UNIVERSIDADE D
COIMBRA

Ana Patrícia Cruz de Matos

Aplicação de Ferramentas Quimiométricas para Categorização de vinhos através dos seus componentes Químicos

Dissertação no âmbito do Mestrado em Química - Controlo da Qualidade e Ambiente, orientada pelo Doutor Fábio Antonio Schaberle e Professor Doutor Alberto António Caria Canelas Pais e apresentada ao Departamento de Química da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Julho de 2021

Faculdade de Ciências e Tecnologia
da Universidade de Coimbra

Aplicação de Ferramentas Quimiométricas para Categorização de Vinhos através dos seus Componentes Químicos

Ana Patrícia Cruz de Matos

Dissertação no âmbito do Mestrado em Química - Controle da Qualidade e Ambiente, orientada pelo Doutor Fábio Antonio Schaberle e pelo Professor Doutor Alberto António Caria Canelas Pais e apresentada ao Departamento de Química da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Julho de 2021

1 2 9 0

UNIVERSIDADE D
COIMBRA

Agradecimentos

Prestes a terminar mais uma etapa na Universidade de Coimbra, não podia deixar de agradecer a todos os que me apoiaram e incentivaram a dar mais e melhor, nunca me deixando desistir. Assim sendo, gostaria de agradecer:

Aos meus orientadores Professor Doutor Alberto Canelas Pais e Doutor Fábio Schaberle pela possibilidade de realizar este trabalho e acima de tudo por todo o apoio, conhecimento transmitido, disponibilidade, ajuda e paciência durante a sua execução.

Aos amigos e colegas de curso que me acompanharam ao longo destes 5 anos, em especial à Catarina Marques, à Carolina Melo, ao Tiago Vicente, à Adelaide Pessoa, ao Leonardo, ao Iúri e ao Alexandre por todos os momentos passados e todo o apoio.

À Andreia Melo por ter sido um dos grandes pilares ao longo destes anos e nunca me ter feito arrepender da minha escolha. És sem dúvida uma das pessoas que quero guardar sempre.

À Soraia Assis por ter confiado nas minhas capacidades e me ter deixado fazer parte de uma equipa única e motivada, que deixa bastantes saudades. Os momentos passados contigo serão sempre recordados.

Aos meus afilhados de praxe, Ana, Diogo, Pedro e Daniela, vocês foram as pessoas que mais me deram ao longo destes anos, tive oportunidade de vos dar a conhecer Coimbra e acima de tudo de vos ver crescer. A maior sorte a todos vocês.

Às minhas meninas do 171 com quem tive a oportunidade de viver ao longo destes anos e sem as quais já não sei viver. Obrigada por todas as conversas, desabafos e apoio. Definitivamente sem vocês o curso não teria sido tão fácil de se fazer.

Aos meus meninos de Viseu, em especial à Inês Antunes, ao David e ao Pedro, pela vossa amizade e pela paciência que tiveram ao longo de todos estes anos quando recusava um café com vocês por falta de tempo por causa da universidade.

Ao meu namorado por ter paciência e acima de tudo coragem de lidar com o meu mau feitio todos os dias. Mas acima de tudo obrigada por teres estado sempre presente quando precisei e por toda a ajuda que me deste. Sem ti nada disto tinha sido igual.

À minha família por todo o apoio, dedicação e por nunca me deixarem desistir. Este percurso não foi só meu também foi vosso. Um obrigado não chega para vos agradecer tudo o que fizeram e vão continuar a fazer por mim.

A todos aqueles com quem tive a oportunidade de me cruzar ao longo deste tempo por Coimbra. E a ti Coimbra por teres sido mais do que aquilo que ambicionava.

"Uma vez Coimbra, para sempre Saudade!"

Índice

Índice de Figuras

Índice de Tabelas

Abreviaturas

a.C Antes de Cristo

AU Approximately Unbiased - Probabilidade aproximadamente imparcial

AD Average Distance - Distância média

ADM Average Distance between Means - Distância média entre médias

APN Average Proportion of Non-overlap - Proporção média de não sobreposição

BP Bootstrap Probability - Probabilidade de Bootstrap

D Índice de Dunn

d.C Depois de Cristo

FOM Figure of Merit - Figura de Mérito

H Estatística de Hopkins

HCA Hierarchical Clustering Analysis - Análise Hierárquica de Agrupamento

PC Principal Component - Componente Principal

PCA Principal Component Analysis - Análise de Componentes Principais

$\mathbf{S_i}$ Largura de Silhueta média

vmp Vizinho mais próximo

Resumo

O vinho é uma bebida que existe há vários milénios. Nas últimas décadas houve um grande aumento no mercado do vinho aumentando o número de produtores e a diversidade de tipos de vinho, resultado das diferenças químicas e físicas dos solos e clima das regiões. Muitas dessas características químicas dos vinhos são colocadas em bases de dados, sendo útil para se perceber a relação que existe entre os vinhos e as suas regiões de origem.

O uso de métodos quimiométricos, como a Análise de Componentes Principais (PCA) e a Análise Hierárquica de Agrupamento (HCA), é fundamental para a análise de bases de dados multivariáveis, isto é, que apresentam múltiplas variáveis, sendo no caso dos vinhos as características químicas presentes. Conjuntamente com o PCA e HCA, é importante o desenvolvimento de métodos para identificação e eliminação de pontos que não façam parte de nenhum grupo, denominados *outliers,* para permitir uma análise de PCA e HCA mais apurada.

O trabalho presente nesta dissertação teve como objetivo a análise multivariada de correlação (estudo de PCA) e similaridade (estudo de HCA) de uma base de dados composta por 178 vinhos provenientes da mesma região de Itália possuindo três tipos de cultivo distintos, assim como o desenvolvimento de um método para limpeza de dados (*outliers*). O interesse principal foi verificar se as características químicas dos vinhos são suficientes para que se conseguisse correlacionar um determinado vinho com o cultivo aplicado.

Os resultados obtidos mostraram que como a utilização das técnicas de PCA, HCA foi possível determinar a formação de grupos pertencentes a um mesmo método de cultivo e que a implementação da eliminação dos *outliers* melhorou a qualidade dos dados na identificação de grupos, sem que alterasse significativamente os resultados finais quando comparados com a análise sem eliminação de *outliers* para o estudo de PCA e de HCA com todas as variáveis, havendo alterações significativas nos resultados quando comparado com o estudo de HCA com as cinco dimensões obtidas do estudo de PCA.

Palavras-Chave: Vinhos, Características Químicas, HCA, PCA, Eliminação do Ruído, Validação

Abstract

Wine is a beverage that has existed for several millenia. In the last decades there was a significant increase in the wine market, which lead to an increase in the number of producers and in the type of wine diversity, resulting from the physical and chemical differences between soils and the production region weather. Plenty of the chemical characteristics of the wine are placed on databases, being that this information is useful to understand the correlation between the wines and their production regions.

The usage of chemometric methods, such as Principal Component Analysis (PCA study) and Hierarchical Cluster Analysis (HCA study), is essential for the analysis of multivariable databases, that is, those that present multiple variables and, in the case of wines the chemical characteristics. Alongside with PCA and HCA, it is important to develop methods to identify and eliminate datapoints that do not belong to any data group, the denominated outliers, to allow a more focused analysis of PCA and HCA.

The work presented in this dissertation aims to perform a multivariable correlation (PCA) and similarity (HCA) analysis in a database composed by 178 wines from the same region of Italy using three distinct culture methods, as well as the development of a data cleaning (outliers) method. The main interest of this work was to verify if the wine chemical characteristics are sufficient to correlate a wine with the applicated culture method.

The obtained results demonstrated that the performance of the PCA and HCA technics allow the formation of groups belonging to the same culture method and that the implementation of noise (outliers) elimination improved the data quality in the identification of groups, without significantly altering the final results when compared to the results obtained for the analysis without outlier removal for the PCA and HCA studies with all the variables. There was however significant variations in the results when compared to those obtained with the HCA study performed with the five variables obtained from the PCA study.

Keywords: Wine, Chemical Characteristics, HCA, PCA, Noise Elimination, Validation

Capítulo 1 - Vinhos

1.1 - História do Vinho

Embora existam algumas dúvidas em relação ao surgimento do vinho, acredita-se que este já exista há alguns milénios, devido à sua associação com a religião com as inúmeras referências bíblicas, o que leva a que se acredite na sua origem primitiva. É também muito associado aos deuses gregos, sendo mesmo considerado o néctar dos deuses, e para além disso existem também documentos de origem egípcia do ano de 2500 antes de Cristo (a.C.) que dão conta do cultivo e produção de vinho. [1,2,3]

Devido às deslocações marítimas e colonizações dos diferentes povos, o vinho viria a ser conhecido um pouco por todo o mundo, perdurando até aos dias de hoje em diferentes regiões e apresentando uma grande variedade de tipos. Relativamente ao cultivo de videiras pela primeira vez na Península Ibérica, pensa-se que tenha ocorrido por volta dos 2000 anos a.C., por mão dos Tartessos. Mais tarde durante o século VII a.C., os Gregos desenvolveram a viticultura na Península Ibérica, dando um destaque principal à produção do vinho. Foram os Celtas, no século VI a.C. que para além de terem transportado consigo diversas variedades de videiras desenvolveram técnicas para armazenamento dos vinhos, como por exemplo o fabrico de barris [1,3]. Atualmente existem diversas variedades de vinho, o que depende, essencialmente, do tipo de casta usada. No entanto supõe-se que a maioria das castas tenham origem na mesma, a *Vitis vinifera*, (Figura 1.1) diferindo entre si pela cor, tamanho, formato da baga, composição do sumo, tempo de maturação e resistência que apresentam para doenças. [1,4,5]

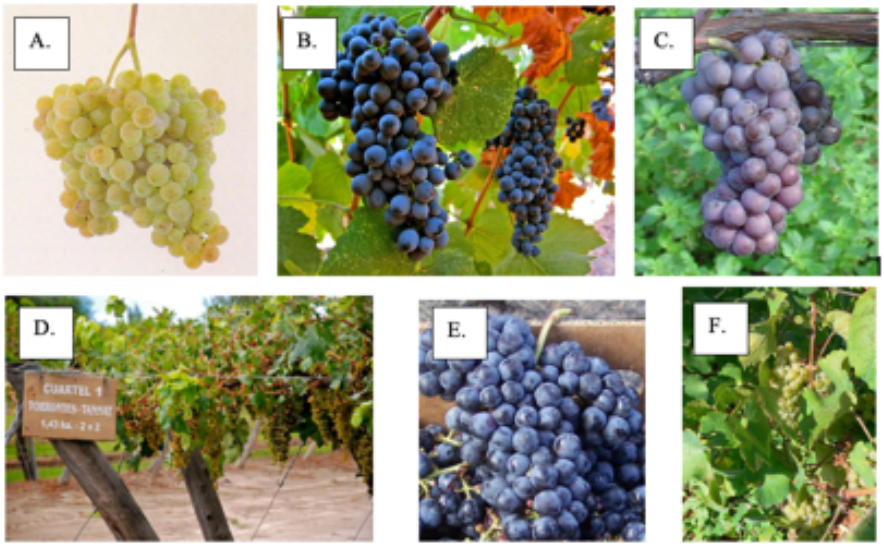


Figura 1.1: Diversas variedades da *Vitis vinifera*. A) Alvarinho, Portugal; B) Touriga nacional, Portugal; C)Pinot gris, França; D) Torrontés, Espanha; E) Sangiovese, Itália; F) Malvasia, Grécia.

No ano 15 a.C. os romanos invadiam a Península Ibérica, o que fez com que houvesse uma modernização do vinho nesta zona através do aumento da variedade de castas existentes bem como com o

aperfeiçoamento das técnicas usadas para o cultivo destas. Tal acontecimento levou a que o cultivo nesta região aumentasse dada a necessidade de enviar vinho para Roma, que já não conseguia responder à procura de vinhos naquela região.[1,6,7] Anos mais tarde, por volta do século VI depois de Cristo (d.C.) viria a acontecer outro marco importante para a produção do vinho com a expansão do Cristianismo, que levou a que o vinho tivesse uma expansão devido à sua associação ao "culto religioso", dado ter-se tornado indispensável no ato da comunhão. [1]

Durante o século XII o vinho era o principal produto exportado, no entanto existem documentos que comprovam a importância deste e da vinha em Portugal ainda antes de 1143. Existindo nessa altura diversas zonas de cultivo de vinhas que mais tarde com a conquista de D. Afonso Henriques viriam a ser expandidas, na qual a maioria dos produtores de vinha oriundos do clero, pois o vinho nessa altura era muito associado a cerimónias religiosas. Devido ao elevado cultivo de vinhas, havia uma elevada produção de vinhos, o que foi responsável pelo aumento da exportação no século XIV. Também os descobrimentos levaram a que se tenham conhecido novos métodos de conservação do vinho, tais como o seu envelhecimento. [1,8,9]

Em Portugal continental e Ilhas, existem atualmente 14 regiões que são grandes produtoras de vinhos (Figura 1.2), tendo sido a região do Porto e Douro demarcada no século XVIII por ordem de Marquês de Pombal. No século XIX, a região do Douro foi atacada por uma praga, que rapidamente se alastrou ao resto das regiões vínicas, ficando apenas a região de Colares (Sintra) a salvo desta praga, devido ao cultivo em terrenos de areia. [1,9]

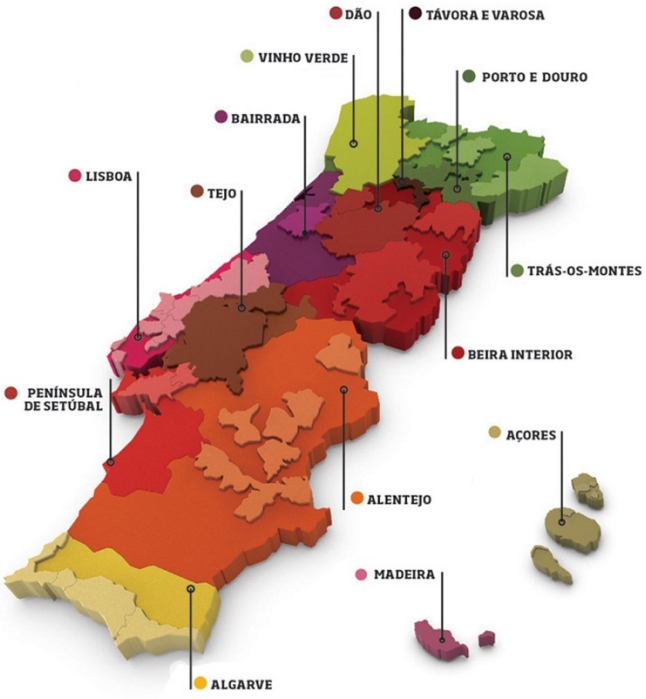


Figura 1.2: Regiões Vinícolas Portuguesas. (retirado de [10])

1.1.1 - Produção de Vinhos

Antes de se começar a produzir um vinho, é necessário saber-se qual o tipo de vinho que se pretende obter, pois tal como é possível observar-se na Figura 1.3 consoante o tipo de vinho que se quer produzir: branco, rosé ou tinto, as etapas usadas para a produção dos vinhos são diferentes. [11]

No entanto, como o vinho já existe há alguns milénios, o processo de produção tem vindo a sofrer alterações, desde o tipo de cultivo usado nas vinhas, a vindima, a prensagem das uvas e até mesmo o armazenamento do vinho.[1,3] Sendo importante salientar que o processo de produção confere diferentes características químicas ao vinho, sendo que as alterações que ocorrem neste processo por interferência natural ou do ser humano vão influenciar essas características, podendo-se assim obter um vinho mais ou menos doce, um vinho com mais ou menos cor, entre outras.

Tendo em conta os passos presentes no fluxograma da Figura 1.3, vai-se apresentar brevemente o que acontece em cada passo.

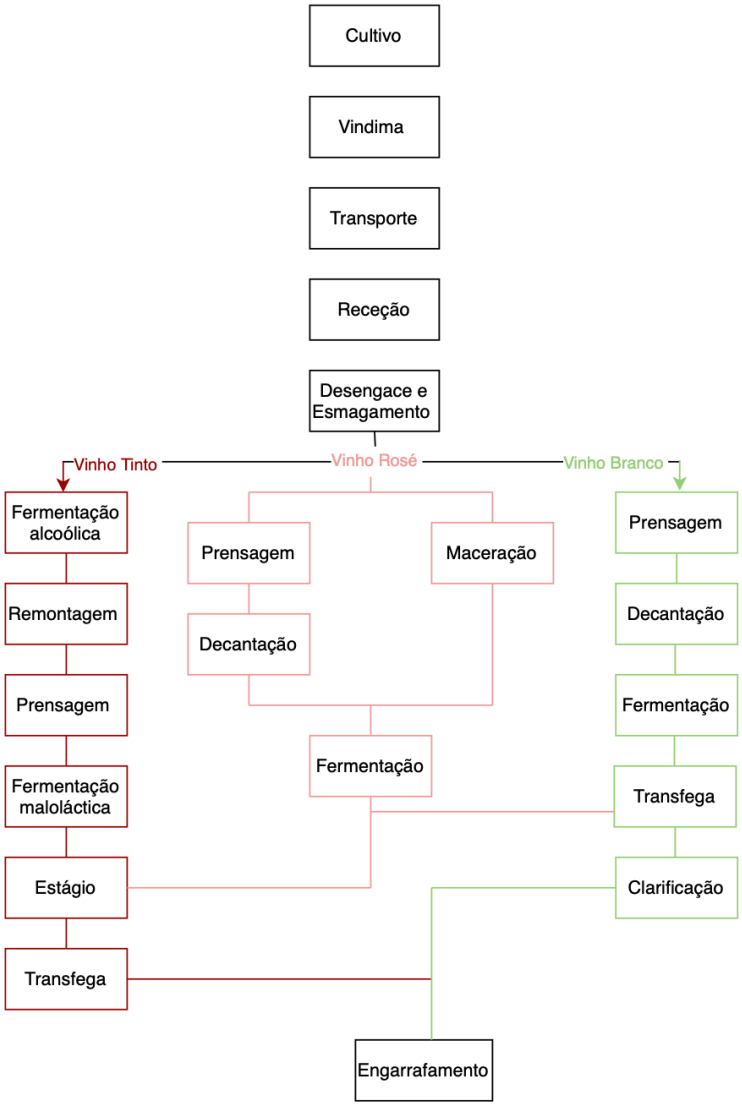


Figura 1.3: Fluxograma da Produção de Vinho: Branco, Rosé e Tinto (adaptado de [11])

1.1.1.1 - Cultivo das Vinhas

O primeiro passo para a produção de vinho consiste na plantação das videiras, sendo estas cultivadas em diversas condições climatéricas, que vão levar a que ocorram alterações químicas nas uvas. Habitualmente, as videiras são plantadas em zonas temperadas, onde a temperatura mínima no Inverno não é inferior a -7 °C e a temperatura média durante a Primavera é de 10 °C, para que seja possível haver o crescimento da videira e da uva. Já para que a maturação da uva ocorra é necessário que a soma de calor (isto é, soma das temperaturas dos dias em que a temperatura média de um dia de Primavera é mais do que 10 °C) seja igual ou superior a 1800 °C, pois quando esta soma é menor que esse valor vamos ter no final açúcar insuficiente e muita acidez no vinho. [3,12] Na Figura 1.4, encontra-se esquematizado o ciclo de formação das uvas.

Conforme existem diversos fatores que influenciam a composição do vinho, vamos ter também fatores que influenciam a maturação, como é o caso da exposição solar, da drenagem do ar, da temperatura do solo e do teor da humidade. Também, a composição do solo vai ter influência na maturação, dado que para além de influenciar a temperatura do solo (uma das principais causas de existir maturação das uvas) vai ter efeito na penetração das raízes, na capacidade de retenção de água e consequentemente na nutrição da videira, sendo que todos estes fatores vão levar a que haja alterações na qualidade do vinho. [1,4,12]

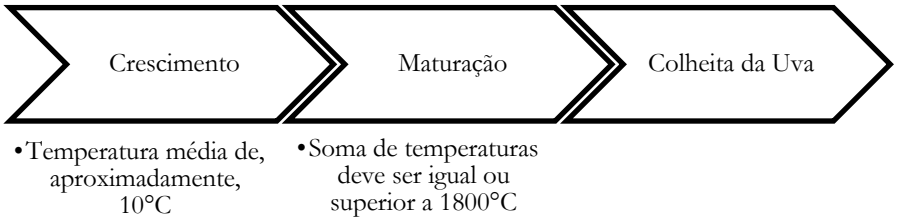


Figura 1.4: Esquema do ciclo de formação da Uva

1.1.1.2 - Vindima

A segunda etapa para que seja possível produzir o vinho é a vindima, que consiste na colheita das uvas. É dada preferências às uvas que se encontram frescas e totalmente maduras, uma vez que é nesta altura que apresentam as melhores características para a produção do vinho. Esta preferência ocorre essencialmente devido ao facto de que, caso as uvas não se encontrem completamente maduras, os vinhos que delas resultem vão possuir um baixo teor de álcool e quando estas já se encontram numa fase de sobre maturação, o vinho resultante vai apresentar um elevado teor de álcool e um baixo teor de ácido, devido à fermentação. [4,13]

1.1.1.3 - Transporte e Receção das Uvas

O transporte das uvas requer um cuidado redobrado, pois de acordo com Fischer et al. [14] e outros autores, as uvas possuem uma carga de bactérias, fungos e vírus que se pode desenvolver consoante as condições em que se encontram, o que leva a que as uvas tenham de permanecer intactas durante o transporte, para evitar o desenvolvimento de micro-organismos. [15,16] Também devido ao tempo quente que

se faz sentir na altura das vindimas é fundamental manter as uvas intactas até chegarem à adega pois caso contrário devido ao calor podem-se desencadear reações de fermentação [11]

Após o transporte, as uvas chegam às adegas, onde irá ocorrer o resto do processo da produção do vinho.

1.1.1.4 - Desengace e Esmagamento das uvas

O primeiro passo após a receção das uvas, é proceder-se ao desengace das uvas, que não é mais do que a separação das uvas do caule. Tal procedimento é realizado, devido ao facto de este possuir uma tendência para amargar o vinho, o que é então indesejável. [17] Para além de eliminar o sabor desagradável que o caule confere ao vinho, este processo é fundamental, dado que vai facilitar a libertação do sumo, uma vez que ao se retirar o caule vai haver o rompimento das uvas. [18] Este processo pode ser feito de duas maneiras: com recurso a pessoas, sendo a retirada feita manualmente com as mãos ou com recurso a máquinas. Habitualmente existe a preferência pelo uso de pessoas, dado que assim é possível selecionar-se as uvas que apresentam mais qualidade e que apresentam as características desejadas. [17,18]

Após o desengace do caule, pode-se então passar ao esmagamento das uvas, sendo este processo pode, novamente, ser realizado por pessoas ou máquinas, existindo atualmente máquinas capazes de realizar os dois passos em simultâneo. Em grandes indústrias de vinho dá-se preferência ao uso de máquinas, mais concretamente trituradoras, no entanto, produtores particulares ou pequenas indústrias, acabam por usar métodos que já vêm sendo usados há muitos anos, como é o caso do esmagamento em lagares, usando para isso os pés. Este método ainda é usado por estes produtores, pois para além de possuir custos menores acredita-se que o vinho através deste processo adquire outras propriedades e qualidades. [3,19]

Desta etapa obtém-se o mosto das uvas, que não é mais do que a mistura do sumo, cascas e sementes das uvas. [19]

1.1.1.5 - Prensagem

A prensagem é o passo que se segue no caso de se querer produzir vinho branco pois caso o objetivo seja produzir vinho tinto é necessário proceder-se à fermentação alcoólica e à remontagem antes deste passo. Sendo que o vinho branco produzido poderá ser também usado em vinhos rosé, visto que este é a mistura entre vinho branco e vinho tinto.

Esta etapa de produção vai permitir a separação das cascas e das sementes do sumo sendo este passo fundamental para a produção de vinho branco, pois é na casca da uva que se encontram as substâncias responsáveis por fornecer cor ao vinho. [3,17,20] Este processo pode ser realizado através do recurso a prensas, sendo que a prensa de Willmes é característica para a produção de vinho branco. Muitas vezes são também usadas prensas pneumáticas, dado que estas permitem que o processo ocorra em diferentes pressões e não existe a destruição das grainhas, que possuem na sua constituição os taninos agressivos que são responsáveis pela alteração do sabor no vinho. [3,21,22] No final deste processo vamos ter o sumo, usado para a produção

de vinhos brancos e rosé, separados do resto. A massa final resultante deste passo pode ser posteriormente usada como destilador de aguardente vínica. [3]

1.1.1.6 - Decantação

A decantação apenas é realizada no caso de o objetivo final ser a produção de vinhos brancos. Isto deve-se essencialmente à realização do passo seguinte, a fermentação, pois apenas o mosto/sumo é fermentado. No caso dos mostos brancos estes são turvos, apresentando partículas em suspensão que caso não ocorra este processo vão dificultar a fermentação e levar a que os aromas do vinho sejam alterados. Este ocorre em cuba ou depósitos, onde os sólidos em suspensão se depositam no fundo. [3,11]

1.1.1.7 - Fermentação

A fermentação é uma das etapas mais importantes para a produção do vinho, pois é a partir desta etapa que o mosto/sumo passa a vinho. No entanto, é também o processo que mais controlo requer, visto que qualquer alteração indesejável leva a que o vinho resultante apresente características diferentes das esperadas. [1,3,23] Os principais requisitos que devem existir neste processo são: a supressão do crescimento de microrganismos indesejáveis, a presença de um número adequado de leveduras desejáveis, bem como uma nutrição adequada para o crescimento destas, o controlo da temperatura para a prevenção de calor excessivo e a prevenção de oxidação. [3]

Ainda dentro da fermentação, é possível dividir-se este processo em dois: a fermentação alcoólica e a fermentação malolática, sendo que a primeira ocorre em todos os tipos de vinho e a segunda ocorre apenas nos vinhos tintos. [23,24]

1.1.1.7.1 - Fermentação Alcoólica

Neste tipo de fermentação ocorre a conversão do açúcar em álcool e dióxido de carbono (CO_2) por ação de leveduras, e dá-se assim a passagem do mosto a vinho. Embora o mosto contenha leveduras, muitas vezes, de forma que seja possível controlar-se melhor o processo, recorre-se ao uso de leveduras externas, como é o caso da Saccharomyces. [3,21]

Tal como referido, um dos aspetos importantes na fermentação é o controlo da temperatura, dado que a sobrevivência das leveduras depende da temperatura a que estas se encontram, sendo que não deve ser inferior a 10 ° C nem superior a 30 °C. No entanto estas vão variar de acordo com o tipo de vinho que se pretende produzir [22,25]

A realização desta etapa nos vinhos tintos pode levar a que seja necessário realizar uma outra etapa, que é a remontagem. Deve-se ao facto de a transformação do açúcar originar CO_2, o que leva a que partículas presentes neste tipo de vinho flutuem, havendo assim a necessidade de se misturar. Este passo é bastante benéfico, dado que favorece a distribuição das leveduras.[11]

$$C_6H_{12}O_6 \xrightarrow{\textbf{Leveduras}} \mathrm{CO_2} + \mathrm{H_2O} + \mathrm{Etanol}$$

1.1.1.7.2 - Fermentação Malolática

A fermentação malolática tem como objetivo promover a estabilidade biológica do vinho, diminuir a sua acidez em torno de 0.1 a 0.3%, sendo fundamental para o sabor do vinho. Apesar disto, esta ocorre apenas em vinhos tintos, sempre após a fermentação alcoólica. [24,25] Este tipo de fermentação dá-se por bactérias lácticas, que são responsáveis pela transformação de ácido málico em ácido láctico. [23,24,25]

$$C_4H_6O_5 \xrightarrow[\text{Lácticas}]{\text{Bactérias}} CO_2 + C_3H_6O_3$$

1.1.1.8 - Maceração

A maceração ocorre quando é pretendido obter-se vinhos tintos para a produção do vinho rosé. Este processo consiste no contacto entre o mosto e as uvas durante o tempo que se desejar. Este passo vai ter implicações ao nível da cor e do sabor. Após o tempo de contacto desejado, separa-se a parte sólida do mosto e este último segue para o processo de fermentação. [11]

1.1.1.9 - Trasfega

A trasfega ocorre tanto para o vinho branco, como o tinto, e mesmo o rosé, no entanto, dependendo do tipo de vinho, vai ocorrer em fases diferentes. A trasfega consiste na passagem do vinho de uma cuba/barril para outro, havendo assim a separação entre o vinho e os resíduos acumulados no fundo. [11,26]

1.1.1.10 - Estágio

O estágio, também conhecido como envelhecimento, ocorre para os vinhos tintos. Estes são deixados a envelhecer em barris de carvalho, o que vai permitir que ao contactar com algum ar adquira características mais "amadeiradas", ganhando assim outro sabor. Durante este processo vai existir a diminuição da acidez do vinho, dando-se também a clarificação e estabilização. [3,11]

1.1.1.11 - Clarificação

Este processo é realizado para os vinhos brancos, dado que para os vinhos tintos este ocorre também durante o envelhecimento. Durante este processo há outros seis processos envolvidos: o refinamento, onde vai ocorrer a adição de substâncias que auxiliam no aclaramento do vinho; a filtração de células de leveduras e bactérias; a centrifugação; a refrigeração de modo a impedir a crescimento das leveduras e aumento do CO_2; a troca iónica que tornará o vinho mais solúvel e o aquecimento para precipitação das proteínas. [3,11]

1.1.1.12 - Engarrafamento

O último passo da produção de vinho é o engarrafamento. Este consiste na distribuição do vinho por garrafas, sendo posteriormente colocada uma rolha habitualmente de cortiça. No caso de grandes produtores de vinhos, é neste passo que se dá também a rotulagem e selagem da garrafa. [3,11]

1.2 - Diferentes Tipos de Vinhos

Os vinhos podem ser de diversos tipos, podendo haver aqueles que são resultado da fermentação do sumo de uva ou que são derivados de outras frutas. [3] Dentro dos vinhos resultantes das uvas, existem diversos tipos de vinho, tais como o vinho tinto, o vinho branco, o vinho verde, o vinho espumante, entre outros, diferindo entre si pela espécie de casta usada e também pelos processos de produção que sofrem, entre outros aspetos. Estes vinhos podem ser agrupados em 2 grandes grupos, os vinhos naturais, quando apresentam um teor de álcool entre os 9 e os 14 % e os vinhos de sobremesa que tem um teor de álcool entre os 15 e os 21%. [2] Os vinhos são rotulados tendo em conta a região em que são produzidos, a maturidade da fruta, a variedade da uva, bem como o tipo de vinho e o ano de colheita, sendo estas características que permitem diferenciá-los pela cor, doçura e aroma. [3]

1.3 - Características dos Vinhos

Para se proceder à classificação dos diferentes tipos de vinho é tido em conta quais os parâmetros necessários para se considerar que temos um vinho. Esses parâmetros são: conter álcool etílico, açúcar, ácidos, aldeídos, ésteres, aminoácidos, minerais, vitaminas, compostos aromatizantes, entre outros. Sendo importante salientar que estes devem obedecer apresentar valores dentro de intervalos definidos.[2]

Para além dos parâmetros necessários para que se possa perceber que estamos perante um vinho é também necessário entender-se quais as características químicas que alteram de vinho para vinho. No entanto, como existem diversas características que são responsáveis pelas alterações nos vinhos, vai-se fazer referência apenas às que se seguem.

1.3.1 - Ácido Málico

O ácido málico encontra-se presente nas uvas, mais concretamente na casca, sementes e caules. Os valores de ácido málico vão nos fornecer informações sobre a fermentação malolática, pois quanto menor for o seu valor maior a percentagem que foi transformada com a fermentação malolática. Além disso, vai ser possível também perceber-se qual o tipo de vinho, dado que os vinhos tintos vão possuir maiores quantidades, dado que a fermentação deste tipo de vinhos é feita sem que haja a separação do mosto da casca e sementes. Ao haver uma diminuição do ácido málico vamos ter alterações na acidez do vinho, devido à sua transformação em ácido lático. [23,24,27]

1.3.2 - Álcool

O álcool do vinho é formado na fermentação alcoólica quando se dá a passagem do açúcar a álcool, pelo que este dependerá da quantidade de açúcar presente nas uvas. Este constituinte do vinho contribui

para as sensações sensoriais do vinho, sendo que para além disso ainda influencia a viscosidade, o corpo do vinho e as perceções gustativas, de acidez, doçura, aroma e sabor. [28,29] No que diz respeito ao teor alcoólico dos vinhos este varia entre os 8.5 e 9.5% para os vinhos verdes, à volta dos 12% para vinhos brancos, entre os 11 e 14% para os vinhos tintos e entre os 19 e 20% para o vinho do Porto, podendo estes valores sofrer alterações de acordo com a região de cultivo.

1.3.3 - Cinza

A cinza, neste contexto, não é mais do que o conteúdo mineral presente nos mostos e respetivos vinhos, que resulta da combustão do extrato seco (resíduos de evaporação dos vinhos e mostos). [30] Nesse podemos encontrar cálcio, sais de magnésio, potássio, ácido sulfúrico, fosfórico, hidroclórico e carbónico. Para além de todos estes constituintes pode ainda haver vestígios de cloro, fluor, cobre, ferro e manganês. Os valores da cinza vão permitir perceber se o vinho é ou não autêntico, dado que caso estes sejam muito baixo ou muito altos é sinónimo de que não se trata de um vinho adulterado. [31,32]

Os valores que a cinza apresenta, em média, no vinho são de 1.2 a 3 g/L, enquanto nos mostos esse valor é de 3 a 5 g/L. [31] Relativamente ao tipo de vinho, existem valores mínimos tabelados pelo Instituto da Vinha e do Vinho, que para os vinhos brancos e rosados é de 1.6 g/L e para os tintos é de 1.8 g/L. [33] Estes valores diferem entre si devido aos processos de clarificação.

1.3.3.1- Alcalinidade da cinza

A alcalinidade da cinza vai fornecer informação relativamente à quantidade de ácidos orgânicos presentes no vinho na forma de sal. [30] Tal acontece devido à presença de ácido fracos, que na combustão são convertidos nos carbonatos correspondentes. [32]

Os valores da alcalinidade da cinza, vão permitir verificar a presença de ácidos na forma livre, mas também permitem verificar se existe ou não adulteração dos vinhos. [32,34]

1.3.4 - Diluição de Vinhos pelo coeficiente OD280/OD315

A diluição de vinhos usando o fator OD280/OD315, a densidade ótica das proteínas a 280 nm a dividir pela densidade ótica a 315 nm, permite que se determine a concentração de proteínas em vinhos diluídos. Isto vai ser útil para determinar o teor de proteína existente em cada vinho. [35]

1.3.5 - Total de fenóis

Os fenóis são compostos importantes, tanto para os vinhos tinto como para os vinhos brancos. Esta importância deve-se ao facto de estes serem responsáveis pela qualidade das uvas e vinhos. [36,37] Este tipo de compostos varia de acordo com diversos fatores, tais como, o clima, o cultivo, a fermentação, a natureza do solo, entre outros. [38]

A distribuição dos compostos fenólicos nos vinhos brancos e tintos é diferente, sendo apresentando os vinhos brancos menores quantidades. Essa diferença deve-se essencialmente ao facto de as uvas nos vinhos brancos não serem esmagadas juntamente com o caule, casca e grainhas, contrariamente ao que

acontece com o vinho tinto. [39] A principal função destes compostos é serem antioxidantes, sendo que a sua atividade está diretamente relacionada com a sua estrutura química. [37,40]

Os fenóis podem se dividir em dois grandes grupos: os fenóis flavonóides e os fenóis não flavonóides.

1.3.5.1 - Flavonóides

Os flavonóides constituem o maior grupo de compostos polifenólicos, estando estes diretamente associados às propriedades organoléticas do vinho, tais como o sabor, a cor e o aroma. [41,42] Dentro deste grupo destacam-se: os taninos hidrolisáveis e condensados (proantocianidinas) e as antocianinas. [37] Tal como a quantidade de fenóis, a quantidade de flavonóides varia de acordo com as condições ambientais e com o amadurecimento. [43]

A sua presença na uva ocorre, exclusivamente, nas partes sólidas (casca, grainhas e caule), sendo que a sua transferência para o mosto e vinho ocorre devido à maceração. É por este motivo que as quantidades nos vinhos brancos são inferiores as dos vinhos tintos, pois o processo de maceração ocorre apenas nos vinhos tintos. [44,45] No entanto, nos vinhos brancos quando existe um intervalo de tempo considerável entre a colheita e a prensagem das uvas a quantidade de flavonóides será maior do que quando esse intervalo é curto. [46]

1.3.5.2 - Não flavonóides

A este grupo de não flavonóides, vão pertencer todos os compostos polifenólicos que não se encontrem no grupo anterior, como é o caso dos ácidos hidroxibenzóicos e dos estilbenos. [41,42] Os não flavonóides encontram-se distribuídos pela casca e polpa das uvas, havendo uma diminuição das suas quantidades, há medida que existe o amadurecimento das uvas e também na fermentação. [47]

Este tipo de constituintes é conhecido por realçar e estabilizar a cor dos vinhos tintos, sendo que contribuem também para o sabor do vinho e, no caso do resveratrol, contribui para as atividades biológicas.[48]

1.3.6 - Cor

A cor é um parâmetro importante para a aceitação de um vinho, dado que esta permite avaliar a sua qualidade, visto que é o primeiro parâmetro sensorial. [49] Esta é influenciada pela composição fenólica, mais concretamente, pelas antocianinas, que vão variar de acordo com diversos fatores, sendo um deles o cultivo. Como este tipo de composto apresenta uma grande reatividade química com o oxigénio, a cor do vinho vai estar a sofrer alterações constantes. [49,50]

As antocianinas são o principal composto responsável pela cor do vinho, o que faz com que sejam causadoras das alterações de cor que ocorrem durante o envelhecimento. [51] A cor não se altera exclusivamente devido às antocianinas, as alterações ocorrem também devido às características químicas e físicas dos vinhos e as condições do meio, sendo estas condicionadas pelo pH e temperatura. [52,53]

Dentro da cor do vinho podem-se estudar dois parâmetros: a intensidade da cor e a tonalidade

1.3.6.1 - Intensidade da cor

Ao observar-se a intensidade do vinho é possível obterem-se informações relativamente ao corpo do vinho, dado que quanto mais escuro é mais encorpado será. É ainda possível, verificar qual a idade e a acidez do vinho, uma vez que quanto mais brilhante mais jovem e mais ácido é. [54,55]

A alteração da intensidade do vinho, deve-se à reatividade dos compostos fenólicos. [49] Para que não haja alterações na intensidade da cor e esta se preserve é necessário que existam mecanismos de estabilização, tais como a complexação das antocianinas com outras espécies e a conservação de antocianinas em pigmentos derivados mais estáveis. [56]

1.3.6.2 - Tonalidade

Em relação à tonalidade do vinho, esta está relacionada com a cor das uvas, sendo que pode sofrer alterações durante os processos de vinificação e o envelhecimento. Pelo tom do vinho, pode ser possível perceber-se quais os métodos usados e as uvas, dado que as castas possuem cores diferentes entre si. Esta informação permite assim antever quais as características que cada vinho possui a nível de sabor. [54]

1.3.7 - Magnésio

O magnésio presente na uva está diretamente relacionado com o magnésio existente no solo de cultivo. No entanto, este é também influenciado por todos os tratamentos que sejam aplicados às videiras e uvas, podendo o valor ser inferior ao esperado. [57]

Este composto vai estar mais presente em vinhos tintos, dado a sua produção uma vez que existe um maior contacto com a parte sólida da uva, local onde este se encontra. A presença deste nos vinhos tintos, vai ser fundamental para a fermentação malolática, visto que melhora o seu arranque e todo o processo. [58,59] Além disso, vai ainda ser fundamental para diminuir a precipitação do ácido tartárico que contribui para o equilíbrio gustativo do vinho. [59]

1.3.8 - Prolina

A prolina vai ser usada como indicador da maturação, dado que a quantidade desta aumenta brutalmente nesta fase devido a saturação de proteínas nas uvas, o que faz com que não haja a proteólise. No entanto, sendo um aminoácido, será uma fonte de nutrientes para as leveduras úteis para a fermentação.[59]

Esta é um dos constituintes do vinho que mais pode sofrer alteração de um ano para o outro, visto que esta está diretamente relacionada com as formas de cultivo usadas. [59]

1.3.9 - Taninos

Os taninos encontram-se na casca, semente e caule das uvas. São compostos fenólicos que são capazes de formar complexos insolúveis com hidratos de carbono e proteínas, levando à precipitação destes. Podendo ser de dois tipos: hidrolisáveis ou condensados, sendo os últimos os mais predominantes nas uvas e vinhos. [36,60]

A concentração destes compostos no vinho tende a diminuir com o envelhecimento, o que é devido aos processos de oxidação e precipitação que ocorrem. [61]

1.3.9.1 - Proantocianidinas

As proantocianidinas, são vulgarmente conhecidas como taninos condensados. Estas são compostos poliméricos que dão origem às antocianinas, composto principal pela coloração do vinho. [21] Quanto à quantidade destes compostos que é possível encontrar-se nas uvas, depende essencialmente da sua localização, sendo que a sua transferência para o vinho se dá devido aos processos usados na vinificação, como é o caso, do esmagamento, maceração e fermentação. [36,62]

Apesar da sua quantidade depender da sua localização, as proantocianidinas são responsáveis pelas características sensoriais. Desempenhando assim um papel importante no que diz respeito ao envelhecimento do vinho devido as suas capacidades de oxidar, condensar e polimerizar compostos. [35,63]

Capítulo 2 - Objetivo

O trabalho realizado teve como objeto um conjunto de dados quimiométricos, mais concretamente uma base de dados de vinhos. Esta base de dados é constituída por vinhos que diferem entre si pelo método de cultivo das vinhas, pelo que o principal objetivo é verificar se é possível realizar-se a categorização dos vinhos através dos seus componentes químicos recorrendo ao uso de ferramentas quimiométricas.

O principal objetivo desta dissertação é a categorização de vinhos. Para atingir essa categorização foram definidos os seguintes objetivos:

- Análise de um conjunto de dados reais sobre vinhos com 178 vinhos distribuídos por 3 classes (tipo de cultivo) em que foram analisadas 13 características químicas ;
- Aplicação de ferramentas quimiométricas para a categorização dos dados, pela formação de agrupamentos naturais ;
- Melhoramento dos dados para a classificação através do uso da técnica de eliminação de ruído;
- Validação dos resultados obtidos utilizando técnicas de relevantes.

Embora no caso sob estudo a quimiometria tenha sido aplicada a uma base de dados de vinhos, esta pode ser aplicada às mais diversas bases de dados existentes. Atualmente existem inúmeras bases de dados dos mais diversos tipos, nas quais se podem aplicar estudos análogos ao realizado.

Nos dias que correm a maioria das empresas trabalha com bases de dados, nas quais é possível realizar-se estudos semelhantes ao aqui descrito. São abordagens pouco dispendiosas para as empresas, uma vez que todo o *softwares* utilizado é de caracter gratuito. Pelo que pode ser uma mais-valia para essas empresas.

Capítulo 3 - Métodos Quimiométricos

Quando se pretende analisar uma base de dados com diversas variáveis, como é o caso em estudo, é necessário recorrer-se à análise multivariada. Este tipo de análise é frequentemente utilizada quando existem múltiplas variáveis ligadas a um objeto. [64]

A análise multivariada contém diversos métodos de análise, no entanto, neste estudo apenas foram usadas técnicas não supervisionadas, ou seja, não existe uma informação pré-definida nem conhecimento adicional sobre os objetos em estudo. As duas técnicas empregadas foram: a Análise de Componentes Principais e a Análise Hierárquica de Agrupamento.

3.1 - Análise de Componentes Principais (PCA)

A análise de componentes principais (PCA) tem como base a determinação da variância e covariância dos dados e a representação dos dados num novo sistema de eixos (dimensões) determinados pela maior variância matematicamente descrito numa relação de vetores e valores próprios . Este tipo de estudo tem como principal objetivo reduzir a dimensionalidade do conjunto de dados e informar sobre o grau de correlação entre objetos e variáveis, retendo o máximo de informação do sistema possível, sendo que essa informação obtida através da variância dos dados. [65]

A redução do número de dimensões é efetuada a partir da transformação de variáveis originais em novas variáveis, que são denominadas de componentes principais (PC's) ou dimensões. [66] As PCs devem reter o máximo de informação das variáveis originais e ortogonais entre si. As componentes vão ser ordenadas de acordo com a informação que cada uma retém, sendo a primeira a que apresenta uma maior variância dos dados. [67]

Após a aplicação do PCA e redução do número de dimensões é possível fazer a representação dos dados num gráfico de duas ou três dimensões, em que os eixos principais serão correspondentes às PC's com maior peso, ou seja, primeira e segunda (e eventualmente terceira) componentes principais. A informação a retirar será tanto mais significativa quanto maior for a soma das percentagem de recuperação de informação dessas duas PC's. Note-se que a escolha mais fundamentada do número de componentes a reter tem de ser feita com base em critérios específicos (ver adiante).

Uma vez definidos os principais objetivos do PCA é necessário definir-se quais os passos necessários para a realização de um estudo:

1. Selecionar o tipo de matriz a usar, entre a matriz de variância-covariância e de correlação;
2. Determinar o número de componentes principais (PC's), usando o critério de Kaiser, de Pearson e a representação em Screeplot;
3. Analisar e interpretar as representações gráficas dos dados.

3.1.1 - Escolha da matriz

Para que seja possível realizar-se um estudo de PCA e após a análise da base de dados, é necessário escolher-se qual a matriz que se pretende usar. Este passo é essencial dado que é a matriz que vai permitir a transformação dos dados, de modo que seja possível calcular quais os valores próprios (λ) necessários para a escolha das componentes principais, que é o passo que se segue.

Dentro do tipo de matrizes que são possíveis de ser usadas na realização deste estudo, encontra-se a matriz de variância-covariância e a matriz de correlação, esta muitas vezes encarada como resultado de uma normalização.

3.1.1.1 - Matriz de Variância-Covariância

A matriz de variância-covariância, frequentemente conhecida apenas como matriz de covariância, é uma matriz simétrica que apresenta na sua diagonal a variância das variáveis, sendo os termos que se encontram fora da diagonal correspondentes à covariância entre as variáveis. [68,69]

Este tipo de matriz é útil para determinar o número de componentes principais, dado que a variância destas corresponde aos valores próprios e a soma destes é igual à soma da diagonal da matriz. [70]

3.1.1.2 - Matriz de Correlação

Este tipo de matriz é usado quando se pretende evitar a influência de uma ou mais variáveis sobre as outras aquando da determinação do número de componentes principais. Geralmente isso acontece quando a ordem de grandeza dessas variáveis é superior à das restantes. Esta matriz diferencia-se da matriz de covariância, pelo facto de os valores da matriz de variância-covariância serem divididos pelos respetivos desvios padrão resultando em valores de covariância que variam entre 0 e 1 e a variância (diagonal da matriz) ser igual a 1. Tal como a matriz de covariância, também a matriz de correlação é simétrica. [70]

3.1.2 - Determinação das componentes principais

O passo seguinte a realizar num estudo de PCA é a determinação das componentes principais (PC's). A determinação das componentes com maior interesse para o estudo está assente em três critérios, sendo eles: o critério de Kaiser, o critério de Pearson (ou regra dos 80%) e a representação em Screeplot:

3.1.2.1 - Critério de Kaiser

O critério de Kaiser ($\lambda > 1$), é usado única e exclusivamente quando se utiliza a matriz de correlação, dado que apenas com este tipo de matriz a média dos valores próprios é a unidade (a soma dos valores próprios corresponde ao número de variáveis). Com este critério apenas se selecionam as componentes principais que apresentem um valor próprio superior a 1, isto é, apenas se consideram aqueles que apresentem um valor superior à média. [68,69]

3.1.2.2 - Critério de Pearson

O critério de Pearson, também conhecido como regra dos 80% é aplicado aos dois tipos de matrizes explicadas na secção 3.1.1. Este critério tem como base a percentagem de variância explicada por cada dimensão, isto é, a variabilidade que cada dimensão é capaz de recuperar. Assim sendo, este critério fornece

o número de componentes principais que apresentam uma variabilidade total igual ou superior a 80%, retendo por isso as n primeiras dimensões que obedeçam a esse critério. [68]

3.1.2.3 - Representação em Screeplot

O screeplot consiste na representação dos valores próprios em função do número da componente que apresenta contribuição. Quando a contribuição de determinada componente é próxima de 0, esta não aparece na representação do screeplot. Este método é usado para medir a variabilidade dos objetos em função do número de componentes que apresentam contribuição para o sistema. [67]

3.1.3 - Representação gráfica

O último passo da realização de um estudo de PCA é a representação gráfica dos dados, como é exemplo da Figura 3.1. Esta representação é muitas vezes realizada através de um biplot, dado que este permite a representação em simultâneo dos objetos e variáveis a duas dimensões. [71]

Neste tipo de representações é possível, por exemplo, observar-se a distribuição dos objetos (pontos da Figura 3.1) em grupos distintos. Cada grupo, consoante a sua posição terá determinadas características, sendo estas definidas pelas *loadings* (setas da Figura 3.1).

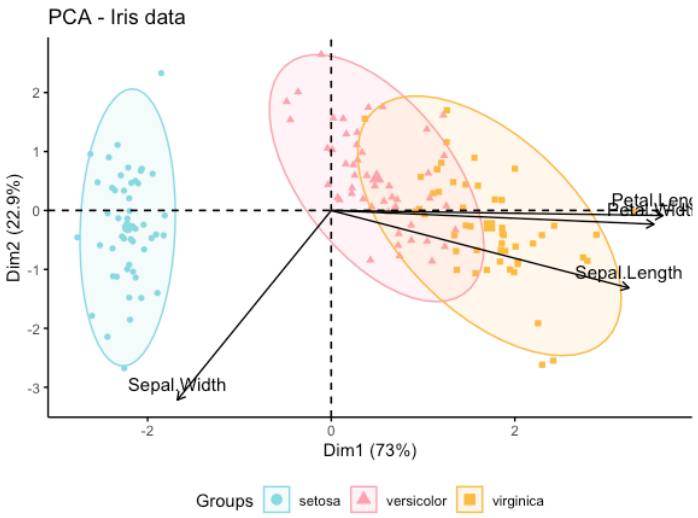


Figura 3.1: Exemplo da representação gráfica do biplot de PCA. As setas a preto correspondem às variáveis e os pontos (círculos, triângulos e quadrados) aos objetos.

3.2 - Análise Hierárquica de Agrupamento (HCA)

A análise hierárquica de agrupamento (HCA) é uma técnica que permite que os objetos da base de dados formem agrupamentos, sendo cada cluster constituído por objetos que apresentam características semelhantes entre si. [72] A formação de clusters pode ter como base dois modos: i) o modo aglomerativo, em que cada objeto é um cluster e, posteriormente, são agrupados conforme a sua distância; ii) e o modo divisivo, em que se parte de um único e se termina com grupos isolados contendo objetos similares. O modo mais frequentemente utilizado é o aglomerativo, que é o que foi aplicado neste trabalho.[73,74] Tal como

referido, os grupos são formados tendo em conta as semelhanças entre eles, sendo estas avaliadas através do cálculo da distância entre objetos e entre os clusters. [75,76]

3.2.1 - Cálculo da distância

Estabelece a distância entre os objetos, tendo como objetivo avaliar a semelhança/dissemelhança entre amostras.

A package *dist*[77] do R fornece 9 maneiras distintas de se proceder ao cálculo das distâncias, tais como a distância (de):

- Binária, também conhecida como distância de Jaccard, dada pela distância entre dois vetores binários, sendo o seu valor dado por 1 menos as características comuns a ambos. Apesar da designação dada no pacote, esta distância é uma de muitas aplicáveis a dados binários; [78]
- Canberra, dada pela soma da diferença absoluta entre dois vetores a dividir pela soma dos valores absolutos de cada vetor, sendo frequentemente usada quando existe um grande número de ocorrências vazias, isto é, quando os dados são dispersos tendo em conta a origem; [79,80]
- Euclidiana, dada pela distância mínima entre dois pontos, podendo ser usada para medir a distância geométrica no espaço multidimensional. É a métrica mais utilizada, mas é necessário ter em atenção que os objetos devem possuir escalas semelhantes; [78,80]
- Manhattan, dada pela distância absoluta entre os dois vetores, o que faz com que seja dada pelo somatório da diferença absoluta entre dois vetores; [80]
- Máxima, tal como o próprio nome indica é dada pela distância máxima entre dois vetores, sendo esta dada pelo maior valor da diferença entre os dois vetores; [80]
- Minkowski, que fornece a distância como sendo a norma de um vetor espacial quando se realiza uma generalização da distância Euclidiana e de Manhattan; [81]
- Kendall, surge a partir do coeficiente correlação de Kendall, permitindo assim avaliar qual o número de desacordos entre os pares de duas permutas. Diz-se que um par é concordante quando a classificação de ambos os elementos concordam uma com a outra, por exemplo $x_i > x_j$ e $y_i > y_j$ e um par é discordante quando as classificações não concordam, por exemplo, $x_i > x_j$ e $y_i < y_j$; [82,83]
- Pearson, usa o coeficiente de correlação de Pearson, sendo este definido como sendo a covariância de duas variáveis a dividir pelo seu desvio padrão. O cálculo da distância será igual a 1 menos o coeficiente de correlação; [84]
- Spearman, é semelhante à distância de Pearson, dado que o coeficiente de Spearman não é mais do que a aplicação do coeficiente de Pearson aos dados transformados por classificação. Este método difere do anterior pelo facto de que permite também o estudo da relação não linear para além do estudo da relação linear. [85]

De modo a facilitar a interpretação do cálculo das distâncias, encontra-se na Tabela 3.1 as equações necessárias para o cálculo de cada tipo de distância, bem como a divisão dos 9 métodos entre os dois grupos: a distância baseada na similaridade e a baseada na correlação.

Tabela 3.1:Cálculo da distância entre objetos.

Tipo de Distância	*Fundamento*	*Representação*	
Binária	Baseadas na similaridade	$d_{bin(x,y)} = 1 - \frac{\sum_i \min(x_i{}^{a}, y_i{}^{b})}{\sum_i \max(x_i, y_i)}$	(3.1)
Canberra		$d_{can(x,y)} = \sum_{i=1}^{n} \frac{\lvert x_i - y_i \rvert}{\lvert x_i \rvert + \lvert y_i \rvert}$	(3.2)
Euclidiana		$d_{euc(x,y)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$	(3.3)
Manhattan		$d_{man(x,y)} = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$	(3.4)
Máxima		$d_{máx(x,y)} = máx(x_i - y_i)$	(3.5)
Minkowski		$d_{mink(x,y)} = \sqrt[p^{c}]{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p}$	(3.6)
Kendall	Baseadas na correlação	$d_{kend(x,y)} = 1 - \frac{n_c{}^{d} - n_d{}^{e}}{\frac{1}{2}n^{f}(n-1)}$	(3.7)
Pearson		$d_{cor(x,y)} = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x}^{g})(y_i - \bar{y}^{h})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$	(3.8)
Spearman		$d_{spear(x,y)} = 1 - \frac{\sum_{i=1}^{n}(x'_i{}^{i} - \bar{x'}^{j})(y'_i{}^{k} - \bar{y'}^{l})}{\sqrt{\sum_{i=1}^{n}(x'_i - \bar{x'})^2 \sum_{i=1}^{n}(y'_i - \bar{y'})^2}}$	(3.9)

[a] x_i corresponde a um vetor de valores reais de tamanho n
[b] y_i corresponde a um vetor de valores reais de tamanho n
[c] p corresponde a um fator configurável
[d] n_c corresponde ao número de pares concordantes
[e] n_d corresponde ao número de pares discordantes
[f] n corresponde ao tamanho de x e y
[g] $\bar{x}$ corresponde ao valor médio dos valores reais de x
[h] $\bar{y}$ corresponde ao valor médio dos valores reais de y
[i] x'_i corresponde a um vetor de valores classificados de tamanho n
[j] $\bar{x'}$ corresponde ao valor médio dos valores classificados de x
[k] y'_i corresponde a um vetor de valores classificados de tamanho n
[l] $\bar{y'}$ corresponde ao valor médio dos valores classificados de y

3.2.2 - Método de ligação

Outro parâmetro a ter em consideração quando se pretende proceder efetuar um estudo de HCA, é a escolha do método de ligação, que não é mais do que o modo como vai ocorrer a ligação de um objeto a um agrupamento já existente ou a ligação de um agrupamento a outro de acordo com a semelhança. [86]

Relativamente aos métodos de ligação é possível, tal como para o cálculo da distância, enumerar diversos tipos comuns, como a ligação "single", que determina a distância entre grupos como sendo a distância entre os objetos mais próximos de cada um [87]; a ligação "average", em que a distância entre grupos

é dada pela média da distância de todos os pontos de um grupo aos pontos do outro [87]; a ligação "complete", que, contrariamente à ligação "single" vai fornecer uma distância entre conjuntos como sendo a distância dos pontos mais afastados [87]; a ligação "Ward", em que a distância é fornecida pela soma dos quadrados das distância entre todos os objetos de cada grupo e o centróide [87]; a ligação "centroid", em que a distância entre grupos é dada pela distância entre o centróide de cada grupo [88]; a ligação "median", que se trata de um caso particular do centróide, onde a distância é dada pela distância ponderada entre os seus centróides [89] e a ligação "McQuitty", onde a distância depende da combinação dos grupos e não da observação individual dos clusters.[90] Sendo possível, na Tabela 3.2 observar-se a representação esquemática de cada um dos tipos de ligação, facilitando assim a sua interpretação.

Tabela 3.2:Representação dos tipos de ligação entre agrupamentos.

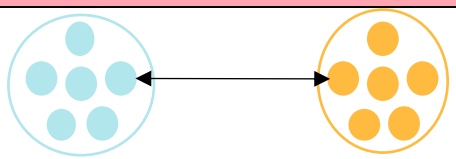
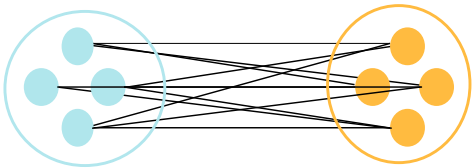
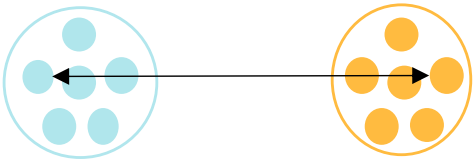
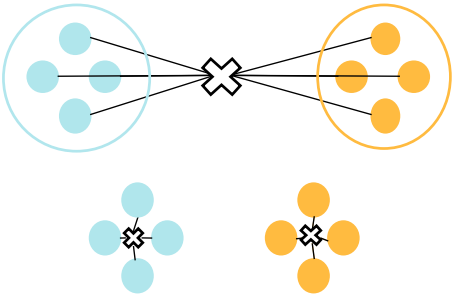
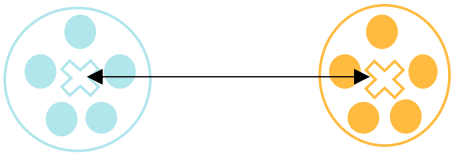
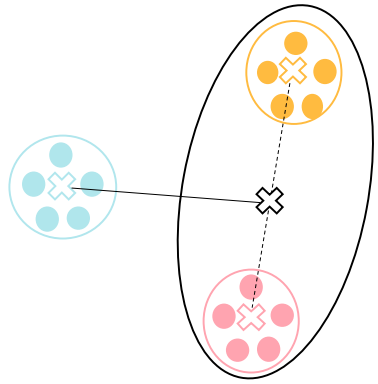
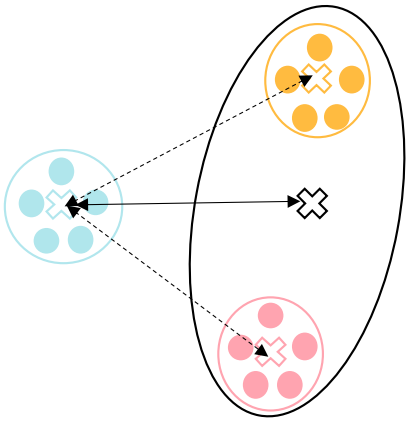
Tipo de Ligação	*Representação*
"single"	
"average"	
"complete"	
"Ward"	
"centroid"	

Tabela 3.2:Representação dos tipos de ligação entre agrupamentos (continuação)

"median"	
"McQuitty"	

3.2.3 - Representação Gráfica

O passo final do estudo de HCA é a representação gráfica do dendrograma, sendo essa representação diferente consoante o método do cálculo de distância e o tipo de ligação escolhido. Os dendrogramas auxiliam a interpretação da formação de grupos, dado que, tal como é possível observar na Figura 3.2, é constituído por linhas, encontrando-se estas ligadas entre si de acordo com a semelhança que os objetos ou grupos apresentam entre si. As linhas de união (linhas verticais) serão representadas com o valor da distância entre os objetos, pelo que quanto menor for o valor desta distância, maior é a semelhança dos grupos.

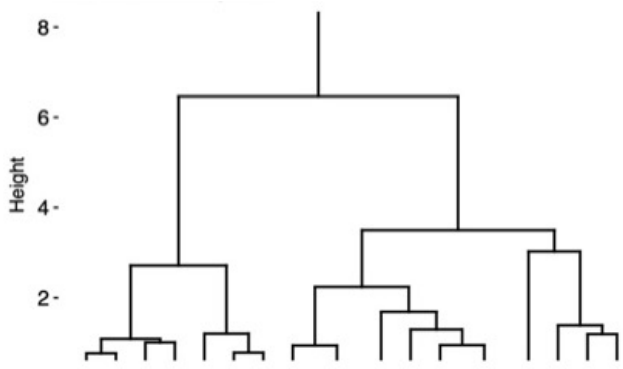


Figura 3.2: Exemplo de representação de um dendrograma.

Capítulo 4 - Validação dos Agrupamentos

A validação tem como objetivo medir a qualidade dos resultados. Neste caso concreto, a validação vai avaliar os clusters formados no estudo de PCA e HCA. O processo de validação deve ser realizado quando existem diversos dados, sendo estes provenientes de bases de dados, uma vez que estas podem possuir outliers, que vão influenciar negativamente os resultados.

De acordo com Kassambara [72], o processo de validação possui cinco passos:

1. Avaliação da tendência da formação de agrupamentos;
2. Determinação do número ideal de agrupamentos;
3. Validação estatística de agrupamentos;
4. Determinação dos melhores algoritmos de clustering;
5. Cálculo do valor de prova para o clustering hierárquico.

4.1 - Avaliação da tendência de agrupamentos

O primeiro passo a ser realizado quando se pretende fazer a validação de grupos é a avaliação da tendência de formação de agrupamentos. Esta avaliação é fundamental, pois os clusters que se formam podem ser formados aleatoriamente, sem que haja qualquer relação entre os objetos que os constituem. Assim sendo, este passo é fundamental para se perceber se a análise hierárquica de agrupamentos é passível de ser feita ou não.

A realização deste passo requer a análise dos clusters formados pelo PCA, pelo que quando o estudo de HCA ocorre após a sua realização e os clusters obtidos se encontram já validados é possível passar para o segundo passo, pois não existe a necessidade de se repetir novamente a representação dos clusters. Todavia, caso o estudo de PCA não tenha sido realizado é necessário que se comparem os clusters obtidos quando se realiza a representação do biplot dos dados em estudo com dados selecionados aleatoriamente, a partir do conjunto de dados originais e espera-se que os dados selecionados não formem qualquer cluster e os dados originais formem clusters significativos.

A avaliação pode ser feita, a partir de dois métodos: o método visual e o método estatístico.

4.1.1 - Método Visual

O método visual permite a avaliação da tendência a partir da observação de gráfico, tal como o que é apresentado na Figura 4.1. Para a obtenção deste tipo de imagens, é necessário o cálculo da matriz de distâncias, que permite perceber qual a diferença entre cada objeto. É de salientar que os objetos que pertencem ao mesmo cluster se encontram próximos, podendo assim comprovar-se se os clusters são significativos ou não.

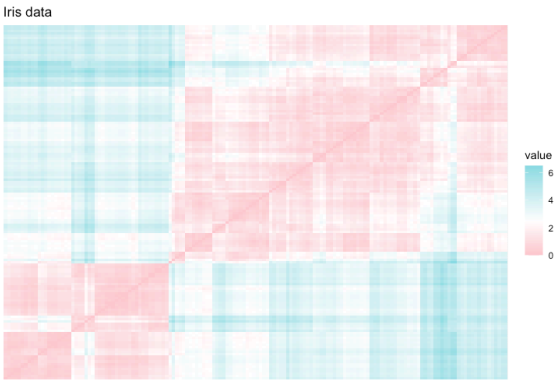


Figura 4.1: Representação de uma matriz de dissimilaridade. A cor-de-rosa temos uma baixa dissimilaridade (alta similaridade) e a azul uma alta dissimilaridade (baixa similaridade).

Outro método visual que permite a avaliação da tendência de agrupamento é o uso da matriz de correlação obtida por análise de PCA, sendo que nos dará pistas sobre a semelhança entre os objetos. Na Figura 3.2 encontra-se representada uma matriz de correlação.

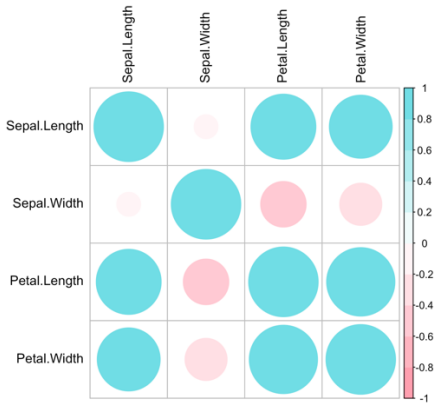


Figura 4.2: Representação de uma matriz de correlação. Os círculos a azul apresentam uma correlação positiva e a rosa uma correlação negativa, o tamanho dos círculos será tanto maior quanto maior for o valor que a correlação apresenta.

4.1.2 - Método Estatístico

Relativamente ao método estatístico, realiza-se a partir da estatística de Hopkins. Esta vai avaliar a tendência a partir da probabilidade que cada um dos conjuntos de dados possuí para que seja gerado aleatoriamente a partir de uma distribuição uniforme dos dados.

A estatística de Hopkins (H) pode ser determinada através do uso da Equação (4.1), onde x_i representa a distância entre o ponto p_i e o seu vizinho mais próximo p_j, sendo estes pertencentes aos dados sem qualquer tratamento e y_i representa a distância entre o ponto q_i e o seu vizinho mais próximo q_j, , pertencente ao conjunto de dados gerados aleatoriamente.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} \qquad \text{Equação (4.1)}$$

Se H é aproximadamente igual a 0.5, o somatório de x_i e y_i é semelhante, o que faz com que os dados apresentem uma distribuição uniforme, e assim tenham sido gerados aleatoriamente. O ideal neste

tipo de método é que H seja diferente de 0.5, para que os clusters formados não tenham sido gerados de maneira aleatória.

4.2 - Determinação do número ideal de clusters

O passo que se segue na validação é a determinação do número ideal de clusters, que embora seja um método subjetivo que depende do particionamento e do método usado para que se medir a semelhança entre objetos é um passo útil, pois fornece ao utilizador uma ideia de qual o número de clusters que se espera obter.

A determinação do número ideal de agrupamentos pode ser feita usando dois tipos de métodos: métodos diretos e métodos indiretos. Os métodos diretos são usados para a otimização de dois critérios: a soma dos quadrados do cluster, usando-se para isto o método do cotovelo, e da silhueta média, através do uso do método de silhueta. Dentro do métodos indiretos, encontram-se essencialmente testes estatísticos, como é o caso da estatística de lacunas, que são usados para que seja possível comparar a evidência com a hipótese nula. Para além dos métodos já referidos existem mais 30 métodos que podem também ser utilizados, sendo a análise destes 30 métodos/índices feita de uma só vez: o número ideal de agrupamentos será aquele que mais vezes for sugerido pelos 30 índices.

4.2.1 - Métodos do Cotovelo

O método do cotovelo, tal como já foi referido é usado para a otimização da soma dos quadrados totais dentro do agrupamento. Este método é frequentemente representado através de um gráfico da soma dos quadrados do clusters em função do número de agrupamentos, sendo o número ideal dado, geralmente, pela dobra no gráfico (Figura 4.3). Essa dobra representa o número ideal pois a partir desse clusters não existem grandes alterações na soma dos quadrados. Este método por vezes é ambíguo, pelo que muitas vezes se usam como alternativa os dois métodos que se seguem.

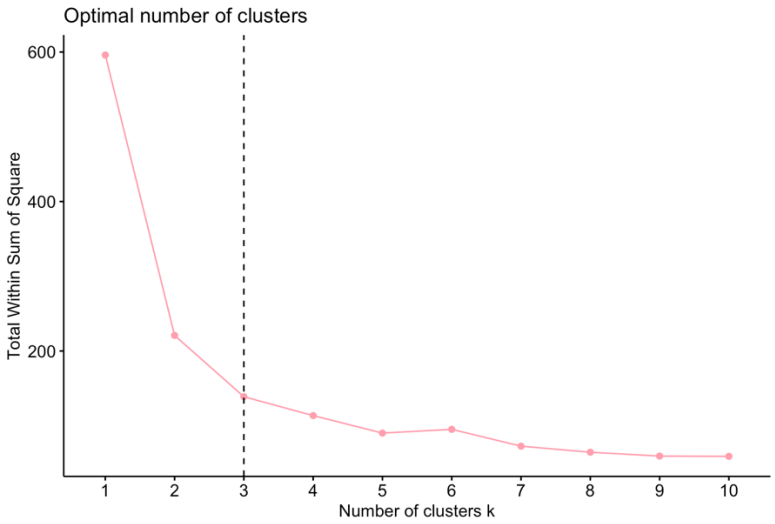


Figura 4.3: Exemplo da representação do método do cotovelo.

4.2.2 - Métodos da silhueta

O método da silhueta mede a qualidade dos agrupamentos, ou seja, determina se os objetos se encontram no agrupamento correto ou não, através do cálculo da distância entre os objetos de cada cluster. Tal como o método do cotovelo, também este método pode ser representado graficamente (Figura 4.4). A partir do gráfico o número ideal de clusters será aquele para o qual existe uma maximização da largura da silhueta média (S_i).

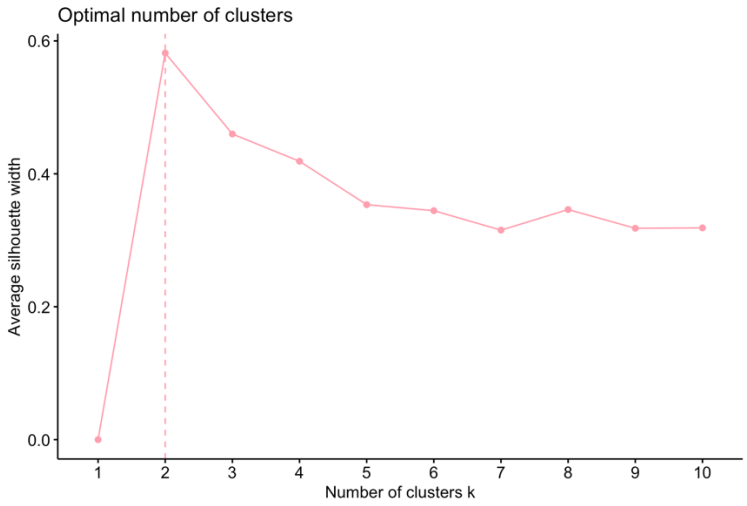


Figura 4.4: Exemplo da representação do método de silhueta.

4.2.3 - Métodos da estatística de lacunas

O método da estatística de lacunas compara a variação total intracluster para os diferentes clusters com os valores esperados quando existe uma distribuição de referência dos dados. Sendo possível observar-se essa distribuição através de gráficos, como o apresentado na Figura 4.5. Através deste é possível observar-se que o número de clusters ideal é aquele em que existe a maximização do valor da estatística de lacuna, o que fornece a informação de esse agrupamento se encontra longe de uma distribuição aleatória.

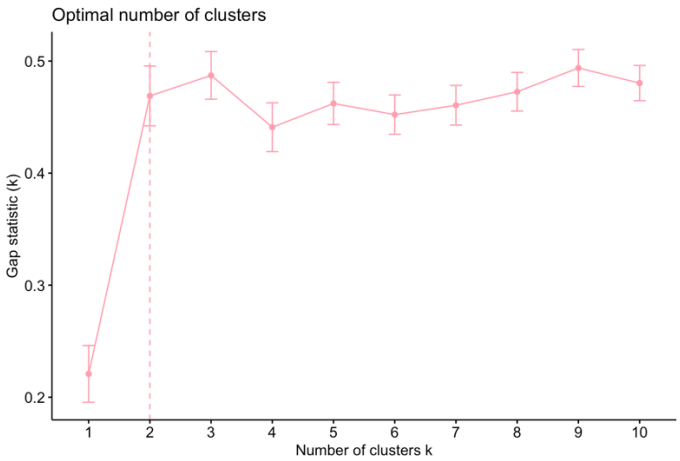


Figura 4.5: Exemplo da representação do método da estatística de lacunas.

Este método deve ser frequentemente preferido quando comparado com os métodos de cotovelo e silhueta pois estes medem apenas uma característica do agrupamento enquanto o método da estatística de lacunas fornece um procedimento estatístico de modo que seja possível combater essas falhas.

4.3 - Validação estatística de agrupamentos

O passo que se segue no processo de validação é a validação estatística de agrupamentos. Esta validação é importante pelo facto de se verificar se existem ou não a presença de padrões aleatórios, sendo estes indesejados. Para além disso, avalia-se a qualidade dos resultados do algoritmo de clustering.

A validação estatística de clusters pode ser classificada de três formas: validação interna, validação externa e validação relativa.

4.3.1 - Validação de cluster interna

A validação interna usa as informações internas do processo do clustering para avaliar a qualidade de uma estrutura do cluster sem qualquer referência a informações externas. Para além disso, pode ser usado para estimar o número de clusters e o algoritmo mais adequado, sem qualquer recurso a dados externos. Dentro da validação de clusters interna existem três medidas úteis para se perceber qual a distribuição dos clusters, são elas: a compactação, a conetividade e a separação.

A compactação vai medir a proximidade entre os objetos que constituem o mesmo agrupamento e quanto mais baixa for, mais compacto será o agrupamento. A conectividade, esta corresponde ao espaço em que os objetos são colocados no mesmo agrupamento que os seus vizinhos mais próximos. O valor da conectividade deve ser minimizado para que não exista a presença de padrões aleatórios. Quanto à separação, passa pela verificação de se um agrupamento se encontra bem separado dos restantes. Atendendo a estes três parâmetros é possível usar o coeficiente de silhueta e o índice de Dunn (ver adiante) para se fazer o estudo de validação interna, podendo também determinar-se através destes processos qual o número ideal de clusters.

4.3.1.1 - Coeficiente de silhueta

O coeficiente de silhueta permite que seja medido se um objetos se encontra no cluster correto e fornece ainda uma estimativa da distância média entre os clusters. Tal como o método de silhueta, também o estudo do coeficiente de silhueta permite que os dados sejam representados num gráfico (Figura 4.6), onde estão representados os objetos de cada grupo e qual a distância a que cada objeto se encontra do cluster mais próximo.

Para a interpretação do gráfico é necessário ter-se em consideração que quanto mais próximo de 1 for a largura média da silhueta (S_i), mais bem agrupados estão os objetos nos clusters; no caso de este valor ser próximo de 0, vão existir objetos que estão entre dois agrupamentos, não sendo possível definir-se qual a mais correta atribuição; quando o seu valor é negativo, é porque existem objetos que foram colocados no cluster errado.

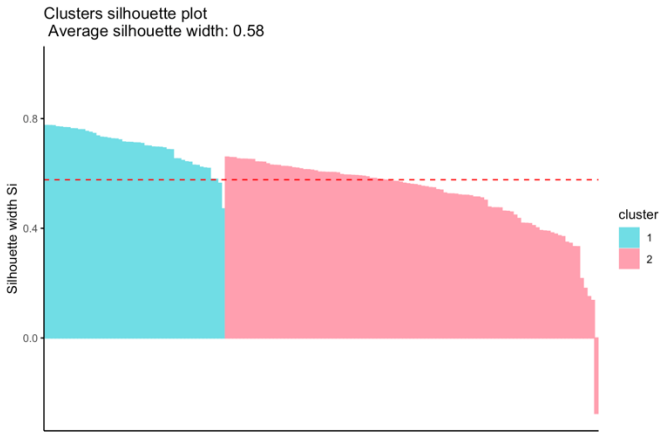


Figura 4.6: Exemplo do gráfico do coeficiente de silhueta.

4.3.1.2 - Índice de Dunn

O índice de Dunn (D) define a relação entre a distância mínima de separação, entre clusters e a distância máxima de compactação, dentro dos clusters. Este índice pode ser calculado pela Equação (4.2).

$$D = \frac{distância\ mínima\ de\ separação}{distância\ máxima\ de\ compactação} \qquad \text{Equação (4.2)}$$

No que diz respeito aos valores que de D, quanto maior este for mais compactos e bem separados serão os clusters, podendo assim fazer-se uma diferenciação de clusters rapidamente.

4.3.2 - Validação de cluster externa

Neste tipo de validação, existe a comparação dos resultados de uma análise de clusters com resultados conhecidos externamente. Esta vai medir se os rótulos dados a cada um dos clusters correspondem aos que são conhecidos. Recorre-se ao uso deste tipo de validação quando se pretende selecionar o algoritmo de agrupamento correto, dado que se conhece qual o número de clusters esperado devido a referências externas.

4.3.3 - Validação de cluster relativa

Este tipo de validação avalia a estrutura dos clusters através da variação de diversos parâmetros para o mesmo algoritmo. Tal acontecimento, leva a que seja frequentemente usada para a determinação do número ideal de clusters.

4.4 - Determinação do melhor algoritmo de clustering

O quarto passo a ser realizado num processo de validação é a determinação do melhor algoritmo de clustering. Para que seja possível fazer-se essa determinação recorrem-se a dois tipos de medidas: medidas internas e medidas de estabilidade.

4.4.1 - Medidas internas

As medidas internas recorrem ao uso de informações intrínsecas aos dados para avaliar a qualidade do agrupamento. Das medidas internas a serem aplicadas neste estudo fazem parte as medidas internas já referidas aquando da validação de clusters interna, que são: a conetividade, o coeficiente de silhueta e o índice de Dunn.

Em relação ao melhor algoritmo de clustering, este ocorrerá quando existe uma minimização do valor de conectividade, para a determinação por esta medida e uma maximização do valor do coeficiente de silhueta e do índice de Dunn.

4.4.2 - Medidas de estabilidade

Relativamente às medidas de estabilidade, estas avaliam a consistência de um grupo através da comparação com os agrupamentos obtidos após a remoção das colunas, sendo esta feita uma de cada vez. Dentro deste tipo de medidas destacam-se:

1. A proporção média de não sobreposição (APN), vai medir a proporção média de objetos não colocados no mesmo clusters, agrupados com base nos dados completos e com base nos dados após a remoção de uma coluna;
2. A distância média (AD),vai medir a distância média entre os objetos que se encontram no mesmo cluster quando não existe remoção de uma coluna e quando existe;
3. A distância média entre médias (ADM), consiste na medição da distância média entre os centro dos clusters para os objetos com base em todos os dados e quando há remoção de uma coluna;
4. A figura de mérito (FOM) está relacionada com a variância intracluster, medindo a variância média da coluna eliminada quando o agrupamento é baseada nas colunas não excluídas.

O melhor algoritmo de clustering para cada uma das medidas de estabilidade ocorre quando estas medidas são minimizadas.

4.5 - Cálculo do valor de prova (p-value) para o clustering hierárquico escolhido

O quinto e último passo, consiste no cálculo dos valores de prova dos clusters formados, usando o método hierárquico escolhido. Este estudo vai permitir analisar os clusters formados, de modo a entender se estes se formam aleatoriamente, devido a erros de amostragem e presença de ruído, ou não.

Os clusters formados, tem dois tipos de valores de prova, sendo que ambos variam entre 0 (os clusters são formados aleatoriamente) e 100 (os clusters formados são estáveis):

1. Os valores de prova aproximadamente imparciais (AU), que resultam de uma reamostragem de bootstrap multiescalar;

2. Os valores de prova de bootstarp (BP), que corresponde à frequência com que se identifica determinado cluster quando se realizam diversas cópias do bootstrap.

É importante referir que entre os dois tipos o que é frequentemente usado para o estudo dos valores de prova é o AU. De modo a tornar o estudo mais percetível e se verificar mais facilmente se os clusters são estáveis ou surgem por acaso, os valores de prova são representados através de dendrogramas, encontrando-se na Figura 4.7 representado um exemplo.

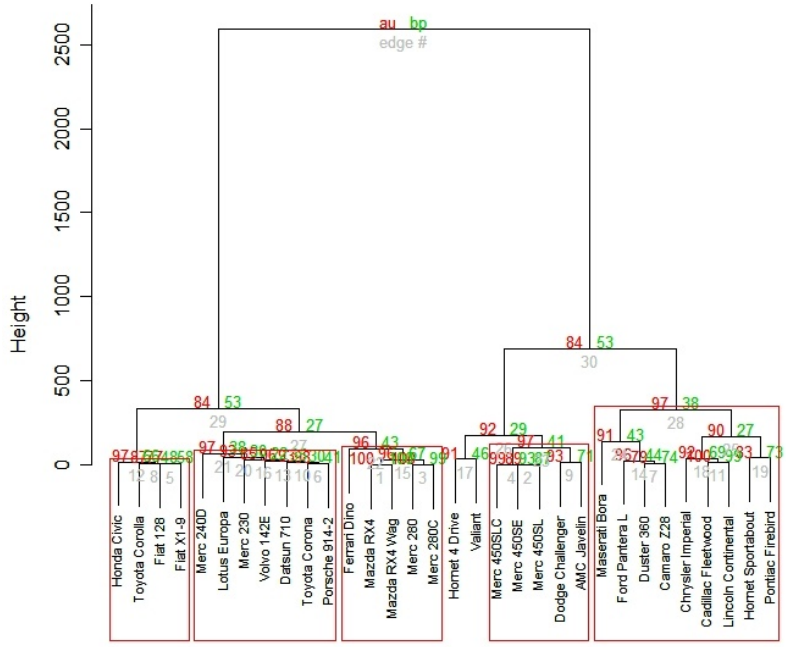


Figura 4.7: Exemplo de um dendrograma com os valores de prova.

Na figura 4.7 é possível observar-se a presença de retângulos, que apenas se representam quando o valor de prova AU do cluster formado é igual ou superior a 95%, visto que apenas os clusters que apresentam esta percentagem são fortemente suportados pelos dados. No entanto, é necessário ter em consideração que associado a cada valor de prova existe um erro associado, encontrando-se na Figura 4.8, um gráfico de um gráfico do valor de prova em função do erro padrão.

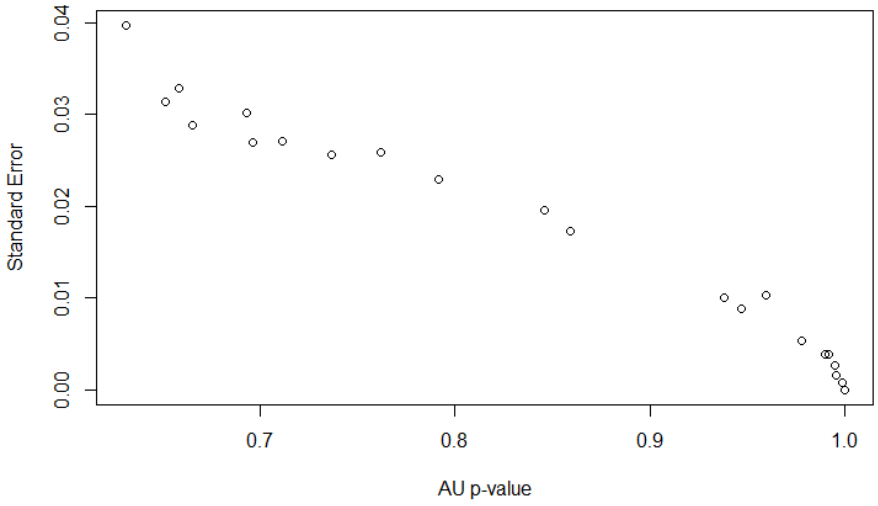


Figura 4.8: Exemplo de um gráfico do erro padrão em função do valor de prova.

Dado que os valores de prova apresentam um erro associado, o valor de prova apresentado será o valor médio, sendo necessário calcular o intervalo de valores entre os quais o valor de prova pode variar. No entanto é necessário considerar-se que apenas se pretende, pelo menos, um intervalo de confiança de 95%, pelo que o cálculo desse intervalo será dado pela Equação (4.3), em que o $\bar{x}$ corresponde ao valor médio da amostra que será o valor apresentado como valor de prova. O SE representa o erro padrão associado a cada valor de prova, sendo retirado de gráficos como o apresentado na Figura 4.8.

$$Limite\ superior\ a\ 95\% = \bar{x} + (SE \times 2) \qquad \text{Equação (4.3)}$$

$$Limite\ inferior\ a\ 95\% = \bar{x} - (SE \times 2)$$

Capítulo 5 - Métodos

5.1- R e RStudio

Neste trabalho utilizou-se a linguagem R (v4.0.3), assente sobre a interface RStudio, na versão 1.3.1093. Este conjunto é frequentemente usado para o tratamento e visualização de dados, sendo usado em diversos trabalhos que requerem estatística computacional. [91]

Quando analisado visualmente o programa, observa-se a presença de 4 janelas distintas, como as apresentadas na Figura 5.1., o painel 1 é onde se encontra o script, que contêm os comandos do código necessário para o estudo. No painel 2 encontram-se armazenadas todas as variáveis e dados resultantes da leitura do código. No painel 3 todos os inputs e output do código. Por fim, o 4º painel é o que apresenta mais funções, isto é, para além de ser possível observar-se a representação gráfica dos dados em estudo é ainda possível observar-se todos os documentos presentes no computador e ainda selecionar os que são de interesse.

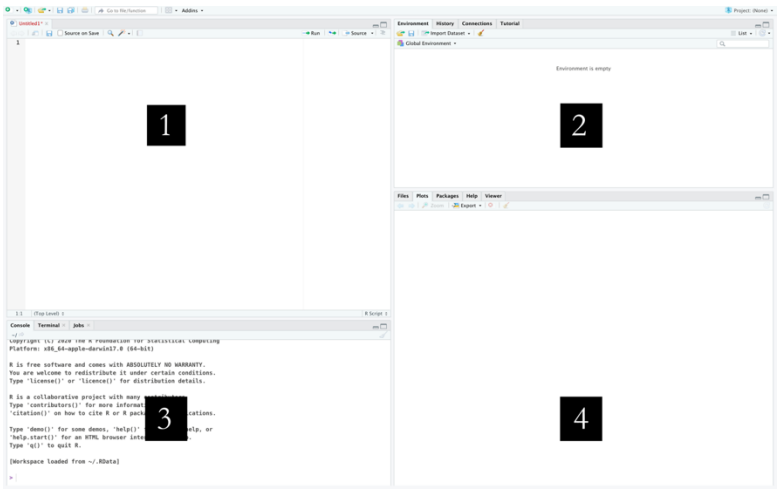


Figura 5.1: Representação dos painéis de trabalho do programa RStudio.

Uma vez que o programa R permite uma enorme variedade de aplicações, possui diversas bibliotecas ou pacotes (packages) que serão úteis para o utilizador, dado que possuem funções e dados específicos para aquele tipo de estudo, sendo por isso necessário o utilizador pesquisar qual o tipo de bibliotecas que existe para aquele estudo e como funcionam. [92]

5.2- Base de Dados

Para a realização deste trabalho recorreu-se à base de dados de vinhos que se encontra em: http://archive.ics.uci.edu/ml/datasets/Wine. A base de dados é constituída por linhas e colunas, sendo que as linhas correspondem aos objetos, que são os 178 vinhos, e as colunas correspondem às variáveis ou atributos, que correspondem às 14 características que foram estudadas para cada vinho.

Para se perceber melhor a constituição da base de dados é necessário ter-se em consideração que apesar de todos os vinhos serem proveniente da mesma região de Itália, é possível dividi-lo em três classes, em que cada classe corresponde a um tipo de cultivo, sendo esta características considerada física, dado que dependem da zona de cultivo. Quanto as restantes variáveis, estas são consideradas características químicas pois só é conhecido o seu valor após a sua medição e vão variar de vinho para vinho. Na Tabela 5.1 é possível observarem-se quais as características físicas e químicas estudadas.

Tabela 5.1: Divisão das variáveis em características físicas e químicas.

Características Físicas	*Características Químicas*
Classes	**Álcool**
Classe 1: 59 vinhos	**Ácido Málico**
Classe 2: 71 vinhos	**Alcalinidade da Cinza**
Classe 3: 48 vinhos	**Cinza**
	Magnésio
	Total de Fenóis
	Flavonóides
	Fenóis não flavonóides
	Proantocianidinas
	Intensidade da Cor
	Tonalidade
	Diluição de vinhos OD280/OD315
	Prolina

5.3 - Métodos Quimiométricos

O estudo e análise da base de dados referida anteriormente ocorreu através do recurso a métodos quimiométricos, tais como, a análise de componentes principais (PCA), a análise hierárquica de clusters (HCA) e a eliminação de ruído. No entanto, antes de que apresentarem os passos exclusivos a cada um dos métodos é importante referir que existem passos comuns, tais como a remoção do histórico do programa RStudio, a leitura dos dados e a instalação de dependências, sendo apresentadas na Figura 5.2 as funções usadas para cada um dos passos.

```
1 rm(list=ls()) # remoção do histórico do R
2 vinhos<-read.csv("vinhos1.csv",header=T,sep=";")# leitura da base de dados
3 #installed.packages("FactoMineR","factoextra")# dependências para o PCA
4 #installed.packages("factoextra","dendextend")# dependências para o HCA
5 #installed.packages("Hmisc","spatstat","fields","scatterplot3d")#dependências para a eliminação de ruído
```

Figura 5.2: Linhas de código comuns para a realização dos diferentes estudos .

A primeira linha de código presente na Figura 5.2, é relativa a remoção do histórico do R, sendo que este passo é facultativo. Sendo utilizado para evitar efeitos de corridas anteriores ou de versões entretanto modificadas.

O próximo passo é a leitura da base de dados, descrita na secção 2 deste capítulo. Sendo que para que fosse possível a leitura desta por parte do programa existiu a necessidade de se converter num ficheiro do tipo ".csv", tendo-se recorrido ao uso da função *read.csv*.

O último passo desta primeira fase de estudo é a instalação de dependências, que tal como referido anteriormente, apresentam diversas funções e dados para a realização de estudos concretos. Posto isto, para a realização do estudo de PCA foram instalados os pacotes "FactoMineR" e "factoextra", para o de HCA as dependências "factoextra" e "dendextend" e por fim para a eliminação do ruído forma instaladas as dependências "Hmisc", "spatstat", "fields" e "scatterplot3d".

5.3.1 - Análise de Componentes Principais (PCA)

O primeiro método quimiométricos usado para o estudo da base de dados foi a análise de componentes principais (PCA). Para a realização desta análise, recorreu-se ao algoritmo apresentado na Figura 5.3. Neste é possível observar-se que o estudo foi dividido em duas partes: o estudo do PCA, onde foram executados todos os passos necessários para a determinação das PC's e a validação de resultados, que permitiu verificar se os resultados obtidos eram válidos ou não.

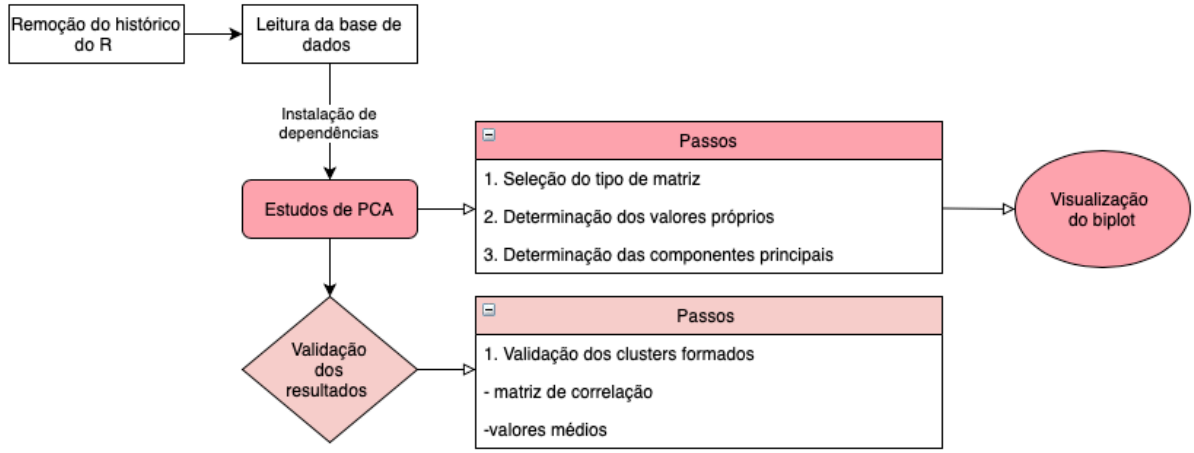


Figura 5.3: Algoritmo aplicado no estudo de PCA.

De acordo, com a Figura 5.3, os passos necessários a ter em consideração são: seleção do tipo de matriz, de entre os dois tipos referidos na secção 3.1.1, sendo esta escolha influenciada pela necessidade ou não de normalização dos dados; a determinação dos valores próprios de cada variável e a determinação do número de componentes principais, sendo a partir destas possível verificar quais as que apresentam maior contribuição para a descrição do sistema. Após a determinação dos passos enumerados anteriormente, é possível proceder-se à representação do biplot.

De modo a facilitar a enumeração das funções usadas para a realização dos passos enumerados, encontra-se a Figura 5.4, em que estão disponíveis as linhas de código executadas.

```
library(factoextra)
library(FactoMineR)
PCA_vinhos<-PCA(vinhos[,-1],  #ficheiro com a remoção da primeira coluna
                graph=F, #não queremos fazer a representação do gráfico
                scale=T) #normalizar os dados
val_proprios<-get_eig(PCA_vinhos) #cálculo dos valores próprios
fviz_eig(PCA_vinhos,addlabels=T,ylim=c(0,40))# representação do screeplot
var<-get_pca_var(PCA_vinhos) #variáveis ativas
head(var$contrib,13)# tabela com a contribuição das variáveis para as principais componentes
fviz_contrib(PCA_vinhos,choice="var",axes=1)# representação gráfica das contribuições das variáveis nas PC's
fviz_pca_biplot(PCA_vinhos,
                col.ind=vinhos$class,# objetos apresentam cores diferentes consoante a classe
                palette=c("#8bdbe3","#fea4b0","#ffbc40"), #cores das diferentes classes
                addEllipses = T, # adicionar as elipses
                label="var", # legenda das variáveis
                col.var="black", # cor das variáveis
                repel=T, # para evitar cortes de palavras
                legend.title="classe") # título da legenda
```

Figura 5.4: Linhas de código para a execução do estudo de PCA.

Tal como já referido, existiu a necessidade de se instalar pacotes que, uma vez instalados, implicam a sua importação para que seja possível aceder às funções que proporcionam.

Para a realização do estudo do PCA, começou-se pelo uso da função *PCA* [93]. Esta função armazena informações como os valores próprios para cada variável, a contribuição que cada uma apresenta para as dimensões/componentes principais, o desvio padrão das distribuições, entre outros. Foi necessário remover a primeira coluna da base de dados, dado que esta se encontra na forma não numérica e a classe é uma variável dependente. Outro aspeto importante a ter em conta, é o facto de ter sido necessário escalar os dados (*scale=T*), dividindo a matriz de variância-covariância pelos desvios padrão. Tal aconteceu devido à existência de variáveis com ordens de grandeza muito superiores às restantes, o que levaria a que houvesse assim uma maior contribuição destas do que das restantes. Assim sendo, dado que existe a normalização dos dados, a matriz que foi usada foi a matriz de correlação.

O passo que se realizou de seguida foi a determinação dos valores próprios para cada variável, sendo que se recorreu à função *get_eig*[94] para se obter esta informação. Ainda com base nos valores próprios e na variância cumulativa , também obtida pela função *PCA* e *get_eig* procedeu-se à representação do screeplot, tendo-se usado a função *fviz_eig*. [94]

Dado que, um dos principais objetivos do estudos de PCA é a redução do número de dimensões, foi necessário proceder-se à determinação do número de componentes principais. Para isso através do recurso à função *get_pca_var* [95].obtiveram-se as variáveis ativas para o estudo, tendo sido possível obter-se uma tabela com as contribuições para cada uma das dimensões consideradas principais.

Uma vez que já haviam sido encontradas as componentes principais, os valores próprios e a percentagem cumulativa, procedeu-se à representação da informação a duas dimensões, através do recurso a *fviz_pca_biplot* [96].

5.3.1.1 - Validação dos dados de PCA

O último passo da realização do estudo de PCA consiste na validação dos dados, tal acontece devido ao facto de se estar a trabalhar com uma base de dados, e os clusters formados serem resultado da presença

de ruído ou até mesmo serem formados aleatoriamente. De modo que fosse possível verificar-se tal acontecimento procedeu-se à validação dos clusters formados para as variáveis e dos formados para os objetos.

No caso das variáveis (características dos vinhos), o estudo foi realizado através do recurso à matriz de correlação, tendo-se seguido o código presente na Figura 5.5.

```
#install.packages("corrplot","pvclust","parallel")
library(corrplot)
library(pvclust)
library(parallel)
corr<-cor(scale(vinhos[,-1]))# obter a matriz de correlação
col1<-colorRampPalette(c("#fea4b0","#FFFFFF","#8bdbe3"))# cores da representação da matriz
corrplot(corr,method="circle",type="lower",tl.col="black",col=col1(20))# representação gráfica da matriz
set.seed(2)# para que os valores sejam sempre os mesmos cada vez que se corre o programa
pv.var<-pvclust(vinhos,# para que seja possível calcular os valores de prova para as variáveis
                method.hclust="ward.D2",# método de ligação
                method.dist="euclidean",# distância medida
                nboot=1000)# número de replicações do bootstrap
plot(pv.var,
     hang=-1,# ajustar a altura das ligações no dendrograma
     cex=0.8,# tamanho da identificação das variáveis
     cex.pv=1,# tamanho dos valores de au e bp
     print.pv=T,# presença dos valores de prova
     print.num=F,# ausência da ordem de ligação
     float=0.01)# posição dos valores de prova em cada um dos agrupamentos
pvrect2(pv.var,
        alpha=0.95,# valor limite para se aceitar um valor de prova
        pv="au",# qual o elemento que serve de base para os valores de prova
        type="geq",# permite considerar os valores de prova iguais ou superiores a alfa
        border="#5badeb",# cor do retângulo formado
        lwd=2,# largura da linha
        lower_rect=9)# limite inferior dos retângulos
seplot(pv.var)# obter o gráfico do desvio padrão em função do valor de prova
```

Figura 5.5: Linhas de código para o processo de validação dos clusters formados tendo em conta as variáveis.

À semelhança do estudo de PCA, existiu a necessidade de se instalar o pacote que permite que seja estudada a correlação e o valor de prova. Uma vez obtida a matriz de correlação, devido ao facto de a interpretação gráfica com cores e círculos de tamanhos distintos facilitar a interpretação procedeu-se à sua representação gráfica. Como alternativa à matriz de correlação podia ter sido executada a matriz de dissimilaridade, uma vez que esta permite que se verifique, em vez da similaridade entre as variáveis a dissimilaridade.

Como referido na secção 4.5, o processo de validação termina com cálculo do valor de prova dos clusters de variáveis que se formam, de modo que fosse possível verificar se estes se formaram aleatoriamente ou não. No entanto, esta análise apenas foi realizada para as variáveis, dado que a função *pvclust* [97] apenas permite a formação de agrupamentos usando as colunas (variáveis). Sendo importante salientar, que para este cálculo é necessário saber-se qual o método de ligação que se pretende usar no estudo de HCA, que será referido mais à frente, de entre os presentes na função *hclust*. [85]

Após o cálculo dos valores de prova existiu a necessidade de se proceder à representação do dendrograma usando para isso a função *plot*. Dado que, apenas quando os valores de prova de AU são iguais ou superiores a 95% os clusters resultantes são estáveis e suportados pelos dados, existiu a necessidade de se verificar quais os clusters que corresponderiam a estas características, tendo sido usada a função *pvrect2*[98].

Foi ainda necessário proceder ao cálculo do erro padrão associado a cada um dos valores, visto que são valores calculados a partir de uma base de dados, tendo-se recorrido para isso à função *seplot*. [99]

Uma vez analisados os resultados para os clusters formados pelas variáveis, resta analisar e validar os clusters formados pelos objetos. Para este processo de validação, recorreu-se ao código presente na Figura 5.6.

```
df <- vinhos[,2:14]
distr_media<-aggregate(df,list(vinhos$classe),mean)# valores médios para cada variável, tendo em conta a classe
grp<-apply(distr_media[,-1],# valores médios
           MARGIN=c(1,2),# obter uma matriz
           FUN=as.numeric)# transformar os dados em dados numéricos
barplot(grp,# matriz de dados
        ylab="[log10Y]",# legenda o eixo dos yy
        log="y",# necessidade de usar o logaritmo devido às diferenças na ordem de grandeza das variáveis
        beside=T,# representação das colunas
        col=c("#8bdbe3","#fea4b0","#ffbc40"))# cores das colunas
legend("right",inset=c(-0.45,0),# posição da legenda no gráfico
       ncol=1,# legenda com apenas uma coluna
       title="classe",# título da legenda
       legend=distr_media$Group.1,# dados usados para a legenda
       fill=c("#8bdbe3","#fea4b0","#ffbc40"),# cores apresentadas na legenda
       bty="l",# contorno da legenda
       cex=1)# tamanho da letra
```

Figura 5.6: Linhas de código para o processo de validação dos clusters formados tendo em conta os objetos.

No que diz respeito à validação dos clusters formados pelos objetos (vinhos), começou-se por agrupar os vinhos das diferentes classes, calculando-se os valores médios para cada uma das características químicas estudadas. Para esse passo recorreu-se à função *aggregate* [100], com o parâmetro *mean*. Dessa agregação resultou um dataframe, que não é mais do que um conjunto de dados de diferentes tipos , tendo sido necessário passar os dados a uma matriz numérica através do uso da função *apply*. O passo que se seguiu foi a representação dos valores médios para cada variável, usando-se a função *barplot* [101]. Dado que as colunas apresentam diferentes cores existiu a necessidade de colocar uma legenda no gráfico para que fosse possível identificar a que classe correspondia cada uma das cores. De salientar que o código usado permite obter um gráfico com todas as variáveis, no entanto dado que este estudo se procede depois da formação dos clusters usando PCA, é possível fazer três gráficos, onde cada um contêm as variáveis que se encontram em cada um dos clusters obtidos no estudo de PCA.

Após a obtenção dos dados relativos à correlação entre variáveis e os valores médios de cada variável em cada classe procedeu-se à verificação de os resultados obtidos se encontravam de acordo com os esperados.

5.3.2 - Análise Hierárquica de Agrupamento (HCA)

O segundo estudo realizado foi a análise hierárquica de agrupamento, encontrando-se na Figura 5.7 o algoritmo seguido para a obtenção de resultados. Há semelhança do que se sucedeu para o estudo de PCA também este estudo se dividiu em duas partes, a obtenção de resultados e a validação destes.

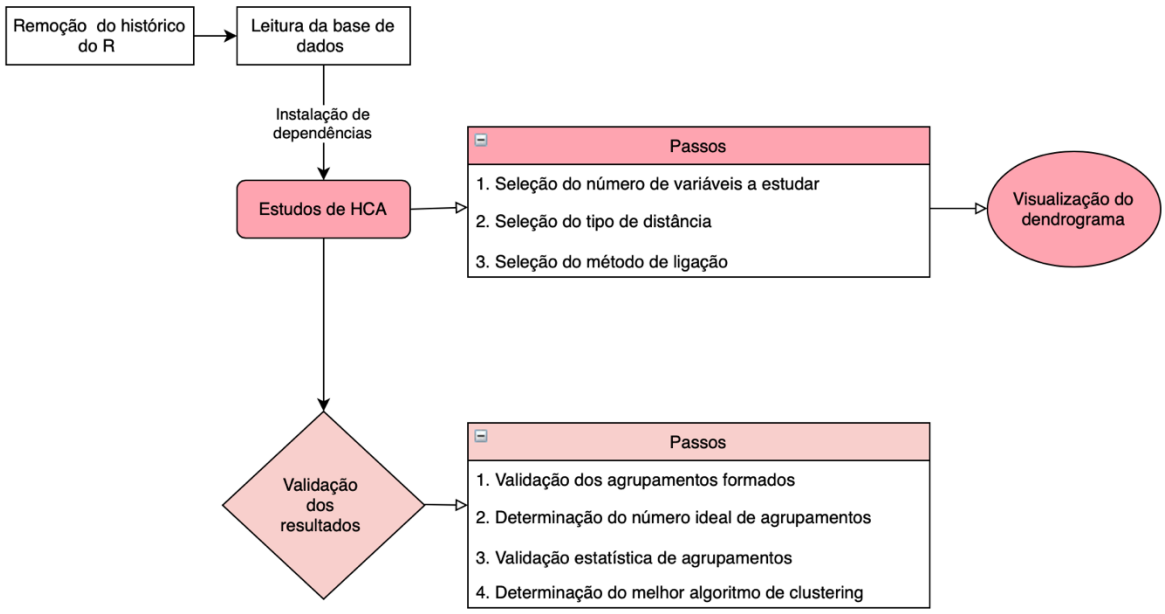


Figura 5.7: Algoritmo aplicado no estudo de HCA.

Relativamente às linhas de código deste processo encontram-se, na Figura 5.8 as linhas relativas ao estudo de HCA, onde foi possível a obtenção de resultados. Dado que este estudo foi realizado após a análise das componentes principais, existiu a necessidade de se comparar os resultados obtidos quando apenas se consideraram as PC's e quando se usou todas as variáveis independentes, ou seja, todas as variáveis que correspondem às características químicas estudadas.

```
6  library(factoextra)
7  library(dendextend)
8  vinhos<-vinhos[,-1]# retirar a primeira coluna
9  vinhos<-vinhos[,c(3,4,6,8,11)]# seleção das variáveis consideradas as PC's
10 vinhos<-scale(vinhos)# normalizar os dados
11 dist <- dist(vinhos,method="euclidean")# calcular a matriz de distância entre os objetos
12 hc <- hclust(dist, method="ward.D2")# calcular a dissimilaridade entre objetos
13 fviz_dend(hc,
14           #k = 3,# número de agrupamento esperado
15           cex = 0.45,# tamanho da identificação dos objetos
16           lwd = 0.8,# tamanho da linha
17           k_colors = c("#ffbc40","#8bdbe3","#fea4b0"), # cor dos agrupamentos formados
18           color_labels_by_k = F,# colocar a identicação dos objetos da cor dos agrupamentos a que pertencem
19           rect = T, # presença dos retângulos
20           rect_border = c("#ffbc40","#8bdbe3","#fea4b0"),# cor dos retângulos em torno dos agrupamentos
21           rect_fill = T, #sombreado dos retângulos
22           lower_rect=1) #limite inferior dos retângulos
```

Figura 5.8: Linhas de código para a realização do estudo de HCA.

Através da análise da Figura 5.8, é possível verificar que tal como aconteceu para o estudo de PCA, existiu a necessidade de importar as dependências para que fosse possível ter-se acesso às funções. Tal como referido, de modo a que fosse possível verificar-se qual a influência das variáveis na formação de agrupamentos, o primeiro passo foi a escolha das variáveis a usar (linhas 8 e 9 da Figura 5.8), de notar que pelo facto de se usar uma técnica não supervisionada, como é o caso do HCA, existiu a necessidade de

retirar a coluna correspondente à *classe* que é uma variável dependente e que caso seja usada faz com que se saiba de antemão quais os grupos que vão ser formados, deixando assim de ser uma técnica não supervisionada.

Após a seleção das variáveis a usar (5 PC's ou todas as variáveis independentes), à semelhança do que aconteceu para o estudo de PCA, procedeu-se à normalização dos dados, para se evitar a influência das variáveis com maior ordem de grandeza no cálculo da distância. Para a medição das distância entre os objetos, recorreu-se à função *dist* [77], onde se definiu como forma de cálculo a distância euclidiana, atendendo ao tipo de dados que possuímos. O passo que se seguiu foi a escolha do método de ligação entre objetos/ grupos, que permitiu calcular a dissimilaridade entre objetos. Uma vez calculadas a distância e a dissimilaridade entre objetos foi possível realizar-se a representação gráfica dos dados, mais concretamente, a representação do dendrograma, tendo-se usado para isso a função *fviz_dend* [102], havendo uma série de parâmetros que foram necessários ter em consideração, tais como, definir o número de agrupamentos que se espera, tendo sido neste caso definido um *k=3*, dado que os dados originais apresentam três classes distintas, sendo assim do nosso interesse foi verificar se essas classes distintas eram possíveis de ser observadas quando se procedeu à representação dos grupos. Além desse parâmetro, de modo que fosse mais percetível a representação, foi necessário atribuir a presença de retângulos (*rect=T*), que nos permitiram diferenciar os grupos formados.

5.3.2.1 - Validação dos dados de HCA

Tal como aconteceu para o estudo de PCA, também o estudo de HCA requereu a realização do processo de validação dos dados. Analisando a Figura 5.7 verifica-se que este processo foi constituído por 4 passos fundamentais, sendo de seguida apresentadas as linhas de código que foram necessárias para cada uma das etapas.

Para o processo de validação, também existiu a necessidade de se instalarem pacotes, encontrando-se na Figura 5.9 as bibliotecas que usadas para a realização do processo.

```
24 installed.packages("Nbclust","fpc","clValid","pvclust","parallel")
25 library(factoextra)
26 library(clustertend)
27 library(NbClust)
28 library(fpc)
29 library(clValid)
```

Figura 5.9: Linhas de código referentes à instalação das dependências necessárias para a validação dos resultados de HCA.

Procedeu-se depois à realização do processo de validação, encontrando-se nas Figura 5.10 a 5.14 as linhas de código utilizadas para cada um dos passos descritos na Figura 5.7.

```
#1ºpasso da validação: validação dos clusters formados
fviz_pca_ind(prcomp(vinhos),# visualização dos clusters formados através da realização do PCA
             title="PCA",# título do gráfico obtido
             geom="point",# representação dos objetos através de pontos
             palette=c("#8bdbe3","#fea4b0","#ffbc40"),# cor dos pontos
             legend="bottom")# legenda
hopkins(vinhos,n=nrow(vinhos)-1)# estatística de Hopkins
fviz_dist(dist(vinhos),# dissimilaridade entre objetos
          show_labels = F,# ausência da identificação dos objetos
          gradient=list(low="#8bdbe3",mid="#FFFFFF",high="#fea4b0"))# cores usadas na representação da matriz
```

Figura 5.10: Linhas de código para a validação dos clusters formados.

O primeiro passo do processo de validação foi a validação dos agrupamentos formados. Dado que o estudo de HCA ocorreu após a realização do estudo de PCA, não existiu a necessidade de se validar os clusters formados, pois os clusters aqui validados são resultantes do estudo de PCA, que já se encontravam validados aquando da realização deste estudo. No entanto, na Figura 5.10 é possível observar-se os passos necessários à validação caso o estudo de HCA se tivesse realizado primeiro. Além disso, os passos que se encontram descritos na imagem 4.10 poderiam substituir os passos apresentados nas Figuras 5.5 e 5.6, evitando-se assim o estudo da matriz de correlação e dos valores médios de cada variável . É de salientar que o valor da estatística de Hopkins é suficiente para se perceber se os clusters formados são ou não significativos, dado que, como referido na secção 4.1.2, um valor de H próximo de 0.5 indica que os clusters formados foram gerados aleatoriamente, não sendo por isso significativos, esperando-se assim valor diferente deste.

```
#2ºpasso da validação: determinação do número ideal de clusters
fviz_nbclust(vinhos,# dados a usar
             kmeans,# método hierárquico usado
             method="wss",# método usado no estudo (wss=cotovelo,silhouette=silhueta e gap_stat=lacunas)
             linecolor = "#fea4b0")# cor da linha
  geom_vline(xintercept =3 ,linetype=2)# representação de uma linha a tracejado com base no número de clusters esperados
nb<-NbClust(vinhos,# dados a usar
            diss=NULL,# não usar a matriz de dissimilaridade
            distance="euclidean",# distância usada para calcular a dissimilaridade
            min.nc = 2,# número mínimo de clusters a formar
            max.nc=15,# número máximo de clusters a formar
            method="ward.D2")# método de ligação entre objetos
 fviz_nbclust(nb)# para a representação dos resultados obtidos
```

Figura 5.11: Linhas de código utilizadas para a determinação do número ideal de clusters.

O passo que se seguiu no processo de validação, foi a determinação do número ideal de agrupamentos (Figura 5.11) , tendo-se recorrido ao uso da função *fviz_nbclust*,[103] usando os métodos *"wss"* (método do cotovelo), *"silhouette"* (método de silhueta) e *"gap_stat"* (método de lacunas), onde o número ideal é representado pela linha a tracejado presente nos gráficos resultantes. Para além do uso de gráficos que permitem a determinação do número ideal de clusters através do estudo dos método enumerados, foi ainda possível estudar-se qual o número ideal usando 30 índices distintos[104], recorrendo-se para isso à função *NbClust* [105]. Uma vez que esta função permite analisar 30 índices em simultâneo, existiu a necessidade de se representar essa informação através de um gráfico de barras. Para essa representação recorreu-se novamente ao uso da função *fviz_nbclust*, onde o gráfico resultante apresenta a frequência da formação dos clusters em função do número que se forma.

```
#3ºpasso da validação: validação estatística dos clusters
hc<-eclust(vinhos,# dados a utilizar
           "hclust",# função de clustering a usar
           hc_metric = "euclidean",# método de cálculo da dissimilaridade entre os objetos
           hc_method="ward.D2",# método de aglomeração usada
           graph=F)# não queremos a representação do gráfico
fviz_dend(hc,
          show_labels = F,# não mostrar a identificação dos objetos
          palette=c("#8bdbe3","#fea4b0","#ffbc40"),# cor do dendrograma
          as.ggplot=T)# representação do dendrograma
fviz_silhouette(hc,
                palette=c("#8bdbe3","#fea4b0","#ffbc40"),# cores dos agrupamentos formados no gráfico
                ggtheme=theme_classic())# tema de fundo
sil<-hc$silinfo$widths# informação sobre os objetos: cluster a que pertencem, qual o vizinho mais próximo e largura de silhueta
sil_neg_linha<-which(sil[,"sil_width"]<0)# fornece o nº da linha em que se encontram os objetos que estão mal posicionados
sil[sil_neg_linha, ,drop=F]# definir quais os objetos que se encontram atribuídos a clusters diferentes do esperado
hc_stats<-cluster.stats(dist(vinhos),hc$cluster)# cálculo de diversos parâmetros, ex: tamanho de cada cluster, índice de Dunn,etc
```

Figura 5.12: Linhas de código utilizadas para a validação estatística dos clusters.

O terceiro passo de validação requer a validação estatística dos agrupamentos (Figura 5.12) , tendo-se recorrido ao uso da função *eclust* [106], para que fosse possível recolher todas as informações relativas ao clusters que se formam, salientando o número e tamanho dos clusters. Uma vez recolhidas essas informações foi possível proceder-se à representação do dendrograma tendo-se usado a função *fviz_dend*, tendo servido para se verificar quantos clusters foram formados, de modo que fosse possível comparar os resultados com os obtidos no estudo. O passo que se seguiu, foi o uso da função *fviz_silhouette* [107], que permitiu que se obtivesse a distribuição dos objetos tendo em conta o coeficiente de silhueta, sendo que este permitiu verificar se existiam objetos que se encontravam mal classificados, sugerindo qual o grupo ideal tendo em conta a distância aos vizinhos mais próximos, caso os objetos se encontrassem mal classificados, usou-se para isso a função *sil* [108]. O último passo realizado neste passo de validação, permite que se estudem diversos parâmetros dos diferentes clusters formados, *cluster.stats* [109], tendo sido útil para se comparar os agrupamentos formados.

```
#4ºpasso:escolha do melhor algoritmo de clustering
clmetodos<-c("hierarchical","kmeans","clara","agnes","pam") # criação de um vetor com os métodos que se pretendem avaliar
int<-clValid(vinhos,
             nClust= 2:5, # nº de clusters a serem formados e avaliados
             clMethods=clmetodos,# métodos a avaliar
             validation="internal",# validação feita a partir de medidas internas
             maxitems =178,# número máximo de objetos
             metric="euclidean",# modo do cálculo da distância
             method="ward")# método de ligação escolhido
summary(int)# fornece o valor ideal para cada uma das medidas internas(índice de Dunn, conectividade e coeficiente de silhueta)
estab<- clValid(vinhos,
             nClust=2:5,
             clMethods=clmethods,
             validation="stability",# validação feita a partir de medidas de estabilidade
             maxitems =178,
             metric="euclidean",
             method="ward")
optimalScores(estab)# fornece o valor ideal para cada uma das medidas de estabilidade (APN,AD,ADM e FOM)
```

Figura 5.13: Linhas de código utilizadas para a escolha do melhor algoritmo de clustering.

O último passo do processo de validação realizado foi a escolha do melhor algoritmo de clustering/método de estudo (Figura 5.13). Para este estudo existiu a necessidade de se definir previamente quais os métodos de estudos que se pretendem analisar, que neste caso foram os métodos que se encontram em *clmetodos* da figura 5.13 (hierárquico, K-médias, CLARA, AGNES e PAM). Uma vez definidos os métodos de estudo, procedeu-se ao estudo das medidas internas e de estabilidade (descrita na secção 4.4),

através do uso da função *clValid* [110], tendo sido determinado qual o melhor método e o número de clusters esperados usando o método escolhido para cada uma das medidas estudadas.

5.3.3 - Eliminação de Ruído

O último estudo a ser executado no âmbito do estudo da base de dados apresentada na seção 4.2 foi a verificação de existência de ruído e consequente eliminação. Para isso, seguiram-se os passos apresentados no algoritmo da Figura 5.14.

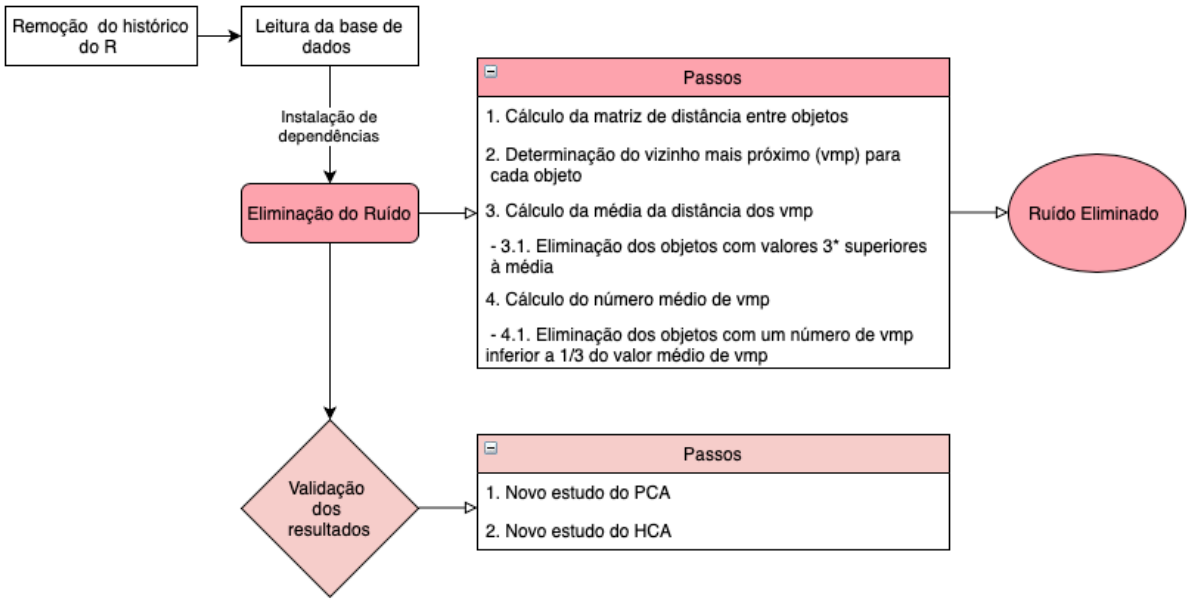


Figura 5.14: Algoritmo para o estudo da eliminação do ruído.

À semelhança dos estudos realizados anteriormente serão apresentadas as linhas de código usadas para a execução do estudo de eliminação do ruído (Figura 5.15 a 5.17). Para além da eliminação de objetos que poderiam ser considerados outliers foi também estudada a influência da eliminação do ruído quando esta ocorre usando apenas 3 das PC's, as 5 PC's e todas as variáveis, sendo de salientar que as PC's consideradas são as determinadas no estudo de PCA, realizado anteriormente.

```
df1_vinhos<-vinhos[,c(4,8,11)]# selecionar as três PC's
vinhos$classe<-as.factor(vinhos$classe)# cria um fator com as classes  para a representação gráfica
df1_vinhos<-scale(df1_vinhos)# normalizar os valores das variáveis
df2_vinhos<-vinhos[,c(3,4,6,8,11)]# selecionar as cinco PC's
df2_vinhos<-scale(df2_vinhos)
df3_vinhos<-vinhos[,-1]# selecionar todas as colunas, com execção da primeira (classe)
df3_vinhos<-scale(df3_vinhos)
```

Figura 5.15: Linhas de código para a determinação das variáveis a usar.

Tal como referido, para além de se proceder à eliminação do ruído estudou-se também a influência do número de variáveis neste tipo de estudo. Para isso, após a leitura da base de dados procedeu-se à escolha das variáveis, tendo-se realizado os passos enumerados na Figura 5.15. Sendo de salientar o passo da

normalização e a criação de um fator com as classes (“um”,”dois”,”três”) que foi fundamental para a representação da eliminação do ruído num gráfico 3D.

```
14  xall<-ppx(data=df1_vinhos)# criar uma estrutura de dados multidimensional
15  x<-xall
16  dist_mat<-rdist(x)# cálculo da matriz distância
17  remover<-cbind()# criar um vector vazio para fazer a eliminação
18  vmp<-nndist.ppx(x,k=1)# calcular o vizinho mais próximo para cada objeto
```

Figura 5.16: Linhas de código usadas para a inicialização do estudo.

O passo que se seguiu para que no final fosse possível realizar-se o estudo, foi a criação de um estrutura multidimensional, através do uso da função *ppx* [111], devido aos dados usados possuírem diversas variáveis, o que leva a que tenham de existir várias dimensões (uma dimensão por cada variável). Após a obtenção da estrutura multidimensional foi possível proceder-se ao cálculo da matriz de distâncias, tendo-se recorrido à função *rdist* [112]. Como um dos principais objetivos era a determinação do vizinho mais próximo (vmp) de cada objeto, recorreu-se à função *nndist.ppx* [113] para a sua determinação. O último passo realizado na inicialização do estudo foi a criação de um vetor vazio (função *cbind* [114]), no qual foram sendo adicionados os objetos considerados outliers.

```
20 repeat{ # repetir o processo as vezes que forem necessárias até haver a eliminação de todos os outliers
21   vmp<-nndist.ppx(x,k=1)# calcular o vizinho mais próximo para cada objeto
22   med_vmp<-mean(vmp)# calcular a média das distâncias ao vmp
23   print(med_vmp)
24   lista1<-(which(vmp>=3*med_vmp,arr.ind=T))# identificação dos objetos com distâncias superiores a 3*med_vmp
25   lista<-which(dist_mat<3*med_vmp & dist_mat>0,arr.ind=T)# identificação dos objetos com distâncias inferiores a 3*med_vmp
26   lin_x<-nrow(x$data)# calcular o número de linhas que constituem a matriz x
27   print(lin_x)
28   nmed_vmp<-nrow(lista)/lin_x# cálculo do número médio de vmp
29   print(nmed_vmp)
30   identify<-table(lista)/2# determinação do número de vmp para cada objeto
31   identify<-as.matrix(identify)# criar uma matriz com os números de vmp para cada objeto
32   nomes<-rownames(identify)# denominação dos objetos
33   remover<-which(identify<nmed_vmp/3)# determinação dos objetos que possuem 1/3 de nmed_vmp
34   remover<-as.numeric(nomes[remover])# transformar em númerico os objetos identificados para remover
35   if(length(lista1)!=0){remover=c(remover,lista1)}# junção das duas condições 3*med_vmp e 1/3*nmed_vmp
36   llr<-length(remover)# tamanho da variável obtida do passo anterior
37   print(llr)
38   xnovo<-x# criação de uma nova estrutura multidimensional(xnovo) que é igual a x
39   if(llr==0){break}# quebra do ciclo quando llr igual a 0
40   if(llr!=0){# quando diferente de 0 o ciclo prossegue
41     xnovo<-x[-remover,  ]# eliminação dos objetos removidos no dataframe de xnovo
42     size.xnovo<-nrow(xnovo)# número de linha do dataframe xnovo
43     dist_mat<-dist_mat[-remover,-remover]# substituição da matriz distâncias inicial atendendo a remoção dos outliers
44   }
45 }
46 xall<-data.frame(xall)
47 xnovo<-data.frame(xnovo)
```

Figura 5.17: Linhas de código usadas no processo de eliminação de *outliers*.

A Figura 5.17 são as linhas de código usadas num dos principais passos realizados, visto que são apresentadas as linhas necessárias para a execução do ciclo que permitiu a eliminação dos *outliers*, sendo importante referir que o ciclo apenas terminou quando deixaram de existir objetos a eliminar. Para isso, foi necessário calcular a média da distância dos vmp (função *mean* [115]). Uma vez calculada foi se verificar quais os objetos que apresentam uma distância ao vmp duas, três, quatro ou cinco vezes superior à média (linha 24 da Figura 5.17), o cálculo dos objetos aconteceu de forma isoladas, de modo a que fosse possível verificar-se qual o critério que seria mais adequado ao estudo. O passo seguinte foi a determinação dos objetos que

apresentavam uma distância inferior a duas, três, quatro ou cinco vezes a média e uma distância superior a 0, sendo este último importante dado que os valores estudados são obtidos a partir de uma matriz de distância onde valores presentes na diagonal são iguais a 0 (linha 25 da Figura 5.17). Uma vez determinados todos os objetos que obedeciam a essa condição procedeu-se a determinação do número médio de vmp sendo que os objetos que foram eliminados apresentavam um número de vmp inferior a um terço do número médio de vmp (linha 33 da Figura 5.17). No final da eliminação, obteve-se uma nova estrutura multidimensional (*xnovo*), que quando passada a dataframe, permitiu verificar o número de objetos eliminados.

Uma vez que já se tinham encontrando os objetos considerados *outliers*, procedeu-se à representação dos resultados obtidos quando apenas foram estudadas apenas 3 das PC's. Essa representação teve como objetivo verificar a localização dos pontos eliminados, dado que se esperava que estes se encontrassem nas zonas de menor densidade de pontos, ou seja, na periferia dos conjuntos formados.

Após a finalização deste estudo, de forma a validar os resultados obtidos procedeu-se novamente aos estudo de PCA e HCA realizados anteriormente, de modo que fosse possível verificar-se quais as influências que o ruído apresentava nos resultados obtidos.

Capítulo 6 - Resultados e Discussão

6.1-Análise das Componentes Principais (PCA)

A base de dados utilizada ao longo deste projeto é uma base que contém amostras de vinhos, todos eles originários da mesma região de Itália. Contudo, apesar da mesma origem, esses vinhos diferem na forma de cultivo em que foram produzidos com características químicas diferentes. No sentido de avaliar se, de facto, essas diferenças de cultivo são suficientes para influenciarem as características de modo que as amostras de cada cultivo sejam identificáveis, a base de dados foi submetida à análise de componentes principais (PCA). Espera-se que havendo influência do cultivo, o PCA aponte para, pelo menos, a formação de três grupos distintos correspondendo a cada um dos métodos de cultivo.

A base de dados é composta por 178 amostras de vinhos, chamadas de objetos do estudo, sendo cada objeto caracterizado por 13 variáveis, que correspondem às características químicas estudadas de cada objeto. Na Tabela 6.1, encontram-se algumas entradas da base de dados, em que a primeira coluna corresponde aos objetos (número da amostra do vinho) e a primeira linha às variáveis. A segunda coluna é a classe (corresponde a uma dada área de cultivo) à qual corresponde cada amostra, sendo as restantes as características químicas estudadas.

A análise de componentes principais tem como objetivo obter correlações entre as variáveis e, em consequência, entre os objetos e diminuir a dimensionalidade do problema. Se existirem correlações, e dependendo do grau dessas (maior ou menor correlação) é possível reduzir o número de dimensões (inicialmente cada variável corresponde a uma dimensão), permitindo projetar os dados em menos dimensões do que as 13 iniciais. A correlação entre as variáveis pode ser positiva, sendo que os valores do coeficiente variam entre 0 e 1 e as duas variáveis apresentam comportamentos idênticos, isto é, quando uma aumenta ou diminui o seu valor, a outra sofre a mesma alteração; negativa, com valores variando entre -1 e 0, e as variáveis têm comportamentos opostos, o que leva a que quando existe um aumento ou diminuição haja o efeito contrário na outra variável; ou nula, que ocorre quando o coeficiente é igual a 0 e não vai existir qualquer relação linear entre as duas variáveis. Essa nova base dimensional permite analisar e retirar conclusões de um problema que inicialmente não permitiria obter informações dado à sua complexidade. Assim, o principal objetivo desse estudo é a redução de dimensões, com uma consequente diminuição do número de variáveis, tornando assim o estudo passível de ser realizado.

Previamente à PCA, é necessário verificar se os valores que as amostras apresentam para as variáveis apresentam todos a mesma ordem de grandeza, pois caso tal não se verifique vai existir influência dos valores com maior ordem de grandeza sobre os outros, levando a que os resultados não sejam corretamente avaliados. De modo a realizar esta verificação, escolheu-se uma amostra aleatória e observaram-se os valores de cada variável como mostrado no gráfico da Figura 6.1.

Tabela 6.1: Exemplos de objetos e variáveis da base de dados.

Obj.	Classe	Álcool	Ácido Málico	Cinza	Alcalinidade da cinza	Magnésio	Total de fenóis	Flavonóides	Fenóis não flavonóides	Proantocianidinas	Intensidade da cor	Tonalidade	Diluição de Vinhos OD280/OD315	Prolina
1	um	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
2	um	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
60	dois	12.37	0.94	1.36	10.6	88	1.98	0.57	0.28	0.42	1.95	1.05	1.82	520
61	dois	12.33	1.1	2.28	16	101	2.05	1.09	0.63	0.41	3.27	1.25	1.67	680
132	três	12.86	1.35	2.32	18	122	1.51	1.25	0.21	0.94	4.1	0.76	1.29	630
133	três	12.88	2.99	2.4	20	104	1.3	1.22	0.24	0.83	5.4	0.74	1.42	530

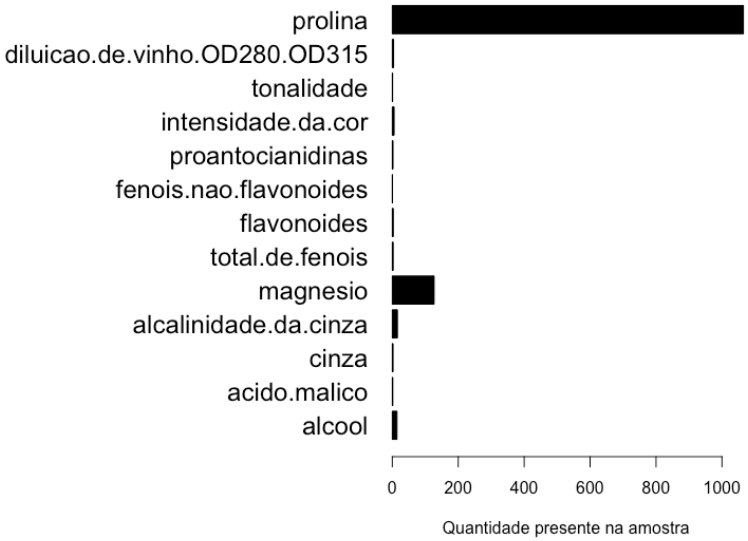


Figura 6.1: Quantidade de cada uma das variáveis numa amostra.

Com base na Figura 6.1, verifica-se que os valores da prolina são superiores aos restantes em, aproximadamente, 3 ordens de grandeza, o que nos indica que será necessário proceder à normalização dos dados, que tem como objetivo transformar todos os valores das variáveis para que fiquem com a mesma ordem de grandeza. É certo que ao normalizar um conjunto de dados parte da informação é perdida, mas essa ação é necessária para se avaliar os dados em conjunto. Para efetuar a normalização (Z) é necessário que se calcule a média (μ) e o desvio padrão (σ) para cada uma das variáveis (X), como descrito na Equação (6.1). Na Tabela 6.2, encontram-se as mesmas amostras da base de dados presente na tabela 6.1, mas com os valores normalizados.

$$Z=\frac{X-\mu}{\sigma} \qquad \text{Equação (6.2)}$$

Um dos objetivos deste estudo é perceber se a distribuição dos objetos se encontra de acordo com aquilo que se esperava, isto é, se a distribuição dos objetos resulta em três agrupamentos distintos dado que se têm três tipos de cultivo. No entanto, de modo a perceber-se qual a influência que o tipo de cultivo tem relativamente aos valores de cada variável, é necessário realizar-se primeiro o estudo das loadings. As loadings representam o peso que as variáveis originais têm no sistema de eixos transformado, o que faz com que existam 13 loadings, uma vez que existem 13 variáveis. A representação das loadings ocorre num referencial que apresenta os quatro quadrantes e uma circunferência de raio 1. Este tipo de representação vai também permitir que se visualize a relação que existe entre as componentes (apenas a 1ª e 2ª dimensão serão representadas graficamente) e as variáveis, dado que quanto mais perto da extremidade do círculo maior é a relação e quanto mais perto do centro menor a relação com a dimensão. Na Figura 6.2, observa-se que os flavonóides apresentam uma grande correlação com a dimensão 1 e a intensidade da cor com a dimensão 2. Ainda com este tipo de representação é possível verificar-se a formação de clusters de variáveis, o que se deve essencialmente à correlação que existe entre estas, e destas, com a dimensão onde se encontram. Pela análise da Figura 6.2, verifica-se que as 13 variáveis se dividem em 3 clusters. O cluster 1 correlaciona-se negativamente com os clusters 2 e 3, devido a encontrar-se do lado contrário a estes quando analisada a 1ª dimensão . Relativamente aos clusters 2 e 3, estes correlacionam se positivamente entre si, pois encontram-se do mesmo lado do referencial quando analisada a 1ª dimensão. No entanto verifica-se uma exceção, a intensidade da cor (pertencente ao cluster 3) correlaciona-se positivamente com o cluster 1, visto que se encontra no 2º quadrante. Quando estudamos a 2ª dimensão , apesar de todos os clusters apresentarem variáveis nos 1º e 2º quadrantes, o cluster 2 vai se correlacionar negativamente com os restantes, tal facto deve-se à representação de 3 das 5 variáveis que o compõem se encontrarem no 4º quadrante, que apresenta uma correlação negativa com o 1º e 2º. Passando à análise da constituição dos clusters, é de salientar o facto de a cinza pertence ao cluster 1 e não ao 3, embora se encontre entre as variáveis deste último, deve-se ao facto de a alcalinidade da cinza se relacionar com esta e também ao facto de apenas termos duas dimensões (1ª e 2ª) representadas graficamente.

Tabela 6.2: Exemplo de entradas da base de dados com os valores normalizados.

Obj.	Álcool	Ácido Málico	Cinza	Alcalinidade da cinza	Magnésio	Total de Fenóis	Flavonóides	Fenóis não flavonóides	Proantocianidinas	Intensidade da cor	Tonalidade	Diluição de Vinhos OD280/OD315	Prolina
1	1.514	-0.561	0.231	-1.166	1.909	0.807	1.032	-0.658	1.221	0.251	0.361	1.843	1.010
2	0.246	-0.498	-0.826	-2.484	0.018	0.567	0.732	-0.818	-0.543	-0.292	0.405	1.110	0.963
60	-0.777	-1.249	-3.669	-2.664	-0.822	-0.503	-1.461	-0.658	-2.046	-1.341	0.405	-1.115	-0.721
61	-0.826	-1.107	-0.315	-0.807	0.088	-0.392	-0.940	2.155	-2.063	-0.771	1.279	-1.326	-0.212
132	-0.149	0.585	0.122	0.151	0.298	-1.590	-0.810	-0.979	-1.329	0.147	-0.951	-1.678	-0.689
133	-0.235	-0.024	0.122	1.349	-0.122	-1.829	-0.940	-0.738	-1.329	0.277	-1.301	-1.763	-0.593

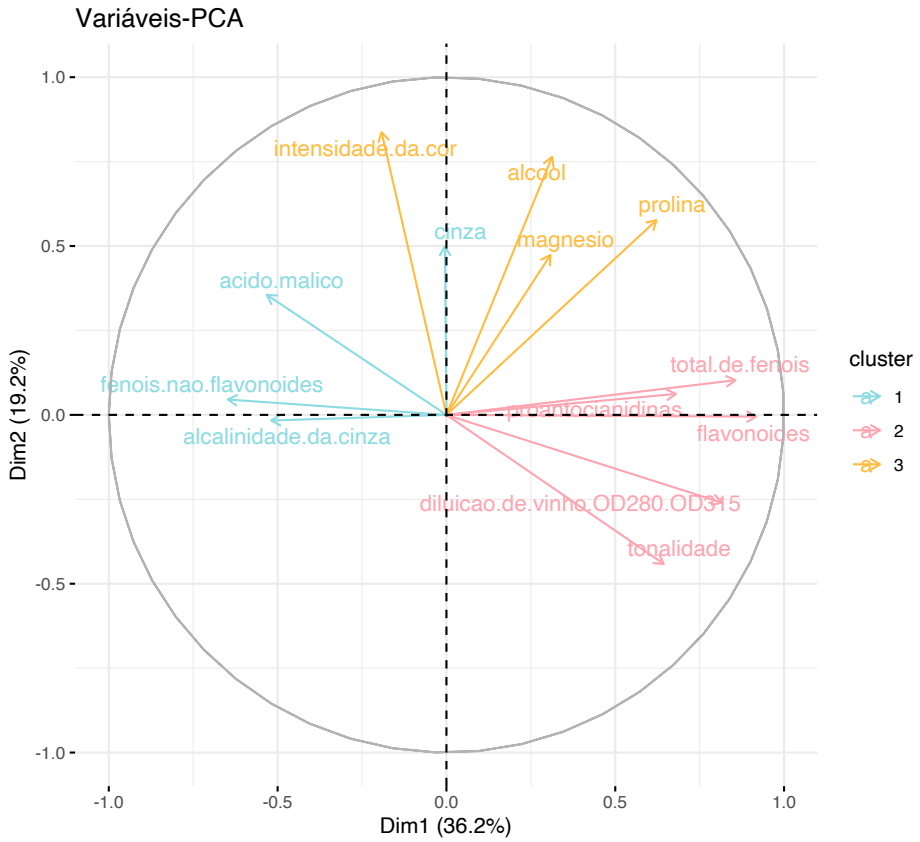


Figura 6.2: Representação das loadings (variáveis) em duas dimensões indicando a formação de grupos.

Visto que já se realizou o estudo da distribuição das loadings, pode-se, a seguir, discutir a distribuição dos objetos ou scores. Tal como já referido, espera-se que esta distribuição resulte na formação de três agrupamentos distintos, dado que a base de dados original contempla três classes distintas. Espera-se que as três formas de cultivo levem a que haja características químicas mais marcantes num grupo do que em outros. Para se estudar a distribuição dos objetos, recorre-se ao biplot, em que se tem a representação dos objetos através de pontos e das variáveis através de vetores. Este é um gráfico de dispersão, pois permite o estudo da relação entre variáveis que vai permitir que a informação dos dados (variáveis e objetos) seja representada graficamente (Figura 6.3).

Ao analisar-se a Figura 6.3, observa-se a existência de três grupos, em que cada um deles corresponde a uma das classes definidas pela base de dados. A classe 1 tem uma maior distribuição no primeiro quadrante (positivo relativamente ao eixo vertical), em que é possível verificar-se a presença do cluster 1 e 3 da Figura 6.2, isto leva a que as variáveis destes clusters sejam as mais características desta classe. A classe 2 apresenta uma distribuição no 3º e 4º quadrantes (negativo relativamente ao eixo vertical), estando por isso associada a valores intermédios para o cluster 2, dado que 3 das cinco variáveis pertencem ao 4º quadrante. Por fim, a classe 3, apresenta uma distribuição maioritariamente no 2º quadrante, onde se encontra o cluster 1 da Figura 6.2, levando a que este tipo de cultivo seja associado às variáveis deste grupo.

Figura 6.3: Representação das loadings ou variáveis (setas) e dos scores ou objetos (pontos).

De acordo com o critério de 80%, discutido na seção 2.1.2 impunha-se que a perda de informação não fosse superior a 20%. Observando-se a Figura 6.2 e 6.3 verifica-se que tal não se sucede, dado que se tem uma recuperação de informação de 55.4%, havendo assim uma perda de 44.6%. No entanto considera-se que esta percentagem de recuperação, para o estudo em questão, é suficiente para que seja possível retirarem se conclusões.

O critério de 80% [67] vai inspecionar a percentagem de variância que cada componente é capaz de explicar, em que o número de dimensões a usar será igual ao número de dimensões necessárias para que se obtenha uma descrição de pelo menos 80% da variância. Por outro lado, o screeplot (Figura 6.4) consiste na representação dos valores próprios, que medem a magnitude da variabilidade dos objetos, em função do número de componentes que apresentam contribuição para o sistema. Além deste critério, e uma vez que houve a necessidade de se normalizarem os dados, existe o critério de Kaiser [67], que vai olhar exclusivamente para os valores próprios para cada dimensão (Tabela 6.3), permitindo a exclusão de todos aqueles que apresentem um valor inferior a 1. Essa característica deve-se ao facto de que quando se realiza a soma de todos os valores o seu resultado deve ser igual ao número de variáveis, pelo que se todas as componentes fornecessem a mesma informação, a sua média seria 1. Sendo assim, os valores que sejam superiores à unidade encontram-se acima da média.

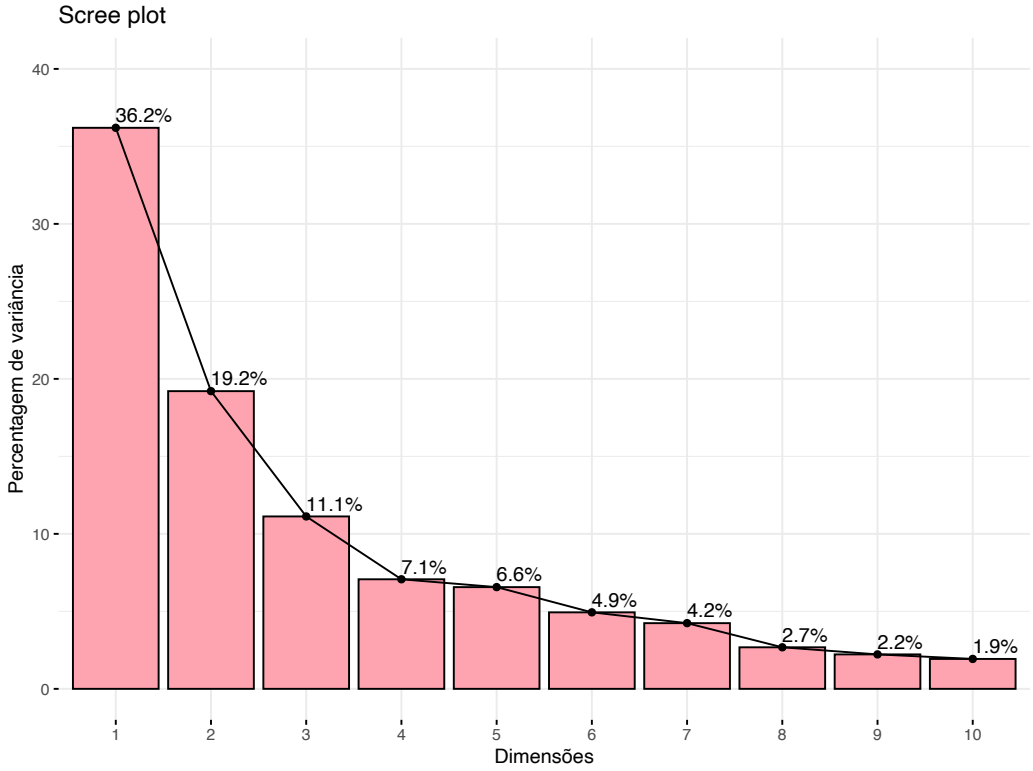


Figura 6.4: Screeplot dos dados de PCA, tendo representadas as percentagens de variância do sistema.

Tabela 6.3: Tabela dos valores próprios para cada uma das dimensões.

Dimensão (Dim)	Valor Próprio (λ)
Dim 1	4.7058503
Dim 2	2.4969737
Dim 3	1.4460720
Dim 4	0.9189739
Dim 5	0.8532282
Dim 6	0.6416570
Dim 7	0.5510283
Dim 8	0.3484974
Dim 9	0.2888799
Dim 10	0.2509025
Dim 11	0.2257886
Dim 12	0.1687702
Dim 13	0.1033779

Com base na Figura 6.4, verifica-se que para que a percentagem de informação seja de pelo menos 80% são necessárias cinco dimensões; neste caso obtendo-se uma percentagem de variância de 80.2%. Além disso, é possível retirar-se a informação de que apenas 10 dimensões (variáveis) apresentam contribuições para o sistema, dado que apenas se encontram 10 dimensões representadas no screeplot, o que significa que as três últimas dimensão apresentam uma contribuição irrelevante.

Analisando-se a Tabela 6.3, observa-se que apenas os valores próprios das três primeiras dimensões são superiores a um o que, pelo critério de Kaiser anteriormente referido, apenas são necessárias três dimensões para a descrição do sistema.

Visto que os dois critérios fornecem informações distintas, deve-se optar sempre pelo critério que tenha menor perda de informação, que neste caso concreto é o critério dos 80% o que permite concluir que o número de dimensões a usar é cinco. Isto porque quando se verifica a percentagem cumulativa de três dimensões esta é de 66,5%, resultando numa perda de informação de 33,5%, enquanto que com cinco dimensões esta perda é de apenas 19,8%.

Tal como referido anteriormente são necessárias cinco variáveis para que haja a menor perda de informação possível, o que leva a que seja indispensável que se proceda à sua determinação. Para isso recorre-se às contribuições que cada variável contém em cada dimensão. Para este tipo de estudo pode-se utilizar tabelas de contribuições de variáveis, como a Tabela 6.4 ou a representação gráfica, sendo em ambos os casos as conclusões a retirar exatamente as mesmas. Aqui optou-se pelas tabelas, encontrando-se os gráficos correspondentes a essa análise da Figura A1 a A6 no Anexo A.

Tabela 6.4: Contribuição de cada variável nas cinco dimensões.

Variáveis	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Álcool	2.083097	23.391882	4.300755	0.0318848	7.057718
Ácido Málico	6.011695	5.059393	0.792329	28.825117	0.124000
Cinza	4.206853e-04	9.989949	39.215637	4.587117	2.045628
Alcalinidade da cinza	5.727426	0.011215	37.464236	0.370386	0.436960
Magnésio	2.016174	8.9780535	1.709737	12.376083	52.859954
Total de fenóis	1.557572e+01	0.423013	2.136829	3.923107	2.229599
Flavonóides	1.788734e+01	0.001128	2.270504	2.319370	1.188663
Fenóis não flavonóides	8.912201	0.082826	2.902531	4.133130	25.070347
Proantocianinas	9.823804	0.154463	2.233659	15.924612	1.873061
Intensidade da Cor	7.852920e-01	28.089541	1.885299	0.434619	0.584258
Tonalidade	8.803953	7.797226	0.726278	18.298838	3.014200
Diluição de vinhos OD280/OD315	1.415019e+01	2.705899	2.755752	3.390045	1.023354
Prolina	8.222684	13.315408	1.606452	5.385688	2.492256

Da análise da Tabela 6.4, consideram-se como variáveis principais as que tiverem maior contribuição em cada dimensão (especialmente nas primeiras componentes), uma vez que esta nos dá informação sobre a percentagem que cada variável apresenta para as componentes principais. As cinco componentes principais são: os flavonóides, a intensidade da cor, a cinza, o ácido málico e, por fim, o magnésio, encontrando-se de acordo com a sua importância da primeira à quinta dimensão,

respetivamente. Na Figura 6.5, obtida através dos valores cos2 (coordenadas das variáveis no gráfico ao quadrado) e que nos fornece informação sobre a qualidade da representação das variáveis na componente em estudo, estão representadas as principais variáveis para a primeira e segunda dimensões, que vêm comprovar o que foi anteriormente discutido, isto é, que os flavonoides são a principal variável na dimensão 1 e a intensidade da cor a principal variável na dimensão 2.

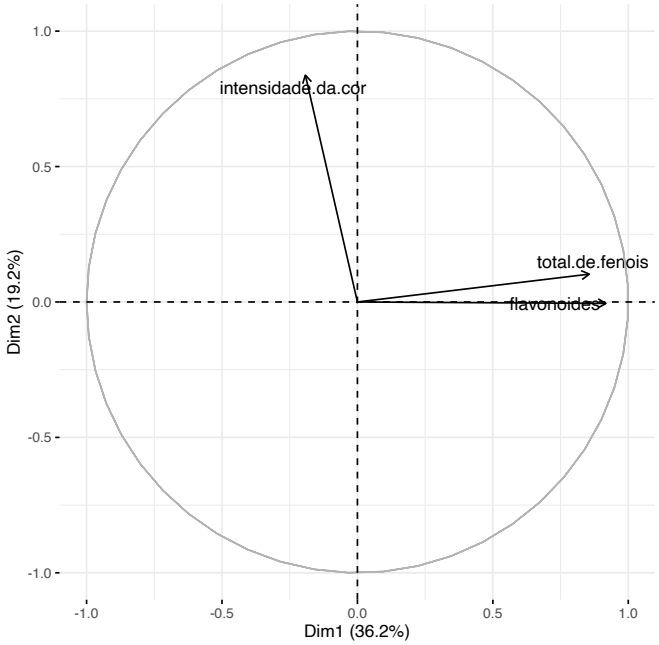


Figura 6.5: Representação das principais variáveis para a primeira e segunda dimensão, usando os valores de cos2.

O último passo a realizar é a validação dos resultados, que serve para verificar se os resultados obtidos se encontram de acordo com os esperados comparativamente à base de dados. Uma vez que se concluiu que as loadings (variáveis) se agrupam em três grupos, existe a necessidade de se validar esse resultado. Para isso utiliza-se os dados de correlação. A correlação tem como objetivo determinar a relação estatística entre duas ou mais variáveis, sendo que pode tomar valores entre -1 a 1, que são denominados de coeficientes de correlação. De acordo com o valor do coeficiente esta pode ser forte, caso apresente valores próximos de -1 ou 1 e fraca quando os valores são próximos de zero, para além disso pode ainda ser positiva, onde tal como referido teremos um comportamento idêntico entre as variáveis; negativa, existindo assim um comportamento contrário entre as variáveis ou nula, onde não existe qualquer tipo de correlação. Este tipo de estudo ocorre tendo por base uma matriz de correlação. Esta matriz, vai conter na sua diagonal uma correlação positiva forte visto que temos a correlação da variável com ela própria, sendo que vai ser a diagonal que vai dividir a matriz na parte superior e inferior, que serão iguais, visto que se trata de uma matriz simétrica. As matrizes de correlação podem ser apresentadas de diversas formas: neste caso, analisou-se a representação gráfica da matriz de correlação usando círculos (Figura 6.6), que sofrem alterações de cor e tamanho consoante o coeficiente de correlação entre variáveis se altera e a matriz de correlação com todos os

valores dos coeficientes entre as variáveis (Tabela A1 no anexo A). Estes dois tipos de representação da matriz de correlação fornecem a mesma informação, uma vez que são representações distintas da mesma matriz, no entanto, pode considerar-se que a matriz de correlação com os valores é complementar da representação gráfica, dado que no caso da representação gráfica não é possível determinar-se com certeza quais os valores do coeficiente de correlação que existe entre duas variáveis, sendo apenas possível verificar o seu intervalo de valores e se é uma correlação forte ou fraca, sendo esta duas informações retiradas pela cor que o círculo apresenta e pelo seu tamanho.

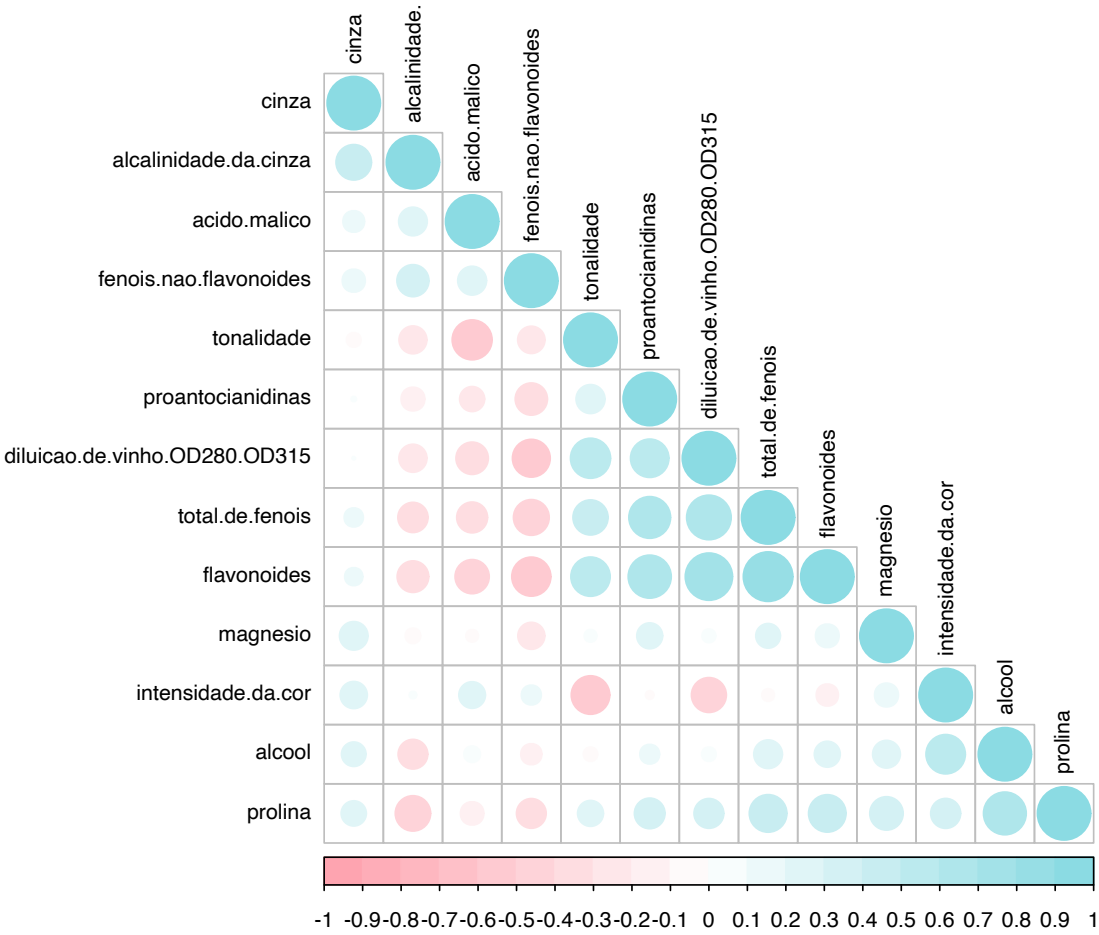


Figura 6.6: Matriz de correlação em que os círculos maiores representam uma maior correlação. As cores: rosa e azul indicam se a correlação é negativa e positiva, respetivamente.

Tal como já referido, a formação de clusters de variáveis deve-se essencialmente à correlação entre estas, sendo que esta pode ser positiva ou negativa. A Figura 6.6, permite que se verifique se os 3 clusters formados se encontram de acordo com o esperado. Para isso, vai-se analisar o tamanho e a cor dos círculos, sendo que quanto maior for o círculo maior será a correlação, sendo de seguida necessário verificar qual a cor, pois caso esta seja num tom azulado teremos uma correlação positiva e se for rosa será uma correlação negativa. Da Figura 6.6, observa-se que as maiores correlações positivas da alcalinidade da cinza são a cinza, o ácido málico e os fenóis não flavonóides, sendo que

estes formam um grupo, confirmando-se assim o cluster 1 identificado na figura 2. Passando ao cluster 2, que apresente na sua constituição os flavonóides, o total de fenóis, a diluição do vinho OD280/OD315, as proantocianidinas e a tonalidade, com base na Figura 6.6 verifica-se que a todos os elementos deste grupo se correlacionam com os flavonóides positivamente. Ainda da análise desta figura verifica-se que a prolina se correlaciona mais positivamente com o álcool, a intensidade da cor, o magnésio, os flavonóides, o total de fenóis e as proantocianidinas, dado que estas três últimas variáveis pertencem ao cluster 2, as restantes irão formar um cluster, no entanto devido a isto espera-se que os clusters dos flavonóides e da prolina se encontrem no mesmo lado do referencial, dado que todas as correlações apresentadas tem por base a dimensão 1. Pode-se também analisar a correlação negativa, sendo que esta será bastante útil para se definir a representação dos clusters no referencial cartesiano, dado que se a correlação é negativa entre duas variáveis de grupos distintos, a sua representação ocorrerá em quadrantes opostos. Analisando assim as correlações negativas da Figura 6.6, verifica-se que a alcalinidade da cinza se correlaciona negativamente com a tonalidade, as proantocianidinas, a diluição de vinhos, os flavonóides, o total de fenóis, o magnésio, a prolina e o álcool, o que faz com que o cluster que contenha esta variável se encontre do lado oposto do cluster 2 e 3, visto que são os clusters formados pelas variáveis com correlação negativa. No entanto, é ainda possível observar-se que das 9 variáveis pertencentes ao cluster 2 e 3, uma delas não apresenta correlação negativa com a alcalinidade da cinza, o que faz com que a intensidade da cor se encontre do lado oposto ao seu restante grupo o que é também verificado pela correlação negativa com a tonalidade, os flavonóides e a diluição de vinhos.

Com base na correlação entre variáveis foi possível verificar que estas se encontravam de acordo com o esperado, no entanto é necessário proceder-se ao cálculo do valor de prova para a formação de clusters com variáveis, descrito na secção 4.5. Os valores de prova para a junção entre cada uma das variáveis encontram-se na figura 6.7, onde a vermelho estão presentes os valores de prova aproximadamente imparciais (AU) e a verde se encontram os valores de bootstrap (BP) que correspondem à frequência com que o cluster é identificado em 1000 repetições. A azul encontram-se retângulos correspondentes aos clusters formados com um valor de AU igual ou superior a 95, sendo assim estes identificam os clusters estáveis, isto é, não formados aleatoriamente.

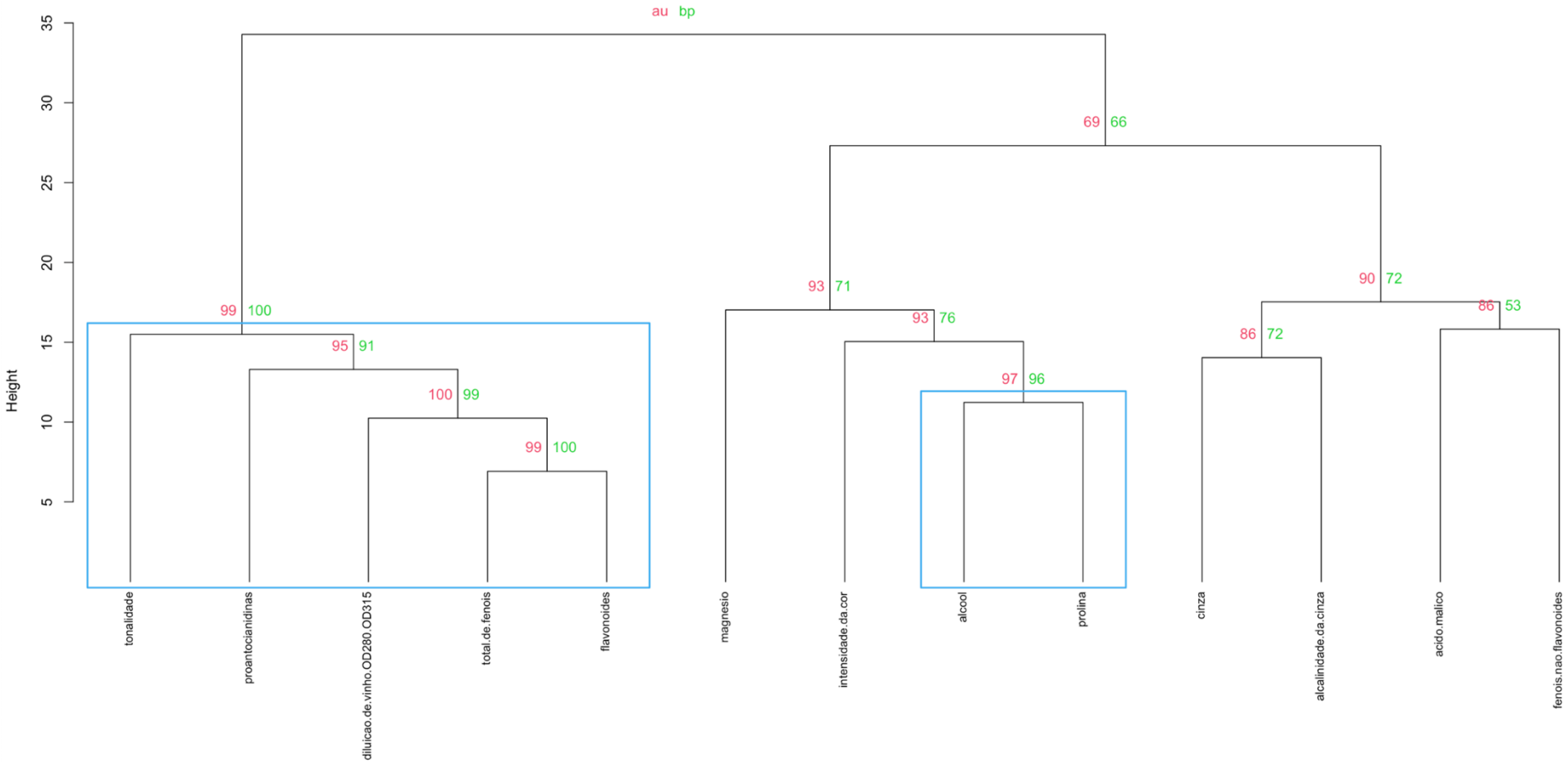


Figura 6.7: Dendrograma com os valores de prova (a vermelho) para cada ligação que se forma.

Analisando os clusters estáveis, constata-se que o cluster que se encontra mais à esquerda é o identificado como cluster 2, sendo o outro cluster identificado como uma fração do cluster 3. No entanto, é necessário ter em consideração que os valores calculados apresentam um erro associado, e que pode levar a alterações nos resultados sendo necessário verificar se isso acontece ou não, encontrando-se na Figura 6.8 o gráfico do erro padrão em função do valor de prova.

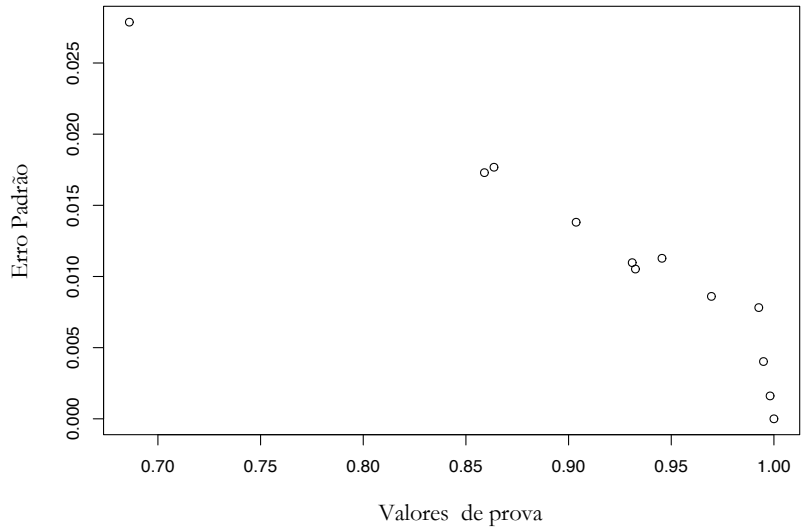


Figura 6.8: Gráfico do erro padrão em função do valor de prova.

Tal como referido existe um erro associado a cada um dos valores de prova existindo a necessidade de se verificar que alterações pode ser provocadas. Assim sendo, atendendo aos valores apresentados nas Figura 6.7 e 6.8 e à explicação, na secção 4.5, para o cálculo do intervalo de valores de prova, calculou-se o intervalo tendo em conta o erro padrão. Visto que quando o valor de prova é igual ou superior a 95%, os clusters formados são suportados pelos dados, foi-se verificar para cada cluster o intervalo de valores entre os quais o valor de prova pode variar. No entanto, estudou-se apenas o intervalo para os valores de prova que se encontravam na última ligação de cada cluster, 99%, 93% e 90%, bem como a ligação entre dois dos clusters, 69%. Ora, tendo em conta os valores do erro padrão, respetivamente, 0.002, 0.011, 0.014 e 0.028, calculou-se o intervalo de valores, tendo-se concluído que apenas o cluster constituído pelo magnésio, intensidade da cor, álcool e prolina vai puder passar de um cluster formado aleatoriamente para um cluster suportado pelos dados, visto que o seu intervalo de valores varia entre 91% e 95% $(0.93 \pm 2 \times 0.011)$. Relativamente à junção de dois clusters, esta junção será meramente aleatória visto que o intervalo de valores será de 63% a 75%. Conclui-se assim que dos três clusters já discutidos apenas um é formado aleatoriamente.

Embora os clusters formados pelas variáveis já se encontrem validados de duas formas distintas, é ainda possível validar estes clusters com base na importância que cada variável apresenta para a produção de vinhos. Tal como já referido o cluster 1 contém a cinza, a alcalinidade da cinza, o ácido málico e os fenóis não flavonóides, sendo estas quatro variáveis responsáveis pelo sabor do vinho [116,117,118]. As variáveis presentes no cluster 2 (flavonóides, proantocianidinas, tonalidade,

diluição de vinhos e total de fenóis) são fundamentais para a obtenção da cor e são também antioxidantes [20,119]. Por fim, a prolina, o álcool, o magnésio e a intensidade da cor, pertencentes ao cluster 3, estão associados à fermentação do vinho [23,24,58]

Com a distribuição das variáveis em grupos validada, é necessário proceder-se à validação da distribuição dos objetos (Figura 6.3), recorrendo-se para isso ao valor médio que cada variável toma em cada uma das classes ou, mais concretamente, em cada uma das áreas de cultivo diferentes. Para que essa validação seja possível é necessário recordar a suas distribuição: a classe 3 apresenta uma distribuição na zona correspondente ao cluster 1, que contém a cinza, o ácido málico, os fenóis não flavonóides e a alcalinidade da cinza; a classe 1 apresenta uma distribuição na zona correspondente ao cluster 3, na qual entram as seguintes variáveis: prolina, magnésio, álcool e intensidade da cor e na zona do cluster 2, onde se encontram as variáveis: tonalidade, diluição de vinhos OD280/OD315, flavonóides, total de fenóis e proantocianidinas; por fim, a classe 2, apresenta baixos valores para os cluster 1 e 3, dado que possuem poucos pontos nos locais caracterizados por estes grupos, sendo que apresenta uma exceção para a alcalinidade da cinza, uma vez que esta está no 3º quadrante e esta classe apresenta uma grande distribuição nesses locais. Para a sua validação, recorreu-se à representação dos valores médios a partir de gráficos de barras. De modo a tornar mais fácil as observações, os gráficos encontram-se divididos de acordo com os clusters formados para as variáveis, estando apresentados na Figura 6.9, os valores médios para o cluster 1, que contém a alcalinidade da cinza, na Figura 6.10, os valores médios são referentes ao cluster 2, que contém os flavonóides e a Figura 6.11, apresenta os valores referentes ao cluster 3, onde está presente a prolina.

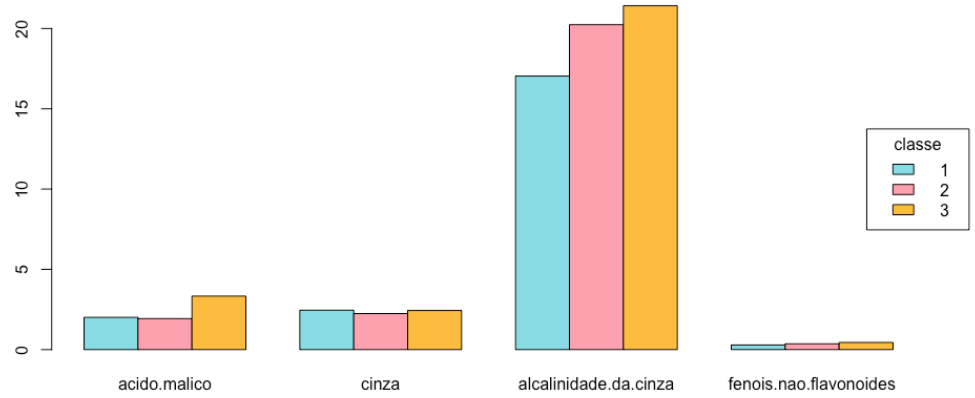


Figura 6.9: Valores médios de cada classe para as variáveis pertencentes ao cluster 1 da figura 6.2.

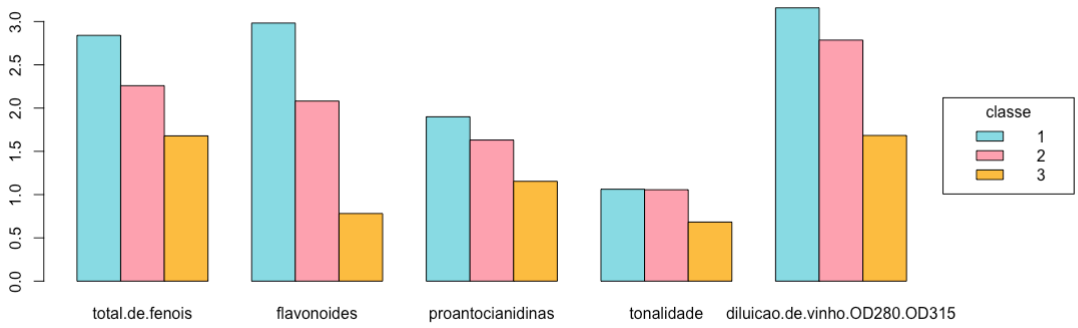


Figura 6.10: Valores médios de cada classe para as variáveis pertencentes ao cluster 2 da figura 6.2.

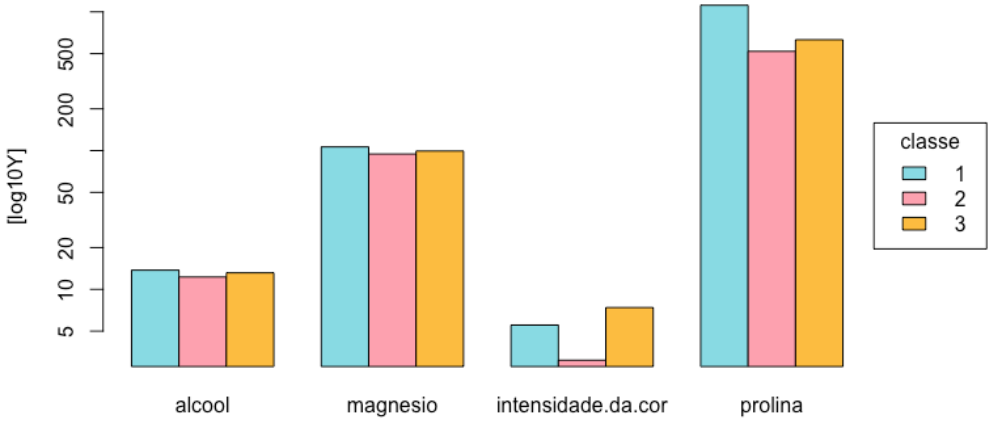


Figura 6.11: Valores médios de cada classe para as variáveis pertencentes ao cluster 3 da figura 6.2.

Da análise dos gráficos dos valores médios de cada classe para as variáveis pertencentes a cada um dos grupos, verifica-se, tal como esperado que a classe 3 apresenta maiores valores para o grupo que contém a alcalinidade da cinza. No que diz respeito à classe 1, esta apresenta valores elevados para os clusters dos flavonóides e da prolina, sendo dado a distribuição dos objetos desta classe ocorrer no 1º e 4º quadrantes, o valor para a intensidade da cor será menor do que para os elementos da classe 3, que apresentam a distribuição dos vinhos na sua grande maioria no 3º quadrante, local onde se encontra a componente desta variável. Relativamente à classe 2, verifica-se que esta apresenta valores mais baixos para as variáveis que se encontram no 2º quadrante, o que é justificado pelo facto de a sua distribuição ocorrer essencialmente no 3º e 4º quadrantes, logo uma vez que a alcalinidade da cinza se encontra no 3º quadrante vai apresentar um valor intermédio, o que acontece também para as variáveis do cluster dos flavonóides dado que das 5 variáveis pertencentes a este grupo 3 se encontram no 4º quadrante, o que leva a que este grupo possua valores intermédias para o grupo dos flavonóides e para a alcalinidade da cinza.

6.2-Análise Hierárquica de Agrupamento (HCA)

Outro estudo realizado foi a análise hierárquica de agrupamento (HCA), em que os diferentes objetos se vão agrupar segundo características comuns. Uma vez que já se realizou o estudo de PCA e se concluiu que cinco componentes são suficientes para se ter uma descrição de 80% do sistema, este estudo vai permitir a comparação dos grupos formados quando se usam apenas as cinco componentes principais obtidas do PCA com os grupos formados quando o estudo é realizado usando todas as variáveis.

Um passo essencial neste tipo de estudos é estipular-se qual o método a ser aplicado para o cálculo da distância, que mede a similaridade entre objetos, e qual o tipo de ligação entre grupos, que é a distância entre um objeto e um grupo ou até mesmo entre grupos já formados. Relativamente ao método usado para o cálculo da distância, tendo em conta os métodos existentes para este cálculo desta descritos na secção 3.2.1, optou-se pela distância euclidiana, dado ser um método aplicado em

vários tipos de dados e que se mostrou robusto no sentido de produzir resultados corretos. A distância euclidiana é definida pela Equação (3.3), em que cada um dos objetos terá um n=5, quando se opta por considerar as 5 primeiras dimensões ou componentes principais e um n=13 no caso de a base de dados possuir todas as componentes, dado que o n corresponde ao número de dimensões.

Antes de se proceder à escolha do tipo de ligação entre grupos é necessário relembrar que o objetivo do uso do HCA é a formação de grupos distintos, em que os objetos que constituem cada grupo possuem características semelhantes entre si e diferentes quando se comparam com as características dos outros grupos. Uma vez que os objetos presentes na base de dados se distribuem em três classes distintas, vai-se verificar se os objetos seguem a formação de agrupamentos naturais, isto é, se os objetos se distribuem de acordo com as classes ou não. Para isso é importante indicar quais os objetos que pertencem a cada uma das classes, bem como quantos objetos é que cada classe possui tendo em conta a base de dados. Essas informações encontram-se na Tabela 6.5.

Tabela 6.5: Distribuição dos vinhos (objetos) de acordo com a classe (tipo de cultivo).

Classe	Objetos	Número de Objetos
1	1-59	59
2	60-130	71
3	131-178	48

De modo avaliar-se qual o melhor método de ligação realizaram-se diversos estudos, usando os diferentes tipo de ligação existentes: “single”, “average”, “complete”, “ward.D2”, “mcquitty”, “centroid” e “median”, como descrito na secção 3.2.2. Entre os dendrogramas obtidos usando estes métodos de ligação apenas o método “ward.D2” permite que se observem três grupos distintos, com tamanhos semelhantes aos esperados, para ambos os estudos, permitindo assim comparar-se as diferenças que existem quando se usam todas as variáveis ou se usam apenas 5 dimensões. Este método de ligação permite o cálculo da distância a partir da soma dos desvios quadrados dos pontos (objetos) aos seus centróides.

Atendendo ao facto de que se pretende verificar quais as diferenças que existem entre os dois dendrogramas obtidos pelo método de ligação “ward.D2” encontra-se representado na Figura 6.12, o dendrograma obtido quando o estudo ocorre tendo por base as cinco componentes principais e na Figura 6.13 o dendrograma obtido a partir do uso de todas as variáveis. No entanto, de modo que seja possível perceber se os grupos são realmente formados pelos objetos que pertencem a cada classe, a constituição de cada cluster encontra-se nas Tabelas 6.6 e 6.7. Relativamente aos restantes dendrogramas usados para a escolha do método de ligação mais adequado encontram-se no anexo B, da figura B1 à B12.

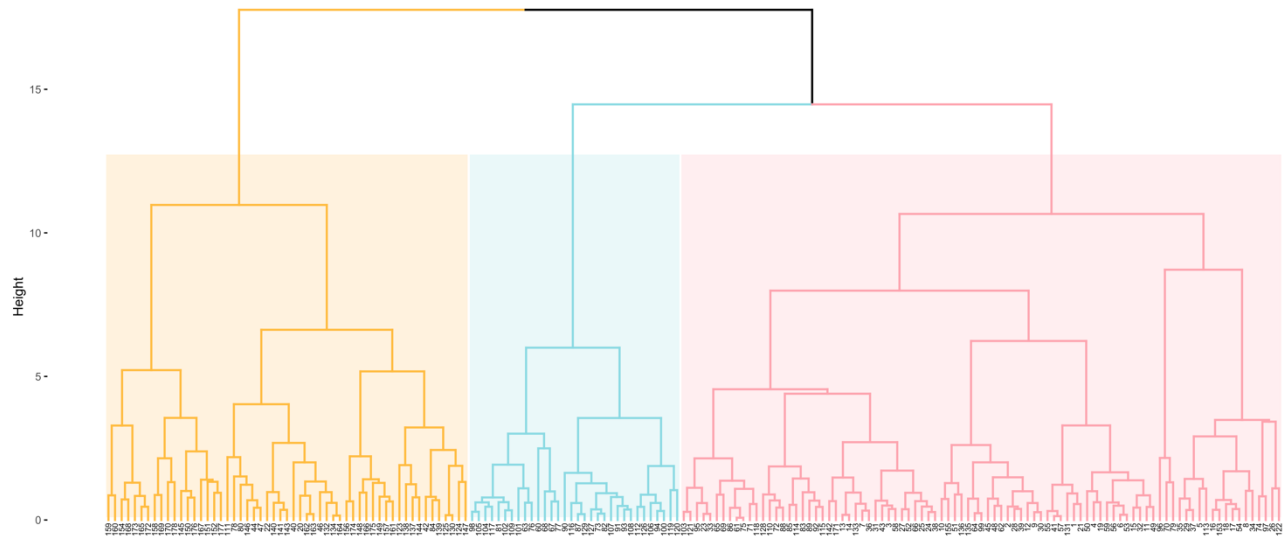


Figura 6.12: Dendrograma baseado nas cinco componentes principais do PCA, com distância euclidiana e o método de ligação "Ward.D2".

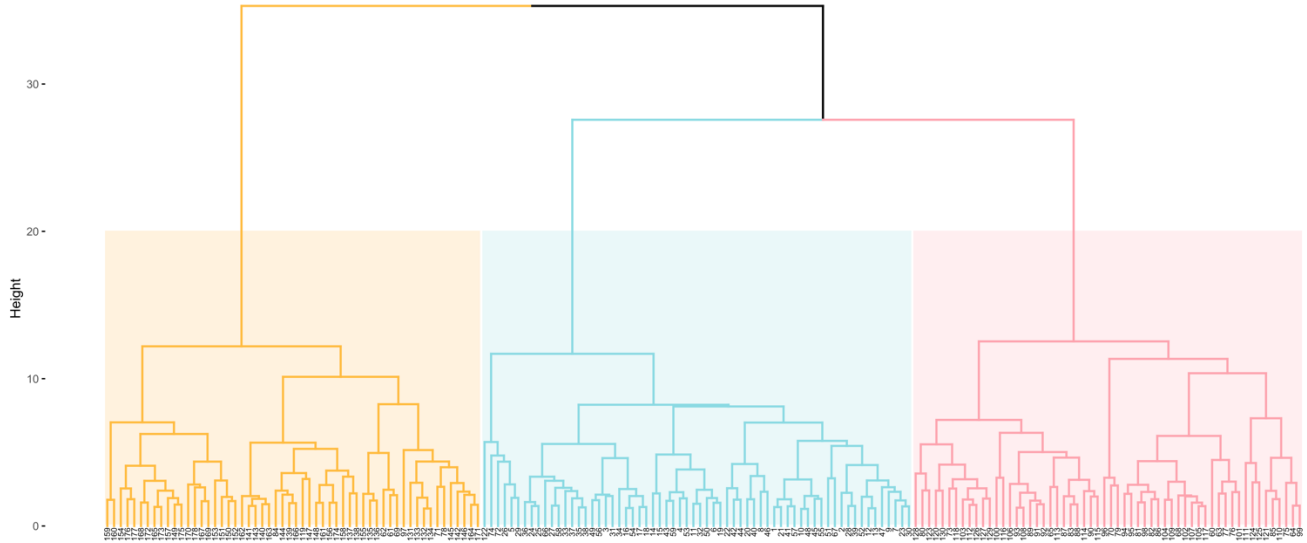


Figura 6.13: Dendrograma de todos os objetos sem aplicar o PCA, com distância euclidiana e o método de ligação "Ward.D2".".

Tabela 6.6: Distribuição dos objetos de acordo com os clusters formados na figura 6.12.

Cluster	Objetos que se encontram no cluster	Classe
Rosa	52 objetos da classe 1 31 objetos da classe 2 8 objetos da classe 3	Classe 1
Azul	32 objetos da classe 2	Classe 2
Laranja	7 objetos da classe 1 8 objetos da classe 2 40 objetos da classe 3	Classe 3

Tabela 6.7: Distribuição dos objetos de acordo com os clusters formados na figura 6.13.

Cluster	Objetos que se encontram no cluster	Classe
Rosa	58 objetos da classe 2	Classe 2
Azul	59 objetos da classe 1 5 objetos da classe 2	Classe 1
Laranja	8 objetos da classe 2 48 objetos da classe 3	Classe 3

Iniciando a análise pelos grupos formados quando se usam apenas as cinco componentes principais obtidas com o PCA (Figura 6.12 e Tabela 6.6) verifica-se que o grupo a laranja possuí 55 objetos, dos quais 40 pertencem à classe 3, o que faz com que seja esta a classe dominante neste grupo, faltando apenas 8 dos objetos que a constituem. Quanto aos restantes 15 objetos, 7 pertencem à classe 1 e 8 à classe 2. No que diz respeito ao grupo a azul, observa-se a presença de 32 objetos, sendo todos eles pertencentes à classe 2. Logo, este grupo representa parte da classe, no entanto faltam mais de metade dos objetos desta que é a classe com maior número de objetos. Relativamente ao grupo que falta, rosa, é o maior grupo deste dendrograma, tendo na sua constituição mais de metade dos objetos (91), em que 52 são referentes a elementos da classe 1, o que faz com que este grupo seja referente a esta classe. De notar que os grupos formados têm apenas em consideração 5 dimensões, as componentes principais definidas no estudo de PCA, o que leva a que os objetos possam ser considerados semelhantes embora pertençam a grupos diferentes devido a apresentarem valores semelhantes para as características estudadas.

O próximo passo a realizar é a análise do dendrograma obtido através do uso de todas as variáveis (Figura 6.13), bem como da Tabela 6.7 onde se encontram descritas as observações feitas para este. Da sua análise verifica-se também a existência de 3 grupos com tamanhos semelhantes ao esperado, tendo como suporte a base de dados original. Começando a análise pelo grupo de cor rosa, verifica-se que este é constituído por 58 objetos, sendo todos eles pertencentes à classe 2, faltando neste grupo 13 dos 71 objetos. O próximo grupo a ser analisado é o de cor laranja, em que dos 56 objetos presentes 48 são pertencentes à classe 3, estando presentes todos os elementos que constituem esta classe. No último grupo, observa-se a presença dos 59 objetos pertencentes à classe 1, bem como 5 objetos pertencentes à classe 2. Com esta

análise, verifica-se que a classe que apresenta os objetos mais dispersos é a classe 2, pois apresenta os objetos em dois grupos distintos.

Da análise dos dois dendrogramas e respetivas tabelas, verifica-se que ambos os estudos apresentam objetos em agrupamentos diferentes dos esperados inicialmente, o que se pode dever à presença de ruído. No entanto, analisando mais ao pormenor os dois dendrogramas verifica-se que o dendrograma obtido apenas com cinco dimensões apresenta mais objetos em classes diferentes das esperadas, pelo que quando se pretende obter resultado mais fidedignos deve-se optar pelo estudo com todas as variáveis sem o uso de PCA.

Tal como aconteceu com o estudo de PCA em que foi necessário realizar-se a validação dos agrupamentos formados pelas variáveis e pelos objetos, é também necessário que se realize a validação dos clusters formados pelo estudo de HCA. A validação dos agrupamentos que se formam é necessária, pois os clusters formados podem ocorrer devido a erros de amostragem ou ruído.

Os passos necessários para a realização da validação encontram-se descrito no Capítulo 4, pelo que serão apenas apresentados os resultados mais relevantes para a validação dos resultados.

Uma vez que os resultados do PCA já se encontram validados, e se verificou que os objetos da base de dados formam clusters significativos (H=0.28), procedeu-se assim à determinação do número ideal de clusters recorrendo ao método do cotovelo, de silhueta e de testes estatísticos. Analisando as figuras correspondentes à representação gráfica de cada um dos métodos anteriormente enunciados (Figura 6.14 a 6.16) verifica-se que o número ideal de clusters para o estudo é de 3.

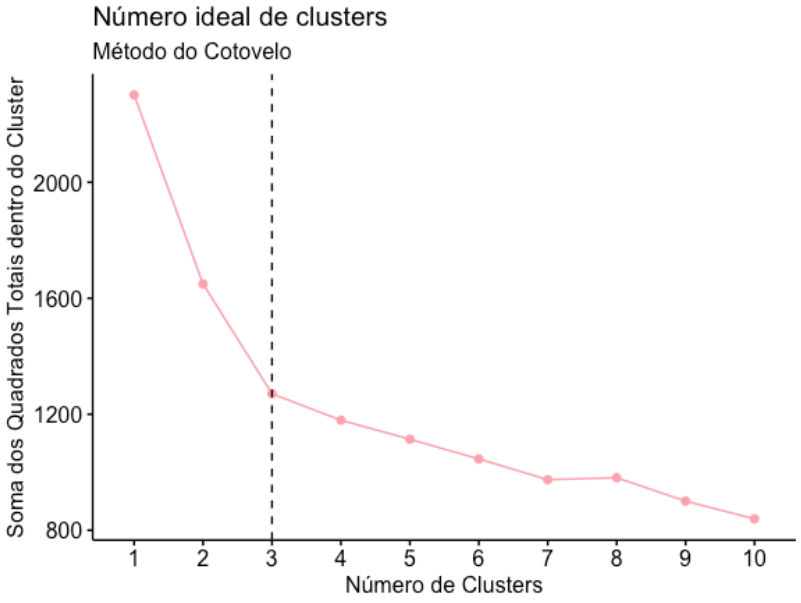


Figura 6.14: Representação gráfica do método do cotovelo, havendo evidência de que o número ideal de clusters é 3.

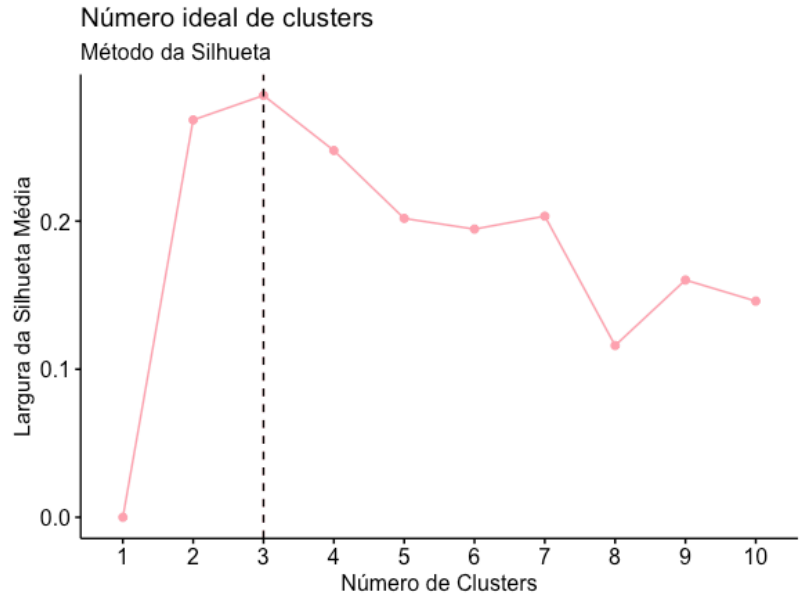


Figura 6.15: Representação gráfica do método de silhueta, havendo evidência de que o número ideal de clusters é 3.

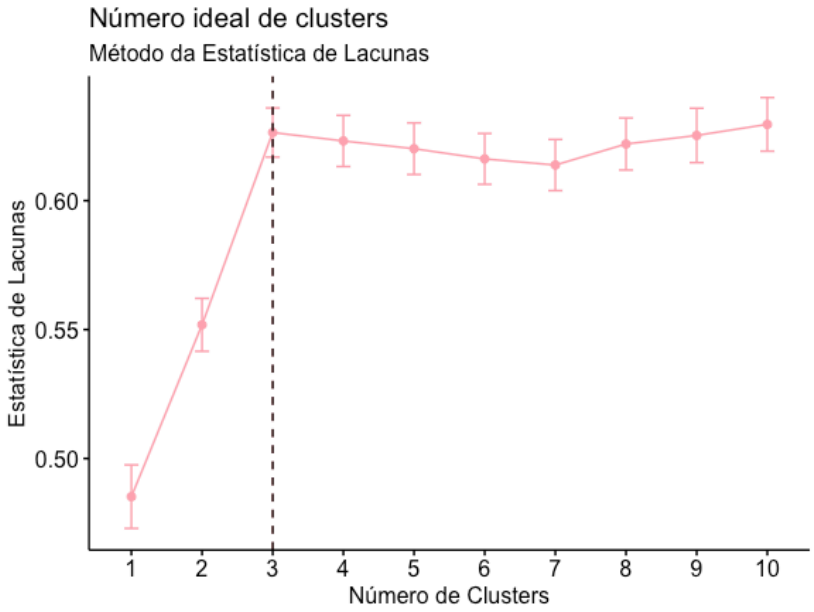


Figura 6.16: Representação gráfica do método da estatística de lacunas, havendo evidência de que o número ideal de clusters é 3.

Além dos métodos enunciados é possível fazer-se a determinação do número ideal recorrendo a função NbClust, em que são estudados 30 índices diferentes. Obteve-se desse estudo os dados representados na Figura 6.17, em que se observa que a maioria dos 30 índice definiu como número ideal de clusters como sendo 3.

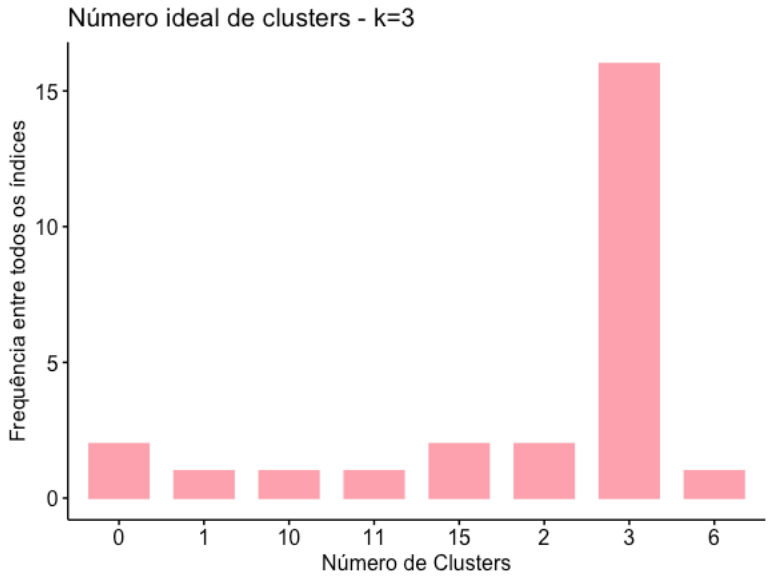


Figura 6.17: Gráfico de frequências para o número ideal de clusters de acordo com os 30 índices**.**

De modo a avaliar-se a qualidade dos resultados recorreu-se ao estudo do coeficiente de silhueta, que se encontra representado graficamente na Figura 6.18. Este tipo de estudo permite verificar se os objetos se encontram bem enquadrados, e além disso é ainda possível obter uma estimativa da distância média entre os clusters vizinhos.

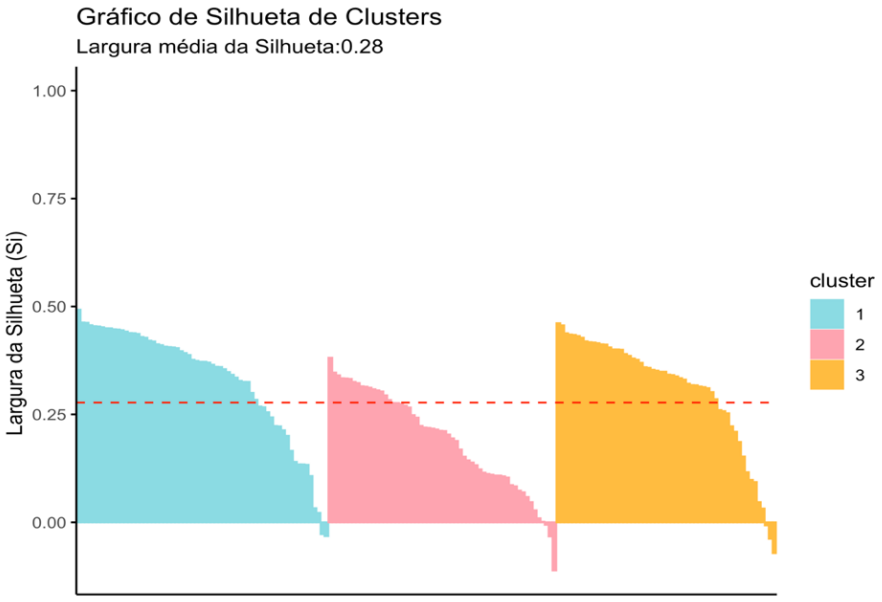


Figura 6.18: Gráfico de Silhueta de Clusters.

Através da análise da Figura 6.18 é possível verificar que existem objetos que estão mal enquadrados, isto é, pertencem a outro grupo diferente do que aquele a que estão atribuídos. Essa análise é possível devido à presença de valores negativos no gráfico. Na Tabela 6.8, encontram-se os objetos que foram considerados como estando no grupo errado, tendo por base o coeficiente de silhueta.

Tabela 6.8: Distribuição dos objetos com uma largura de silhueta negativa.

Objetos	Cluster em que se encontra	Cluster mais provável
66	1	2
67	1	2
75	2	1
96	2	1
99	2	1
61	3	2
78	3	2
97	3	2

Da Tabela 6.8, podemos retirar a informação de que 8 objetos que se encontram atribuídos a clusters que não o esperado de acordo com as suas características. Os 8 objetos apresentados pertencem todos à mesma classe de acordo com a base de dados (classe 2), no entanto apenas 5 desses apresentam como mais provável essa classe.

No processo de validação o passo que se segue consiste na determinação do melhor algoritmo de clustering a usar, isto é, qual o melhor método de estudo a utilizar e qual o número de clusters que se espera obter com esse método. Para isso, recorreu-se ao uso das medidas internas e de estabilidade, descrita na seção 4.4.1. Para estes dois tipos de medidas foram estudados cinco métodos distintos: hierárquico, K-médias, CLARA, AGNES e PAM quando o número de clusters formados varia entre 2 e 5. Os resultados obtidos encontram-se na Tabela 6.9 para o estudo das medidas internas e na 6.10 para o estudo das medidas de estabilidade.

Tabela 6.9: Método de estudo e número de clusters esperados para cada uma das medidas internas.

	Método de estudo	Número de clusters
Conectividade	Hierárquico	2
Coeficiente de Silhueta	K-médias	3
Índice de Dunn	K-médias	3

Tabela 6.10: Método de estudo e número de clusters esperados para cada uma das medidas de estabilidade.

	Método de estudo	Número de clusters
APN	K-médias	3
AD	K-médias	5
ADM	K-médias	3
FOM	K-médias	5

Ao analisar-se a Tabelas 6.9 verifica-se que o número ideal de clusters é 3, dado que duas das três medidas internas sugerem este cluster como sendo o melhor. Relativamente ao algoritmo de agrupamento

(método de estudo) o que apresenta melhor desempenho é o estudo a partir do método K-médias. Analisando a Tabela 6.10, constata-se que o melhor método de estudo para as quatro medidas é o K-médias. No entanto relativamente ao número ideal de clusters existe um empate, dado que duas das medidas tem um melhor desempenho quando apenas existem 3 grupos enquanto outras duas apresentam uma performance melhor quando existem 5 agrupamentos.

6.3-Eliminação do ruído

O último estudo quimiométrico a ser realizado foi a eliminação do ruído. Este foi aplicado apenas no final dos restantes estudos para que se tivesse um conhecimento prévio de possíveis *outliers*, dado que o método usado para a eliminação tem por base a densidade espacial de objetos, tendo sido possível observar a respetiva distribuição destes no estudo de PCA (através do biplot). Para além disso, no estudo de HCA, tendo em conta o gráfico de silhueta de clusters (Figura 6.18) foi também possível identificar possíveis *outliers* de grupos, dado que os objetos com uma silhueta negativa se encontram mal classificados no grupo em que se encontram.

Tal como referido, o principal objetivo deste estudo foi proceder-se à eliminação dos objetos considerados *outliers*. O processo de eliminação foi feito por dois processos: i) procedeu-se à eliminação dos objetos estudando cada uma das classes separadamente e ii) o estudo recaiu sobre todo o conjunto de objetos. Salienta-se que ambos os processos foram efetuados utilizando as dimensões consideradas PC's (apenas as 3 com mais importância e as 5) e todas as dimensões, de modo que fosse também possível verificar qual a influência que o número de dimensões apresentava no estudo, sendo estudado o efeito da normalização dos dados nos resultados.

Para o processo de eliminação foram aplicados dois critérios, que, tal como referido, tiveram como base a densidade dos objetos. Esses critérios foram, a distância média de cada vizinho mais próximos (vmp), Figura 6.19, e o número médio de vmp para cada objeto, Figura 6.20, que se complementam dado que ambos eliminam objetos. No caso da distância média dos vmp são considerados ruído os objetos que apresentam uma distância superior a 3 vezes a distância média, sendo que antes de se considerar esses valores como ruído foram analisados os resultados obtidos quando se considerou uma distância 2, 4 e 5 vezes superiores à distância média. Relativamente ao estudo dos objetos considerados ruído quando analisado o número médio de vmp para cada objeto, foram considerados aqueles que apresentavam um número de pontos inferior a 1/3 do número médio de vmp para cada ponto, não tendo sido realizada qualquer análise alterando esse valor uma vez que se espera que os pontos eliminados sejam o que se encontram na periferia, logo os que apresentam menos densidade, não se devendo baixar muito da metade desses valores.

Resumindo, o que foi referido anteriormente para os critérios definidos os passos que se devem seguir são:

- Cálculo da distância do vmp para cada um dos objetos;
- Cálculo da média das distâncias dos vmp;
- Determinação do número de pontos no raio de 3x a média das distâncias, para cada ponto;

- Cálculo do número médio de pontos do critério acima;
- Determinação dos pontos que apresentam um número de vmp acima de 1/3 do número médio (Figura 6.20);

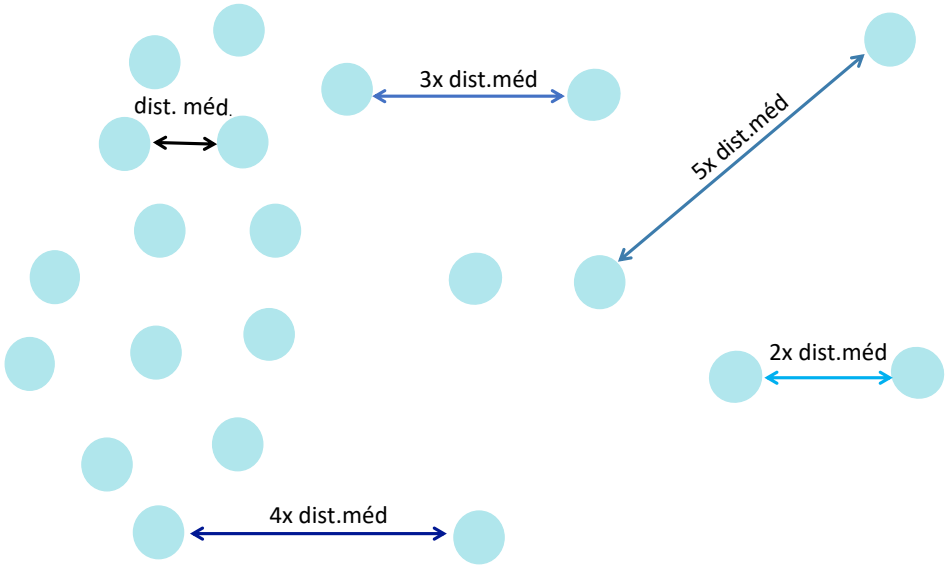


Figura 6.19: Representação do critério de eliminação de pontos que se encontram a uma distância 2, 3, 4 e 5x superior à distância média entre o vmp.

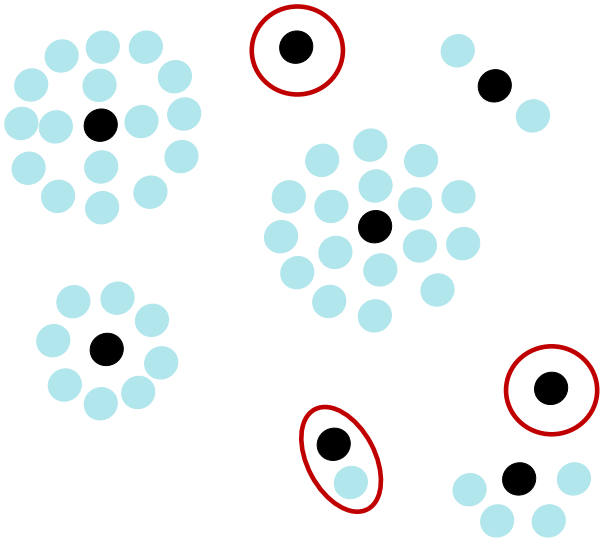


Figura 6.20: Representação do critério de eliminação dos pontos que apresentam 1/3 do número médio de vmp para cada ponto. Considerou-se um ponto central (preto) para cada grupo formado sendo os pontos a azul os vmp desse ponto. Atendendo a que esse número médio de vmp é 6, para o exemplo, será considerado ruído todos os pontos que apresentam um número de vmp inferior a 1/3 de 6, ou seja 2 (grupos com circunferência a vermelho).

Na Tabela 6.11, referente ao estudo usando apenas três dimensões, é possível observar-se quantos objetos são eliminados quando se faz a normalização dos dados e quando esta não é realizada, bem como qual o efeito que existe quando se considera as classes isoladas ou o conjunto como um todo. No entanto, da análise da tabela é apenas possível verificar qual o número de objetos considerado ruído e não a

localização destes pelo que, uma vez que apenas se utilizam três dimensões, realizou-se o estudo da distribuição dos objetos, tendo sido usado para isso gráficos tridimensionais representados na Figura 6.21.

Tabela 6.11: Número de objetos considerados ruído quando se recorre ao uso das três dimensões correspondentes às PC's (flavonoides, intensidade da cor e cinza), com e sem normalização dos dados.

Objetos	Conjunto de dados COM NORMALIZAÇÃO	Conjunto de dados SEM NORMALIZAÇÃO
Classe 1 (1-59)	4	6
Classe 2 (60-130)	15	17
Classe 3 (131-178)	14	2
Todos	38	51

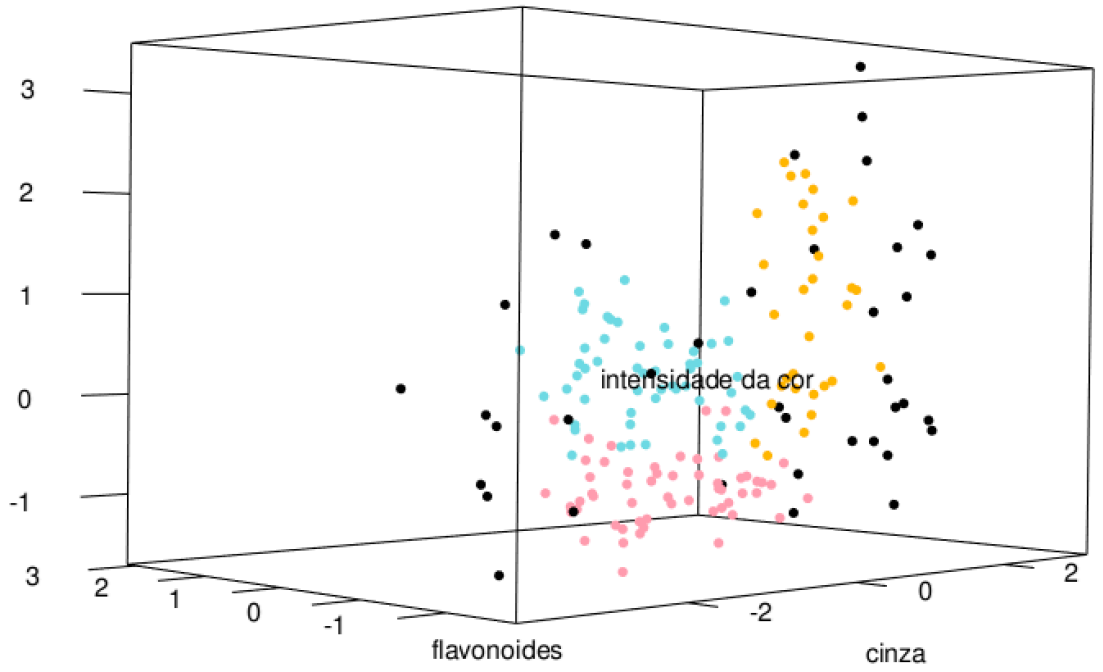


Figura 6.21: Gráfico tridimensional da distribuição dos objetos, quando estes são normalizados. A preto encontram-se os objetos considerados ruído, a azul estão assinalados os objetos pertencentes à classe 1, a cor de rosa os pertencentes à classe 2 e a laranja os pertencentes à classe 3.

A Figura 6.21 vem demonstrar que os objetos eliminados (pontos a preto na figura) se encontram na periferia do conjunto de dados, no entanto ao fazer-se uma análise conjunta dos dados obtidos na Tabela 6.11 e do gráfico verifica-se que o número de objetos eliminados quando se considera todo o conjunto de dados, 38, é diferente da soma dos objetos eliminados individualmente por cada classe, 33, pelo que, antes de se retirarem conclusões relativamente à posição dos objetos eliminados é necessário analisar-se em que zona ocorre a eliminação dos objetos quando considerada apenas uma classe de cada vez. Assim sendo, encontram-se representados na Figura 6.22 os gráficos tridimensionais para a distribuição dos objetos tendo em consideração apenas cada classe individualmente, encontrando-se os seus valores normalizados.

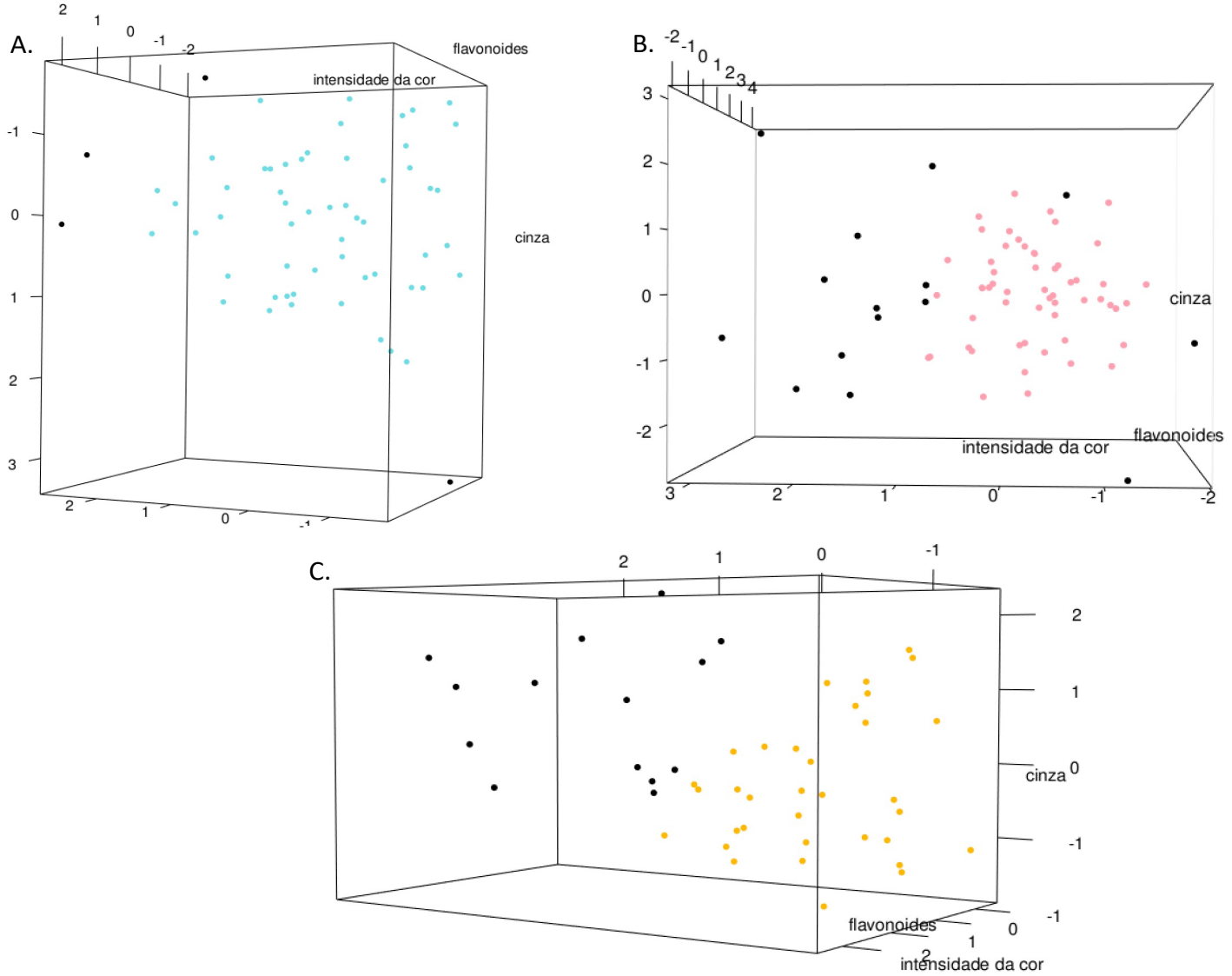


Figura 6.22: Gráfico tridimensional da distribuição dos pontos, encontrando-se a preto os objetos eliminados. A- Classe 1, B- Classe 2 e C- Classe 3.

Com base na figura 6.22 é possível verificar a presença dos objetos eliminados na zona da periferia do conjunto formados pelos objetos de cada classe, o que leva a que se possa assumir que atendendo ao uso dos critérios selecionados os objetos eliminados se encontrem na zona da periferia. Assim sendo, encontram-se de seguida as tabelas relativas ao número de objetos considerados ruído quando se analisam os dados usando as cinco dimensões correspondentes às PC's e todas as dimensões (Tabela 6.12 e 6.13, respetivamente).

Tabela 6.12: Número de objetos considerados ruído quando se recorre ao uso das cinco dimensões correspondentes às PC's (flavonoides, intensidade da cor ,cinza, ácido málico e magnésio), com e sem normalização dos dados.

Objetos	Conjunto de dados COM NORMALIZAÇÃO	Conjunto de dados SEM NORMALIZAÇÃO
Classe 1 (1-59)	8	7
Classe 2 (60-130)	16	29
Classe 3 (131-178)	3	7
Todos	50	78

Tabela 6.13: Número de objetos considerados ruído quando se recorre ao uso de todas as dimensões com e sem normalização dos dados.

Objetos	Conjunto de dados COM NORMALIZAÇÃO	Conjunto de dados SEM NORMALIZAÇÃO
Classe 1 (1-59)	0	35
Classe 2 (60-130)	0	22
Classe 3 (131-178)	0	5
Todos	5	42

Através da análise das tabelas relativas ao número de objetos considerados ruído (6.11,6.12 e 6.13) é possível verificar que o número de dimensões tem influência sobre a eliminação do ruído, uma vez que existe alteração do número de objetos eliminados quando se altera o número de dimensões. Também a normalização tem efeito sofre o ruído, dado que a quantidade de objetos eliminados para o mesmo conjunto de dados tende a ser menor, de maneira geral, quando existe normalização dos dados. Destas tabelas é ainda possível perceber-se que os objetos da classe 2 são os que apresentam mais *outliers*, o que se encontra de acordo com o esperado, tendo por base a Tabela 6.8, onde todos os objetos que se encontram mal classificados pertencem a este grupo, o que sugere uma maior dispersão. Para se comprovar que são efetivamente *outliers,* verificou-se quais os objetos desta classe que foram eliminados, sendo que 5 destes se encontravam entre os objetos eliminados quando a análise recaiu sobre o estudo com três dimensões,

enquanto 4 se encontram entre o ruído do estudo com cinco dimensões, mas apenas 1 é considerado ruído quando a análise recai sobre o estudo com todas as variáveis.

Com os objetos considerados *outliers* eliminados, procedeu-se à realização dos estudos realizados anteriormente, de modo que fosse possível verificar-se a influência do ruído nos resultados dos estudos de PCA e HCA.

Para cada um dos estudos definiu-se quais os dados, bem como o número de variáveis a usar. Assim sendo, para o novo estudo de PCA, tendo em conta o objetivo de determinação das componentes principais, recorreu-se ao uso de todos os objetos usando todas as variáveis, sendo fundamental que estas se encontrem normalizadas. No que diz respeito ao novo estudo de HCA, este foi realizado usando dois conjuntos de dados: o primeiro constituído por todas as variáveis normalizadas, bem como todos os objetos, enquanto o segundo, visto que no estudo inicial se usaram as cinco dimensões correspondentes às principais componentes do estudo de PCA, também neste novo estudo se usaram.

Em relação ao estudo de PCA, foi possível verificar que o ruído apresentava alguma influência, embora esta não fosse significativa, pois ao comparar o gráfico da Figura 6.23 com o gráfico da Figura 6.2, verifica-se que existem loadings que se deslocaram ligeiramente, sendo essa deslocação mais visível no caso da alcalinidade da cinza, do magnésio e dos flavonoides.

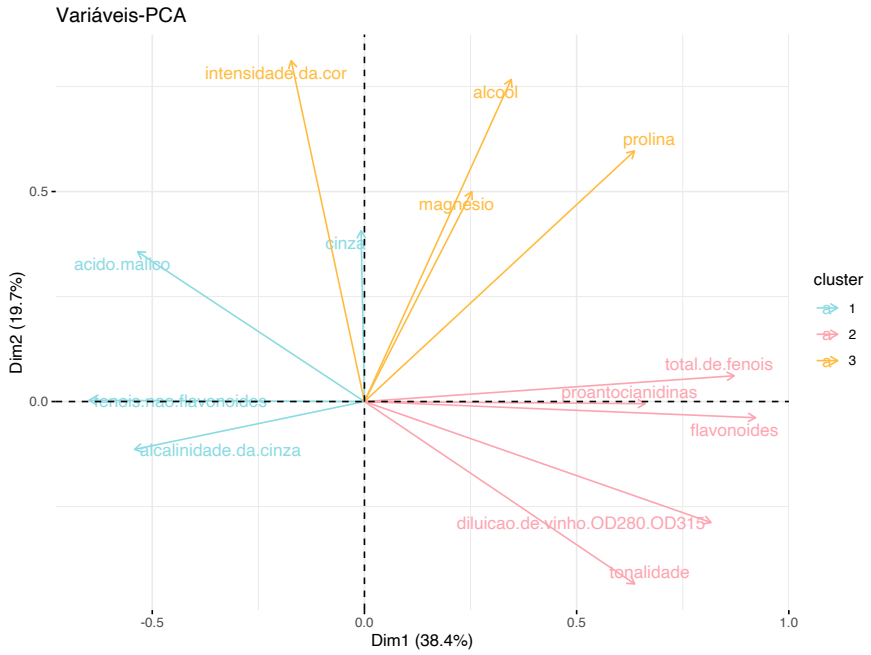


Figura 6.23: Representação das loadings em duas dimensões apontado para a formação de grupos, após a eliminação do ruído.

Outro aspeto também importante é o facto de que com a eliminação dos *outliers*, a duas dimensões o sistema passou a ter uma descrição de 58.1% enquanto anteriormente era de 55.4%, o que leva a que o ruído neste caso influenciasse negativamente a recuperação de informação. No entanto relativamente ao número de PC's necessárias para a descrição de 80% do sistema , continua a ser de 5. A única alteração prende-se com o facto de que a quinta dimensão, deixou de ser o magnésio passando a ser os fenóis não

flavonóides, encontrando-se no Anexo C os gráficos da contribuição das variáveis em cada dimensão bem como o mapa de fatores.

Da comparação do biplot obtido antes da eliminação do ruído (Figura 6.3) e do obtido após a eliminação do ruído (Figura 6.24) é possível verificar que a eliminação dos objetos teve uma pequena influência na movimentação dos conjuntos, o que levou a que o conjunto constituído pela classe 2 não apresente uma sobreposição tão grande quanto a que existia antes da eliminação dos *outliers*. Assim sendo, pode-se concluir que a presença de *outliers* tem influência nos resultados obtidos no estudo de PCA.

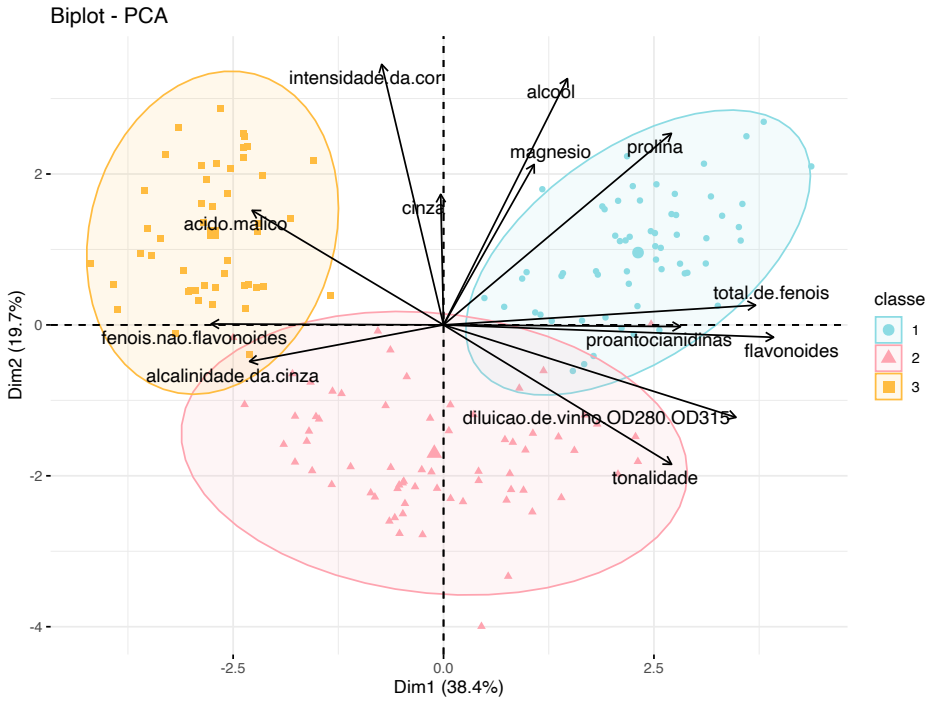


Figura 6.24: Representação das loadings ou variáveis (setas) e dos scores ou objetos (pontos) após a eliminação do ruído.

Neste caso de estudo, não se procedeu ao processo de validação pois os clusters formados pelas variáveis são as mesmos, tendo já estes sido validados no estudo do PCA antes da eliminação do ruído. Em relação à distribuição dos objetos pelos clusters é em todo semelhante à distribuição que já existia anteriormente, o que levou a que não se realizasse novamente o processo de validação.

Um vez estudado o efeito da presença de *outliers* no estudo de PCA, realizou-se o estudo do efeito do ruído nos resultados de HCA. De modo a ser possível verificar qual a influência que a presença de *outliers* tinha sobre os resultados, no estudo usando todas as variáveis e apenas as correspondentes às PC's empregou-se a distância euclidiana e como método de ligação o "Ward.D2", visto que só assim foi possível determinar quais as diferenças que é possível observáveis.

A primeira parte do estudo recorreu ao uso dos dados resultantes da eliminação de ruído quando consideradas todas as variáveis, resultando no dendrograma da Figura 6.25. Tendo por base o dendrograma obtido antes e depois da eliminação do ruído, elaborou-se a Tabela 6.14, em que é possível observar as alterações que a eliminação provocou.

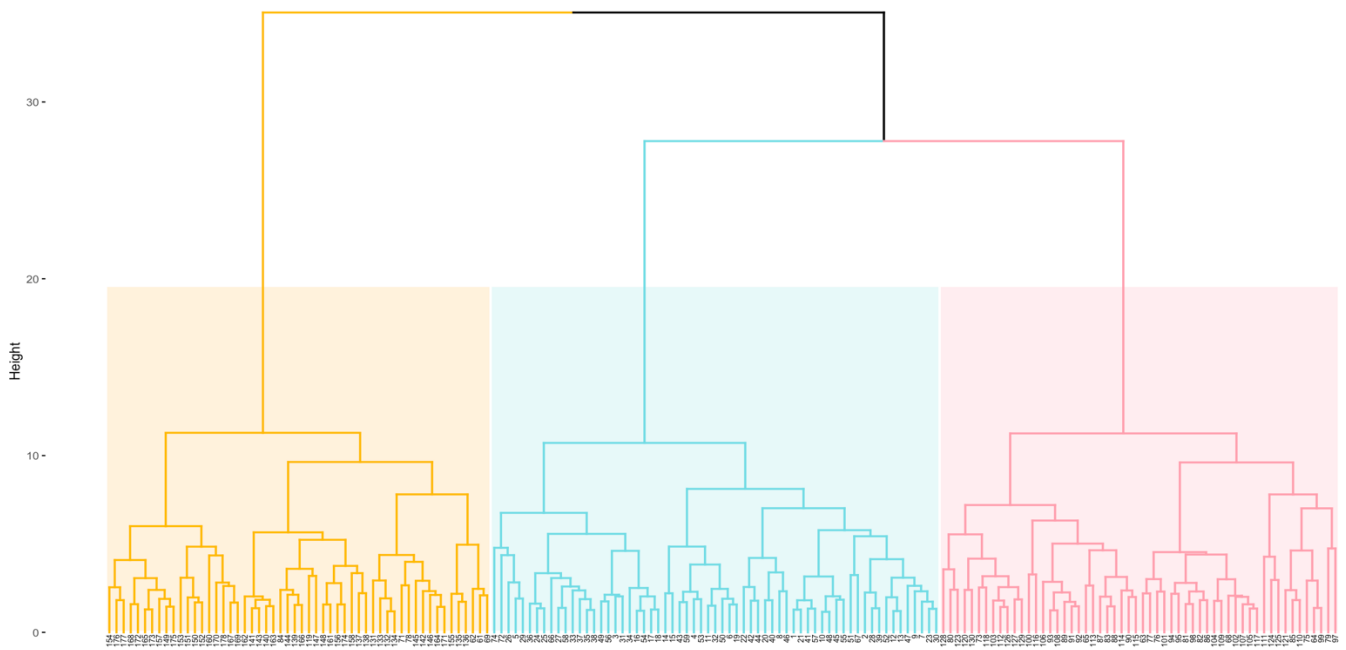


Figura 6.25: Dendrograma obtido para o estudo de todas as variáveis após a eliminação do ruído.

Tabela 6.14: Comparação entre os grupos formados no dendrograma, antes e após a eliminação do ruído

	ANTES DA ELIMINAÇÃO		APÓS ELIMINAÇÃO	
Grupo	**Número de objetos**	**Classe**	**Número de objetos**	**Classe**
Laranja	56	Classe 3	54	Classe 3
Azul	64	Classe 1	63	Classe 1
Rosa	58	Classe 2	56	Classe 2

Com base na Tabela 6.14 é possível ver que as alterações que existiram após a eliminação do ruído são pouco significativas, pois continuam a existir três grupos e, apesar de todos eles terem sofrido uma diminuição do número de objetos, a classe mantém-se a mesma quando comparado o dendrograma antes (Figura 6.11) e depois (Figura 6.25) da eliminação do ruído. O que sugere que o ruído não tem grande influência nos resultados deste estudo. À semelhança do realizado anteriormente no estudo de HCA, procedeu-se à validação dos resultados, mais concretamente dos clusters formados. No entanto, apenas serão apresentados os passos mais importantes, encontrando-se os restantes no Anexo D. Como um dos principais passos de validação é a determinação do número ideal de clusters, na Figura 6.26 encontra-se o gráfico do método do cotovelo e na Figura 6.27 o gráfico de frequência do número de ideal de clusters a formar tendo por base 30 índices, sendo ambos bastante úteis para a determinação do número ideal de clusters.

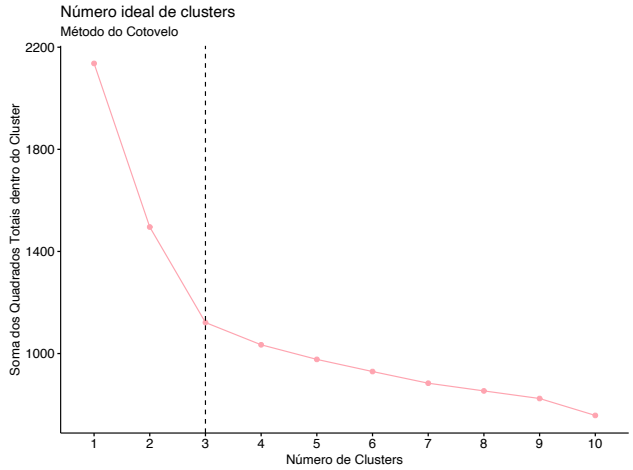


Figura 6.26: Gráfico do método do cotovelo para a determinação do número ideal de clusters no estudo de HCA após a eliminação de *outliers*.

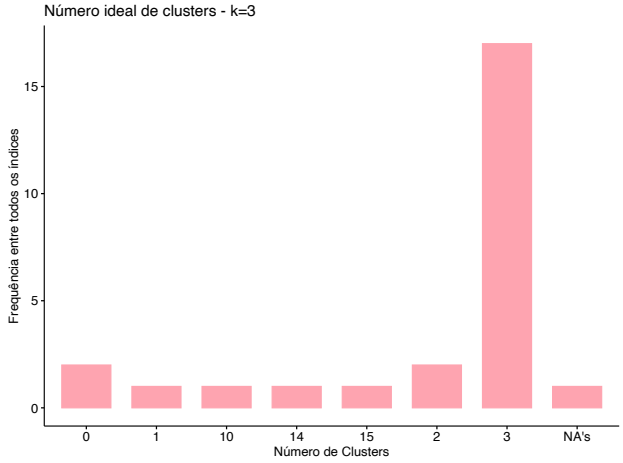


Figura 6.27: Gráfico de frequência do número ideal de clusters a ser formado, tendo em consideração 30 índices.

Da análise das Figura 6.26 e 6.27 verifica-se que para este conjunto de dados o número ideal de clusters a formar é 3, pelo que os resultados obtidos se encontram de acordo com o esperado. Considera-

se então que os resultados são válidos. Concluindo-se assim que o ruído não apresenta grande influência sobre a formação de grupos quando os dados são provenientes de todas as variáveis.

Após o estudo do efeito da eliminação do ruído no estudo de HCA quando utilizadas todas as variáveis, é necessário realizar esse mesmo estudo para quando se consideram as cinco dimensões correspondentes às PC's, obtidas pelo estudo de PCA. Relativamente às cinco dimensões usadas, estas foram determinadas pelo estudo de PCA, o que leva a que as dimensões a usar após a eliminação do ruído sejam distintas das usadas anteriormente. Assim sendo, existe a necessidade de se verificar quantos objetos eliminados quando as dimensões em estudo são: os flavonoides, a intensidade da cor, a cinza, o ácido málico e os fenóis não flavonoides. O estudo da eliminação de *outliers* tendo com base nessas dimensões encontra-se na Tabela 6.15.

Tabela 6.15: Número de objetos considerados ruído quando se recorre ao uso das cinco dimensões correspondentes às PC's (flavonoides, intensidade da cor ,cinza, ácido málico e fenóis não flavonoides), com e sem normalização dos dados.

Objetos	Conjunto de dados COM NORMALIZAÇÃO	Conjunto de dados SEM NORMALIZAÇÃO
Classe 1 (1-59)	12	9
Classe 2 (60-130)	11	22
Classe 3 (131-178)	3	6
Todos	76	77

Através da análise da Tabela 6.15 é possível verificar que, à semelhança do ocorrido para os restantes estudos de eliminação de ruído, a soma do número de objetos eliminados tendo em conta cada classe individualmente, 26, é diferente do número total de objetos,76. Neste caso, é necessário determinar qual o número de objetos eliminados em cada classe tendo em conta todos os objetos, pois facilitará a interpretação dos resultados, assim sendo, na Tabela 6.16 encontra-se o número de objetos eliminados em cada classe, com e sem normalização.

Tabela 6.16: Número de objetos considerados ruído para cada classe, tendo em conta o número de objetos eliminados quando considerados todos os dados e as cinco dimensões correspondentes às PC's.

Objetos	Classe 1	Classe 2	Classe 3
COM NORMALIZAÇÃO	7	22	47
SEM NORMALIZAÇÃO	11	19	47

Com base na Tabela 6.16 é possível verificar que o grupo denominado como classe 3, foi quase todo considerado ruído (apenas resta um objeto), pelo que se verifica que neste caso o ruído apresenta uma influência grande sobre os resultados. Os resultados obtidos sugerem que para estas cinco dimensões o grupo denominado de classe 3 apresente uma baixa densidade de pontos, pelo que são considerados ruído. Em seguida, na Figura 6.28, encontra-se o dendrograma obtido após a eliminação do ruído para dados normalizados, que servirá para comprovar a influência que o ruído apresenta.

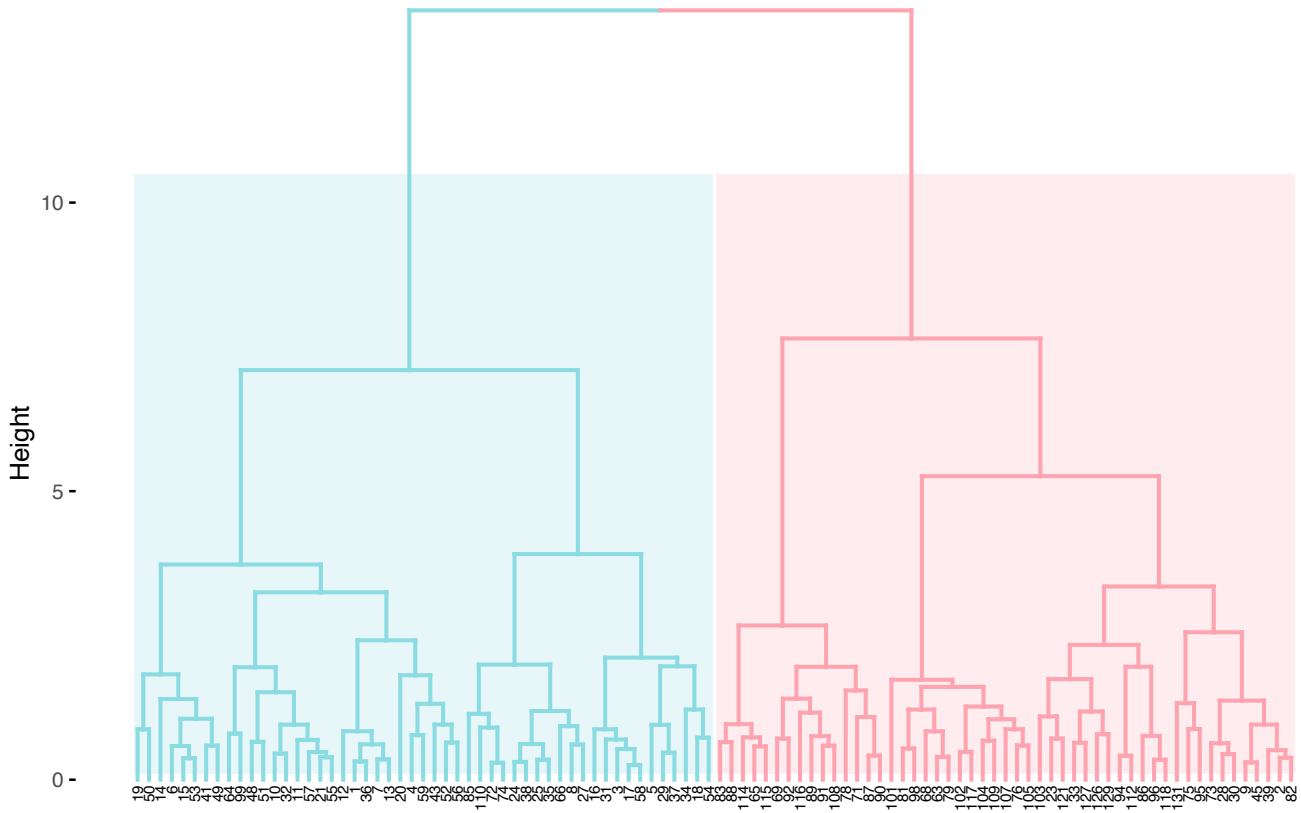


Figura 6.28: Dendrograma obtido para o estudo das cinco dimensões correspondentes às PC's, após a eliminação do ruído.

No dendrograma da figura 6.28, é possível verificar os dados apresentados na tabela 6.16. Como a 3ª classe apenas apresentaria um objeto e este, tendo em conta as variáveis é bastante semelhante a objetos da classe 2, o objeto não eliminado encontra-se no grupo cor de rosa. Uma vez que os resultados são diferentes dos esperados, dado que era esperado obter-se três grupos distintos, cada um pertencente a uma classe da base de dados original, procedeu-se à validação dos resultados.

O processo de validação realizado foi igual ao realizado para o estudo antes da eliminação do ruído, destacando-se por isso o gráfico do número ideal de clusters a serem formados (figura 6.29) e o gráfico de frequência do número ideal de clusters, tendo em conta os 30 índices (figura 6.30). Encontrando-se no anexo D, os restantes gráficos obtidos no processo de validação.

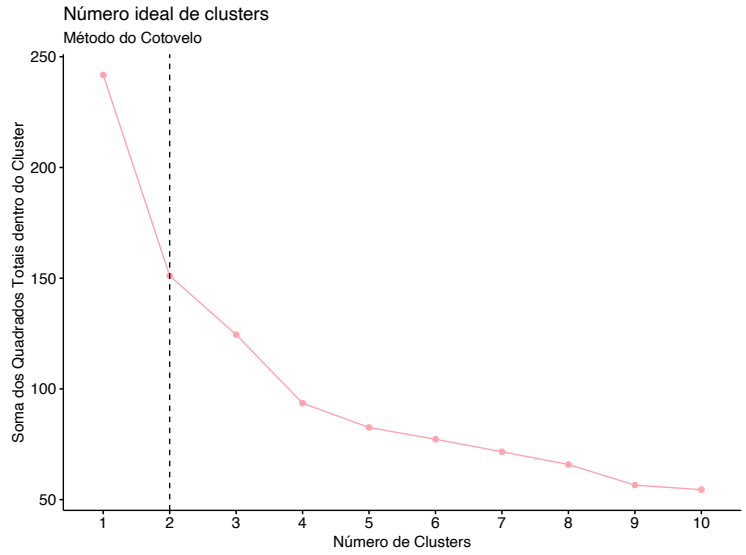


Figura 6.29: Gráfico do método do cotovelo para a determinação do número ideal de clusters no estudo de HCA após a eliminação de outliers.

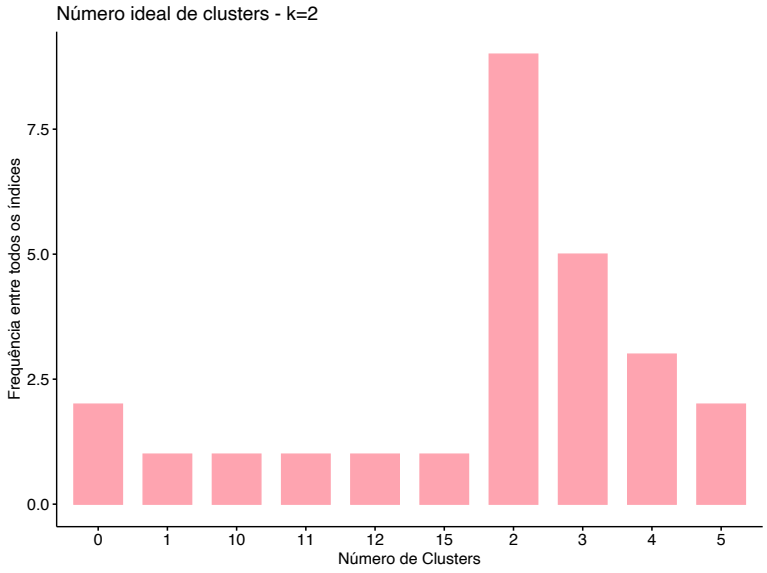


Figura 6.30: Gráfico de frequência do número ideal de clusters a ser formado, tendo em consideração 30 índices.

Com base nas figuras 6.29 e 6.30 é possível verificar que embora não se estivesse à espera de que o ruído eliminasse um grupo completo, os resultados obtidos encontram-se de acordo com o esperado tendo em conta o processo de validação. Este resultado permitiu-nos verificar que a presença de ruído apresenta uma grande influência nos resultados quando o estudo de HCA se realiza tendo em consideração apenas as cinco dimensões resultantes do PCA. Tal pode dever-se ao facto de os objetos do cluster 3 não apresentarem características comuns aos restantes, dado que as dimensões que se estão a estudar não são características deste cluster.

Capítulo 7 - Conclusão

O principal objetivo deste trabalho foi a utilização de ferramentas quimiométricas, como o PCA e o HCA para o estudo e análise de uma base de dados de vinhos. Outro objetivo foi o desenvolvimento e teste de uma rotina para a eliminação de ruído em múltiplas dimensões. A base de dados escolhida é constituída por 178 vinhos provenientes de três zonas distintas de Itália.

Os estudos de PCA mostraram uma tendência dos dados em formar três grupos distintos, o que se encontra de acordo com os três tipos de cultivo das uvas. Atendendo às características químicas dos vinhos é possível diferenciar qual o método de cultivo usado, dado que quando o vinho é proveniente do cultivo denominado de classe 1 vai haver uma maior quantidade de prolina e uma maior intensidade da cor; quando este pertence à classe 3 vamos ter um valor maior da alcalinidade de cinza. A classe 2 vai ser mais difícil de identificar, dado que as quantidades presentes são, na sua maioria intermédias quando comparadas com as restantes. No entanto, se estivéssemos perante um vinho de classe desconhecida seria possível determiná-la.

Os estudos de HCA, tal como os estudos de PCA, mostraram uma tendência para a formação de três grupos distintos, independentemente do estudo ser realizado com as 5 primeiras dimensões do PCA ou com todas elas, pelo que se conclui que o estudo de HCA realizado antes ou após o estudo de PCA dá resultados bastante idênticos. No entanto, com este estudo foi também possível concluir que quanto maior o número de variáveis que se analisa, melhor é a distribuição dos objetos, isto é, menos objetos se encontram em grupos diferentes do esperado.

Quanto à eliminação do ruído os resultados encontram-se de acordo com o esperado, dado que os objetos eliminados se encontram na periferia do conjunto, tal como previsto. No entanto, o ruído apresenta uma maior influência sobre o estudo de HCA com cinco dimensões do que do PCA e HCA com todas as variáveis. No estudo de PCA após a eliminação do ruído, os resultados melhoraram ligeiramente, passou-se de uma recuperação de informação do sistema de 55% para 58%, enquanto no estudo de HCA após a eliminação contendo apenas as cinco dimensões obtidas pelo PCA existiu a remoção de um grupo completo, o que faz com que o ruído tenha uma grande influência neste caso sendo essa negativa. Em relação ao estudo de HCA após a eliminação usando todas as variáveis não existiram alterações significativas pelo que neste caso o ruído não teve grande influência.

Em suma, as ferramentas quimiométricas são úteis para a análise e estudo de bases de dados, no caso deste trabalho de vinhos, mas também de doenças, de espécies, entre muitas outras. Uma das funcionalidades das ferramentas quimiométricas é permitir a classificação das principais características que levam à existência de diferenças entre as classes, permitindo diferenciar os vinhos, pelo que podem ser aplicadas na indústria vinícola. Para além disso, com o recurso a estas ferramentas é ainda possível identificar e eliminar outliers melhorando as mais diversas bases de dados existentes, levando assim a que os resultados sejam mais fidedignos.

Referências

[1] Ministério da Agricultura. (2016). Instituto da Vinha e do Vinho. A Vinha e o Vinho Em Portugal. https://www.ivv.gov.pt/np4/47/ [acesso a 3/12/2020]

[2] Amerine MA, Kunkee R, Ough KCS, Singleton VL and Webb AD (1980). The technology of wine making (4th Ed). Avi Pub Co. ISBN:9780870553332

[3] Amerine, M. A. (2020). Wine. Encyclopaedia Britannica. https://www.britannica.com/topic/wine [acesso a 28/10/2020]

[4] Ministério da Agricultura. (2016). Instituto da Vinha e do Vinho. Património Vitícola. https://www.ivv.gov.pt/np4/33/ [acesso a 28/10/2020]

[5] Estreicher, S. K. (2004). Wine the Past 7.400 Years.

[6] INEGI/Mercatura. (2009). Infovini - O portal do Vinho Português. História. http://www.infovini.com/pagina.php?codNode=18094#tab2 [acesso a 3/12/2020]

[7] Cabral, C. (2006). Porto: Um Vinho sua Imagem- Um passeio pelos rótulos da coleção Soromenho (1o Ed). ISBN: 9788529301068

[8] Wine Tourism in Portugal. (2013). Sobre Portugal. https://www.winetourismportugal.com/pt/sobre-portugal/ [acesso a 10/12/2020]

[9] D'Avillez, V., Andrade, I., Guilherme, R., Páscoa, J., Malheiro, P., & Alves, F. (2006). *A Vinha e o Vinho em Portugal - Viticultura Geral*. (2006) ISBN:978989200442

[10] Turiventos - Turismo e Eventos (2015). *Regiões Vinícolas de Portugal*. Mapa Vinícola de Portugal/Enoturismo. https://turiventos-turismoeventos.blogs.sapo.pt/regioes-vinicolas-de-portugal-20156 [acesso a 10/12/2020]

[11] INEGI/Mercatura. (2009). Infovini - O portal do Vinho Português. Vinho Tranquilo. http://www.infovini.com/pagina.php?codNode=18100 [acesso a 10/12/2020]

[12] INEGI/Mercatura. (2009). Infovini - O portal do Vinho Português. Cultura Da Vinha. http://www.infovini.com/pagina.php?codNode=18099 [acesso a 10/12/2020]

[13] Dias, J. P. (2006). Fases da Maturação da Uva. Centésimo Curso Intensivo de Vinificação, 1–8.

[14] Fischer D. Craig, W. A. B. H. (1990). Reducing transportation damage to grapes and strawberries. Journal of Food Distribution Research, 21, 193–202.

[15] Chitarra, M. I. F. (2005). Pós-colheita de frutas e hortaliças: fisiologia e manuseio (U. F. de Lavras (Ed.); 2o Ed).

[16] Thompson, James; Viegneault, C. et al. (2009). Transportation of fresh horticultural produce. Research Signpost, 2, 1–24.

[17] LMH Wines. (2020). Vinificação e o Processo de Produção Do Vinho. https://lmh-wines.pt/vinificacao-e-o-processo-de-producao-do-vinho/ [Acesso a 10/12/2020]

[18] João Portugal Ramos. (n.d.) João Portugal Ramos. Processo de Vinificação. https://www.jportugalramos.com/pt/vinificacao/ [acesso a 3/12/2020]

[19] Ciani, M., & Comitini, F. (2019). Chapter 4 - Use of Non-Saccharomyces Yeasts in Red Winemaking. In Red Wine Technology (pp. 51–68). https://doi.org/10.1016/B978-0-12-814399-5.00004-9

[20] Meléndez, M. E., Sánchez, M. S., Íñiguez, M., Sarabia, L. A., & Ortiz, M. C. (2001). Psychophysical parameters of colour and the chemometric characterisation of wines of the certified denomination of origin "Rioja." Analytica Chimica Acta, 446(1–2), 157–167. https://doi.org/10.1016/s0003-2670(01)01274-0

[21] Herderich, M. J., & Smith, P. A. (2005). Analysis of grape and wine tannins: Methods, applications and challenges. Australian Journal of Grape and Wine Research, 11(2), 205–214. https://doi.org/https://doi.org/10.1111/j.1755-0238.2005.tb00288.x

[22] Statusknowledge, L. (2017). Clube Vinhos Portugueses. Como Se Produz Vinho - Prensagem e Defecação/Sedimentação. https://www.clubevinhosportugueses.pt/vinhos/como-se-produz-vinho-prensagem-e-defecacaosedimentacao-3/ [acesso a 3/12/2020]

[23] Chambers, P. J., & Pretorius, I. S. (2010). Fermenting knowledge: the histry of Winemaking, Science and Yeast Research. Embo Reports, 11(12), 914–920.

[24] Rizzon, Luiz; Zanuz, Mauro; Miele, A. (1997). Efeito da Fermentação Maloláctica na composição do vinho tinto. Ciência Rural, 27(3), 497–500.

[25] Winepedia. (2016). Fermentação Alcoólica e Malolática. https://www.wine.com.br/winepedia/sommelier-wine/fermentacao-alcoolica-e-malolatica/ [acesso a 18/12/2020]

[26] Pato, O. (1998). O Vinho - Sua Preparação e Conservação (C. Editora (Ed.); 04ed.).

[27] Grizzo, A. (2018). Veja como funciona a fermentação malolática. Revista ADEGA, 150.

[28] Fischer, U., & Noble, A. C. (1994). The Effect of Ethanol, Catechin Concentration, and pH on Sourness and Bitterness of Wine. American Society for Enology and Viticulture, 45(1), 6–10.

[29] Jordão, A., Vilela, A., & Cosme, F. (2015). From Sugar of Grape to Alcohol of Wine: Sensorial Impact of Alcohol in Wine. Beverages, 1(4), 292–310. https://doi.org/10.3390/beverages1040292

[30] Rizzon, L. A., & Gatto, N. M. (1987). Características Analíticas dos Vinhos da Microrregião homogênea vinicultora de Caxias do Sul (MRH 311) - Análises Clássicas. 6, 1–5.

[31] Dikanović-Lučan, Željka ; Palić, Angelina; Hanser, D. (1993). Determination of Ash Content in Wines by the Conductometric Method. Prehrambeno-Technol. Biotechnol, 31, 15–18.

[32] Cecchi, H. (2003). Fundamentos teóricos e práticos em análise de alimentos. Campinas- SP – Brasil: SciELO - Editora da Unicamp. doi:10.7476/9788526814721

[33] Ministério da Agricultura. (2016). Instituto da Vinha e do Vinho. Parâmetros Analíticos - Portugal. [acesso a 18/12/2020]

[34] Rizzon, Luiz Antenor; Miele, Alberto; Scopel, G. (2009). Características analíticas de vinhos Chardonnay da Serra Gaúcha. Ciência Rural, 39(8), 2555–2558.

[35] Bai, X., Wang, L., & Li, H. (2019). Identification of red wine categories based on physicochemical properties. 5th International Conference on Education Technology, Management and Humanities Science, 1443–1448. https://doi.org/10.25236/etmhs.2019.309

[36] Zoecklein, B. W., Fugelsang, K. C., Gump, B. H., & Nury, F. S. (1990). Determination of total phenolics. In Phenolic Compouns and Wine Color (pp. 129–168). Springer,Boston,MA. https://doi.org/978-1-4615-8146-8-7

[37] Mamede, M. & Pastore, G. M. (2004). Compostos Fenólicos Do Vinho: Estrutura E Ação Antioxidante. Boletim Do Centro de Pesquisa de Processamento de Alimentos, 22(2), 233–252. https://doi.org/10.5380/cep.v22i2.1192

[38] Kova, V; Bourzeix, M; Heredia, N; Ramos, T. (1995). Études des cátechines et proantocyanidols de raisins et vins blancs. Rev. Fran. Oen, 7–15.

[39] Frankel, E. N., Waterhouse, A. L., & Teissedre, P. L. (1995). Principal Phenolic Phytochemicals in Selected California Wines and Their Antioxidant Activity in Inhibiting Oxidation of Human Low-Density Lipoproteins. Journal of Agricultural and Food Chemistry, 43(4), 890–894. https://doi.org/10.1021/jf00052a008

[40] Hrazdina, G., Borzell, A. J., & Robinson, W. B. (1970). Studies on the Stability of the Anthocyanidin-3,5-Digucosides. American Journal of Enology and Viticulture, 21(4), 201 LP – 204.

[41] Garrido, J., & Borges, F. (2013). Wine and grape polyphenols - A chemical perspective. Food Research International, 54(2), 1844–1858. https://doi.org/10.1016/j.foodres.2013.08.002

[42] Fernandes, I., Pérez-Gregorio, R., Soares, S., Mateus, N., De Freitas, V., Santos-Buelga, C., & Feliciano, A. S. (2017). Wine flavonoids in health and disease prevention. Molecules, 22(292), 1–30. https://doi.org/10.3390/molecules22020292

[43] Larice, J. L. et al. (1989). Composition anthocyanique des cépages. II-Essai de classification sur trois ans par analyse en composantes principales et étude des variations annuelles de cépages de même provenance.

[44] Cheynier, V. (2006). Flavonoids in wine. In Flavonoids: chemistry, biochemistry and applications (pp. 263–318). CRC Press LLC. https://doi.org/10.1201/9781420039443.ch5

[45] Cheynier, V., Rigaud, J., Souquet, J.-M., Barillere, J., & Moutounet, M. (1989). Effect of pomace contact and hyperoxidation on the phenolic composition and quality of Grenache and Chardonnay wines. American Journal of Enology and Viticulture, 40, 36–42.

[46] Somers, T., & Pocock, K. (1991). Phenolic assessment of white musts : varietal differences in free-run juices and pressings. Vitis: Journal of Grapevine Research, 30(3), 189–201.

[47] Macheix, J., Sapis, J., Fleuriet, A., & Lee, C. Y. (1991). Phenolic compounds and polyphenoloxidase in relation to browning in grapes and wines. Critical Reviews in Food Science and Nutrition, 30(4), 441–486. https://doi.org/10.1080/10408399109527552

[48] Moreno-Arribas, M. V., & Polo, M. C. (2009). Non-flavonoid Phenolic Compounds. In Wine Chemistry and Biochemistry (pp. 509–527). https://doi.org/10.1007/978-0-387-74118-5

[49] Somers, T., & Evans, M. (1986). Evolution of red wines I. Ambient influences on colour composition during early maturation. Vitis, 25(1), 31–39.

[50] Cheynier, V., Prieur, C., Guyot, S., Rigaud, J., & Moutounet, M. (1997). The Structures of Tannins in Grapes and Wines and Their Interactions with Proteins. In Wine (Vol. 661, pp. 8–81). American Chemical Society. https://doi.org/doi:10.1021/bk-1997-0661.ch008

[51] Moreno-Arribas, M. V., & Polo, M. C. (2009). Non-flavonoid Phenolic Compounds. In Wine Chemistry and Biochemistry (pp. 509–527). https://doi.org/10.1007/978-0-387-74118-5

[52] Timberlake, C., & Bridle, P. (1976). The effect of processing and other factors on the colour characteristics of some red wines. Vitis, 15, 37–49.

[53] Cabrita, M. J., Ricardo-da-Silva, J., & Laureano, O. (1999). Os Compostos Polifenólicos Das Uvas E Dos Vinhos. I Seminário Internacional de Vitivinicultura, 61–102.

[54] Daidone, P. (n.d.). Reserva85. O Que a Cor Dos Vinhos e a Sua Intensidade Revelam? https://reserva85.com.br/analise-sensorial-como-provar-um-vinho/o-que-a-cor-dos-vinhos-e-sua-intensidade-revelam-analise-visual/ [acesso a 10/01/2021]

[55] Bakker, J., & Timberlake, C. F. (1997). Isolation, Identification, and Characterization of New Color-Stable Anthocyanins Occurring in Some Red Wines. Journal of Agricultural and Food Chemistry, 45(1), 35–43. https://doi.org/10.1021/jf960252c

[56] Brouillard, R., & Dangles, O. (1994). Anthocyanin molecular interactions: the first step in the formation of new pigments during wine aging? Food Chemistry, 51(4), 365–371. https://doi.org/https://doi.org/10.1016/0308-8146(94)90187-2

[57] Rizzon, L. A., Salvador, M. B. G., & Miele, A. (2008). Teores de cátions dos vinhos da Serra Gaúcha. Ciência e Tecnologia de Alimentos, 28(3), 635–641. https://doi.org/10.1590/s0101-20612008000300020

[58] Bonin, S. (2014). Effects of magnesium ions on both VHG batch and continuous fruit wine fermentations. Journal of the Institute of Brewing,120(4),477–485. https://doi.org/10.1002/jib.170

[59] Togores, J. H. (2010). Tratado de Enologia. Tomo I (Mundi-Prensa (Ed.); 2 Ed).

[60] Bravo, L. (1998). Polyphenols: Chemistry, dietary sources, metabolism, and nutritional significance. Nutrition Reviews, 56(11), 317–333. https://doi.org/10.1111/j.1753-4887.1998.tb01670.x

[61] Cheynier, V., Dueñas-Paton, M., Salas, E., Maury, C., Souquet, J. M., Sarni-Manchado, P., & Fulcrand, H. (2006). Structure and properties of wine pigments and tannins. American Journal of Enology and Viticulture, 57(3), 298–305.

[62] Jerez, M., Touriño, S., Sineiro, J., Torres, J. L., & Núñez, M. J. (2007). Procyanidins from pine bark: Relationships between structure, composition and antiradical activity. Food Chemistry, 104(2), 518–527. https://doi.org/10.1016/j.foodchem.2006.11.071

[63] Silva, Jorge; Rigaud, Jacques; Cheynier, Véronique; Cheminat, Annie; Moutounet, M. (1991). Procyanidin Dimers and Trimers from Grapes Seeds. Phytochemistry, 30(4), 1259–1264.

[64] Rencher, Alvin C.; Chirstensen, W. F. (2012). Methods of Multivariate Analysis. In G. et al. Balding, David; Cressie, Noel; Fitzmaurice (Ed.), *John Wiley & Sons, Inc.* (3rd ed.). https://doi.org/10.1002/978111839

[65] Drab, K., & Daszykowski, M. (2014). Clustering in analytical chemistry. Journal of AOAC International, 97(1), 29–38. https://doi.org/10.5740/jaoacint.SGEDrab

[66] Jolliffe, I.T. (2002). Principal Component Analysis, 2nd Ed. *Springer*, 487.

[67] Davies, A.M.C; Fearn, T. (2004). Back to basics: the principles of principal component analysis. *Spectroscopy Europe*, *16*(6), 20–23.

[68] Jolliffe, I. T. (2002). Choosing a Subset of Principal Components or Variables. In *Principal Component Analysis* (2nd ed., pp. 111–149). Springer.

[69] Morais, J. E., Silva, A. J., Marinho, D. A., Seifert, L., & Barbosa, T. M. (2015). Cluster stability as a new method to assess changes in performance and its determinant factors over a season in young swimmers. *International Journal of Sports Physiology and Performance*, *10*(2), 261–268. https://doi.org/10.1123/ijspp.2013-0533

[70] Araujo, W., & Coelho, C. (2009). Análise de Componentes Principais (PCA)

[71] Kassambara, A. (2017). Practical Guide to Princpal Component Methods in R. In *Multivariate Analysis II* (1st ed.). STHDA.

[72] Kassambara, A. (2017). Practical Guide to Cluster Analysis in R. In Multivariate Analysis I (1st ed.). STHDA.

[73] Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Jounal of the American Statistical Association*, *58*(301), 236–244.

[74] Suzuki, R., & Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, *22*(12), 1540–1542. https://doi.org/10.1093/bioinformatics/btl117

[75]Almeida, J. A. S., Barbosa, L. M. S., Pais, A. A. C. C., & Formosinho, S. J. (2007). Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, *87*(2), 208–217. https://doi.org/10.1016/j.chemolab.2007.01.005

[76] Ferreira, M. (2015). *Quimiometria: Conceitos, Métodos e Aplicações* (Editora Unicamp (ed.)). https://doi.org/10.7476/9788526814714

[77] DataCamp. (n.d.). *dist: Distance Matrix Computation (including Aitchison distance).* https://www.rdocumentation.org/packages/coda.base/versions/0.3.1/topics/dist [acesso a 20/1/2021]

[78] Foley, M. (2019). Unsupervised Learning with HCA.

[79] Lance, G. N., & Williams, W. T. (1966). Computer Programs for Hierarchical Polythetic Classification ("Similarity Analyses"). *The Computer Journal*, *9*(1), 60–64. https://doi.org/10.1093/comjnl/9.1.60

[80] DataCamp. (n.d.). *dist: Distance Matrix Computation*. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/dist [acesso a 12/12/2020]

[81] Frédéric Le Bel. Agglomerative Clustering for Audio Classification using Low-level Descriptors. [Re- search Report] Ircam UMR STMS 9912. 2017. hal-01491270

[82] Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*(1), 81–93.

[83]Piek, A. B., & Petrov, E. (2021). On a Weighted Generalization of Kendall's Tau Distance. *Annals of Combinatorics*, *25*(1), 33–50. https://doi.org/10.1007/s00026-020-00519-y

[84] Puthier, D. (2017). *Distance Metrics and Clustering*. http://pedagogix-tagc.univ-mrs.fr/courses/ASG1/practicals/distances/distances.html [acesso a 12/12/2020]

[85] van Dongen, S., & Enright, A. J. (2012). Metric distances derived from cosine similarity and Pearson and Spearman correlations. http://arxiv.org/abs/1208.3145

[86] DataCamp. (n.d.). *hclust: Hierarchical Clustering*. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust [acesso a 12/12/2020]

[87] Li, Z., & De Rijke, M. (2017). The impact of linkage methods in hierarchical clustering for active learning to rank. SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 941–944. https://doi.org/10.1145/3077136.3080684

[88]Jarman, A. M. (2020). Hierarchical Cluster Analysis : Comparison of Single linkage , Complete linkage , Average linkage and Centroid Linkage Method. 2. https://doi.org/10.13140/RG.2.2.11388.90240

[89] Hintze, J. L. (2007). *User ' s Guide - IV* (pp. 445:1-12). https://www.ncss.com/download/ncss/manuals/

[90] Chen Y, Mao D, Zhang Y and Ouyang Z. Unsupervised gene selection for predicting cell spatial positions in the Drosophila embryo [version 2; peer review: 1 approved, 1 not approved]. *F1000Research* 2021, **9**:124 (https://doi.org/10.12688/f1000research.20446.2)

[91] *R Project*. (n.d.). https://www.r-project.org [acesso a 20/1/2021]

[92] Available CRAN Packages by Name. (n.d.). https://cran.r-project.org/web/packages/available_packages_by_name.html [acesso a 20/1/2021]

[93] DataCamp. (n.d.). PCA: Principal Component Analysis. https://www.rdocumentation.org/packages/FactoMineR/versions/2.4/topics/PCA

[94] DataCamp. (n.d.). Eigenvalue: Extract and Visualize the eigenvalues/ variances of dimensions. https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/eigenvalue[acesso a 20/1/2021]

[95] DataCamp. (n.d.). get_pca: Extract the results for individuals/variables - PCA. https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/get_pca [acesso a 20/1/2021]

[96] DataCamp. (n.d.). fviz_pca: Visualize Principal Component Analysis. https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_pca [acesso a 20/1/2021]

[97] DataCamp. (n.d.). *pvclust: Calculating P-values for Hierchical Clustering*. https://www.rdocumentation.org/packages/pvclust/versions/2.2-0/topics/pvclust[acesso a 20/1/2021]

[98] DataCamp. (n.d.). *pvrect2: Draw Rectangles Around a Dendrogram's Clusters with High/Low P-values*. https://www.rdocumentation.org/packages/dendextend/versions/1.14.0/topics/pvrect2[acesso a 20/1/2021]

[99] DataCamp. (n.d.). *seplot: Diagnostic Plot for Standard Error of p-value*. https://www.rdocumentation.org/packages/pvclust/versions/2.2-0/topics/seplot[acesso a 20/1/2021]

[100] DataCamp. (n.d.). aggregate: Compute Summary Statistics of Data Subsets. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aggregate [acesso a 20/1/2021]

[101] DataCamp. (n.d.). barplot: Bar Plots. https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/barplot [acesso a 20/1/2021]

[102] DataCamp. (n.d.). *fviz_dend: Enhanced Visualization of Dendrogram*. https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_dend [acesso a 20/1/2021]

[103]DataCamp. (n.d.). *fviz_nbclust: Determining and Visualizing the Optimal Number of Clusters*. https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_nbclust[acesso a 20/1/2021]

[104] Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36. https://doi.org/10.18637/jss.v061.i06

[105] DataCamp. (n.d.). *NbClust: NbClust Package for determining the best number of clusters*. https://www.rdocumentation.org/packages/NbClust/versions/3.0/topics/NbClust[acesso a 20/1/2021]

[106] DataCamp. (n.d.). *eclust: Visual enhancement of clustering analysis*. https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/eclust[acesso a 20/1/2021]

[107] DataCamp. (n.d.). *fviz_silhouette: Visualize Silhouette Information from Clustering*. https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_silhouette[acesso a 20/1/2021]

[108] DataCamp. (n.d.). *sil: Silhouette index and plot*. https://www.rdocumentation.org/packages/kmed/versions/0.4.0/topics/sil[acesso a 20/1/2021]

[109] DataCamp. (n.d.). *cluster.stats: Cluster Validation Statistics*. https://www.rdocumentation.org/packages/fpc/versions/2.2-9/topics/cluster.stats [acesso a 20/1/2021]

[110]DataCamp. (n.d.). *clValid: Validate Cluster Results*. https://www.rdocumentation.org/packages/clValid/versions/0.7/topics/clValid[acesso a 20/1/2021]

[111] DataCamp. (n.d.). *ppx: Multidimensional Space-Time Point Pattern*. https://www.rdocumentation.org/packages/spatstat.geom/versions/2.2-0/topics/ppx [acesso a 21/1/2021]

[112] DataCamp. (n.d.). *rdist: An R package for distances*. https://www.rdocumentation.org/packages/rdist/versions/0.0.5/topics/rdist [acesso 21/1/2021]

[113] DataCamp. (n.d.). *nndist.ppx: Nearest Neighbour Distances in Any Dimensions*. https://www.rdocumentation.org/packages/spatstat.geom/versions/2.2-0/topics/nndist.ppx [acesso a 21/1/2021]

[114] DataCamp. (n.d.). *cbind: Combine R Objects by Rows or Columns*. https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/cbind [acesso a 21/1/2021]

[115] DataCamp. (n.d.). *mean: Arithmetic Mean*. https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/mean [acesso a 21/1/2021]

[116] Singleton, V. L. (1987). Oxygen with Phenols and Related Reactions in Musts, Wines, and Model Systems: Observations and Practical Implications. American Journal of Enology and Viticulture, 38(1), 69 LP – 77. http://www.ajevonline.org/content/38/1/69.abstract

[117] Volschenk, H., van Vuuren, H. J. J., & Viljoen-Bloom, M. (2017). Malic Acid in Wine: Origin, Function and Metabolism during Vinification. *South African Journal of Enology & Viticulture*, *27*(2). https://doi.org/10.21548/27-2-1613

[118] Jaarsveld, F. P., Hattingh, S., & Minnaar, P. (2009). Rapid induction of ageing character in brandy products - Part III. Influence of toasting. *South African Journal of Enology and Viticulture*, *30*(1), 24–37. https://doi.org/10.21548/30-1-1421

[119] Allen M (1994) Advanced Oenology. Charles Sturt University

Anexos

Anexo A - PCA

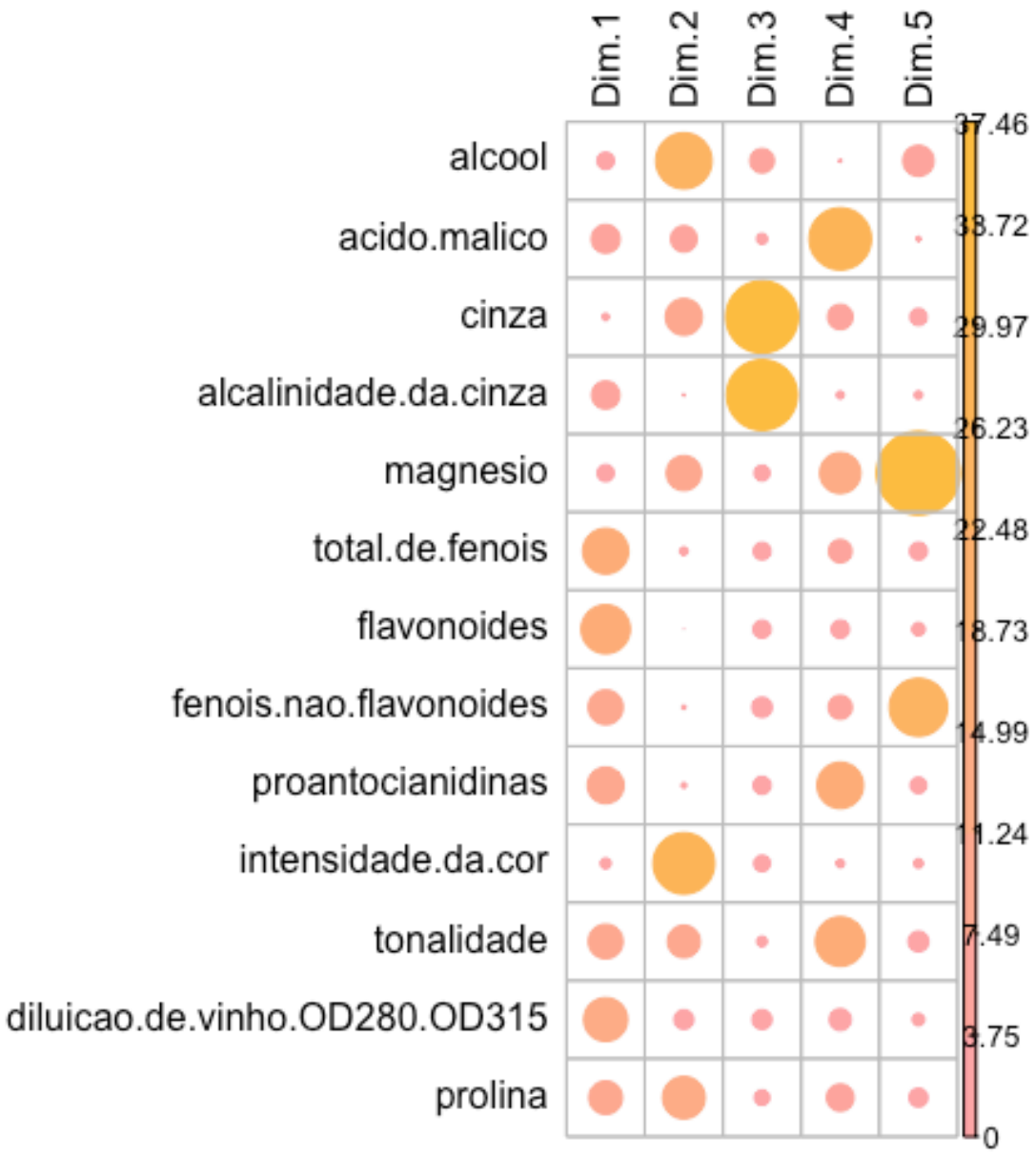


Figura A1: Mapa de fatores da contribuição das variáveis para as cinco componentes principais.

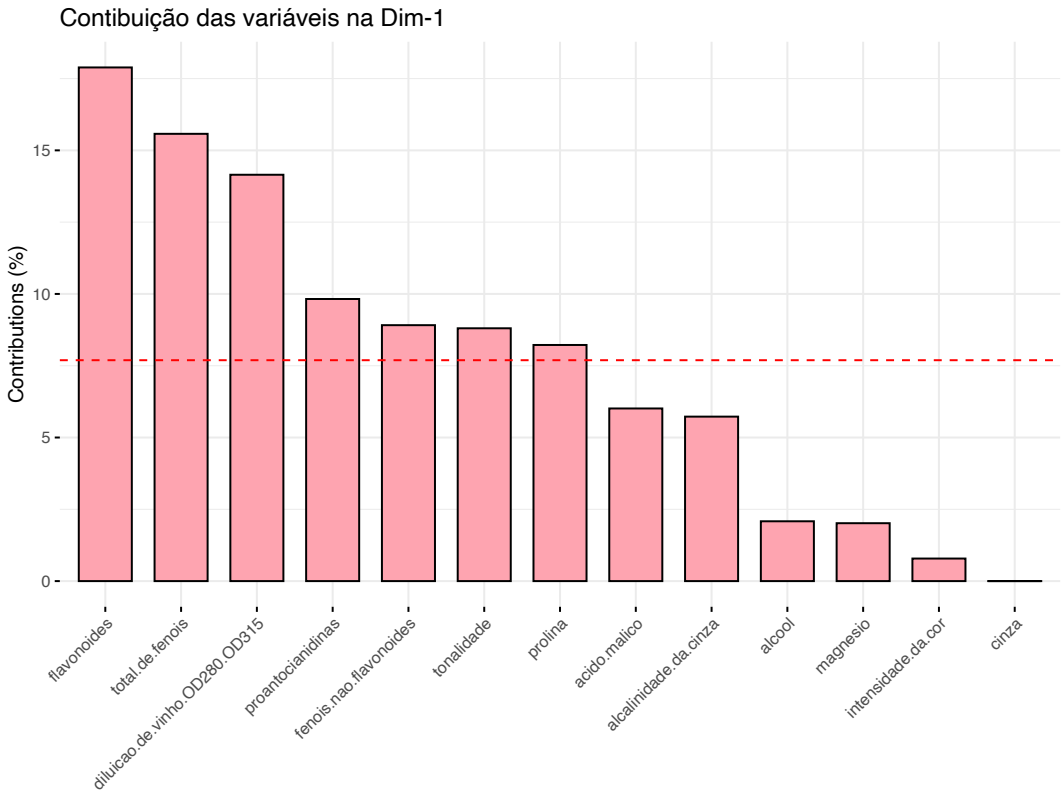


Figura A2: Gráfico das contribuições das variáveis para a primeira dimensão.

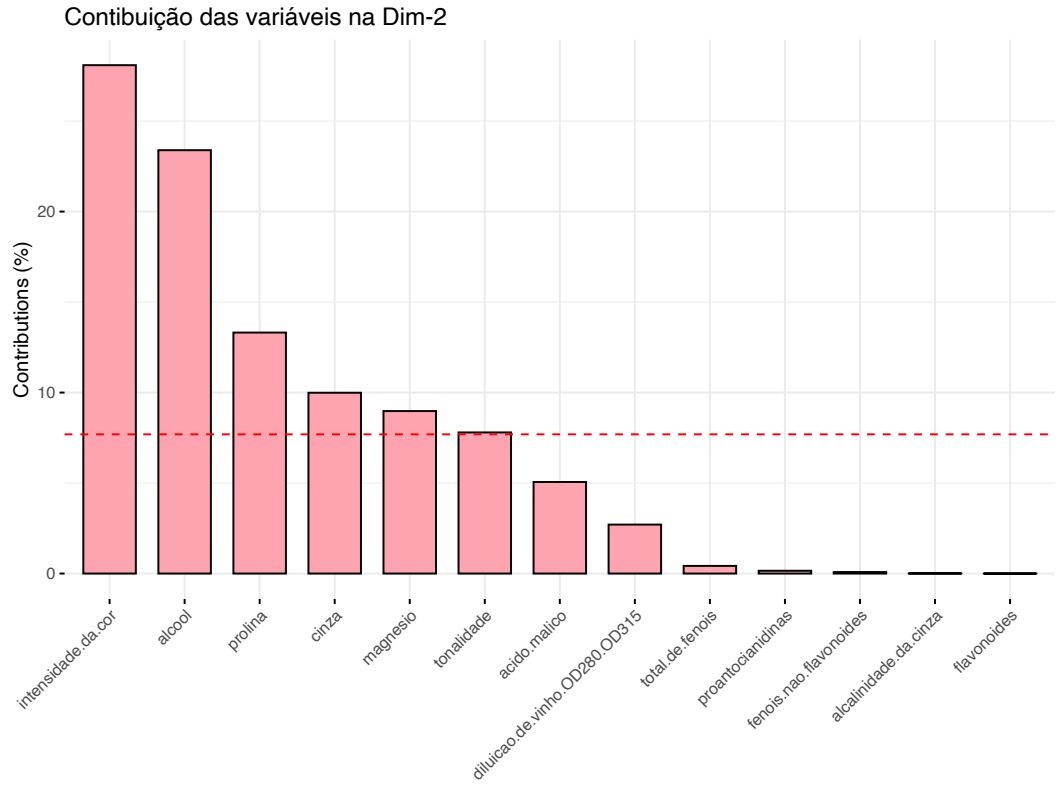


Figura A3: Gráfico das contribuições das variáveis para a segunda dimensão.

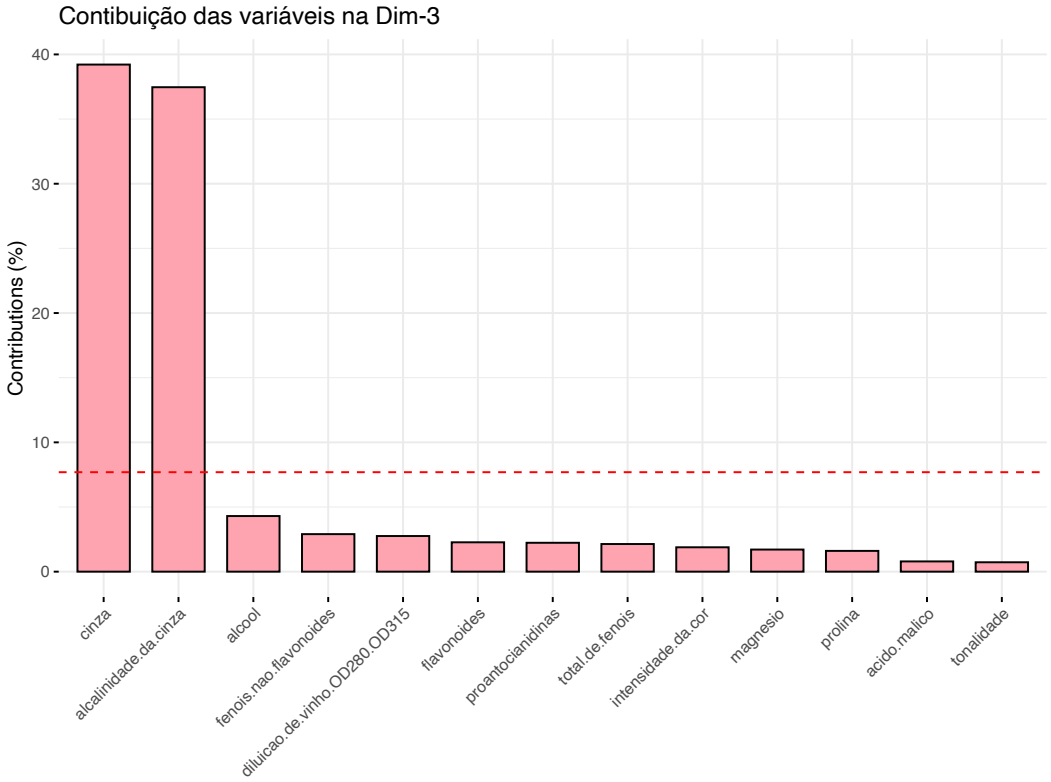


Figura A4: Gráfico das contribuições das variáveis para a terceira dimensão.

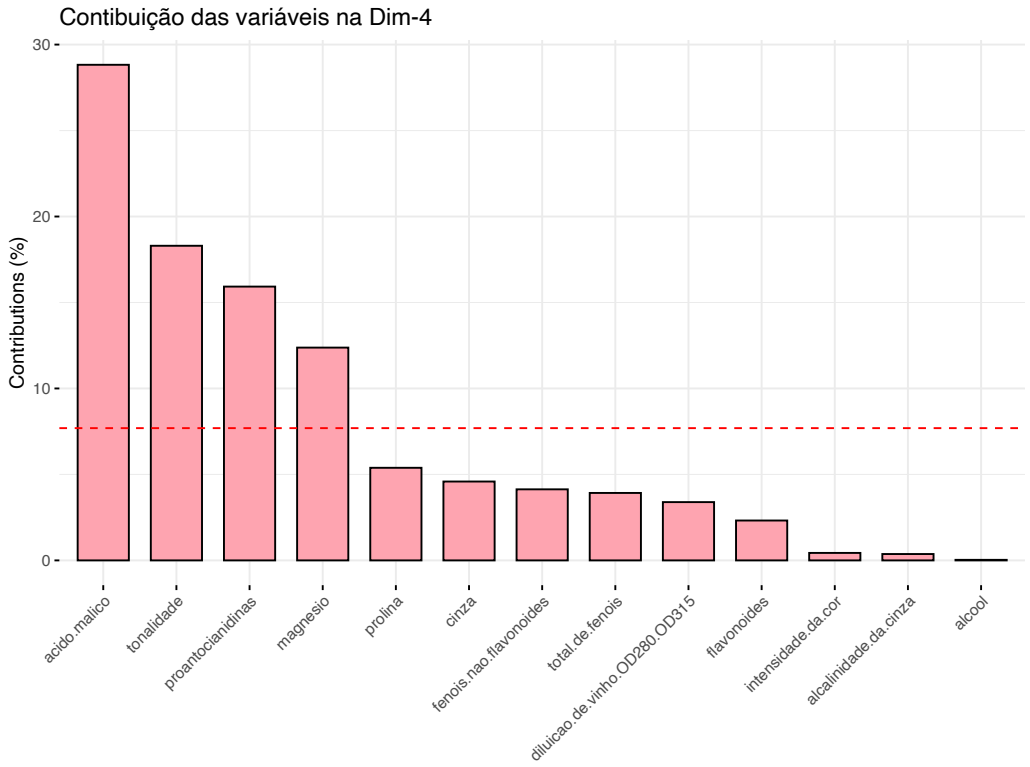


Figura A5: Gráfico das contribuições das variáveis para a quarta dimensão.

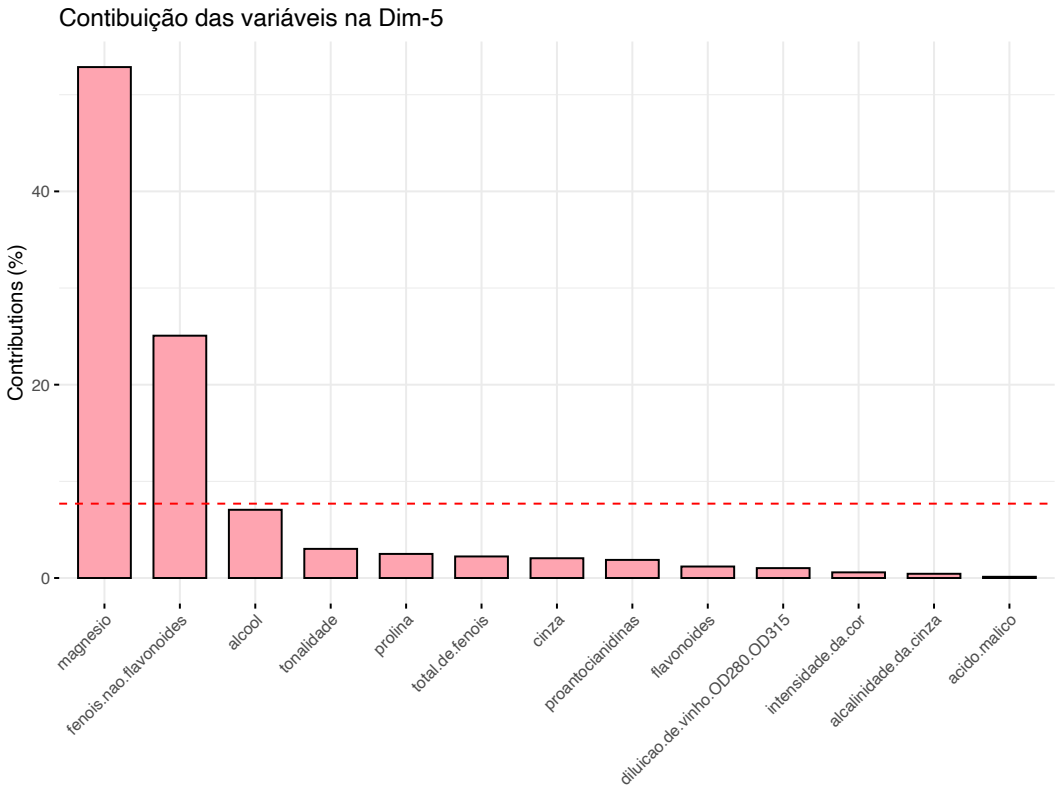


Figura A6: Gráfico das contribuições das variáveis para a quinta dimensão.

Tabela A1: Matriz de correlação entre as variáveis.

	Cinza	Alcalinidade da cinza	Ácido Málico	Fenóis Não Flavonóides	Tonalidade	Proantocianidinas	Diluição de vinhos OD280/OD315	Total de fenóis	Flavonóides	Magnésio	Intensidade da Cor	Álcool	Prolina
Cinza	1	0.44	0.16	0.19	-0.07	0.01	0	0.13	0.12	0.29	0.26	0.21	0.22
Alcalinidade da cinza	0.44	1	0.29	0.36	-0.27	-0.20	-0.28	-0.32	-0.35	-0.08	0.02	-0.31	-0.44
Ácido Málico	0.16	0.29	1	0.29	-0.56	-0.22	-0.37	-0.34	-0.41	-0.05	0.25	0.09	-0.19
Fenóis Não Flavonóides	0.19	0.36	0.29	1	-0.26	-0.37	-0.50	-0.45	-0.54	-0.26	0.14	-0.16	-0.31
Tonalidade	-0.07	-0.27	-0.56	-0.26	1	0.30	0.57	0.43	0.54	0.06	-0.52	-0.07	0.24
Proantocianidinas	0.01	-0.20	-0.22	-0.37	0.30	1	0.52	0.61	0.65	0.24	-0.03	0.14	0.33
Diluição de vinhos OD280/OD315	0	-0.28	-0.37	-0.50	0.57	0.52	1	0.70	0.79	0.07	-0.43	0.33	0.31
Total de fenóis	0.13	-0.32	-0.34	-0.45	0.43	0.61	0.70	1	0.86	0.21	-0.06	0.29	0.50
Flavonóides	0.12	-0.35	-0.41	-0.54	0.54	0.65	0.79	0.86	1	0.20	-0.17	0.24	0.49
Magnésio	0.29	-0.08	-0.05	-0.26	0.06	0.24	0.07	0.21	0.20	1	0.20	0.27	0.39
Intensidade da Cor	0.26	0.02	0.25	0.14	-0.52	-0.03	-0.43	-0.06	-0.17	0.20	1	0.55	0.32
Álcool	0.21	-0.31	0.09	-0.16	-0.07	0.14	0.07	0.29	0.24	0.27	0.55	1	0.64
Prolina	0.22	-0.44	-0.19	-0.31	0.24	0.33	0.31	0.50	0.49	0.39	0.32	0.64	1

Anexo B - HCA

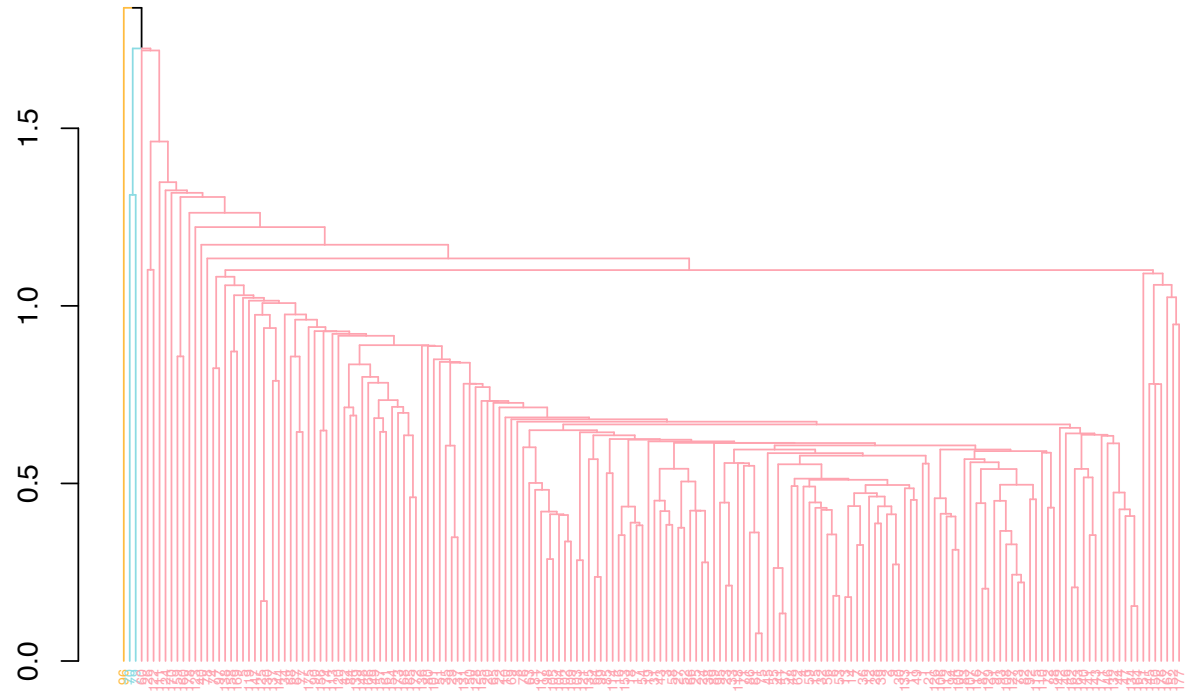


Figura B1: Dendrograma baseado nas cinco componentes principais do PCA, com distância euclidiana e o método de ligação "single".

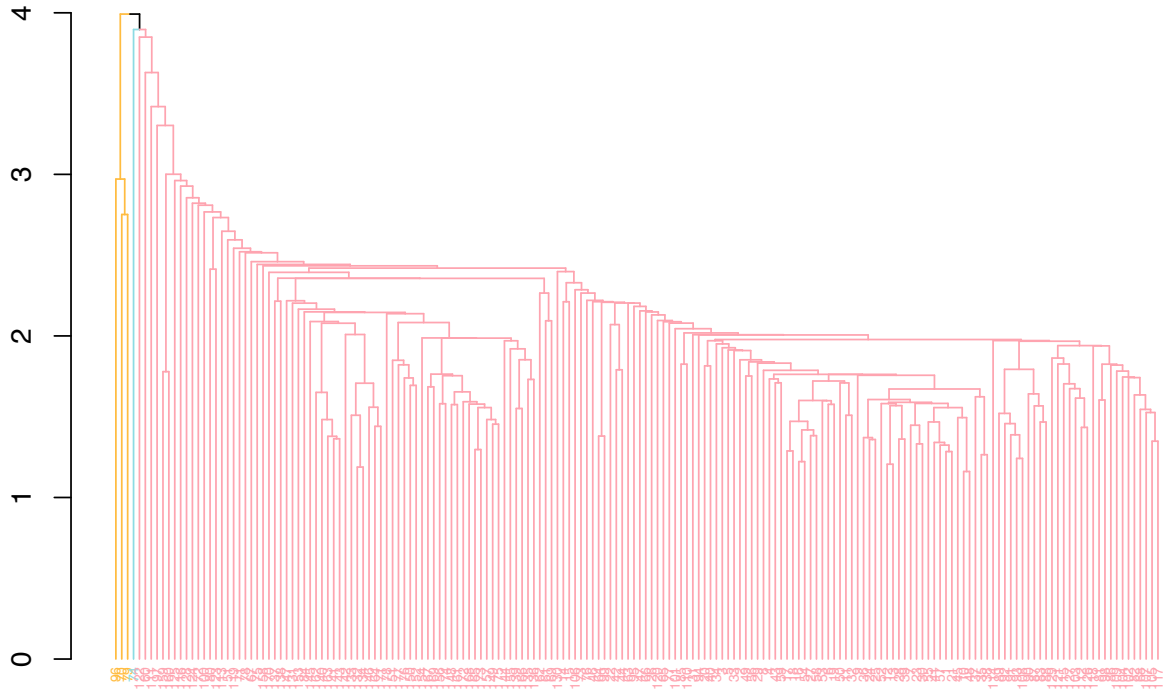


Figura B2: Dendrograma de todos os objetos sem aplicar o PCA, com distância euclidiana e o método de ligação "single".

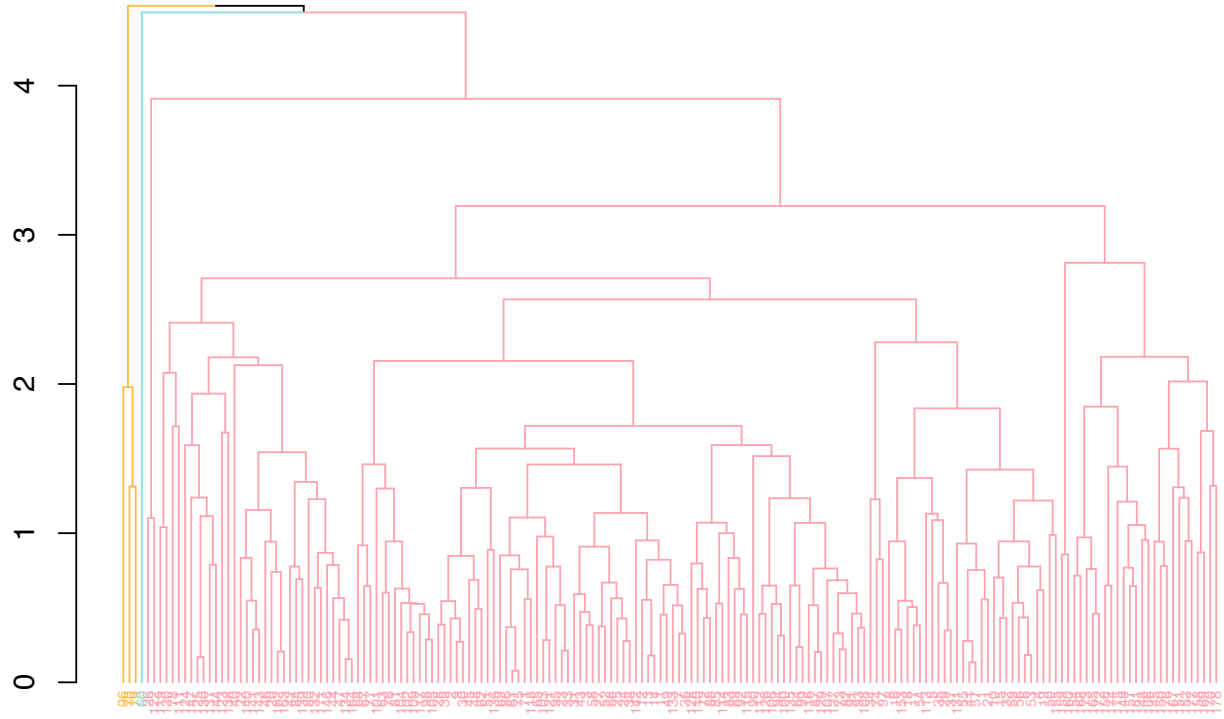


Figura B3: Dendrograma baseado nas cinco componentes principais do PCA, com distância euclidiana e o método de ligação "average".

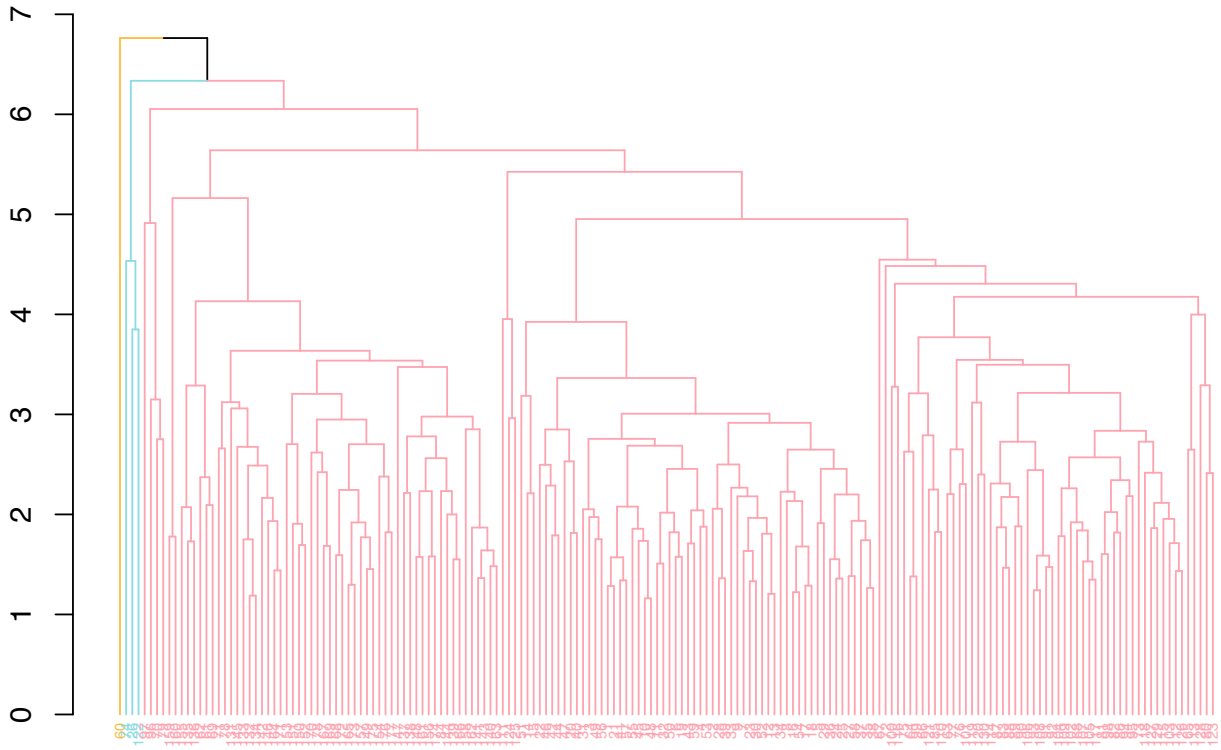


Figura B4: Dendrograma de todos os objetos sem aplicar o PCA, com distância euclidiana e o método de ligação "average".

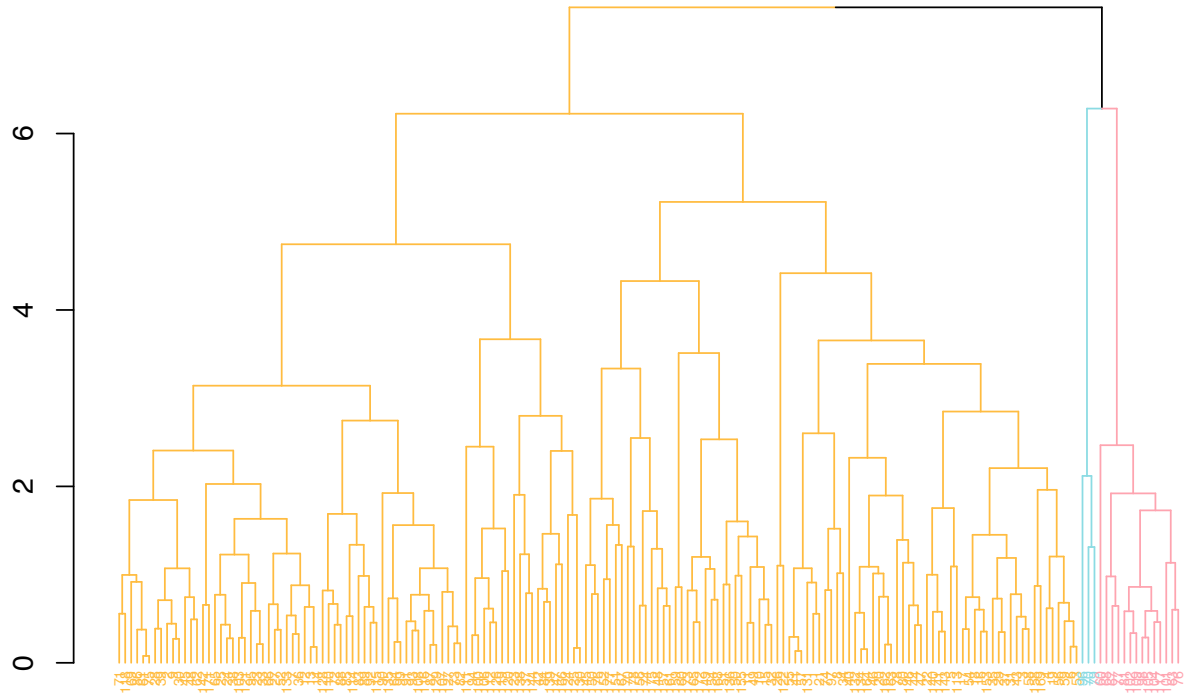


Figura B5: Dendrograma baseado nas cinco componentes principais do PCA, com distância euclidiana e o método de ligação "complete".

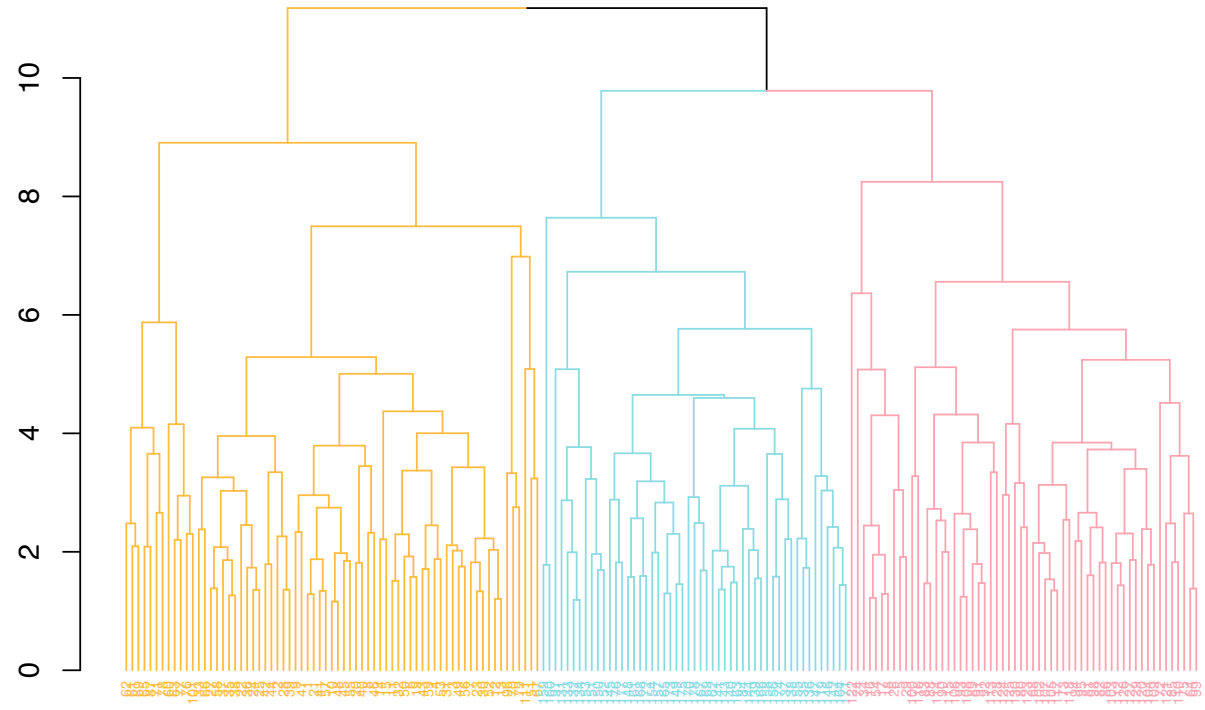


Figura B6: Dendrograma de todos os objetos sem aplicar o PCA, com distância euclidiana e o método de ligação "mcquitty".

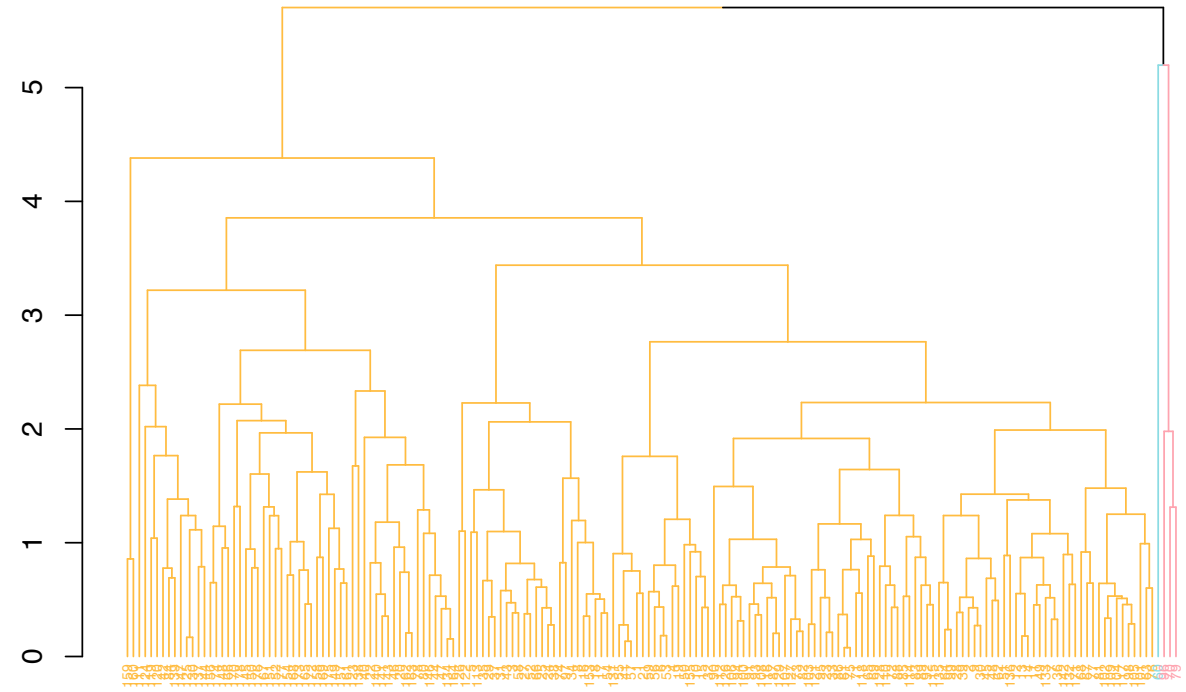


Figura B7: Dendrograma baseado nas cinco componentes principais do PCA, com distância euclidiana e o método de ligação "mcquitty".

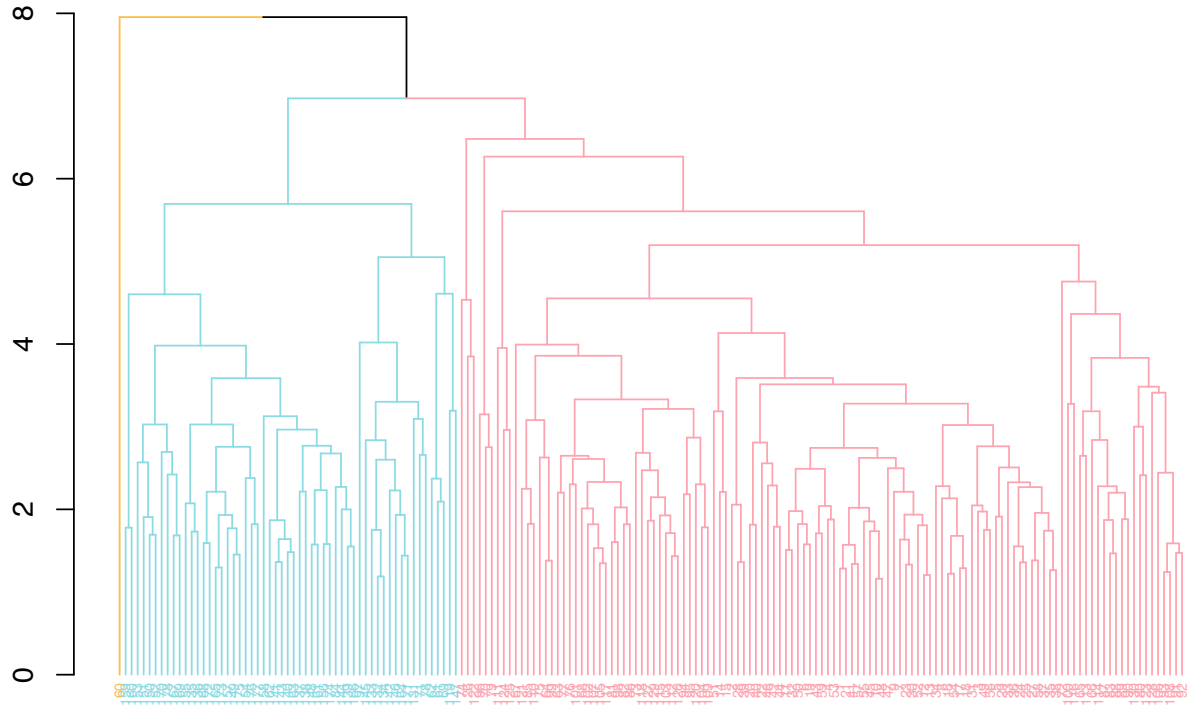


Figura B8: Dendrograma de todos os objetos sem aplicar o PCA, com distância euclidiana e o método de ligação "mcquitty".

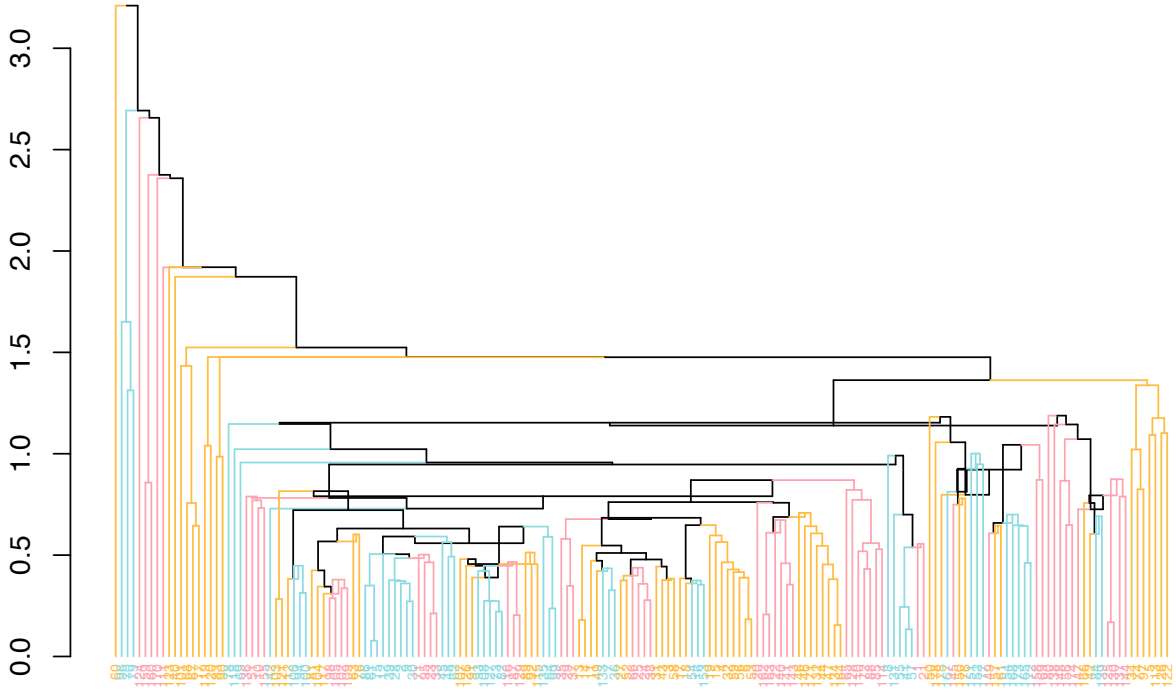


Figura B9: Dendrograma baseado nas cinco componentes principais do PCA, com distância euclidiana e o método de ligação "centroid".

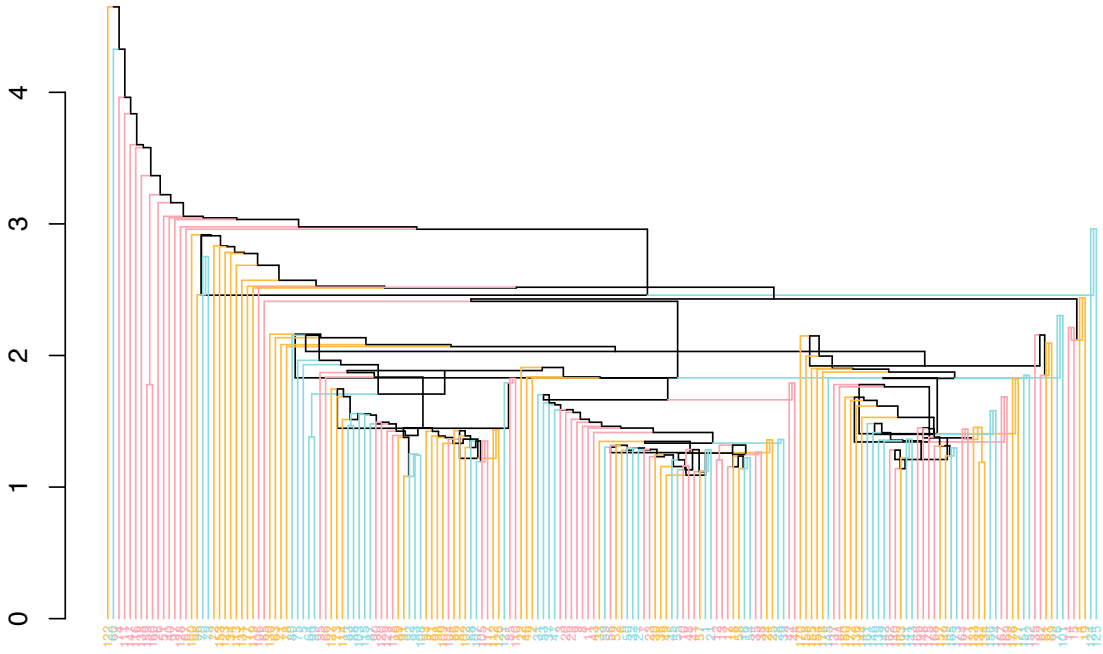


Figura B10: Dendrograma de todos os objetos sem aplicar o PCA, com distância euclidiana e o método de ligação "centroid".

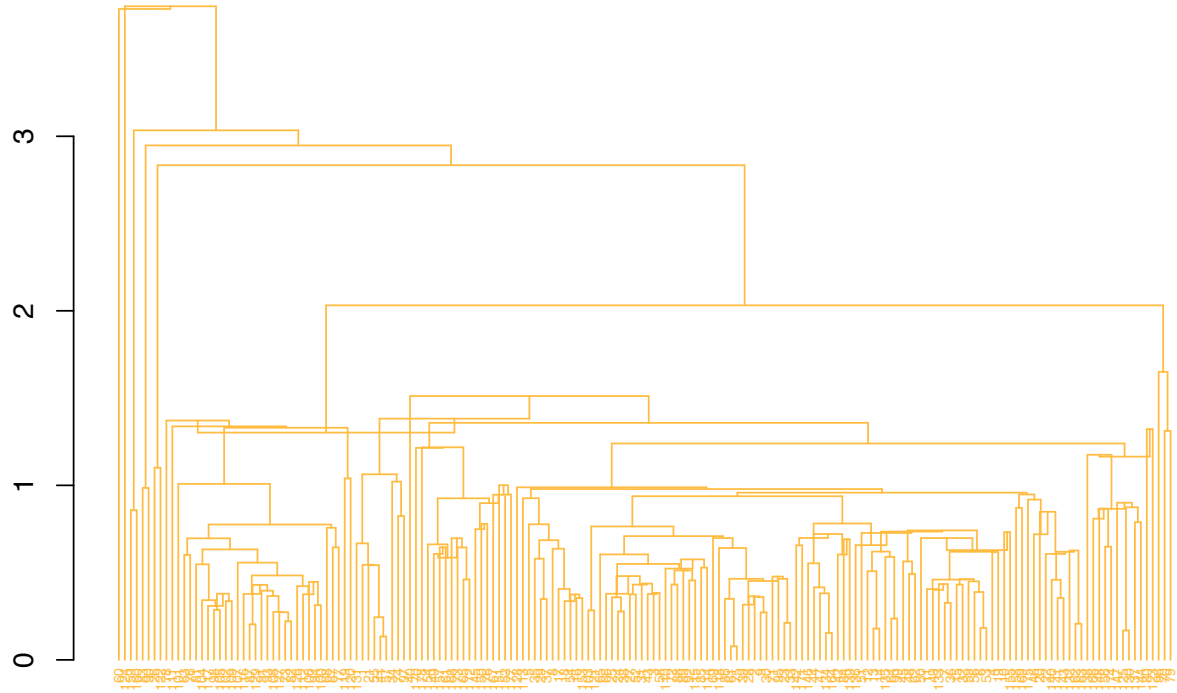


Figura B11: Dendrograma baseado nas cinco componentes principais do PCA, com distância euclidiana e o método de ligação "median".

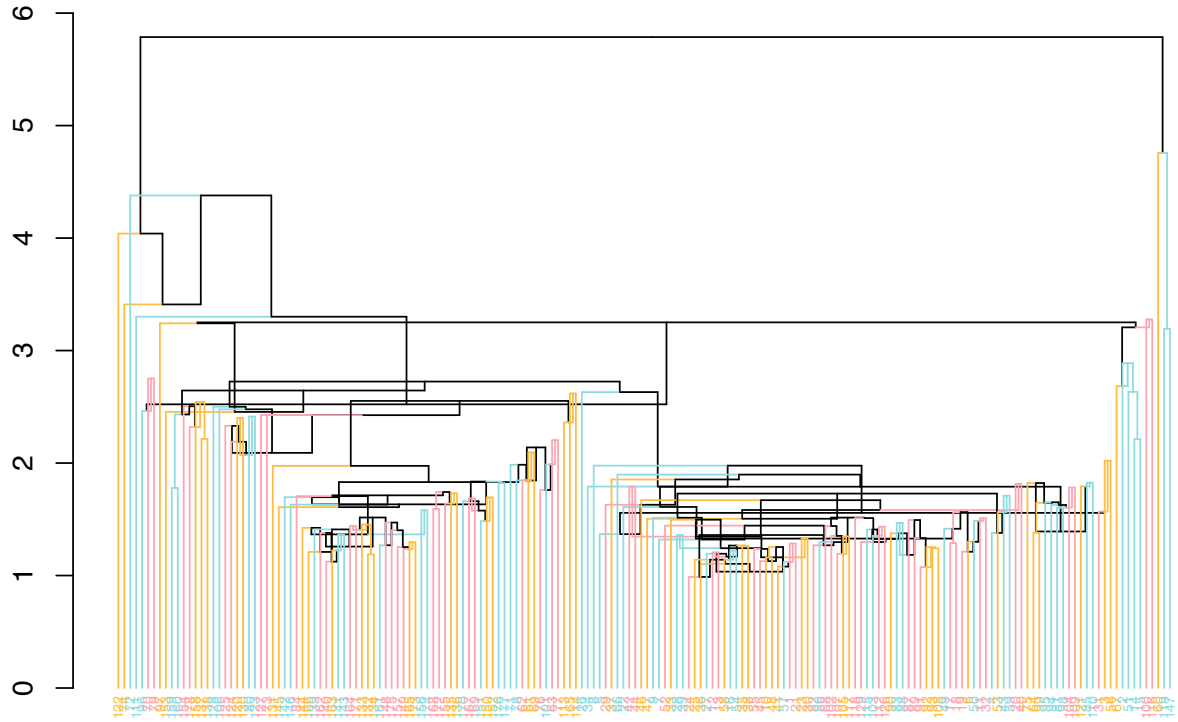


Figura B12: Dendrograma de todos os objetos sem aplicar o PCA, com distância euclidiana e o método de ligação "median".

Anexo C - Eliminação do Ruído: PCA

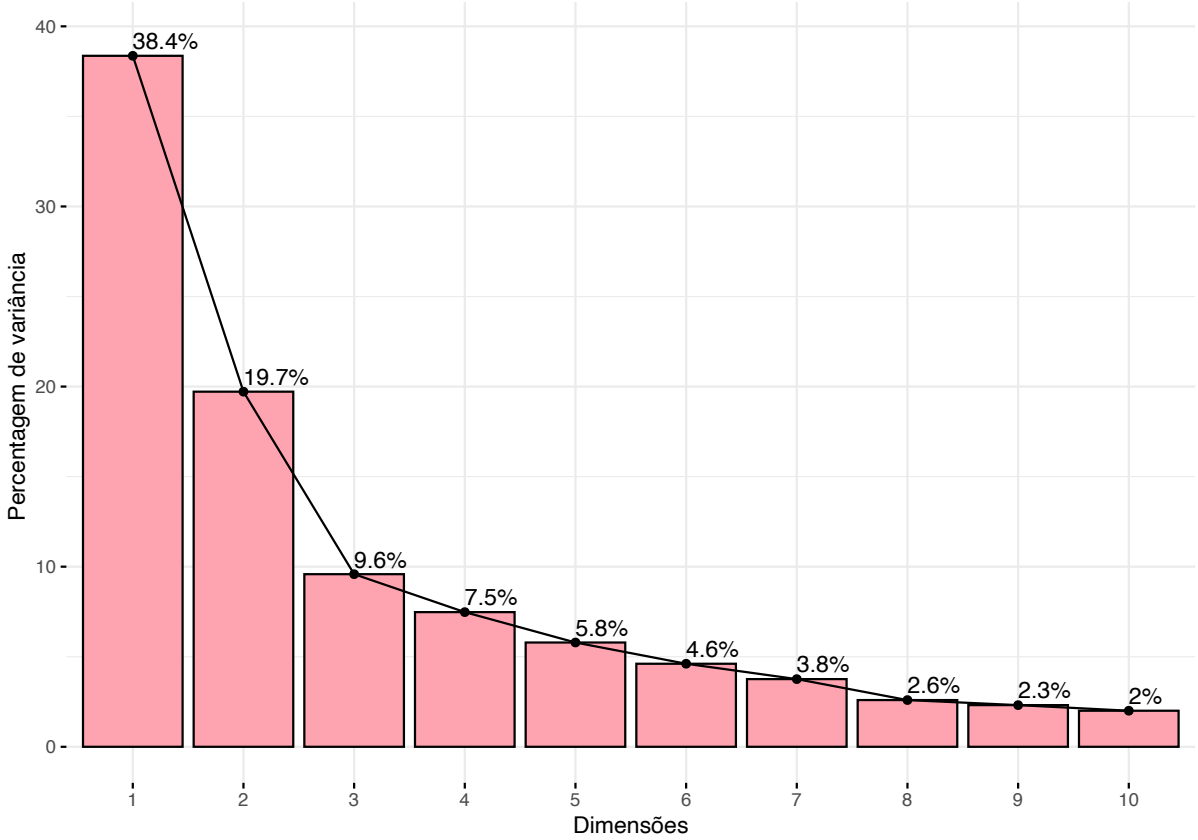


Figura C1: Screeplot dos dados de PCA após a eliminação do ruído, tendo representadas as percentagens de variância do sistema.

Tabela C1: Tabela dos valores próprios para cada uma das dimensões.

Dimensão (Dim)	Valor Próprio (λ)
Dim 1	4.73785289
Dim 2	2.43493398
Dim 3	1.18357584
Dim 4	0.92342995
Dim 5	0.71507556
Dim 6	0.56948411
Dim 7	0.46410560
Dim 8	0.32045914
Dim 9	0.28576875
Dim 10	0.24733235
Dim 11	0.22782328
Dim 12	0.16140105
Dim 13	0.07838505

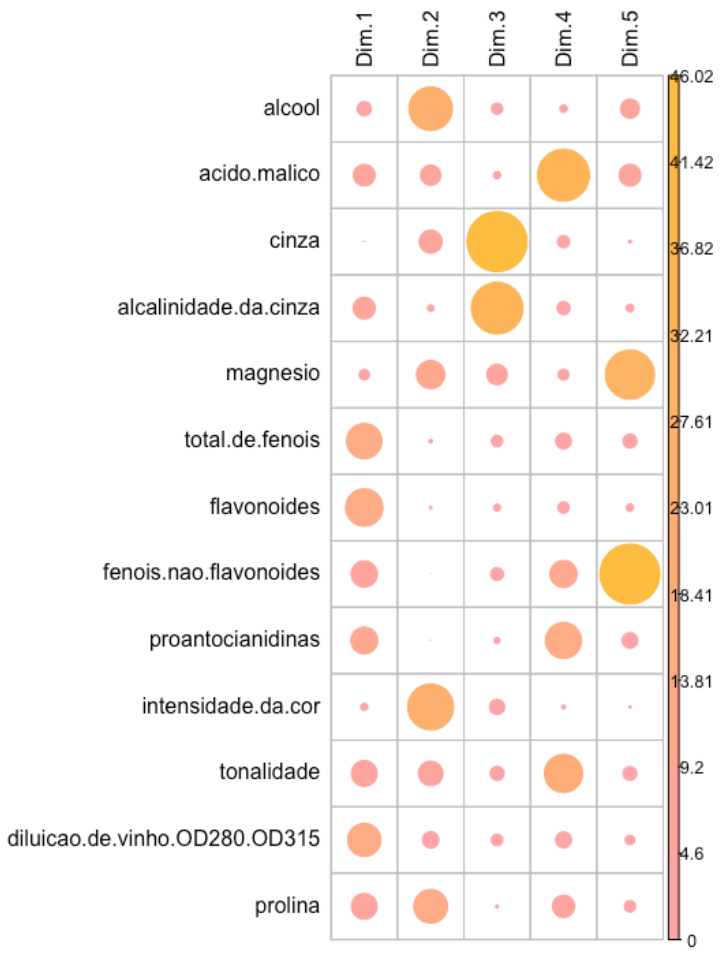


Figura C2: Mapa de fatores da contribuição das variáveis para as cinco componentes principais.

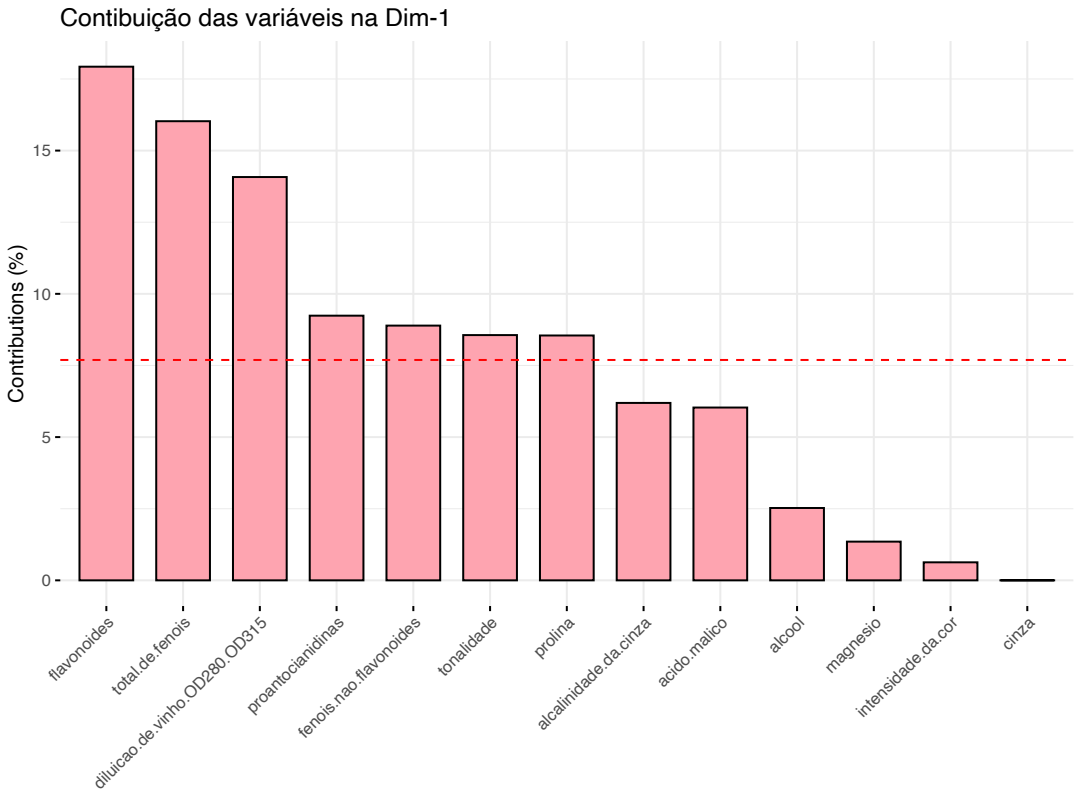


Figura C3: Gráfico das contribuições das variáveis para a primeira dimensão.

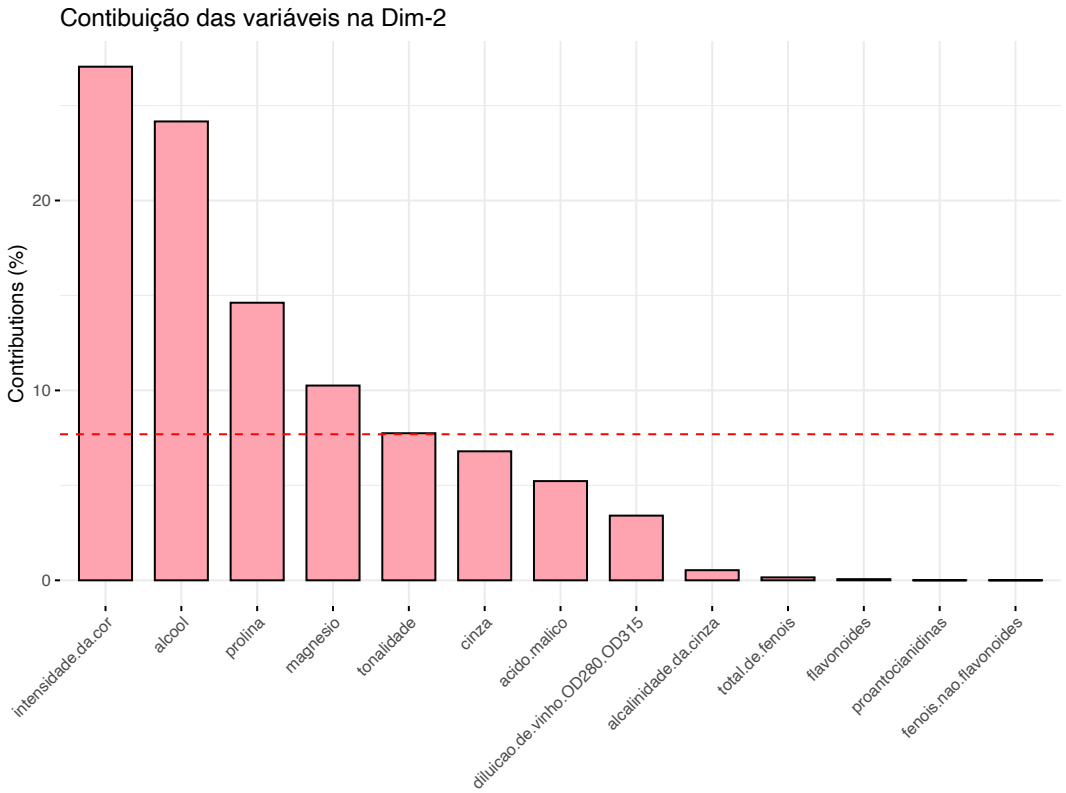


Figura C4: Gráfico das contribuições das variáveis para a segunda dimensão.

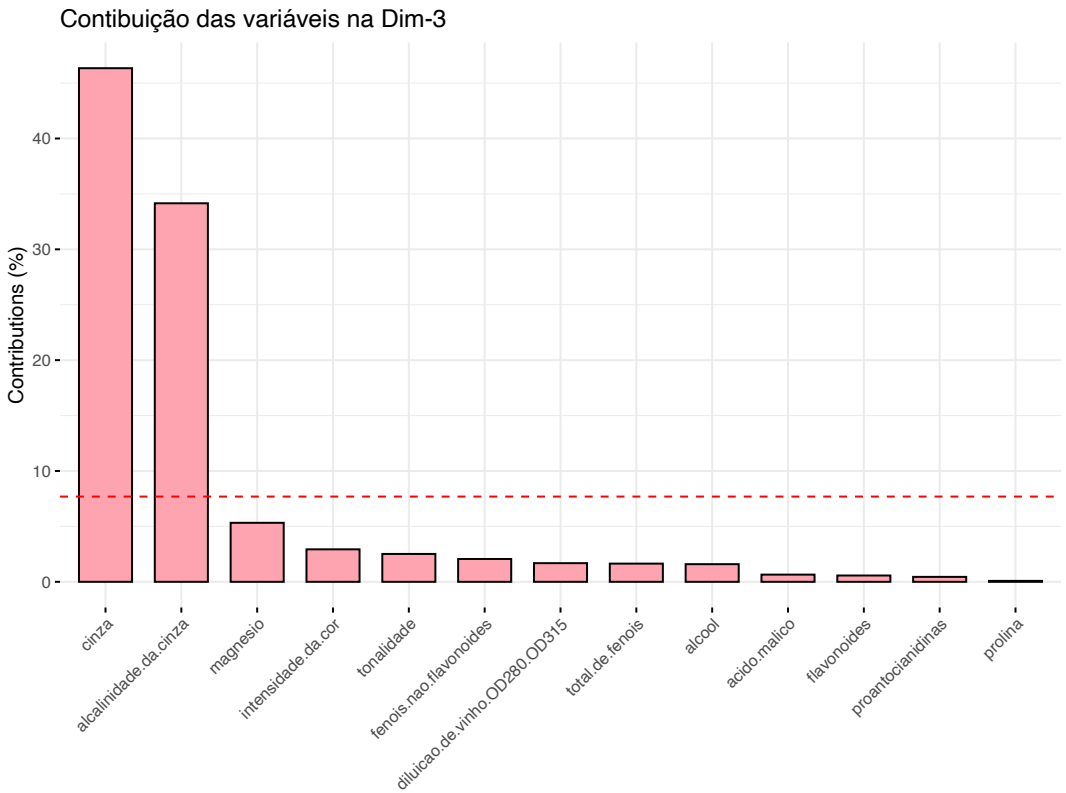


Figura C5: Gráfico das contribuições das variáveis para a terceira dimensão.

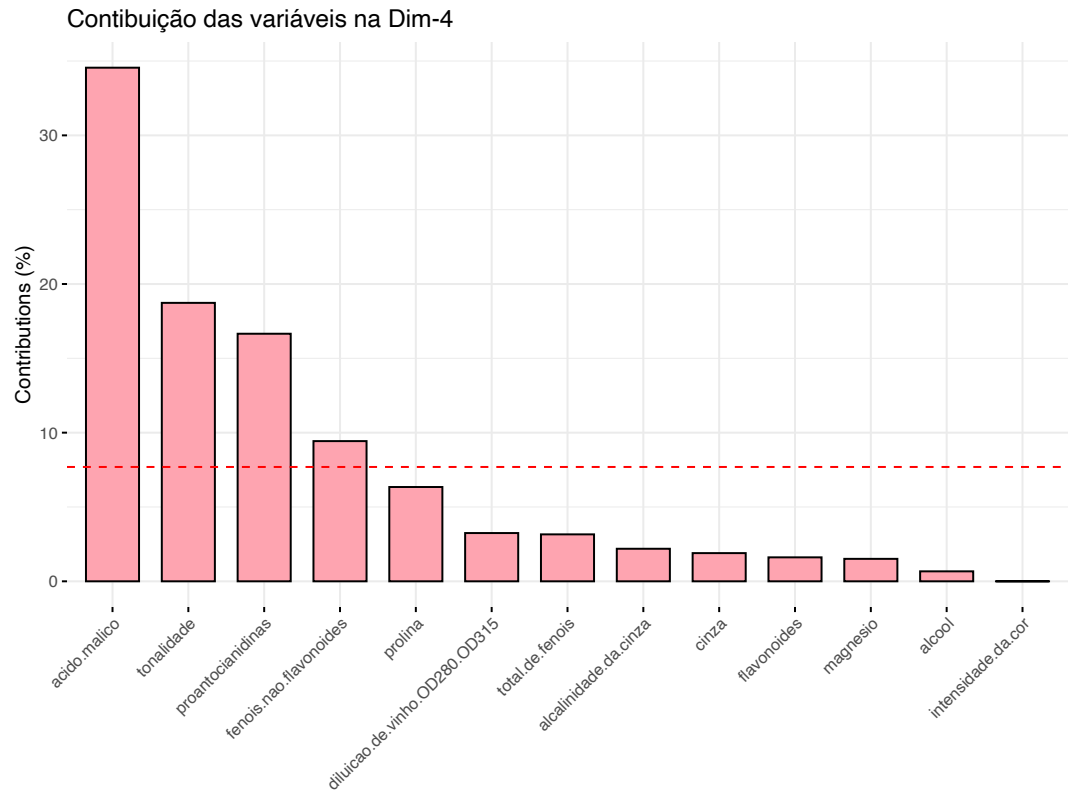


Figura C6: Gráfico das contribuições das variáveis para a quarta dimensão.

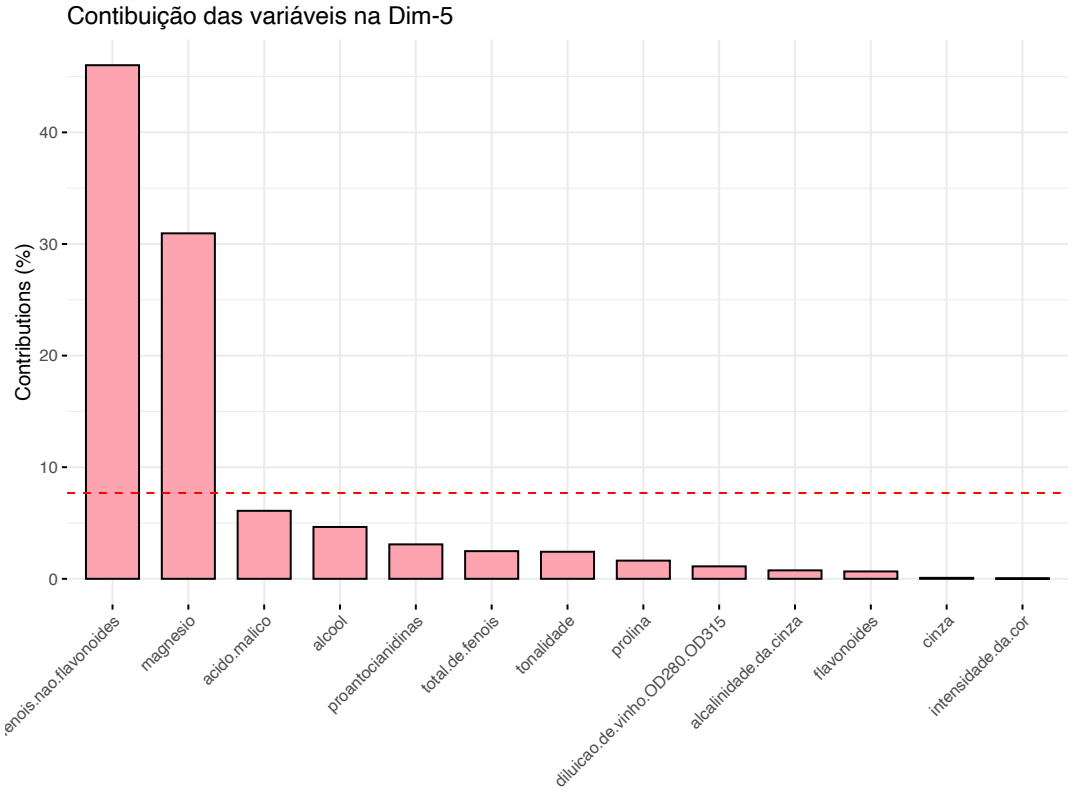


Figura C7: Gráfico das contribuições das variáveis para a quinta dimensão.

Anexo D - Eliminação do Ruído: HCA

Processo de validação 5 variáveis

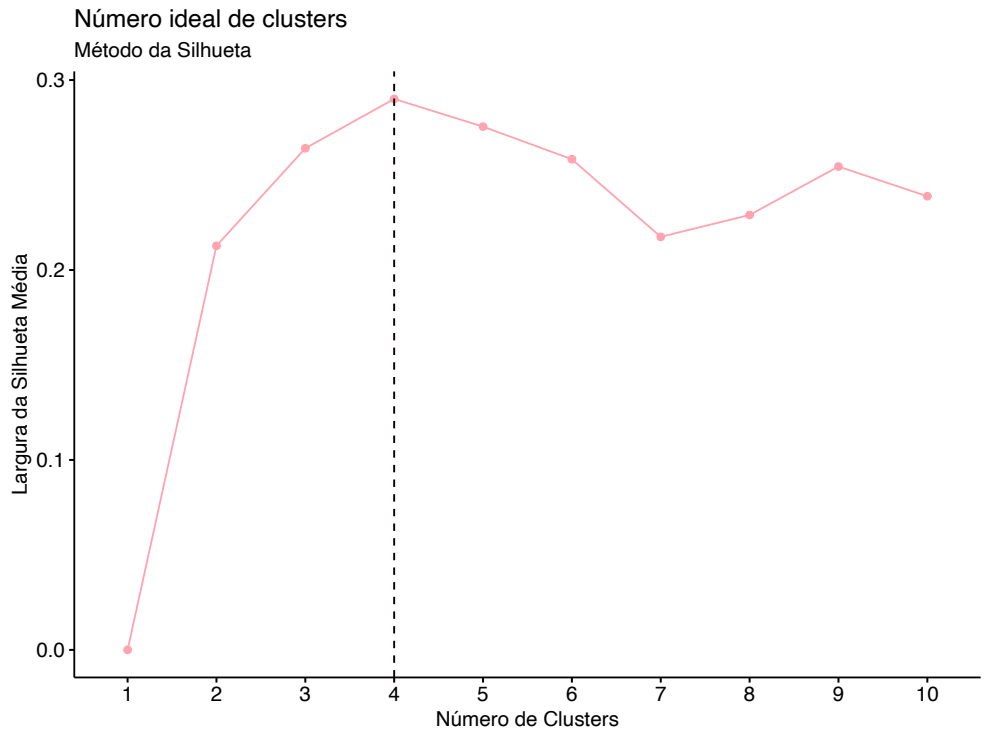


Figura D1: Representação gráfica do método de silhueta, havendo evidência de que o número ideal de cluster é 4.

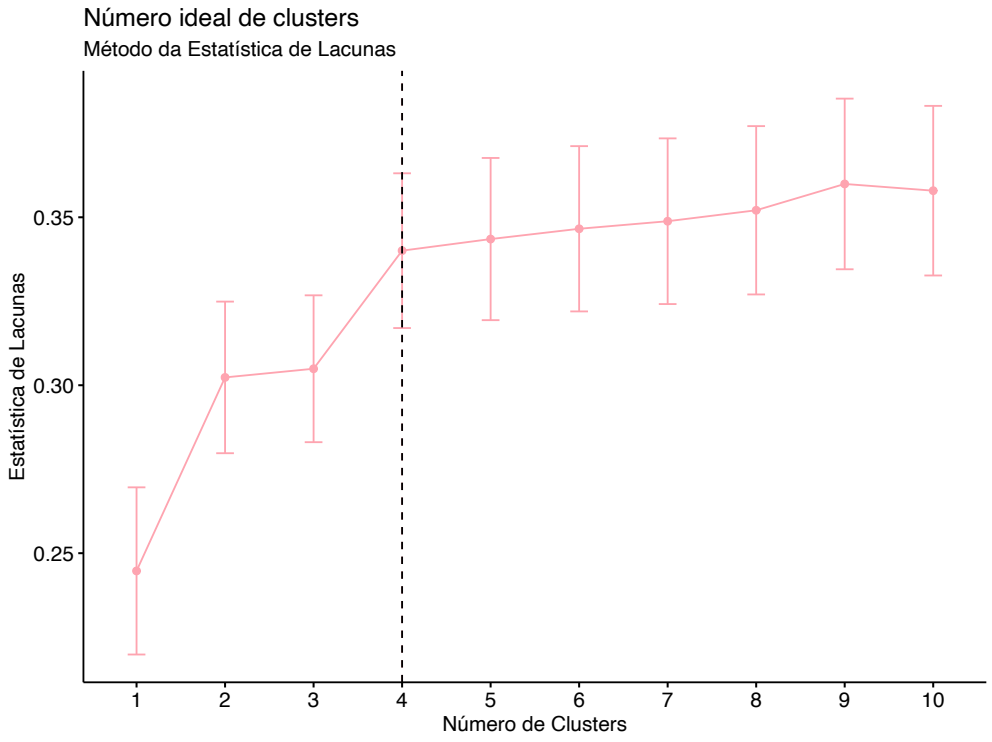


Figura D2: Representação gráfica do método de lacunas, havendo evidência de que o número ideal de cluster é 4.

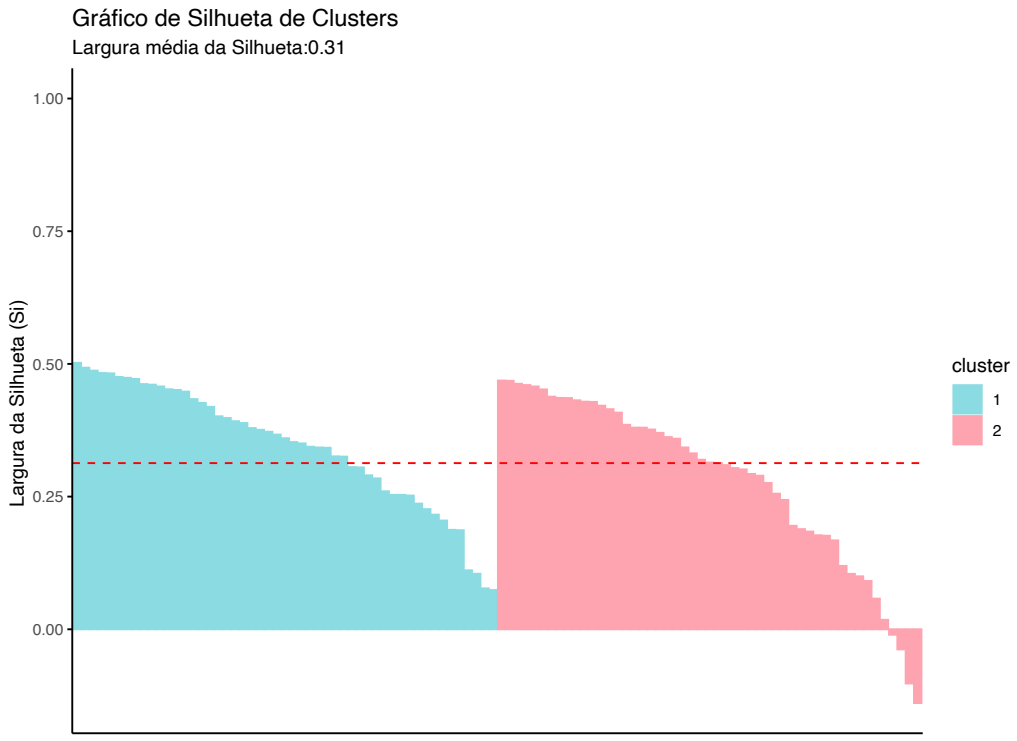


Figura D3: Gráfico da Silhueta de Clusters.

Tabela D1 Distribuição dos objetos com uma largura de silhueta negativa.

Objetos	Cluster em que se encontra	Cluster mais provável
45	2	1
121	2	1
9	2	1
23	2	1

Tabela D2 Método de estudo e número de clusters esperados para cada uma das medidas internas.

	Método de estudo	Número de clusters
Conectividade	Hierárquico	2
Coeficiente de Silhueta	Hierárquico	2
Índice de Dunn	Clara	5

Tabela D3: Método de estudo e número de clusters esperados para cada uma das medidas estabilidade.

	Método de estudo	Número de clusters
APN	Clara	2
AD	Clara	6
ADM	Clara	2
FOM	Clara	6

Processo de validação 10 variáveis

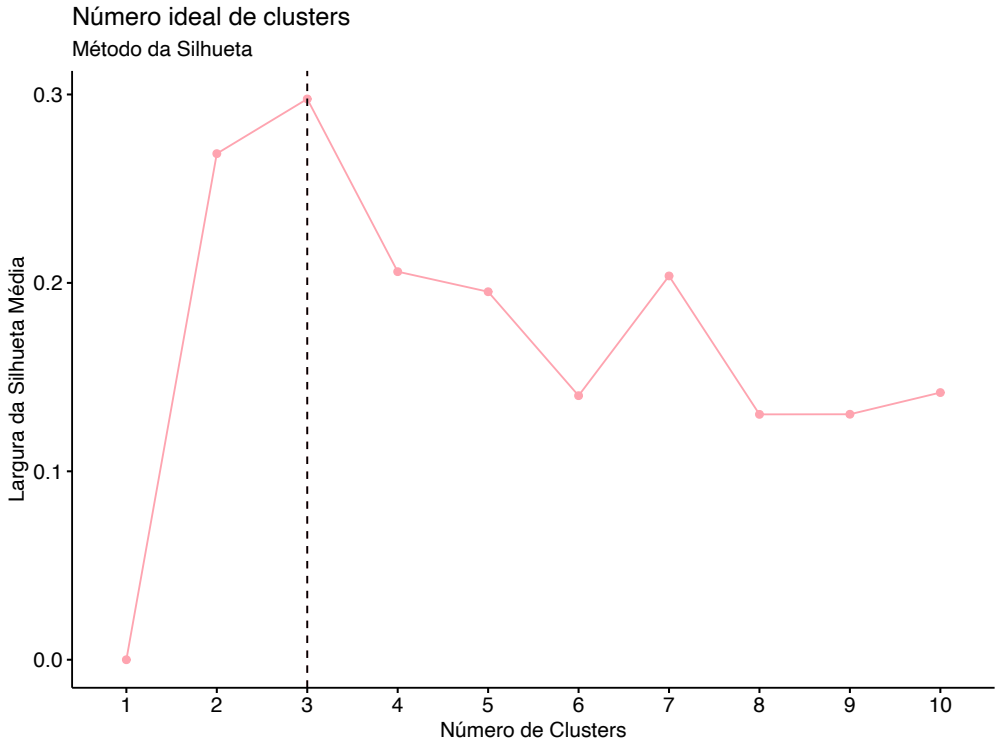


Figura D4: Representação gráfica do método de silhueta, havendo evidência de que o número ideal de cluster é 3.

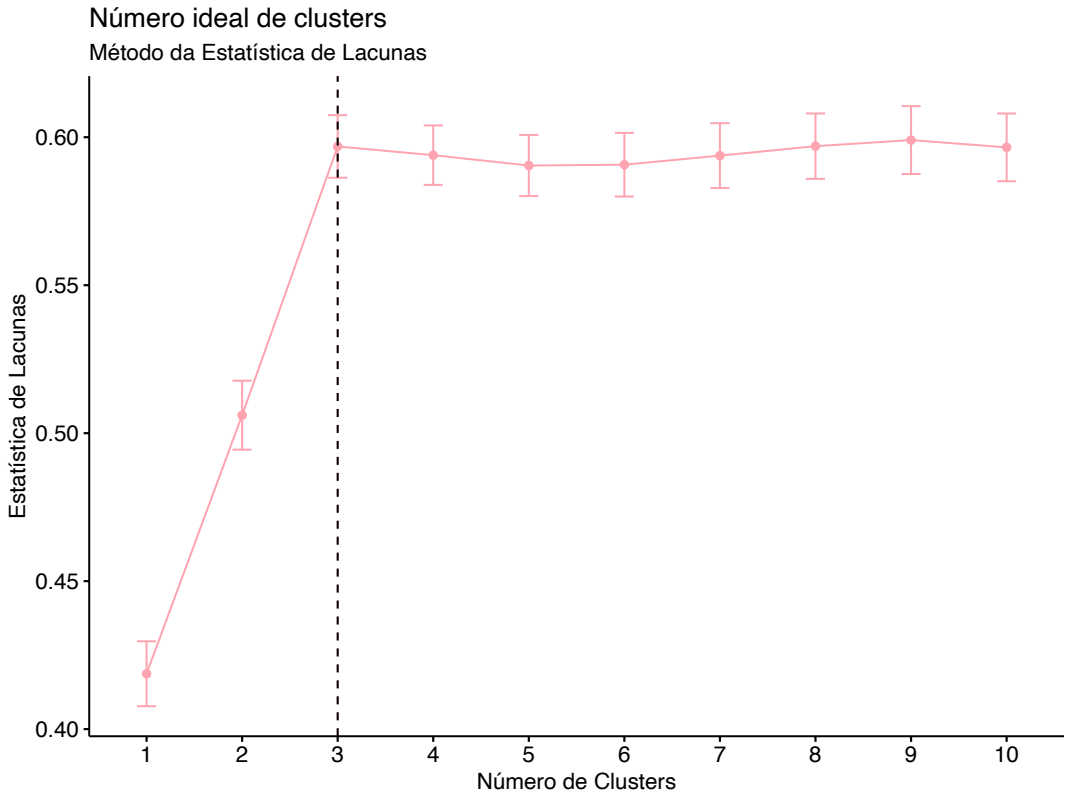


Figura D5: Representação gráfica do método de lacunas, havendo evidência de que o número ideal de cluster é 3.

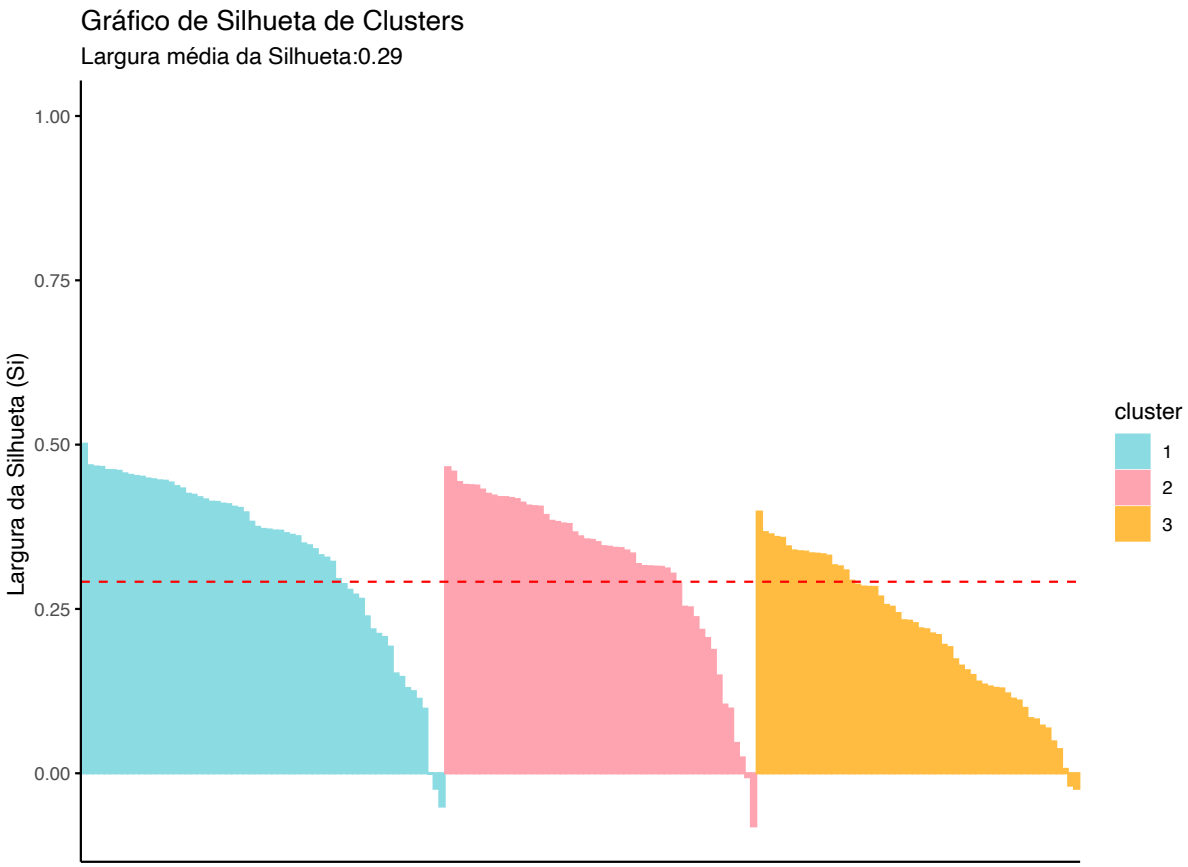


Figura D6: Gráfico da Silhueta de Clusters.

Tabela D4 Distribuição dos objetos com uma largura de silhueta negativa.

Objetos	Cluster em que se encontra	Cluster mais provável
72	1	3
67	1	3
66	1	3
61	2	3
78	2	3
79	3	1
99	3	1

Tabela D5 Método de estudo e número de clusters esperados para cada uma das medidas internas.

	Método de estudo	Número de clusters
Conectividade	Hierárquico	2
Coeficiente de Silhueta	K-médias	3
Índice de Dunn	Hierárquico	6

Tabela D6: Método de estudo e número de clusters esperados para cada uma das medidas estabilidade.

	Método de estudo	Número de clusters
APN	Hierárquico	2
AD	K-médias	6
ADM	K-médias	3
FOM	K-médias	6