 3 *clusters* retidos de acordo com os elementos do solo.

	MO	pH	FF	FG	$P2O5$	$K2O$	Ca	Mg	AzT	Ni	Cr	Cd	N	B	CTC
***CLUSTER* 1**															
Mínimo	2,610	4,930	54,293	12,712	0,905	41,217	249,000	54,000	0,113	2,720	0,208	0,067	1,560	0,390	6,302
Máximo	7,730	5,860	87,289	45,707	3,061	57,227	334,500	79,500	0,282	6,160	0,752	0,091	1,800	0,669	8,686
Mediana	4,360	5,010	72,899	27,101	1,832	49,237	300,000	60,000	0,173	3,840	0,416	0,079	1,620	0,474	7,442
Média	5,186	5,224	71,721	28,279	1,770	48,434	298,500	63,900	0,196	4,352	0,448	0,078	1,662	0,516	7,522
Desvio Padrão	2,343	0,393	13,928	13,928	0,826	7,042	33,423	9,698	0,080	1,705	0,195	0,010	0,107	0,122	0,881
1^o Quartil	3,640	4,970	61,381	17,255	1,218	41,522	286,500	60,000	0,134	2,880	0,416	0,071	1,580	0,428	7,239
3^o Quartil	7,590	5,350	82,745	38,619	1,835	52,967	322,500	66,000	0,281	6,160	0,448	0,084	1,750	0,617	7,941
***CLUSTER* 2**															
Mínimo	8,240	4,240	62,060	10,024	5,195	57,312	421,500	72,000	0,322	3,520	0,352	0,098	2,690	0,681	10,658
Máximo	16,520	5,280	89,976	37,940	22,588	82,963	714,000	120,000	0,518	4,960	0,768	0,122	3,040	0,947	16,704
Mediana	11,870	5,270	68,290	31,710	7,032	78,973	460,500	87,000	0,463	4,480	0,560	0,103	2,900	0,791	10,673
Média	12,210	4,930	73,442	26,558	11,605	73,083	532,000	93,000	0,434	4,320	0,560	0,108	2,877	0,806	12,678
Desvio Padrão	4,150	0,598	14,654	14,654	9,556	13,803	158,818	24,556	0,101	0,733	0,208	0,013	0,176	0,134	3,486
1^o Quartil	10,055	4,755	65,175	20,867	6,113	68,143	441,000	79,500	0,392	4,000	0,456	0,101	2,795	0,736	10,666
3^o Quartil	14,195	5,275	79,133	34,825	14,810	80,968	587,250	103,500	0,490	4,720	0,664	0,113	2,970	0,869	13,688
***CLUSTER* 3**															
Mínimo	6,730	4,660	52,456	36,580	5,794	47,478	298,500	63,000	0,261	6,000	0,144	0,069	2,040	0,576	7,663
Máximo	13,140	5,250	63,420	47,544	15,774	68,661	444,000	87,000	0,482	6,960	0,528	0,082	2,900	0,874	10,875
Mediana	8,775	4,940	56,545	43,455	8,989	65,119	315,750	64,500	0,310	6,180	0,392	0,077	2,460	0,690	7,928
Média	9,355	4,947	57,241	42,759	9,887	61,594	343,500	69,750	0,341	6,330	0,364	0,076	2,465	0,707	8,598
Desvio Padrão	2,856	0,242	5,062	5,062	4,205	9,684	67,561	11,522	0,100	0,436	0,171	0,007	0,352	0,133	1,523
1^o Quartil	7,412	4,855	53,482	39,695	8,103	59,277	308,684	64,125	0,274	6,060	0,276	0,071	2,333	0,615	7,840
3^o Quartil	10,717	5,032	60,305	46,518	10,773	67,436	350,635	70,125	0,377	6,450	0,480	0,082	2,592	0,782	8,687

Observando os resultados obtidos, também se pode verificar que:

- no *cluster* 1, grande parte das variáveis apresentam o menor desvio padrão, nomeadamente a variável MO, $P2O5$, $K2O$, Ca, Mg, AzT, N, B, e CTC e, apenas Ni mostra um desvio padrão maior comparativamente à que tem nos restates *clusters*;
- no *cluster* 2, à exceção das variáveis Ni e N, todas as variáveis têm o maior desvio padrão;
- no *cluster* 3, a variável N apresenta o desvio padrão mais elevado, sendo que a variável pH, FF, FG, Ni, Cr e Cd têm o menor desvio padrão.

Conhecido o número de *clusters*, pretende-se saber se existem diferenças significativas entre eles relativamente às concentrações dos constituintes do solo, às variáveis que traduzem a produtividade da videira e a qualidade do vinho. Para testar a existência destas diferenças recorreu-se de novo ao teste não paramétrico de *Kruskal-Wallis*, uma vez que o pressuposto da normalidade da distribuição dos grupos não se verifica e o número de observações é bastante reduzido. Os resultados estatisticamente significativos derivados da aplicação deste teste, assumindo-se como hipótese nula a não existência de diferenças significativas entre os três grupos relativamente às variáveis indicadas, encontram-se na Tabela 4.15. Também se pode encontrar nesta tabela os valores das medianas de cada variável, para cada um dos grupos.

Tabela 4.15: Resultados do teste de *Kruskal-Wallis* para as variáveis em estudo.

			Medianas		Estatística	Graus de	*Valor-prova*
	Variáveis	*Cluster* 1	*Cluster* 2	*Cluster* 3	de teste	liberdade	
SOLO	MO	4,360	11,870	8,775	6,369	2	**0,041****
	$P2O5$	1,832	7,032	8,989	8,122	2	**0,017****
	$K2O$	49,237	78,973	65,119	6,369	2	**0,041****
	Ca	300,000	460,500	315,750	5,779	2	0,057*
	Mg	60,000	87,000	64,500	4,854	2	0,088*
	AzT	0,173	0,463	0,310	5,804	2	0,058*
	Cd	0,079	0,103	0,077	6,287	2	**0,043****
	N	1,620	2,900	2,460	9,017	2	**0,011****
	B	0,474	0,791	0,690	6,369	2	**0,041****
	CTC	7,442	10,673	7,928	5,814	2	0,058*
VIDEIRA	$Pbruto$	0,140	0,120	0,160	5,108	2	0,078*
MOSTO	$Ac.TOT$	10,500	10,800	9,415	8,590	2	**0,014****
UVAS	$FL1U$	171,250	98,300	152,130	5,070	2	0,079*
	$FL5U$	37,410	33,140	26,985	5,510	2	0,064*
	$FG1U$	4,270	6,990	4,535	7,082	2	**0,029****
	$FG7U$	0,730	1,090	0,675	6,076	2	**0,048****
VINHO	$Nota.Final$	64,140	59,860	58,345	6,600	2	**0,037****

*significativo a 10% **significativo a 5%

Analisando os resultados da Tabela 4.15, conclui-se que há evidência estatística, a um nível de significância $0,05$, para rejeitar a igualdade de valores medianos nos 3 *clusters* nas variáveis do

solo MO, $P2O5$, $K2O$, Cd, N, na variável aromática das uvas $FG1U$, $FG7U$, na variável $Ac.TOT$ das uvas e também na variável $Nota.Final$; a um nível de significância mais elevado, $0,1$, passa a haver evidência estatística para rejeitar também a igualdade de valores medianos das variáveis do solo Ca, Mg, AzT, CTC, na variável $Pbruto$ e das variáveis aromáticas das uvas $FL1U$ e $FL5U$. Relativamente às variáveis do solo pH, FF, FG, Ni e Cr, há evidência estatistica que apresentam concentrações idênticas em todos os *clusters* considerados.

Encontrada diferenças entre os *clusters*, interessa saber quais os grupos que diferem entre si.

Tabela 4.16: Resultados do teste LSD para as variáveis em estudo (**significativo a 5%, *significativo a 10%).

Variável	*cluster*	*cluster*	*Valor-prova*
MO	1	2	**0,008****
		3	**0,051****
	2	3	0,237
$P2O5$	1	2	**0,004****
		3	**0,001****
	2	3	0,717
$K2O$	1	2	**0,008****
		3	**0,051****
	2	3	0,237
Cd	1	2	**0,014****
		3	0,803
	2	3	**0,012****
N	1	2	**0,000****
		3	**0,002****
	2	3	0,072*
B	1	2	**0,008****
		3	**0,051****
	2	3	0,237
$Ac.TOT$	1	2	0,079*
		3	**0,003****
	2	3	**0,000****
$FG1U$	1	2	**0,003****
		3	0,134
	2	3	**0,039****
$FG7U$	1	2	**0,002****
		3	0,586
	2	3	**0,039****
$Nota.Final$	1	2	**0,053****
		3	**0,006****
	2	3	0,321
Ca	1	2	**0,012****
		3	0,408
	2	3	**0,054****
Mg	1	2	**0,026****
		3	0,346
	2	3	0,128
AzT	1	2	**0,013****
		3	0,104
	2	3	0,221
CTC	1	2	**0,011****
		3	0,168
	2	3	0,121
$Pbruto$	1	2	0,129
		3	0,206
	2	3	**0,021****
$FL1U$	1	2	**0,022****
		3	0,287
	2	3	0,132
$FL5U$	1	2	0,183
		3	**0,015****
	2	3	0,242

Para este efeito aplicou-se também o teste não paramétrico de comparações múltiplas LSD. Os resultados obtidos encontram-se na Tabela 4.16.

Feita uma análise aos resultados obtidos, a um nível de significância de $0,05$, há evidência estatística para afirmar que, relativamente à(s) variável(eis):

- MO, $P2O5$, $K2O$, N e B, o *cluster* 1 difere dos *clusters* 2 e 3 apresentando um valor mediano inferior;
- Cd, $FG1U$, $FG7U$ e Ca, o *cluster* 2 difere dos *clusters* 1 e 3 apresentando valores medianos superiores;
- $Ac.TOT$, o *cluster* 3 difere dos *clusters* 1 e 2 apresentando um valor mediano inferior;
- $Nota.Final$ e $FL5U$, o *cluster* 1 difere do *cluster* 3, apresentando um maior valor mediano;
- Mg, CTC, AzT e $FL1U$, o *cluster* 1 difere do *cluster* 2, apresentando um valor mediano inferior relativamente à variável Mg, CTC e AzT e superior em relação a $FL1U$;
- $Pbruto$, o *cluster* 2 difere do *cluster* 3, apresentando um valor mediano inferior.

Tendo como objetivo perceber qual a influência de cada um dos constituintes do solo na qualidade do vinho, determinaram-se os valores das correlações entre estas variáveis. Seguidamente, para testar se esta associação medida pelo valor do coeficiente de correlação é estatisticamente significativa, realizou-se o teste baseado no coeficiente de correlação não paramétrico de *Spearman*, uma vez que as dimensões das amostras são demasiado pequenas e, consequentemente, não se consegue provar a multinormalidade dos dados. Na Tabela 4.17 encontram-se os resultados das correlações das variáveis que são diferentes de zero, ou seja, em que a hipótese nula foi rejeitada a um nível de significância de $0,05$. Analisando os resultados obtidos, observa-se que:

- existe uma correlação negativa entre as variáveis MO, $K2O$, AzT e B com a variável TAP (a diminuição da Matéria Orgânica, de Potássio Assimilável, Azoto Total e Boro leva a um aumento significativa do Teor de Álcool Provável). É no *cluster* 1 que existem as menores concentrações de MO, $K2O$ e B, comparativamente com as restantes *clusters*, pelo que haverá neste terreno vinhos com maior teor de álcool; comparativamente ao *cluster* 2, também o *cluster* 1 apresenta menor concentração de Azoto Assimilável, pelo que contribuirá para que neste *cluster* se produza vinho com maior teor de álcool. Note-se que a variável $K2O$ também apresenta uma correlação estatisticamente significativa positiva com $FL6U$, concluindo-se assim que quanto maior for a concentração de Potássio no solo maior, será a concentração dos compostos aromáticos da família $FL6U$. Sendo o *cluster* 1 o que apresenta menor concentração deste elemento, vai ser onde haverá menor concentração dos compostos de $FL6U$;

Tabela 4.17: Correlações estatisticamente significativas, ao nível de significância de $0,05$, entre as variáveis originais do solo e as variáveis relacionadas com a videira, o mosto, as uvas e a qualidade do vinho.

			MO	FF	FG	$P2O5$	$K2O$	Ca	Mg	AzT	Ni	Cd	N	B	CTC
VIDEIRA	$Pbruto$	Corr						-0,654							
		Valor-prova						0,021							
MOSTO	TAP	Corr	-0,594				-0,636			-0,622				-0,594	
		Valor-prova	0,042				0,026			0,031				0,042	
	$Ac.TOT$	Corr									-0,722	0,644			
		Valor-prova									0,008	0,024			
UVAS	$FL1U$	Corr				-0,594		-0,848	-0,838			-0,658	-0,616		-0,777
		Valor-prova				0,042		0,000	0,001			0,020	0,033		0,003
	$FL2U$	Corr				-0,594		-0,848	-0,838		-0,807	0,715			
		Valor-prova				0,042		0,000	0,001		0,002	0,009			
	$FL5U$	Corr		0,636	-0,636	-0,636							-0,588		
		Valor-prova		0,026	0,026	0,026							0,045		
	$FL6U$	Corr					0,608								
		Valor-prova					0,036								
	$FG1U$	Corr				0.608		0,693	0,682			0,602	0,772		0,664
		Valor-prova				0,036		0,013	0,015			0,038	0,003		0,018
	$FG7U$	Corr											0,582		
		Valor-prova											0,048		
	$Amarelo\_U$	Corr						-0,636	-0,639		0,609	-0,588			
		Valor-prova						0,026	0,025		0,036	0,045			
	$Azul\_U$	Corr						0,707	0,725		-0,609	0,680			0,636
		Valor-prova						0,010	0,008		0,036	0,015			0,026
	$Rend\_sumo\_U$	Corr						-0,608							
		Valor-prova						0,036							
VINHO	$Nota.Final$	Corr		0,678	-0,678	-0,636							-0,736		
		Valor-prova		0,015	0,015	0,026							0,006		

- relativamente ao Magnésio, (Mg), é um constituinte que se encontra correlacionado positivamente com a variável $FG1U$ e com a variável referente à percentagem da cor azul nas uvas, e negativamente com $FL1U$, $FL2U$ e com a cor amarela das uvas, isto é, concentrações reduzidas de Magnésio no solo corresponderão a baixas concentrações de compostos da família $FG1U$ e à diminuição da cor azul nas uvas e corresponderão a altas concentrações dos compostos da família $FL1U$, $FL2U$ e ao aumento da cor amarela nas uvas. É no *cluster* 1, comparativamente ao *cluster* 2, que se verifica menor concentração deste elemento do solo e, portanto, produzindo uvas com menor quantidade de compostos aromáticos da família $FG1U$, e com maior quantidade de compostos aromáticos da família $FL1U$ e $FL2U$; este terreno produz também as uvas com uma menor percentagem de cor azul e uma maior percentagem de cor amarela;

- o elemento cálcio, (Ca), possui as mesmas correlações que a variável Magnésio, ou seja, reduzidas concentrações de Ca no solo corresponderão a baixas concentrações de $FG1U$ e a diminuição da cor azul nas uvas e corresponderão a altas concentrações de $FL1U$, $FL2U$ e ao aumento da cor amarela nas uvas. Sendo o *cluster* 2, o que apresenta maior concentração deste elemento, vai ser o terreno a produzir uvas com maior quantidade de compostos aromáticos da família $FG1U$ e com menor quantidade de compostos aromáticos da família $FL1U$ e $FL2U$; este terreno também produz as uvas com uma maior percentagem de cor azul e uma menor percentagem de cor amarela. Note-se também que a variável Ca é o único constituinte do solo que se encontra correlacionado com o Peso bruto de cachos por videira e com a percentagem do rendimento em sumo: quanto maior for a concentração de cálcio menor vai ser o $Pbruto$ e o $Rend\_sumo\_U$. É no *cluster* 2 que haverá menor produtividade das videiras, no que diz respeito ao peso bruto de cachos por videira e menor percentagem de rendimento em sumo, responsável por este elemento;

- o aumento de FG diminui os compostos da Família $FL5U$ e diminui a avaliação final atribuída às características do vinho. Não havendo diferenças significativas na percentagem de Fração Grosseira nos diferentes *clusters*, a sua influência nos compostos da família $FL5U$ e na avaliação final do vinho nos diferentes *clusters* será a mesma;

- $P2O5$ e N são dois constituintes que estão ligados às famílias dos compostos aromáticos, uma vez que se verifica que o aumento das suas concentrações leva à diminuição da concentrações de alguns compostos voláteis do aroma na fração livre (diminuição de $FL1U$ e $FL5U$ para valores elevados de $P2O5$ e N e diminuição de $FL2U$ apenas para valores elevados de $P2O5$), e ao aumento de concentrações de compostos voláteis do aroma na fração glicosilada (aumento de $FG1U$ para valores elevados de $P2O5$ e N e aumento de $FG7U$ para valores elevados de N). Sendo os valores de concentração do fósforo e de nitratos me-

nores no *cluster* 1, comparativamente aos *clusters* 2 e 3, é neste terreno que haverá maior quantidade de compostos da família $FL1U$, $FL2U$ e $FL5U$ e menor quantidade de $FG1U$ e $FG7U$. Note-se que o aumento destes dois elementos do solo leva a uma diminuição na nota final que classifica os vinhos, pelo que se pode afirmar que o fosfato e os nitratos contribuem para que o cluster 1 produza o vinho que melhor classificação obteve pelos provadores;

- o aumento da capacidade de troca catiónica (CTC) leva a uma diminuição significativa da quantidade de compostos aromáticos da família $FL1U$, a um aumento dos compostos da família $FG1U$ e a um aumento da percentagem da cor azul nas uvas. Verificada a existência de menor CTC no *cluster* 1, comparativamente ao *cluster* 2, conclui-se que esta capacidade contribui para que nesta zona exista uma maior concentração de $FL1U$, uma menor concentração de $FG1U$ e uma produção de uvas em que a percentagem de cor azul é menor;

- Ni e Cd são dois constituintes que influenciam alguns compostos aromáticos, cor do mosto e acidez total das uvas de forma contrária: enquanto o aumento de níquel no solo faz diminuir a $Ac.TOT$ das uvas, os compostos voláteis de aroma da família $FL2U$ (Família de Álcoois do aroma do mosto na fração livre), a percentagem de cor azul nas uvas, e o aumento da percentagem de cor amarela, Cd tem um efeito contrário nestas quatro variáveis; concentrações elevadas de cádmio ainda levam a baixas concentrações de compostos da família $FL1U$ (Família de Compostos em $C6$ do aroma do mosto na fração livre) e ao aumento dos compostos da família $FG1U$ (Família de Composto em $C6$ do aroma do mosto na fração glicosilada). Comparando o *cluster* 2 com os restantes dois, é este que possui maiores concentrações de Cd, pelo que será nesta parcela que este constituinte vai contribuir para que os mostos tenham maior acidez total, as uvas possuam maior quantidade de compostos aromáticos das famílias $FL2U$ e $FG1U$ e maior percentagem de cor azul; o seu aumento contribui também para que neste terreno existam as uvas com menores concentrações de compostos da família $FL1U$ e menor percentagem de cor amarela. Não havendo diferenças significativas na concentração de Ni nos diferentes *clusters*, a sua influência em qualquer variável referida anteriormente será a mesma;

- as variávis pH, Cr, Ca e Mg não apresentam qualquer tipo de correlação com as variáveis responsáveis pela qualidade do vinho.

Agora, vai-se apresentar a Análise de *Clusters* realizada a partir dos fatores retidos resultantes de uma Análise Fatorial (AF).

Com o objetivo de transformar um problema que envolve um elevado número de variáveis originais e correlacionadas num problema com um número reduzido de variáveis, efetuou-se uma Análise Fatorial; com esta análise pretendeu-se verificar se existia um pequeno número de variáveis que fossem responsáveis por explicar grande parte da variabilidade total associada aos dados originais relativos ao solo e, portanto, reduzir assim a dimensionalidade do problema.

Para dar início à realização de uma AF, analisaram-se as relações existentes entre os diferentes constituintes químicos do solo em estudo e, para tal, calcularam-se os valores das correlações de *Spearman* entre cada par de constituintes. Estas correlações encontram-se apresentadas na Tabela 4.18. Nestes cálculos participam todos os *outliers* presentes nos dados.

Feita uma análise à matriz de correlações entre os constituintes do solo, ao contrário do que aconteceu com os dados do Minho, existe uma maior correlação entre estas variáveis. A partir da observação da Tabela 4.18, verifica-se que entre as variáveis AzT e MO, B e MO, CTC e Mg, Ca e Mg, Ca e CTC e AzT e B existe uma correlação quase perfeita e as variáveis $K2O$ e MO, $K2O$ e $P2O5$, Ca e $P2O5$, Mg e $P2O5$, CTC e $P2O5$, AzT e $K2O$, B e $K2O$ encontram-se fortemente correlacionadas. Considera-se ainda existir uma boa correlação $(0,63 < r < 0,72)$ entre as variáveis $P2O5$ e MO, Cd e MO, N e MO, AzT e $P2O5$, Cd e $P2O5$, B e $P2O5$, N e Cd, B e Cd, CTC e Cd, Cd e AzT, N e AzT e B e N, havendo ainda muitas delas que se encontram moderadamente correlacionadas.

Tabela 4.18: Matriz de correlações das variáveis do solo da Galiza.

	MO	pH	FF	$FG4$	$P2O5$	$K2O$	Ca	Mg	AzT	Ni	Cr	Cd	N	B	CTC
MO	1.00														
pH	***-0,65***	1,00													
FF	0,10	-0,08	1,00												
FG	-0,10	0,08	***-1,00***	1,00											
$P2O5$	***0,69***	-0,32	-0,02	0,02	1,00										
$K2O$	***0,82***	-0,45	0,06	-0,06	***0,76***	1,00									
Ca	0,46	-0,02	-0,13	0,13	***0,79***	**0,54**	1,00								
Mg	0,48	-0,09	-0,13	0,13	***0,83***	***0,55***	***0,96***	1,00							
AzT	***0,99***	***-0,55***	0,09	-0,09	***0,71***	***0,83***	**0,51**	***0,52***	1,00						
Ni	0,25	-0,25	0,16	-0,16	0,06	0,17	-0,19	-0,18	0,24	1,00					
Cr	0,47	-0,44	0,45	-0,45	0,06	0,32	0,00	0,01	0,46	0,14	1,00				
Cd	***0,67***	-0,46	0,17	-0,17	***0,71***	***0,60***	***0,62***	***0,63***	***0,64***	-0,10	0,28	1,00			
N	***0,69***	***-0,62***	0,07	-0,07	***0,57***	***0,54***	0,44	0,46	***0,65***	0,24	0,43	***0,71***	1,00		
B	***0,99***	***-0,60***	0,06	-0,06	***0,68***	***0,79***	0,48	0,49	***0,99***	0,27	0,47	***0,64***	***0,67***	1,00	
CTC	***0,51***	-0,07	-0,12	0,12	***0,83***	***0,60***	***0,99***	***0,98***	***0,56***	-0,17	0,03	***0,64***	0,47	***0,52***	1,00

Após se obter um determinante da matriz de correlações muito próximo de zero, $9,14 \times 10^{-10}$, uma estatística de KMO indicadora da existência de uma correlação razoável entre as variáveis e se ter rejeitado a hipótese nula associada ao teste de esfericidade de *Bartlett*, Tabela 4.19 , concluiu-se que a aplicação de uma AF aos dados é apropriada.

Tabela 4.19: Resultados do teste de esfericidade de *Bartlett* e da estatística de KMO.

Estatística de KMO (*Kaiser-Meyer-Olkin*)		0,78
Teste de esfericidade de *Bartlett*	Estatística de Teste	808,242
	Graus de Liberdade	78
	Valor-prova	< 0,001

A AF foi realizada sem a variável FG e sem a variável Ca de forma a que a matriz de correlações se tornasse definida positiva.

A partir do método de componentes principais extraíram-se apenas 4 fatores explicando-se aproximadamente $85,5\%$ da variabilidade total dos dados, notando-se que o valor próprio associado ao quarto fator após a realização de uma rotação *varimax* ainda se encontra com valor superior a 1, como se pode observar pela Tabela 4.20. Além disso, a análise do *scree plot*[5] sugere tal extração e a opinião dos peritos nas áreas em estudo também pesou para decisão da escolha do número de fatores.

Tabela 4.20: *Loadings* dos 4 primeiros fatores com rotação *Varimax* e comunalidades das variáveis em estudo; proporção de variância explicada por cada fator.

Variáveis	Fator				Comunalidades
	1	2	3	4	
MO	0,521	0,771	0,093	0,255	0,939
pH	0,021	-0,893	0,002	-0,062	0,802
FF	-0,022	-0,011	0,954	0,102	0,921
$P2O5$	0,890	0,296	-0,038	0,112	0,894
$K2O$	0,655	0,519	0,070	0,254	0,767
Mg	0,939	0,088	-0,079	-0,156	0,920
AzT	0,577	0,698	0,101	0,278	0,907
Ni	-0,109	0,176	0,083	0,911	0,879
Cr	-0,059	0,606	0,617	-0,032	0,752
Cd	0,643	0,524	0,191	-0,239	0,782
N	0,396	0,724	0,073	0,020	**0,687**
B	0,528	0,744	0,071	0,284	0,918
CTC	0,953	0,098	-0,065	-0,129	0,939
Valor próprio	4,482	4,003	1,380	1,243	
Proporção de variância (%)	34,474	30,790	10,616	9,561	
Proporção acumulada (%)	34,474	65,264	75,880	85,441	

Observando-se agora as correlações entre os vários constituintes do solo e os fatores apresentados na Tabela 4.20, verifica-se que:

- as variáveis que apresentam maior correlação positiva com o 1^{o} fator são as variáveis CTC, Mg, $P2O5$, $K2O$ e Cd;
- relativamente ao 2^{o} fator são as variáveis MO, AzT, N e B que apresentam maior correlação positiva. Este fator ainda apresenta uma forte correlação negativa com a variável pH;

[5] O *scree plot* encontra-se no Anexo IV.

- os constituintes FF e Cr são os que mais influenciam o 3^o fator, com correlação positiva;
- relativamente ao 4^o fator é a variável Ni que apresenta uma forte correlação positiva com este fator.

Observando a comunalidade associada a cada variável original, no geral, os fatores comuns retidos conseguem explicar uma boa proporção de variância de cada variável original, sendo o constituinte N o que mais informação perde, uma vez que os 4 fatores apenas conseguem explicar aproximadamente 69% da variância desta variável.

Analisando agora os *scores* de cada fator extraídos através do método dos mínimos quadrados ponderados, Tabela 4.21, observa-se que os pontos de amostragem do solo $B8$, $B41$ e $B42$ são os que apresentam maiores *scores* para o 1^o fator, o que significa que são nestes pontos que as variáveis $P2O5$, $K2O$, Mg, Cd e CTC apresentam maiores concentrações uma vez que são estas as variáveis que estão mais correlacionadas positivamente com o 1^o fator; $B1$ é o ponto que apresenta menor *score* e, consequentemente, o que apresenta menor concentração destes cincos constituintes do solo.

Tabela 4.21: *Scores* dos quatro primeiros fatores.

Terreno	Fator 1	Fator 2	Fator 3	Fator 4	Terreno	Fator 1	Fator 2	Fator 3	Fator 4
B1	-1,054	0,280	-0,722	0,759	B24	-0,287	-0,225	1,464	-1,548
B2	-0,806	-1,048	0,51	0,300	B25	-0,406	0,957	-1,470	-1,639
B3	-0,443	-0,930	0,266	0,801	B26	-0,533	-0,075	-1,754	-0,414
B4	-0,528	-1,146	0,669	0,524	B27	-0,320	0,195	-0,203	-1,376
B5	-0,645	-0,899	0,501	0,407	B28	-0,242	0,907	0,864	-0,329
B6	0,865	0,647	1,378	1,575	B29	-0,052	3,570	0,715	0,391
B7	-0,510	0,963	-0,835	0,537	B30	0,196	1,886	0,608	-1,012
B8	2,591	-0,082	-0,028	1,968	B31	-0,269	0,015	1,640	-0,998
B9	-0,398	0,545	-0,621	1,678	B32	-0,725	0,005	1,775	-0,578
B10	-0,249	0,597	-0,235	1,559	B33	-0,670	-0,360	0,675	-1,421
B11	-0,603	0,395	-1,533	0,622	B34	-0,061	-0,854	1,065	-0,850
B12	0,981	-0,742	0,789	0,639	B35	-0,676	-0,530	,038	-1,421
B13	0,507	-0,742	-1,441	0,939	B36	0,091	-0,338	-1,371	-0,490
B14	-0,251	0,352	-1,550	1,324	B37	0,072	-1,507	-,795	-0,986
B15	-0,411	-1,045	-1,080	0,703	B38	-0,049	-0,276	-1,013	-1,574
B16	-0,750	-0,487	1,383	0,738	B39	0,253	0,197	1,869	-0,138
B17	-0,220	-1,285	0,059	0,328	B40	0,456	0,906	0,644	-0,023
B18	-0,220	-1,114	1,126	0,939	B41	1,990	-0,650	-0,298	-1,136
B19	-0,809	-0,570	0,306	1,103	B42	4,760	-0,213	-0,050	-0,546
B20	-0,532	-0,868	0,020	0,943	B43	-0,238	2,172	-0,494	-0,633
B21	-0,568	1,264	0,165	0,731	B44	0,450	-0,276	-0,756	-0,632
B22	0,481	1,178	-0,368	0,023	B45	0,007	-0,482	-0,370	-1,043
B23	-0,174	-0,285	-1,544	-0,744					

O 2^o fator encontra-se mais correlacionado positivamente com as variáveis MO, AzT, N e B e negativamente correlacionado com o pH, pelo que são nos pontos de amostragem que apresentam maiores *scores*, $B21$, $B22$, $B29$, $A30$ e $B43$ que existe maior concentração de MO, AzT, N e B e menor concentração de pH; para este fator os *scores* mais baixos observam-se em $B2$, $B4$, $B15$, $B17$, $B18$ e $B37$ e como tal são nestes pontos que se encontram as menores concentrações de MO, AzT, N e B e as maiores concentrações de pH.

Relativamente ao 3^o fator, os *loadings* mais elevados encontram-se nas variáveis FF e Cr

pelo que as concentrações destes constituintes são mais altas nos pontos de amostragem $B6$, $B16$, $B18$, $B23$, $B24$, $B31$, $B32$, $B34$ e $B39$, por apresentarem os *scores* mais elevados, e mais pequenas em $B11$, $B13$, $B14$, $B15$, $B25$, $B26$, $B36$ e $B38$ (*scores* mais pequenos).

Quanto ao 4^{o} fator, este apresenta *scores* mais elevados em $B6$, $B8$, $B9$, $B10$, $B14$ e $B19$ o que indica que são nestes pontos que a concentração de Ni é mais elevada e mais baixa nos pontos $B24$, $B25$, $B27$, $B30$, $B33$, $B35$, $B38$, $B41$ e $B45$.

Encontrados assim os 4 fatores extraídos a partir da AF e os respetivos *scores*, procedeu-se de seguida a uma Análise de *Clusters* com o objetivo de identificar na parcela em estudo zonas com características semelhantes em relação às variáveis do solo. Note-se que, tal como aconteceu no estudo realizado com as variáveis originais, a partir do cálculo de médias amostrais, os 45 *scores* de cada fator foram reduzidos a 12, que representarão as diversas quantidades observadas de cada fator nas parcelas $BSR\,1$ a $BSR\,12$.

Para esta análise foram de novo utilizados vários métodos hierárquicos pelo algoritmo aglomerativo (método de ligação completa, de ligação média, de ligação simples, método de *Ward* e do centróide), obtendo-se agrupamentos distintos quando utilizado o método *Ward*, método de ligação completa e método de ligação média, o que sugere que não existe uma clara estrutura de grupos subjacentes aos dados. Perante tais resultados, permaneceu a dúvida sobre o método que se deveria escolher.

Após a formação dos *clusters* a partir destes três métodos, representaram-se os gráficos que mostram a distância para a junção entre dois grupos, para averiguar quantos *clusters* se deverim considerar em cada um dos métodos, Figura 4.11.

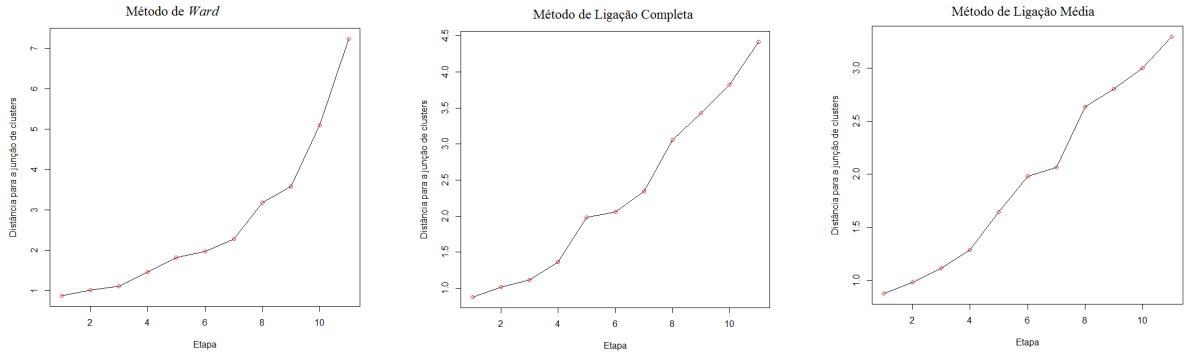


Figura 4.11: Critério para obter o número de *clusters*.

Analisando os gráficos, segundo o método de ligação completa e de ligação média pararecer razoável reterem-se 4 *clusters*, sendo que o gráfico associado ao método *Ward* sugere a retenção de apenas 3 *cluters*. A Figura 4.12 apresenta o mapeamento dos 4 grupos obtidos para cada um dos três métodos referidos anteriormente.

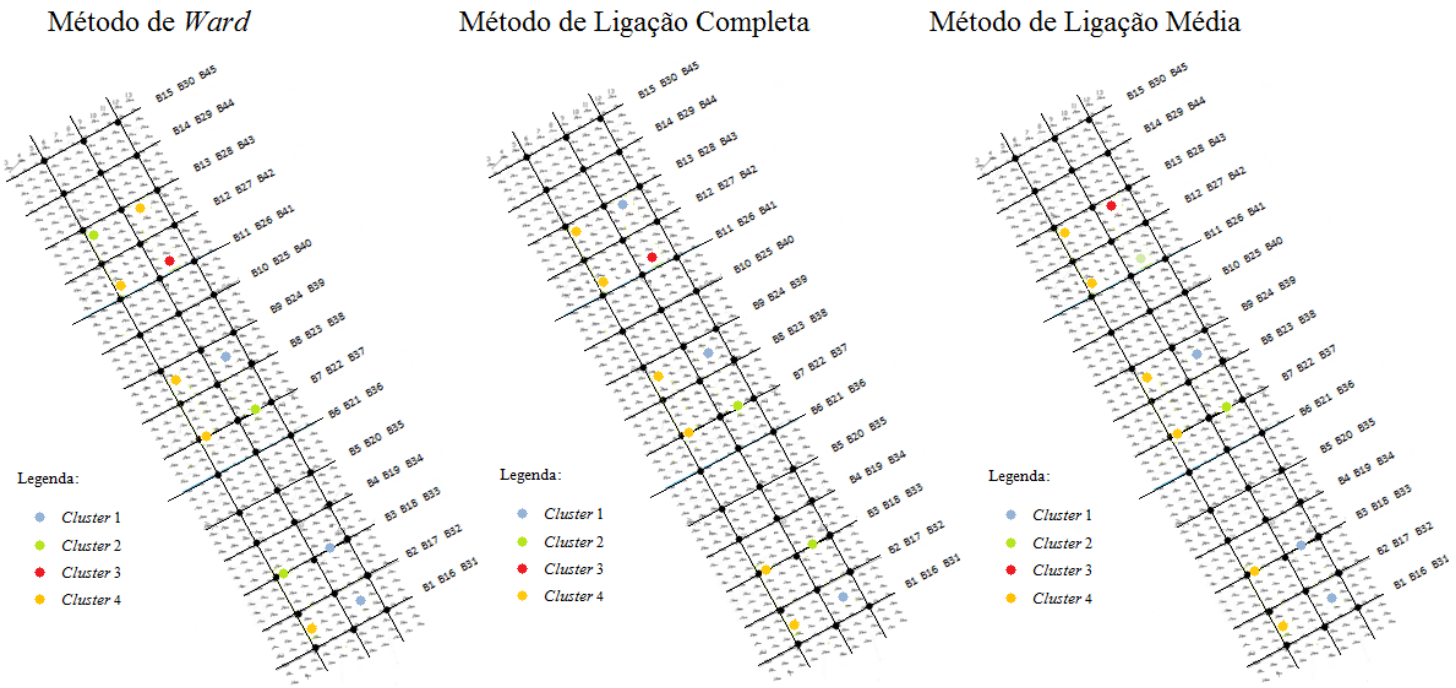


Figura 4.12: Distribuição dos 4 grupos na parcela em estudo de acordo com os resultados obtidos a partir do método *Ward*, método de ligação completa e método de ligação média.

Os padrões de agrupamentos obtidos nesta análise, que são diferentes aos encontrados a partir da Análise de *Clusters* realizada com as variáveis originais do solo, foram discutidos e concluiu-se que nenhum destes agrupamentos faria sentido, face ao conhecimento deste terreno por parte dos peritos nas áreas em estudo[9].

Perante tal opinião, decidiu-se não dar continuidade ao estudo a partir destes resultados.

[9]Os padrões de agrupamentos obtidos quando retidos apenas 3 *clusters*, também se mostraram todos diferentes, pelo que a conclusão retirada foi a mesma.

Capítulo 5

Conclusão

Este trabalho resultou da necessidade de se identificar quais as variáveis físico-químicas do solo que influenciaram a produtividade das videiras e a qualidade das uvas, do mosto e, consequentemente, dos vinhos numa parcela da Estação Vitivinícola Amândio Galhano, no concelho de Arcos de Valdevez, e numa parcela localizada na Bodega Santiago Ruiz, em *Tomiño* na Região da Galiza. Os dados analisados reportaram-se apenas ao ano de 2011.

De forma a concretizar os objetivos estabelecidos para o estudo, uma primeira ideia foi aplicar uma Análise de Regressão Múltipla aos dados, uma vez que se pretendia estudar a relação entre uma variável resposta - variável referente à videira, à uva/mosto ou ao vinho, e várias variáveis explicativas - variáveis referentes às características do solo. Devido ao número muito reduzido de observações que a maioria das variáveis resposta apresentavam, apenas foi possível realizar-se este tipo de análise com duas variáveis referentes à produtividade da videira: peso bruto de cachos por videira e peso médio de cacho por videira. Após a análise realizada com os dados relativos à quinta da Galiza, verificou-se que os resultados obtidos não foram estatisticamente significativos e da análise dos dados da quinta do Minho resultaram conclusões escassas: apenas o magnésio teve influência estatística nestas duas variáveis resposta. Assim, decidiu-se apenas analisar o caso de aplicação de Análises de Regressão Linear Simples.

A interpretação e discussão dos resultados provenientes da Análise de Regressão Linear Simples foi efetuada com a ajuda de peritos na área de estudo. Concluiu-se que os resultados provenientes deste tipo de análise aos dados da Galiza não tinham interpretabilidade no contexto do problema. Tais resultados podem ser consequência do facto deste estudo ser enviesado, uma vez que a análise referida foi realizada apenas com um total de 12 observações.

Quanto aos resultados provenientes da análise dos dados da quinta do Minho, verificou-se que as variáveis pH, $P2O5$, Ca, Mg, Ni, Cr, Cd, N e CTC influenciam de forma estatisticamente significativa o peso bruto de cachos por videira e as variáveis $P2O5$, Ca, Mg, Cr, N e CTC estão

correlacionadas com o peso médio de cacho por videira. Grande parte destes resultados foram ao encontro do conhecimento que os peritos possuíam, bem como dos resultados provenientes de estudos que foram realizados nesta área. No entanto, nestes estudos não há conhecimento da existência de associações estatisticamente significativas entre as variáveis Ni e Cr, ao contrário do que se verificou neste trabalho.

Contudo, esta análise foi criticada por um outro perito na área da geologia, argumentando não se poder estabelecer uma relação direta entre as concentrações químicas do solo e a produtividade da videira, uma vez que existem vários constituintes no solo, como por exemplo a água disponível, que não se integraram como variáveis neste estudo, e que, sabe-se que influenciam a capacidade de absorção da planta destes nutrientes. Desta forma, as concentrações que vão influenciar a produtividade da videira não são as que existem no solo mas sim as que a planta absorve. Devido a este facto os resultados desta análise podem não ser coerentes com o conhecimento da realidade e, desta forma, esta análise foi excluída do trabalho, encontrando-se apenas uma curta apresentação em anexo (Anexo V).

De seguida, surgiu uma segunda ideia com o intuito de se encontrar e analisar padrões espaciais relativamente às variáveis do solo de cada quinta, iniciou-se uma Análise Estatística Espacial. Todavia, devido ao facto de se tratarem de parcelas com pouca área e com poucas observações, verificou-se que não existia uma grande variação das concentrações dos constituintes ao longo do terreno e, como tal, não havia a possibilidade de analisar/estimar concentrações em locais sem medição e, assim, obter modelos de padrões espaciais.

Para a concretização do objetivo principal desta tese foram, por fim, usadas técnicas de Estatística Multivariada em conjunto com alguns processos de Inferência Estatística (em particular, testes de hipóteses não paramétricos) de modo a lidar com o problema da não normalidade dos dados e do número reduzido de observações.

A Análise de *Clusters* realizada a partir das variáveis originais do solo permitiu obter um agrupamento dos pontos de amostragem em três grupos homogéneos, relativamente às características do solo em ambas as quintas (Minho e Galiza). Com a aplicação do teste não paramétrico de *Kruskal-Wallis* verificou-se que as variáveis do solo da quinta do Minho MO, FF, FG, $P2O5$, Mg, AzT, Ni, Cr, B e CTC apresentaram diferenças significativas entre os diferentes *clusters*. Estas diferenças foram ainda detetadas para as variáveis ácido málico do mosto e para a família de compostos $FL1M$ do mosto. Relativamente à quinta da Galiza, as variáveis do solo MO, $P2O5$, $K2O$, Ca, Mg, AzT, Cd, N, B e CTC são as que apresentaram diferenças significativas entre os três *clusters*. Foram, ainda, encontradas estas diferenças para a variável acidez total do mosto, para as famílias dos compostos $FL1U$, $FL5U$, $FG1U$ e $FG7U$ das uvas e para a variável representativa da nota final atribuída ao vinho. O teste de comparações múltiplas não paramétrico permitiu, ainda,

avaliar em que *clusters* é que a variação destas variáveis foi significativa.

Por último, a produtividade da videira, a qualidade das uvas, do mosto e do vinho foram avaliadas através das suas correlações com as variáveis do solo, identificando-se as zonas da parcela que apresentam um défice/excesso nas concentrações dessas variáveis.

Paralelamente a este estudo, foi aplicada uma Análise de *Clusters* a partir dos fatores resultantes de uma Análise Fatorial, permitindo, assim, a redução da dimensionalidade do problema tendo em consideração as variáveis do solo analisadas. A aplicação desta técnica, aos dados de ambas as quintas, resultou em quatro fatores os quais explicaram aproximadamente 72% da variabilidade total dos dados originais da quinta do Minho e explicaram aproximadamente $85,5\%$ da variabilidade total dos dados originais da quinta da Galiza.

Os resultados obtidos a partir desta Análise de *Clusters* não expuseram uma clara estrutura dos grupos subjacentes aos dados da quinta da Galiza e, portanto, como se obteve resultados com baixo grau de estabilidade ou seja, não muito fiáveis, optou-se por eliminar este estudo do trabalho. Relativamente à quinta do Minho, a Análise de *Clusters* realizada a partir dos fatores anteriormente selecionados, bem como os testes de comparações múltiplas, evidenciaram resultados semelhantes aos obtidos a partir das variáveis originais do solo.

Assim, este trabalho evidenciou um conjunto de indicadores (muitos deles já conhecidos, resultantes de estudos já realizados) 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, assim, para uma maior produtividade da videira e uma maior qualidade da uva e do vinho. Note-se que ao longo deste trabalho foi indispensável optar por determinadas decisões, nomeadamente, a redução do número de observações a partir de cálculos das médias amostrais, ou mesmo a duplicação de observações a partir do método do vizinho mais próximo, de modo a tornar possível a aplicação destes métodos estatísticos, reduzindo, no entanto, a fiabilidade do estudo.

A maior parte destes problemas e limitações encontrados ao longo deste trabalho estão relacionados com o enviesamento do processo de amostragem, não estando adequado para sustentar uma boa base de dados e, consequentemente, bons resultados. Por de trás deste processo está um plano de amostragem intencional: foram escolhidos os pontos de amostragem do solo de modo a formarem uma grelha que revestisse toda a área do terreno. Para além disso, a seleção dos pontos de amostragem para se avaliarem várias variáveis relativas à planta, ao seu fruto e consequentemente ao seu vinho, foram escassos, o que levou a que não houvesse dados suficientes para se aplicarem, mais adequadamente, as metodologias estatísticas.

Foram vários os fatores que levaram a estes resultados, mais precisamente o medíocre processo de amostragem devido à falta de financiamento para se recolherem boas amostras. Contudo,

é de salientar o esforço que os peritos nas áreas em estudo fizeram em ajudar a contornar tais obstáculos, apresentando sempre as suas opiniões sobre o assunto.

5.1 Trabalho Futuro

No decorrer deste estudo foram muitas as limitações encontradas que impossibilitaram a aplicação de diversas metodologias e técnicas estatísticas, que seriam de maior interesse para o desenvolvimento e valorização deste trabalho.

Uma vez que estas limitações devem-se, essencialmente, à base de dados com a qual se trabalhou, como trabalho futuro sugere-se:

- a revisão do processo de amostragem mais adequado para que haja uma melhor recolha das amostras, com um número mais elevado de observações e idêntico para cada variável em estudo, de maneira a que se consigam aplicar metodologias estatísticas, nomeadamente, a Análise de Regressão, obtendo-se, assim, mais resultados e com um grau de fiabilidade maior;
- a introdução no estudo da variável água disponível no solo, uma vez que se trata de um elemento que influencia de forma significativa a absorção dos nutrientes por parte das plantas, podendo, assim, prejudicar ou melhorar a produtividade da videira e a qualidade das uvas, do mosto e do vinho, consoante a sua disponibilidade no solo. Para além disso, seria interessante realizar-se uma Análise de Regressão Múltipla, interagindo esta variável explicativa com cada uma das restantes variáveis do solo, ou seja, construir e validar modelos estatísticos que permitissem descrever a relação existente entre as variáveis solo-vinha-vinho;
- a recolha e estudo de variáveis referentes à climatologia (precipitação, exposição solar, etc.), visto que a sua interação com o fator solo resultam na produção de vinhos de qualidade e, portanto, uma boa avaliação destes fatores em conjunto ajudaria a aumentar e a melhorar a rentabilidade dos sistemas produtivos e a minimizar a degradação do recurso do solo, para garantir a sustentabilidade dos sistemas de produção vitivinícola;
- a realização do mesmo propósito do presente trabalho mas complementado com uma análise foliar, uma vez que as concentrações dos elemento químicos do solo não serão as disponíveis na planta porque as raízes apenas vão absorver parte destes nutrientes, e, portanto, este estudo poderá dar resultados e correlações ainda mais interessantes e credíveis das obtidas no presente trabalho;
- a realização deste estudo em vários anos para assim se poder tirar conclusões fidedignas e até mesmo perceber-se melhor a evolução da qualidade do fruto da videira e, consequente-

mente, do vinho, caso se altere as concentrações de alguns constituintes do solo, a partir do processo de fertilização, consoante as sugestões obtidas a partir de resultados obtidos em estudos anteriores;

- que se efetue este mesmo estudo numa parcela com uma maior área e com mais observações, de modo a que se consiga realizar uma Análise Estatística Espacial para se estudar a distribuição espacial das concentrações de todas as variáveis em estudo, procurando-se, assim, estabelecer padrões espaciais;
- a realização de um estudo mais pormenorizado sobre a influência do níquel (Ni) e do crómio (Cr) na produtividade da videira, visto terem sido apontados como variáveis significativas, quando realizada uma Análise de Regressão Linear simples e não serem conhecidos estudos idênticos ao presente trabalho sobre estes dois elementos.

Bibliografia

[1] Afonso, J. (2009). O Solo da Vinha. *Revista de Vinhos*, http://www.revistadevinhos.pt.

[2] 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.

[3] Araújo, I. (2010). *Estudo da parcela Santiago de Ruiz*. Estudo desenvolvido no âmbito do projeto Prospeg, AGROcontrol, Vila Verde.

[4] Armada, N. (1990). *Caracterização dos Solos da Estação Vitivinícola Amãndio Galhano e sua Relação com a Vinha*. Relatório de Estágio da Licenciatura em Engenharia Agrícola. Universidade de Trás os Montes e Alto Douro, Vila Real.

[5] Athayde, E. (2011). *Estatística R*. Universidade do Minho. Departamento de Matemática, Braga.

[6] Azevedo, J. e Oliveira, J. (2010a). *Cartografia geológica da envolvente Bodega Santiago Ruiz*. Projeto AGROcontrol, Vila Verde.

[7] Azevedo, J. e Oliveira, J. (2010b). *Cartografia geológica da envolvente Quinta Campos de Lima*. Projeto AGROcontrol, Vila Verde.

[8] Bartholomew, D. e Knott, M. (1999). *Kendalls Library of Statistics 7: Latent Variable Models and Factor Analysis*. Hodder Arnold, UK, 2ª edição.

[9] Borges, E. P. (2008). *ABC Ilustrado da Vinha e do Vinho*. MAUAD Editora, Rio de Janeiro, 8ª edição.

[10] Branco, A. J. (2004). *Uma Introdução à Análise de Clusters*. Sociedade Portuguesa de Estatística, Évora.

[11] Brown, B., Hendrix, S., Hedges, D. e Smith, T. (2012). *Multivariate Analysis for the Biobehavioral and Social Sciences - A Graphical Approach*. Wiley, New Jersey.

[12] Cadima, J. (2010). *Apontamentos de Estatística Multivariada, Departamento de Matemática*, Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Lisboa.

[13] Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika*, 18, 23.

[14] Caten, A. (2008). *Aplicação de Componentes Principais e Regressões Logísticas Múltiplas em Sistema de Informações Geográficas para Predição e o Mapeamento Digital de Solos*. Tese de Mestrado em Ciências do Solo, Universidade Federal de Santa Maria, Brasil.

[15] Chatfield, C. e Collins, A. J. (1980). *Introduction to Multivariate Analysis*. Chapman and Hall, Cambridge.

[16] Conover, W. J. (1980). *Pratical Nonparametric Statistics*. Wiley, New York, 2ª edição.

[17] Corporation, I. (2011). *IBM SPSS Statistics para Windows*. IBM Corpotation - SPSS, New York.

[18] Dagnelie, P. (1973). Estatística - Teoria e Métodos, volume 2. *Publicações Europa-América*, São Paulo.

[19] Dubes, R. C. e Jain, A. K. (1988). *Algorithms for Clustering Data*. Prentice Hall, New Jersey.

[20] Everitt, B. e Hothorn, T. (2011). An Introduction to Applied Multivariate Analysis with R. Springer, USA.

[21] Failla, O., Scienza, A., e Brancadoro, L. (1996). Effects of nutrient spray applications on malic and tartaric acid levels in grapevine berry. *Journal Plant Nutrition*, 19, 41-50.

[22] Flores, C. A. (2011). *Influência do solo na tipicidade do vinho*. InfoBibos, Organiação de Eventos Científicos, http://www.infobibos.com.

[23] Friel, C. (2003). *Notes on factor analysis*. Criminal Justice Center: Sam Houston State University, Texas.

[24] Hair, J. F., Anderson, R. E., Tatham, R. L. e Black, W. C. (1995). *Multivariate Data - Analysis with readings*. Prentice Hall, New Jersey, 4ª edição.

[25] Hakstian, A. R., Rogers, W. T. e Cattell, R. B. (1982). The Behavior of Numbers of factor Rules with Simulated Data. *Multivariate Behavioral Research*, 17, 192-219.

[26] Hendrickson, A. E. e White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Mathematical and Statistical Psychology*, 17, 65-70.

[27] Higgins, J. H. (2004). *Introduction to Modern Nonparametric Statistics*. Thomson, Toronto.

[28] Jennrich, R. I. e Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika*, 31, 313-323.

[29] Jobson, J. D. (1992). *Applied Multivariate Data Analysis: Categorical and Multivariate Methods*. Springer, USA.

[30] Johnson, R. A. e Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Prentice Hall, New Jersey, 6ª edição.

[31] Jollife, I. T. (2002). *Principal Component Analysis*. Springer, USA, 2ª edição.

[32] Jordão, A. J. (2007). *Gestão do Solo na Vinha*. Texto elaborado no âmbito do Plano de Acção para a Vitivinicultura da Alta Estremadura.

[33] Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187- 200.

[34] Layon, D. M., Cass, A. e Hansen, D. (2004). The effect of soil properties on vine performance. CSIRO Land and Water, *Technical Report*, 34, Australia.

[35] Levene, H. (1960). *Robust tests for equality of variances*, in I. Olkin (ed.). Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, Standford University Press, Standford, 278-292.

[36] Maciel, A. (2005). *A Pertinência dos Estudos de Microclimatologia para a Prevenção dos Riscos Climáticos num Vinhedo do Entre Douro e Minho*. Tese de Mestrado em Gestão dos Riscos Naturais, Faculdade de Letras da Universidade do Porto, Porto.

[37] Melo, G., Basso, A., Furini, G., Bortoli, L., Lopes, A. e Brunetto, G. (2008). Efeitos de diferentes níveis de boro no crescimento da videira em dois solos. *Resumos do XII Congresso Brasileiro de Viticultura e Enologia - Anais*. Embrapa, Brasil.

[38] Milton, P. L. B. (2003). *Análise de Estatística de dados geológicos*. UNESP, São Paulo, 2ª edição.

[39] 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, Lisboa.

[40] Mota, T. e Garrido, J. (2001). *Implementação da Vinha. Castas, Porta enxertos, Sistemas de condução e Plantação*. Manual técnico, CVRVV-EVAG, Arcos de Valdevez.

[41] Murteira, B., Ribeiro, C. A., Silva, J. A. e Pimenta, C. (2002). *Introdução à Estatística*. McGraw-Hill, Lisboa.

[42] Norman, G. e Streiner, D. (2000). *Biostatistics: the bare essentials*. B. C. Decker, London, 2ª edição.

[43] Pacheco, C. (1999). *Contribuição para o estudo da fertilização da vinha: influência da fertilização azotada, fosfatada e potássica na produção e na qualidade dos mostos da casca loureiro na região demarcada dos vinhos verdes*. Tese de Doutoramento em Engenharia Agronómica. Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Lisboa.

[44] Pearson, E. e Hartley, H. (1972). *Biometrika Tables for Statisticians*, volume 2. Cambridge University Press, Cambridge.

[45] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Versão 2.15.2. R Foundation for Statistical Computing, Austria. http//www.R-project.org.

[46] Reis, E. (2001). *Estatística Mutivariada Aplicada*. Edições Sílabo, Lisboa, 2ª edição.

[47] Reis, E., Melo, P., Andrade, R. e Calapez, T. (1997). *Estatística aplicada*, volume 2. Edições Sílabo, Lisboa.

[48] Shapiro, S. e Francia, R. (1972). Approximate analysis of variance test for normality. *Journal of the American Statistical Association* 67, 215-216.

[49] Siegel, S. (1975). *Estatística não-paramétrica (Para as Ciências do Comportamento)*. McGraw-Hill, São Paulo.

[50] Siegel, S. e Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 2ª edição.

[51] Silva, A. (2011). *Análise estatística multivariada no estudo da relação de variá veis de um solo residual granítico com a cultura da vinha, Caso da casta vinhão*. Tese de Mestrado em Estatística de Sistemas, Departamento de Produção e Sistemas, Escola de Engenharia da Universidade do Minho, Guimarães.

[52] Siqueira, H., Lima, L., Santana, M., Silva, J. e Silva, E. (2008). Aplicação de cloreto de cálcio em pré-colheita na conservação da uva Vênus. *Resumos do XII Congresso Brasileiro de Viticultura e Enologia - Anais*. Embrapa, Brasil.

[53] Somers, T. C. (1977). *Le rapport entre les teneurs en potasse de la vendange et al qualité relative des vins rouges australiens*. Symposium Internacional sur la Qualité de la Vendange, OIV, África do Sul.

[54] Spiegel, M. R. (1993). *Estatística*. McGraw-Hill, São Paulo, 2ª edição.

[55] Statistical Package for the Social Sciences (SPSS) (2011). Versão 20.0. [Computador *software*]. IBM SPSS Estatistics, Chicago.

[56] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236-244.

[57] Zaballa, O., Garcia-Escudeiro, E., Chavarri, J. B., Medrano, H. e Arroyo, M. C. (1997). Influence of vine irrigation (Vitis vinifera L.) on potassium nutrition. III Simpósio Internacional de Nutrição Mineral de folha caduca Árvores de Fruto, *Acta Horticulturae* 448, 49-73.

Anexos A

Anexo I

Concentrações dos compostos voláteis do aroma do mosto da casta Vinhão

Tabela A.1: Concentrações médias (μg/l) dos compostos voláteis da fração livre do aroma do mosto da casta Vinhão em função do $4-nonanol$.

	B1			B2			B3		
	N	C	P	N	C	P	N	C	P
Compostos em C_6									
(E)-2-hexenal	100,43	90,26	116,26	50,01	88,23	35,17	64,69	128,53	108,28
1-hexanol	458,43	568,61	441,97	580,85	573,04	618,59	622,26	505,43	450,98
(Z)-3-hexeno-1-ol	10,24	13,41	11,78	13,31	11,72	14,11	14,92	12,39	11,50
(E)-2-hexeno-1-ol	103,29	38,90	112,40	56,85	112,29	65,86	193,41	131,59	183,74
(Z)-2-hexeno-1-ol	12,48	10,69	11,44	19,20	10,03	7,83	13,35	10,11	13,52
TOTAL	**684,87**	**721,88**	**693,85**	**720,22**	**795,31**	**741,55**	**908,62**	**788,05**	**768,01**
Álcoois									
3-metil-3-buteno-1-ol	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
(Z)-2-penten-1-ol	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
3-metil-2-buteno-1-ol	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1-octeno-3-ol	3,91	6,96	6,04	5,48	5,76	7,05	6,05	6,00	4,10
álcool benzílico	13,01	12,88	22,42	22,59	20,69	31,13	19,12	21,28	19,37
2-feniletanol	57,25	164,38	75,24	246,86	93,56	139,57	100,89	76,78	55,07
2-fenoxietanol	0,00	1,59	0,00	2,61	0,00	1,37	0,00	0,00	1,21
TOTAL	**74,17**	**185,81**	**103,71**	**277,54**	**120,00**	**179,12**	**126,06**	**104,06**	**79,76**
Álcoois monoterpénicos									
linalol	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
4-terpineol	43,70	78,19	28,67	57,78	38,73	39,12	96,77	67,94	37,10
nerol	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
geraniol	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,89
TOTAL	**43,70**	**78,19**	**28,67**	**57,78**	**38,73**	**39,12**	**96,77**	**67,94**	**39,99**
Fenóis voláteis									
salicilato de metilo	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
vanilina	0,92	2,16	1,22	1,28	0,22	2,44	1,90	1,20	1,08
acetovanilona	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
zingerona	0,48	1,31	1,09	1,11	0,00	2,39	1,01	1,02	0,96
TOTAL	**1,39**	**3,47**	**2,31**	**2,39**	**0,22**	**4,83**	**2,91**	**2,22**	**2,04**
Compostos carbonilados									
benzaldeído	19,14	42,98	25,18	17,88	24,00	15,44	5,39	30,77	15,21
feniletanal	0,00	0,00	0,00	0,00	1,13	0,00	0,00	3,04	0,00
TOTAL	**19,14**	**42,98**	**25,18**	**17,88**	**25,14**	**15,44**	**5,39**	**33,81**	**15,21**

Tabela A.3: Concentrações médias (μg/l) dos compostos voláteis da fração glicosilada do aroma do mosto da casta Vinhão em função do $4 - nonanol$.

	B1			B2			B3		
	N	C	P	N	C	P	N	C	P
Composto em C_6									
1-hexanol	40,53	45,04	45,17	16,27	56,42	42,17	37,70	41,47	47,54
(Z)-3-hexeno-1-ol	3,20	3,62	4,15	1,17	3,76	3,51	2,97	2,91	3,65
(E)-2-hexeno-1-ol	20,56	21,74	24,86	7,00	22,90	20,84	15,93	18,00	24,58
TOTAL	**64,30**	**70,39**	**74,18**	**24,43**	**83,07**	**66,53**	**56,60**	**62,38**	**75,78**
Álcoois									
3-metil-3-buteno-1-ol	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
3-metil-2-buteno-1-ol	5,87	3,71	3,87	0,00	4,09	2,65	2,85	3,18	3,75
1-octeno-3-ol	1,83	1,80	2,10	1,70	2,90	2,05	1,30	4,53	0,62
1-octanol	2,65	2,34	2,01	0,00	4,89	2,22	1,89	3,05	2,83
1-feniletanol	2,19	1,96	1,38	0,00	2,41	1,73	1,43	3,60	2,47
álcool benzílico	135,47	112,28	115,35	43,02	128,88	73,99	95,69	110,20	102,41
2-feniletanol	128,10	121,94	97,10	39,05	125,15	70,05	104,63	125,76	88,71
TOTAL	**276,11**	**244,03**	**221,80**	**83,77**	**268,32**	**152,69**	**207,79**	**250,32**	**200,80**
Álcoois monoterpénicos									
linalol	0,96	3,16	0,86	1,01	1,86	1,43	1,42	1,63	0,80
4-terpineol	2,82	8,42	2,08	1,51	6,20	1,81	4,03	8,67	2,83
?-terpineol	0,56	1,50	0,48	0,33	0,93	0,69	0,94	1,39	0,64
nerol	0,80	1,02	0,83	0,00	1,39	0,69	0,99	1,37	0,91
geraniol	6,14	5,90	4,19	2,67	7,26	4,33	4,95	8,89	5,45
TOTAL	**11,27**	**19,98**	**8,44**	**5,52**	**17,64**	**8,96**	**12,33**	**21,96**	**10,62**
Óxidos e dióis monoterpénicos									
óxido furânico de linalol, trans-	1,61	4,23	1,64	1,92	3,18	1,77	2,34	3,34	2,11
óxido furânico de linalol, cis-	6,76	12,17	6,75	4,70	12,94	4,59	7,14	14,77	8,12
óxido pirânico de linalol, trans-	0,70	1,63	0,61	0,41	1,50	0,56	0,60	2,05	0,90
óxido pirânico de linalol, cis-	1,06	2,92	0,78	0,59	2,49	0,67	1,13	3,86	1,12
(E)-8-hidroxilinalol	1,85	2,55	1,60	0,00	2,86	1,49	1,81	4,51	1,70
(Z)-8-hidroxilinalol	3,35	3,27	2,37	1,18	5,46	2,72	2,52	5,78	3,27
ácido gerânico	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
TOTAL	**15,33**	**26,77**	**13,76**	**8,80**	**28,42**	**11,79**	**15,53**	**34,31**	**17,23**
Norisoprenóides em C_{13}									
3,4-dihidro-3-oxo-actinidol I	1,44	2,40	1,01	0,57	1,03	0,89	1,38	3,88	1,38
3,4-dihidro-3-oxo-actinidol II	1,10	1,17	0,60	0,57	0,78	0,92	1,30	2,53	1,12
3,4-dihidro-3-oxo-actinidol III	1,50	2,73	1,40	0,80	2,33	1,55	2,22	5,14	1,88
3-hidroxi-β-damascona	10,14	15,32	5,18	6,50	12,28	8,96	8,59	30,50	9,61
3-oxo-α-ionol	3,98	5,92	3,24	1,79	4,16	3,14	5,45	12,42	4,68
3-hidroxi-7,8-dihidro-β-ionol	1,93	3,69	1,87	0,00	3,24	1,66	2,20	9,04	2,44
4-oxo-7,8-dihidro-β-ionol	1,96	2,89	1,80	0,55	1,84	1,72	1,97	6,37	2,16
3-oxo-7,8-dihidro-α-ionol	4,20	5,97	3,72	1,87	4,30	5,36	5,10	12,20	5,00
3-hidroxi-7,8-dehidro-β-ionol	3,18	4,93	2,96	0,00	2,79	2,09	2,84	10,06	2,67
vomifoliol	5,94	10,16	2,59	1,62	4,03	3,52	4,64	18,21	5,47
TOTAL	**35,37**	**55,17**	**24,37**	**14,26**	**36,78**	**29,82**	**35,69**	**110,33**	**36,41**
Fenóis voláteis									
salicilato de metilo	14,94	6,27	11,27	5,70	17,58	9,86	10,86	17,17	17,17
guaiacol	1,89	1,53	1,05	0,87	1,97	1,94	1,63	2,57	2,00
4-vinilguaiacol	13,52	7,10	2,99	4,81	3,18	4,31	13,08	12,12	12,87
4-vinilfenol	5,45	2,13	1,09	0,53	1,95	1,42	6,32	5,28	3,66
vanilina	2,99	3,41	2,28	0,00	2,76	1,81	2,35	5,65	2,67
vanilato de metilo	1,83	0,83	0,00	0,00	1,16	1,21	1,13	2,61	1,48
acetovanilona	7,09	4,01	3,71	0,94	4,92	4,63	6,00	5,46	4,75
3,4-dimetoxifenol	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
zingerona	2,85	2,15	1,53	0,37	1,96	1,68	2,15	3,33	2,28
álcool 3,4,5-trimetoxibenzílico	1,90	1,37	1,31	0,83	1,27	1,07	1,39	2,48	1,05
2,5-dihidroxibenzoato de metilo	14,82	11,37	10,81	1,03	13,99	4,43	16,11	54,25	8,74
3,4,5-trimetoxifenol	3,21	3,93	1,05	0,51	2,62	1,05	2,11	5,42	1,90
TOTAL	**70,49**	**44,11**	**37,09**	**15,59**	**53,36**	**33,42**	**63,15**	**116,34**	**58,55**
Compostos carbonilados									
benzaldeído	3,95	4,11	5,63	2,69	4,85	2,83	2,54	2,31	2,96
TOTAL	**3,95**	**4,11**	**5,63**	**2,69**	**4,85**	**2,83**	**2,54**	**2,31**	**2,96**

Anexos B

Anexo II

Concentrações dos compostos voláteis do aroma das uvas da casta Alvarinho

Tabela B.1: Concentrações médias (μg/l) dos compostos voláteis da fração livre do aroma das uvas da casta Alvarinho.

	BSL					
	1-2	3-4	5-6	7-8	9-10	11-12
Compostos em C_6						
(E)-2-hexenal	65,64	76,06	91,04	111,37	101,18	64,62
1-hexanol	53,66	33,42	38,78	29,14	46,67	35,63
(Z)-3-hexeno-1-ol	14,42	13,00	15,45	11,91	11,15	9,81
(E)-2-hexeno-1-ol	35,81	22,73	41,90	24,30	16,93	15,77
(Z)-2-hexeno-1-ol	1,73	1,40	2,17	1,30	1,50	1,00
TOTAL	**171,25**	**146,62**	**98,30**	**178,01**	**177,43**	**126,83**
Álcoois						
3-metil-3-buteno-1-ol	7,90	4,93	7,20	5,53	6,44	6,11
(Z)-2-penten-1-ol	1,83	0,00	9,97	6,44	6,93	7,38
3-metil-2-buteno-1-ol	6,50	0,00	0,00	0,00	0,00	0,00
1-octeno-3-ol	0,44	2,47	2,47	2,20	2,66	1,78
1-octanol	0,58	0,43	0,63	0,55	0,53	0,45
álcool benzílico	26,50	26,23	22,22	15,28	15,16	16,43
2-feniletanol	42,04	24,92	29,73	23,10	24,69	22,29
2-fenoxietanol	0,43	0,47	0,58	0,61	0,90	0,99
TOTAL	**86,23**	**59,45**	**72,80**	**53,71**	**57,32**	**55,42**
Álcoois monoterpénicos						
linalol	0,90	0,99	0,97	0,90	1,01	0,90
citronelol	0,53	0,28	0,76	0,89	1,13	0,77
nerol	0,99	1,41	1,16	1,14	1,54	1,28
geraniol	9,90	10,69	9,20	7,90	9,58	7,49
TOTAL	**12,32**	**13,36**	**12,10**	**10,83**	**13,26**	**10,44**
Óxidos e dióis monoterpénicos						
óxido pirânico de linalol, trans-	5,80	5,43	10,00	7,20	14,63	4,37
óxido pirânico de linalol, cis-	0,57	0,63	0,63	0,57	1,20	0,53
3,7-dimetilocta-1,5-dieno-3,7diol	4,48	1,88	2,80	1,28	2,07	0,55
3,7-dimetilocta-1,7-dieno-3,6diol	0,00	0,00	0,00	0,00	0,00	0,00
(E)-8-hidroxilinalol	0,00	0,00	0,00	0,00	0,00	0,00
(Z)-8-hidroxilinalol	2,90	0,90	0,80	0,73	1,27	0,43
ácido gerânico	0,00	0,00	0,00	0,00	0,00	0,00
TOTAL	**13,75**	**8,85**	**14,24**	**9,78**	**19,17**	**5,88**
Fenóis Voláteis						
salicilato de metilo	0,00	0,00	0,00	0,88	0,00	0,00
eugenol	0,74	0,97	0,73	0,90	1,17	1,13
vanilina	4,07	2,99	2,96	2,11	2,63	2,00
vanilato de metilo	1,58	0,44	1,06	0,74	0,99	1,02
acetovanilona	0,00	0,06	0,00	0,00	0,00	0,00
TOTAL	**6,40**	**4,45**	**4,76**	**4,63**	**4,78**	**4,15**
Ácidos gordos voláteis						
Ácido Hexanóico	4,30	18,44	2,38	4,43	5,49	1,91
TOTAL	**4,30**	**18,44**	**2,38**	**4,43**	**5,49**	**1,91**
Compostos carbonilados						
benzaldeído	3,26	2,34	2,76	1,88	1,88	1,91
feniletanal	46,46	24,36	30,37	35,52	26,24	23,94
TOTAL	**49,72**	**26,71**	**33,14**	**37,41**	**28,12**	**25,85**

Tabela B.3: Concentrações médias (μg/l) dos compostos voláteis da fração glicosilada do aroma das uvas da casta Alvarinho.

	BSL					
	1-2	3-4	5-6	7-8	9-10	11-12
Compostos em C_6						
1-hexanol	3,16	4,32	4,89	3,08	3,18	3,48
(Z)-3-hexeno-1-ol	0,54	1,03	1,11	0,57	0,68	0,62
(E)-2-hexeno-1-ol	0,51	1,03	1,00	0,62	0,52	0,59
TOTAL	**4,21**	**6,39**	**6,99**	**4,27**	**4,39**	**4,68**
Álcoois						
3-metil-3-buteno-1-ol	0,00	0,00	0,00	0,00	0,00	0,00
3-metil-2-buteno-1-ol	0,90	1,48	0,99	0,74	0,86	0,91
1-octeno-3-ol	0,48	0,48	0,48	0,33	0,24	0,32
1-octanol	0,42	0,32	0,74	0,74	0,45	0,39
1-feniletanol	0,56	0,70	0,65	0,73	0,61	0,76
álcool benzílico	13,84	45,14	31,86	27,12	24,64	25,45
2-feniletanol	15,43	22,38	21,76	22,53	19,06	19,31
TOTAL	**31,64**	**70,50**	**56,48**	**52,19**	**45,86**	**47,14**
Álcoois monoterpénicos						
linalol	12,22	10,88	13,48	23,19	13,59	9,49
HO-trienol	1,15	1,55	1,80	1,93	1,46	1,06
?-terpineol	0,43	0,51	0,60	0,80	0,59	0,52
citronelol	0,00	0,00	0,34	0,34	0,33	0,29
nerol	1,06	1,45	1,58	2,05	1,29	1,49
geraniol	5,21	6,11	6,68	10,87	6,85	6,88
TOTAL	**20,07**	**20,50**	**24,47**	**39,17**	**24,11**	**19,73**
Óxidos e dióis monoterpénicos						
óxido furânico de linalol, trans-	11,64	15,71	15,35	16,22	14,35	11,84
óxido furânico de linalol, cis-	9,84	9,98	8,42	10,06	12,91	8,87
óxido pirânico de linalol, trans-	6,50	8,07	7,62	9,13	6,81	5,28
óxido pirânico de linalol, cis-	3,24	2,90	2,03	2,37	2,53	2,33
3,7-dimetilocta-1,5-dieno-3,7-diol	20,48	25,89	23,80	28,40	21,09	19,34
hidrato de linalol	0,25	0,29	0,33	0,00	0,11	0,00
3,7-dimetilocta-1,7-dieno-3,6-diol	0,76	0,93	0,98	1,83	0,85	0,82
8-hidroxi-6,7-dihidrolinalol	1,59	1,96	1,23	2,00	1,14	1,69
(E)-8-hidroxilinalol	4,37	5,23	4,62	6,38	4,94	4,76
(Z)-8-hidroxilinalol	11,19	14,45	16,77	23,30	13,72	12,40
ácido gerânico	0,00	0,00	0,73	1,84	0,87	1,54
TOTAL	**69,85**	**85,41**	**81,87**	**101,53**	**79,31**	**68,86**
Norisoprenóides em C_{613}						
3,4-dihidro-3-oxo-actinidol I	0,64	1,00	0,63	0,95	0,71	0,73
3,4-dihidro-3-oxo-actinidol II	0,48	0,64	0,67	0,80	0,58	0,53
3,4-dihidro-3-oxo-actinidol III	0,74	0,90	0,88	1,02	0,78	0,71
3-hidroxi-β-damascona	5,63	0,00	0,79	0,98	1,02	0,00
3,4-dihidro-3-oxo-actinidol IV	0,65	7,39	6,36	7,07	5,54	5,99
megastigma-7-ene-3,9-diol	0,65	1,05	0,92	1,44	0,97	0,85
3-oxo-α-ionol	6,42	8,95	7,82	12,76	8,47	7,81
3-hidroxi-7,8-dihidro-β-ionol	0,42	0,57	0,47	1,05	0,66	0,70
4-oxo-7,8-dihidro-β-ionol	0,57	0,95	0,67	1,45	0,91	0,70
3-oxo-7,8-dihidro-α-ionol	2,27	3,90	3,62	5,87	3,56	4,19
3-hidroxi-7,8-dehidro-β-ionol	2,25	3,22	2,75	3,18	2,58	2,88
vomifoliol	3,14	4,31	2,97	5,77	4,73	5,27
TOTAL	**23,86**	**32,88**	**28,55**	**42,33**	**30,51**	**30,35**
Fenóis voláteis						
salicilato de metilo	0,76	4,04	2,56	2,39	1,28	1,11
guaiacol	0,22	0,25	0,23	0,45	0,35	0,26
eugenol	1,52	2,58	2,27	1,60	1,21	1,73
4-vinilguaiacol	0,99	1,35	1,07	6,53	3,43	15,61
4-vinilfenol	0,31	0,00	0,60	1,94	1,55	1,37
vanilato de metilo	5,41	6,57	7,17	10,03	5,71	6,78
acetovanilona	0,72	1,24	0,88	1,46	0,82	1,22
3,4-dimetoxifenol	0,00	0,00	0,98	0,00	0,00	0,00
zingerona	0,16	0,24	0,34	0,51	0,43	0,45
álcool 3,4,5-trimetoxibenzílico	0,34	0,71	0,53	0,78	0,61	0,85
2,5-dihidroxibenzoato de metilo	1,55	5,06	3,30	2,79	0,99	3,57
3,4,5-trimetoxifenol	0,66	2,06	1,75	1,34	0,90	1,17
TOTAL	**12,64**	**24,10**	**21,68**	**29,83**	**17,28**	**34,12**
Compostos carbonilados						
benzaldeído	0,58	1,02	1,09	0,73	0,65	0,70
TOTAL	**0,58**	**1,02**	**1,09**	**0,73**	**0,65**	**0,70**

Anexos C

Anexo III

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

Figura C.1: Ficha de Prova descritiva dos vinhos.

Anexos D

Anexo IV

Scree Plots

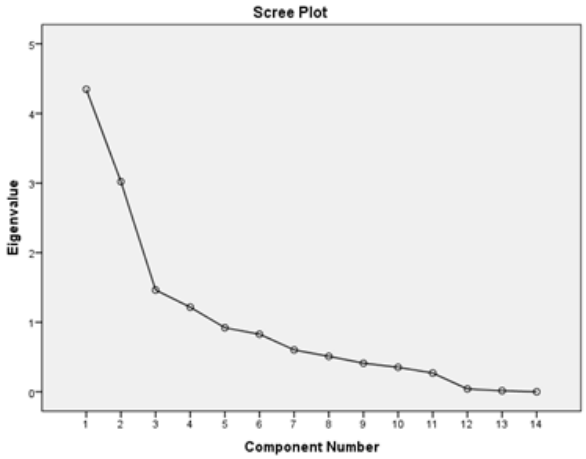


Figura D.1: *Scree Plot* associado aos dados do Minho.

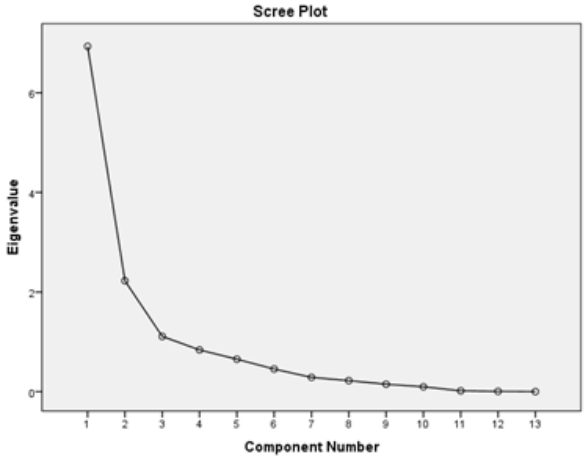


Figura D.2: *Scree Plot* associado aos dados da Galiza.

Anexos E

Anexo V

Análise de Regressão Linear Simples

Em diversas áreas da Ciência surge frequentemente a necessidade de explicar como determinada(s) variável(eis) irá(ão) afetar outra(s), ou seja, a necessidade de investigar a natureza das relações entre fenómenos.

A análise de regressão e a análise de correlação permitem estudar tais relações que possam existir na população: a análise de correlação mede a intensidade da relação linear entre duas variáveis e a análise de regressão permite, através de uma equação, explicar como uma variável pode ser estimada (ou predita) a partir de outra(s).

Modelo de Regressão Linear Simples

Para qualquer estudo de análise de regressão linear simples é necessário ter uma amostra de duas variáveis de n observações,

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n). \tag{E.1}$$

Inicialmente uma representação gráfica dos dados, diagrama de dispersão, pode sugerir a existência ou não de correlação linear entre as duas variáveis. No caso da análise do diagrama de dispersão sugerir uma relação linear entre as variáveis e o valor de correlação entre as variáveis, r_{xy}, for elevado, pode-se ajustar uma reta de regressão (um modelo) ao dados que melhor traduza a relação entre estas duas variáveis. O modelo de regressão linear simples pode ser definido como

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \;\; i = 1, \ldots, n \tag{E.2}$$

com $\varepsilon_i \sim N(0, \sigma^2)$ e $cov(\varepsilon_i, \varepsilon_j) = 0$, $i, j = 1, \ldots, n$; $i \neq j$, onde:

- y_i : observação da variável resposta (ou variável dependente) no indivíduo i;
- x_i : observacão da variável independente no indivíduo i;
- β_0 : ordenada na origem (parâmetro desconhecido do modelo);
- β_1 : declive (parâmetro desconhecido do modelo);
- ε_i : erro aleatório associado à observação da resposta no indivíduo i.

Pressupostos do Modelo

É necessário que se verifique os seguintes pressupostos do modelo:

- $E(\varepsilon_i) = 0$, $i = 1, \ldots, n$,: o erro é uma variável aleatória de média 0, ou seja, espera-se que em média os valores de ε_i, com $i = 1, \ldots, n$, sejam nulos para todas as observações;
- $var(\varepsilon_i) = \sigma^2$, $i = 1, \ldots, n$, (homocedasticidade): o erro é uma variável aleatória com variância constante o que significa que a variância do erro é a mesma em todas as observações;
- ε_i's são variáveis aleatórias independentes, ou seja, $cov(\varepsilon_i, \varepsilon_j) = 0$ $(i \neq j; i, j = 1, \ldots, n)$;
- ε_i segue uma distribuição Normal, $\varepsilon_i \sim N(0, \sigma^2)$.

Método de Estimação de Parâmetros: Método dos Mínimos Quadrados (MMQ)

Um dos métodos mais utilizados para estimar os parâmetros do modelo é conhecido pelo Método dos Mínimos Quadrados (MMQ). Este método consiste em determinar os valores para os parâmetros da reta β_0 e β_1, de forma a minimizar a soma dos quadrados dos erros.

Interpretação dos Parametros Estimados

Encontrados e testados os coeficientes de regressão, procede-se à interpretação destes a fim de retirar conclusões à cerca do estudo. Observando a Figura E.1, verifica-se que β_0 (ordenada na origem da reta de regressão) representa o valor esperado de Y quando o valor da variável explicativa é nulo; β_1 (declive da reta) representa a variação do valor esperado de Y por cada incremento unitário na variável explicativa.

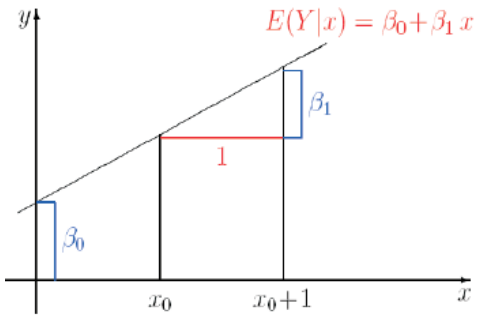


Figura E.1: Gráfico ilustrativo da interpretação dos coeficientes de regressão.

Avaliação da Qualidade do Modelo de Regressão

Após estimar os parâmetros do modelo é necessário avaliar a qualidade de ajustamento do modelo aos dados.

É usual utilizar-se o Coeficiente de Determinação, representado geralmente por R^2. Este coeficiente é dado pelo rácio $R^2 = SQR/SQT$ e pode ser interpretado como a proporção da variabilidade total em y que é explicada pela variável independente x. Desta forma, pode-se reescrever R^2 como

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}. \tag{E.3}$$

onde:

- SQR: Soma de Quadrados da Regressão;
- SQT: Soma de Quadrados Total;
- SQE: Soma de Quadrados dos Resíduos.

Adicionalmente, pode ser mostrado que

$$R^2 = [cor(y, x)]^2. \tag{E.4}$$

Portanto, se $R^2 = 0$ significa que o modelo não é adequado aos dados; se pelo contrário se tem $R^2 = 1$ o ajustamento é perfeito.

Análise dos Resíduos

A análise de um modelo de regressão é completada com a realização de uma Análise de Resíduos. Os resíduos $\hat{\varepsilon}_i$, com $i = 1, 2, \ldots, n$, podem ser utilizados para detetar possíveis violações dos pressupostos do modelo, assim como, detetar a existência de algumas observações que poderão ser responsáveis por um ajustamento menos bom.

A verificação dos pressupostos do modelo é feita de modo geral com recurso a representações gráficas, envolvendo também um teste à normalidade dos resíduos, o teste de *Shapiro-Wilk*.

Análise e Discussão dos Resultados Obtidos

Relativamente à aplicação da análise de regressão linear aos dados em estudo, devido ao número muito reduzido de observações que a maioria das variáveis apresentaram, apenas foi possível realizar-se este tipo de análise com duas variáveis resposta referentes à produtividade da videira: $Pbruto$ e $Pcacho$.

Após a análise realizada com os dados relativos à quinta da Galiza, e de acordo com a interpretação e discussão dos resultados obtidos com a ajuda de peritos nas áreas em estudo, concluiu-se que não faziam qualquer sentido. Tais resultados podem ser consequência do facto deste estudo ser enviesado, uma vez que a análise referida foi realizada com um total de 12 observações.

Por este motivo, as Tabelas E.1 e E.2 apenas apresentam os resultados obtidos na análise de regressão linear simples efetuada aos dados da quinta do Minho, cuja dimensão da amostra era de 45 observações. Note-se também que no decorrer desta análise, alguns pontos da amostra identificados como pontos influentes, *outliers* e/ou *high-leverage points* foram eliminados com o objetivo de melhorar a qualidade de ajustamento do modelo.

Tabela E.1: Resultados significativos (nível de significância de 0,05) obtidos na Análise de Regressão Simples, considerando como variável dependente o $Pbruto$.

Variável	Observações	Regressão Linear Simples				Shapiro.Wilk
Explicativa	Retiradas	Estimativas	$\hat{\sigma}$	Valor-prova	R^2	Valor-prova
pH	1 10 15 23 31 32 39	$\hat{\beta}_0$ = -16,53	8,45	0,06	0,21	0,27
		$\hat{\beta}_1$ = 4,34	1,47	0,01		
$P2O5$	1 10 11 16 29 32	$\hat{\beta}_0$ = 4,75	1,04	5,80e-05	0,30	0,56
		$\hat{\beta}_1$ = 0,09	0,02	5,00e-04		
Ca	1 23 32 41 42	$\hat{\beta}_0$ = 3,56	1.37	0,01	0,30	0,47
		$\hat{\beta}_1$ = 0,01	0.00	0,00		
Mg	1 20 32	$\hat{\beta}_0$ = -2,48	2,28	0,28	0,40	0,47
		$\hat{\beta}_1$ = 0,15	0,03	1,82e-05		
Ni	3 16 32	$\hat{\beta}_0$ = 13,58	1,32	2,02e-12	0,27	0,30
		$\hat{\beta}_1$ = -1,21	0,33	0,00		
Cr	16 17 32	$\hat{\beta}_0$ = 5,92	1,03	1,31e-06	0,21	0,42
		$\hat{\beta}_1$ = 6,50	2,09	0,00		
Cd	1 10 15 29 32	$\hat{\beta}_0$ = 1,43	1,97	0,47	0,27	0,40
		$\hat{\beta}_1$ = 64,94	18,10	0,00		
N	10 15 29 41	$\hat{\beta}_0$ = -1,60	2,75	0,56	0,29	0,43
		$\hat{\beta}_1$ = 8,02	2,10	0,00		
CTC	1 23 32 42	$\hat{\beta}_0$ = 2,66	1,54	0,09	0,31	0,36
		$\hat{\beta}_1$ = 0,54	0,13	0,00		

Tabela E.2: Resultados significativos (nível de significância de 0,05) obtidos na Análise de Regressão Simples, considerando como variável dependente o $Pcacho$.

Variável	Observações	Regressão Linear Simples				Shapiro.Wilk
Explicativa	Retiradas	Estimativas	$\hat{\sigma}$	Valor-prova	R^2	Valor-prova
$P2O5$	1 10 11 24 29	$\hat{\beta}_0$ = 0,14	0,014	7,13e-13	0,14	0,43
		$\hat{\beta}_1$ = 0,00	0,00	0,02		
Ca	1 23 24 41 42	$\hat{\beta}_0$ = 1,16e-01	1,47e-02	2,69e-09	0,32	0,08
		$\hat{\beta}_1$ = 1,22e-04	3,03e-05	0,00		
Mg	1 20 24	$\hat{\beta}_0$ = 0.06	0,03	0,03	0,35	0,16
		$\hat{\beta}_1$ = 0,00	0,00	7,41e-05		
Cd	1 10 15 24 29	$\hat{\beta}_0$ = 0,08	0,02	0,00	0,36	0,97
		$\hat{\beta}_1$ = 0,82	0,18	8,87e-05		
N	1 10 24 41	$\hat{\beta}_0$ = 0,11	0,02	5,31e-05	0,14	0,18
		$\hat{\beta}_1$ = 0,04	0,02	0,02		
CTC	1 10 24 32 36	$\hat{\beta}_0$ = 0,11	0,01	1,69e-11	0,51	0,26
		$\hat{\beta}_1$ = 0,01	0,00	7,84e-07		

Pela observação da Tabela E.1 verifica-se que:

- por cada unidade de pH, em média o $Pbruto$ aumenta $4,342\ kg$;

- por cada unidade de $P2O5\ (ug/g)$, em média o $Pbruto$ aumenta $94\ g$;

- por cada unidade de $Ca\ (ug/g)$, em média o $Pbruto$ aumenta $11\ g$;

- por cada unidade de $Mg\ (ug/g)$, em média o $Pbruto$ aumenta $154\ g$;

- por cada unidade de $Ni\ (ug/g)$, em média o $Pbruto$ diminui $1,2093\ kg$;

- por cada unidade de $Cr\ (ug/g)$, em média o $Pbruto$ aumenta $6,498\ kg$;

- por cada unidade de $Cd\ (ug/g)$, em média o $Pbruto$ aumenta $64,942\ kg$;

- por cada unidade de $N\ (ug/g)$, em média o $Pbruto$ aumenta $8,023\ kg$;

- por cada unidade de $CTC\ (m.e./100g)$, em média o $Pbruto$ aumenta $536\ g$.

Observa-se ainda que o coeficiente de determinação varia entre $0,21$ e $0,40$, o que indica que os modelos se ajustam razoavelmente aos dados.

Pela observação da Tabela E.2 verifica-se que:

- por cada unidade de $P2O5\ (ug/g)$, em média o $Pcacho$ aumenta $0,7453\ g$;

- por cada unidade de $Ca\ (ug/g)$, em média o $Pcacho$ aumenta $1,22\ g$;

- por cada unidade de $Mg\ (ug/g)$, em média o $Pcacho$ aumenta $2\ g$;

- por cada unidade de $Cr\ (ug/g)$, em média o $Pcacho$ aumenta $83\ g$;

- por cada unidade de $N\ (ug/g)$ em média o $Pcacho$ aumenta $44\ g$;

- por cada unidade de $CTC\ (m.e./100g)$, em média o $Pcacho$ aumenta $6\ g$.

Observa-se ainda que o coeficiente de determinação varia entre $0,14$ e $0,51$, o que indica que os modelos se ajustam razoavelmente aos dados.

De salientar que todo este processo foi acompanhado com uma Análise de Resíduos, verificando-se sempre os pressupostos exigidos para a realizaçãoo desta análise.