INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Departamento de Engenharia Química

Estudo de uma rede de monitorização da qualidade do ar

RAQUEL MARIA GAIÃO BACALHAU
(Licenciada em Saúde Ambiental)

Trabalho Final de Mestrado para obtenção do grau de Mestre
em Engenharia da Qualidade e Ambiente

Orientadores: Professora Doutora Célia Maria da Silva Fernandes (ISEL)
Professor Doutor Paulo José Raimundo Ramos (ISEL)

Júri:
Presidente: Professora Doutora Isabel Maria da Silva João (ISEL)
Vogais: Professora Doutora Sandra Maria da Silva Figueiredo Aleixo (ISEL)
Professora Doutora Célia Maria da Silva Fernandes (ISEL)

Setembro de 2023

INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Departamento de Engenharia Química

Estudo de uma rede de monitorização da qualidade do ar

RAQUEL MARIA GAIÃO BACALHAU
(Licenciada em Saúde Ambiental)

Trabalho Final de Mestrado para obtenção do grau de Mestre
em Engenharia da Qualidade e Ambiente

Orientadores: Professora Doutora Célia Maria da Silva Fernandes (ISEL)
Professor Doutor Paulo José Raimundo Ramos (ISEL)

Júri:
Presidente: Professora Doutora Isabel Maria da Silva João (ISEL)
Vogais: Professora Doutora Sandra Maria da Silva Figueiredo Aleixo (ISEL)
Professora Doutora Célia Maria da Silva Fernandes (ISEL)

Setembro de 2023

Estudo de uma rede de monitorização da qualidade do ar

Raquel Maria Gaião Bacalhau

Copyright

Agradecimentos

O desenvolvimento desta tese permitiu me evoluir pessoal como profissionalmente, contudo foi uma longa jornada constituída por desafios. Foi um processo de angústia, trabalho árduo e por fim alegria. A sua realização contou com a ajuda de diversas pessoas, das quais agradeço:

À Professora Doutora Célia Maria Da Silva Fernandes e ao Professor Doutor Paulo José Raimundo Ramos pelo apoio, disponibilidade, conselhos e dedicação nas orientações prestadas. Fico extremamente grata por terem partilhado tanto o material técnico como o seu vasto conhecimento nesta área de estudo e me terem incentivado ao longo do percurso. Foi um enorme privilégio ter sido orientada por ambos.

Aos meus colegas de mestrado pelo companheirismo, amizade e espírito de equipa transmitido durante a minha formação académica.

Por último e mais importante, aos meus pais, por todo o apoio incondicional, pelo altruísmo de quererem o meu sucesso e pela força e sustentabilidade financeira. Aproveito também para agradecer todo o suporte e amor que me deram em casa.

A todos, os meus profundos e sinceros agradecimentos!

Parte I

Resumo/Abstract

Resumo

Uma Rede de Medição da Qualidade do Ar (RMQAr) é composta por várias Estações de Medição (EM). Em cada uma destas EM é recolhida informação sobre os principais poluentes presentes na sua zona de implantação. A otimização da RMQAr é um fator essencial para garantir o bom funcionamento das EM, pelo que é importante avaliar a existência de EM redundantes. A sua existência pode permitir a transferência das mesmas para novas localizações, alargando a área monitorizada, sem aumentar o custo com a sua manutenção. Neste trabalho pretende-se identificar zonas envolventes às EM da área metropolitana de Lisboa, com comportamentos semelhantes de poluição do ar, por um dos poluentes presentes e contribuir para a identificação de possíveis EM redundantes, usando técnicas de Estatística Multivariada.

Palavras-chave: Dióxido de azoto; Material particulado; Redes de monitorização da qualidade do ar; Estatística multivariada.

Esta página foi intencionalmente deixada em branco.

Abstract

An Air Quality Measurement Network (RMQAr) is composed of several Measurement Stations (EM). In each of these MS, information is collected on the main pollutants present in their area of implantation. Optimizing the RMQAr is an essential factor to ensure the proper functioning of the EM, so it is important to assess the existence of redundant EM. Their existence may allow their transfer to new locations, expanding the monitored area, without increasing the cost of maintenance. In this work, we intend to identify areas surrounding MS in the metropolitan area of Lisbon, with similar air pollution behaviors, by one of the pollutants present, and contribute to the identification of possible redundant MS, using Multivariate Statistics techniques.

Keywords: Nitrogen dioxide; Particulate matter; Air quality monitoring networks; Multivariate statistics.

Esta página foi intencionalmente deixada em branco.

Lista de Figuras

Lista de Tabelas

Esta página foi intencionalmente deixada em branco.

Lista de Acrónimos

AC Análise de Clusters, página 11

ACP Análise de Componentes Principais, página 11

AIC *Akaike Information Criterion*, página 12

APA Agência Portuguesa do Ambiente, página 2

$ATMIS$ Sistema de recolha e processamento de dados de qualidade do ar, página 2

$CCDR$ Comissões de Coordenação e Desenvolvimento Regional, página 1

DRA Direções Regionais de Ambiente das Regiões Autónomas dos Açores e Madeira, página 2

EM Estações de Monitorização, página 6

$EMQAr$ Estações de Monitorização da Qualidade do Ar, página 1

H_0 Hipótese Nula, página 21

H_1 Hipótese Alternativa, página 21

KMO Kaiser-Meyer-Olkin, página 26

KSL Kolmogorov-Smirnov com correção de Lilliefors, página 21

OMS Organização Mundial de Saúde, página 1

$REAR$ Regime de Emissões para o Ar, página 5

REI Regime de Emissões Industriais, página 5

$RMQAr$ Rede de Medição da Qualidade do Ar, página 1

$TEAR$ Título de Emissões para o Ar, página 5

UE União Europeia, página 5

VL Valor-Limite, página 2

Elementos químicos

Ar Árgon, página 3

CO Monóxido de Carbono, página 2

CO_2 Dióxido de Carbono, página 3

H_2O Água, página 3

N Azoto, página 3

NO Monóxido de Azoto, página 3

NO_x Óxidos de Azoto, página 3

NO_2 Dióxido de Azoto, página 3

O_2 Oxigénio, página 3

O_3 Ozono, página 3

PM Material Particulado, página 3

$PM_{2.5}$ Material Particulado com diâmetro aerodinâmico inferior a 2.5 micrómetros, página 4

PM_{10} Material Particulado com diâmetro aerodinâmico inferior a 10 micrómetros, página 3

Unidades de medida

mg/m^3 Miligramas por metro cúbico, página 2

$\mu g/m^3$ Microgramas por metro cúbico de volume de ar, página 2

Índice

Esta página foi intencionalmente deixada em branco.

Capítulo 1

Introdução

1.1 Rede de medição da qualidade do ar

Devido ao rápido crescimento populacional e à urbanização nas últimas décadas, a poluição do ar no meio urbano tornou-se uma das principais preocupações ambientais. Os efeitos diretos na saúde pública e no ambiente levaram a um aumento dos esforços para prevenir e controlar a poluição atmosférica (Gokce et al., 2020).

Há mais de 60 anos que a Organização Mundial da Saúde (OMS) se dedica a esta problemática, estimando-se que, mundialmente, cerca de sete milhões de pessoas morram prematuramente a cada ano devido à poluição do ar (Roser, 2021).

Com o intuito de avaliar a qualidade do ar do país, recorreu-se a Redes de Medição da Qualidade do Ar (RMQAr) que são constituídas por Estações de Monitorização da Qualidade do Ar (EMQAr) (QualAR, 2020). Estas permitem obter as concentrações dos poluentes atmosféricos, sendo que para cada poluente existe um conjunto de locais de medição, cuja localização obedece a um conjunto de requisitos (APA, 2021a).

Antes de descrever estas estações é importante diferenciar os conceitos de "zona" e "aglomeração", definidos no Decreto-Lei n.º 102/2010, de 23 de setembro:

- **Zona**: "área geográfica de caraterísticas homogéneas, em termos de qualidade do ar, ocupação de solo e densidade populacional delimitada para fins de avaliação e gestão da qualidade do ar."
- **Aglomeração**: "zona que constitui uma conurbação caraterizada por um número de habitantes superior a 250 000 ou em que o número de habitantes se situe entre os 250 000 e os 50 000 e tenha uma densidade populacional superior a 500 hab/km^2."

As EMQAr devem ser colocadas, de modo a fornecer dados sobre as zonas e aglomerações, onde a população possa estar direta ou indiretamente exposta a elevadas concentrações de poluentes e, em zonas e aglomerações que são representativas da exposição da população em geral. A sua quantidade depende do tamanho da população e dos níveis de poluição, sendo que as suas localizações estão definidas na Diretiva da União Europeia 2008/30/EC (Pires et al., 2009).

Estas zonas são delimitadas, periodicamente, a nível nacional, através dos resultados obtidos na avaliação da qualidade do ar e podem ser consultadas na base de dados do sistema de informação QualAR (APA, 2021b).

Em Portugal, estas estações são geridas e operadas pelas Comissões de Coordenação e Desenvolvimento Regional (CCDR) da região onde se inserem e pelas Direções Regionais de

Ambiente (DRA) das Regiões Autónomas dos Açores e Madeira e, encontram-se caraterizadas na base de dados no sistema de informação QualAR, através do qual também se pode obter as suas localizações (APA, 2021a).

A tipologia das estações de medição depende das emissões predominantes nas zonas rurais, urbanas e suburbanas onde se inserem bem como, os tipos de exposição da população à poluição atmosférica (APA, 2021a; Mobilizar, s.d.).

- **Estações de tráfego**: Situam-se na proximidade das vias de tráfego intenso. Permitem obter as concentrações máximas dos poluentes resultantes das emissões rodoviárias a que a população está exposta, durante períodos de curta duração.
- **Estações de fundo**: Localizam-se em zonas que não são afetadas diretamente por vias de tráfego ou por outra fonte de poluição próxima. Têm como objetivo avaliar a exposição média da população a concentrações de fundo.
- **Estações industriais**: Encontram-se próximas de zonas industriais, avaliando as concentrações máximas de poluentes emitidos no decorrer das atividades fabris.

Cada estação está equipada com um dispositivo de amostragem que recolhe o ar e o distribuí pelos analisadores, recorrendo-se a métodos de referência ou equivalentes para medir as concentrações de poluentes atmosféricos (APA, 2021a).

As medições efetuadas correspondem às concentrações médias, diárias ou anuais, dos poluentes presentes na atmosfera e são expressas em microgramas por metro cúbico de volume de ar $\left(\mu g/m^3\right)$ com exceção, do monóxido de carbono (CO) que é medido em miligramas por metro cúbico $\left(mg/m^3\right)$ (Mobilizar, s.d.).

Os dados provenientes do local de medição são enviados, de hora a hora, para os servidores localizados nas CCDR e nas DRA, que centralizam toda a informação, sujeita a uma primeira validação automática efetuada pela aplicação ATMIS (Sistema de recolha e processamento de dados de qualidade do ar). É da responsabilidade de um operador analisar e verificar se as concentrações excedem os limiares de informação e alerta definidos para os diversos poluentes (APA, 2021a).

De seguida, os dados são enviados para o sistema de informação QualAR, sediada na Agência Portuguesa do Ambiente (APA), sob forma de concentrações médias horárias e de um índice de qualidade do ar para as diversas zonas sendo, posteriormente, disponibilizados ao público (APA, 2021a; QualAR, 2020).

A Figura 1.1 descreve, graficamente, as etapas supramencionadas.

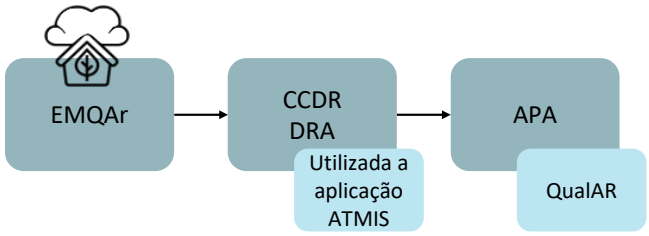

Figura 1.1: Etapas de uma Rede de Monitorização de Qualidade do Ar.

Para proteger a saúde humana efetua-se a análise de conformidade legal através de cálculos estatísticos, que permitem comparar os valores-limite (VL) legislados e identificar a existência de zonas em excedência aos valores estabelecidos (QualAR, 2020).

Assim, a avaliação da qualidade do ar, ou seja, a sua monitorização, possibilita a aquisição de conhecimento, para que posteriormente sejam implementadas medidas de gestão, com o intuito de reduzir os níveis dos diversos poluentes na atmosfera (APA, 2021a).

1.2 Importância dos estudos da qualidade do ar e os impactes na saúde humana

A atmosfera é uma camada gasosa que envolve a Terra e tem cerca de 1000 quilómetros de extensão, sendo composta por seis camadas (troposfera, a estratosfera, a mesosfera, a termosfera, a ionosfera e a exosfera) e constituída por cerca de 78% de azoto (N), 21% de oxigénio (O_2) e 1% de árgon (Ar), vapor de água (H_2O) e dióxido de carbono (CO_2) (APA, 2021c).

As substâncias emitidas para esta camada podem ter um maior ou menor impacte na qualidade do ar, consoante a composição química, a concentração, as reações químicas e físicas e a topografia do local (APA, 2021c). Adicionalmente, as condições meteorológicas determinam a forma como os poluentes se comportam no ar uma vez que, o vento horizontal, a estabilidade da atmosfera, a altura da fonte emissora e a variação da temperatura e pressão com a altura, dispersam os poluentes (Gomes, 2022).

A poluição atmosférica e o consequente fator de risco para a saúde humana e para o ambiente levaram ao aumento dos esforços para prevenir e controlar este fenómeno (Pires et al., 2009).

Atualmente, o material particulado (PM), o dióxido de azoto (NO_2) e o ozono (O_3) são geralmente reconhecidos como os três poluentes que mais influenciam a saúde humana. A gravidade do impacto da exposição prolongada e dos picos de exposição a estes poluentes varia, desde os danos causados ao sistema respiratório até à morte prematura (EEA, 2020a).

Assim, para esta dissertação será referido o NO_2 e material particulado em suspensão na atmosfera com diâmetro aerodinâmico inferior a 10 micrómetros por metro cúbico (PM_{10} em $\mu g/m^3$).

1.2.1 Dióxido de azoto

Os óxidos de azoto (NO_x) são um grupo de gases reativos compreendendo, compostos de azoto e oxigénio nomeadamente, o monóxido de azoto (NO) e o dióxido de azoto (NO_2), sendo estes considerados os poluentes mais significativos para a troposfera (Pires et al., 2008a).

Estes poluentes podem ser emitidos por processos industriais, comerciais, residenciais, energéticos e de fabrico que envolvam a utilização de azoto (Pires et al., 2008b). Eventos naturais como, processos biológicos anaeróbios ocorrentes no solo e na água, e ainda a atividade vulcânica também contribuem para esta emissão (APA, 2021d; Pires et al., 2008b).

Contudo, nas zonas urbanas, os transportes marítimos e rodoviários com motores a explosão ou combustão são a principal fonte de NO_x, representando mais de 80% das emissões dos transportes (EEA, 2020b).

As contribuições das diferentes fontes e setores de emissão para as concentrações de NO_x no ambiente dependem não só da quantidade de poluente emitida, das condições meteorológicas, mas também das condições de emissão como, a altura dos pontos de emissão e da distância até ao local recetor (EEA, 2020b).

Devido à elevada instabilidade, a molécula de NO_x reage, rapidamente, com o oxigénio, formando NO_2 (APA, 2021d). Assim sendo, o NO_2 é um gás castanho-avermelhado ou um líquido amarelo-escuro com odor forte e irritante, oxidante e extremamente tóxico, que resulta

da queima de combustível fóssil, especialmente no tráfego automóvel e no setor industrial (APA, 2021d; WHO, 2021). Quando presente na atmosfera pode contribuir para a formação de ozono (O_3) e dar origem a ácido nítrico e a nitratos orgânicos, afetando, negativamente, o ambiente e o ecossistema marinho através das chuvas ácidas e consequentemente à eutrofização dos lagos e rios. As reações que ocorrem entre a luz solar e os poluentes, nomeadamente, NO_2 e hidrocarbonetos, provocam uma névoa avermelhada na atmosfera, deixando o material particulado (PM) e o O_3 ao nível do solo. Este fenómeno designa-se de "*smog*" fotoquímico e é mais comum estar presente em cidades com climas ensolarados, quentes e secos, contudo, através da ação do vento, também pode afetar as áreas envolventes (Rani et al., 2011). Embora, o "*smog*" fotoquímico não seja muitas vezes visível, leva a irritações no aparelho respiratório e nos olhos, e em casos de elevadas concentrações de poluentes oriundos da industrialização e dos veículos automóveis pode ser fatal (Rani et al., 2011).

Os efeitos da exposição a este poluente ainda não são totalmente conhecidos, embora diminuam a resistência a infeções pulmonares e aumentem o risco de doenças graves, como edema agudo do pulmão.

Em contrapartida, a exposição frequente a níveis elevados pode causar distúrbios respiratórios, provocando, em crianças e grupos de risco como os asmáticos, uma maior tendência para problemas respiratórios como, enfisema pulmonar, bronquites, traqueítes e cancro (Castro et al., 2013; Silva, 2019). Continuamente, pode provocar lesões cáusticas na pele, irritação nos olhos e na garganta e desencadear danos no sistema nervoso central e nos tecidos (APA, 2021d).

1.2.2 Material Particulado

As partículas em suspensão encontradas na atmosfera podem ser substâncias minerais e/ou orgânicas no estado sólido ou líquido (Pires et al., 2008a). Se o PM for libertado diretamente para a atmosfera é considerado primário enquanto que se se formar conjuntamente com outros poluentes, a partir de reações químicas ocorrentes na camada gasosa, é designado de secundário (APA, 2021e).

O PM_{10} pode ser emitido, por fontes naturais, ou seja, erupções vulcânicas, atividade sísmica, incêndios florestais, intrusões de massas de ar contendo partículas e poeiras com origem nos desertos do Norte de África, mas também por fontes antropogénicas englobando, os processos industriais, comerciais, residenciais, construção, transportes e atividades agrícolas que envolvam a combustão (APA, 2021e; Pires et al., 2008a).

Uma vez que as PM são consideradas poluentes atmosféricos históricos é de elevado interesse monitorizar as suas concentrações. Para tal, deve-se ter em consideração o seu transporte por ação do vento ou chuva, provocando um aumento das concentrações em locais distantes da fonte (Pires et al., 2008a).

Devido às suas reduzidas dimensões, estes poluentes são os que mais influenciam negativamente a saúde humana na Europa, estando associados tanto a problemas respiratórios e cardiovasculares, como à morbilidade e mortalidade infantil (Pires et al., 2008a). Embora, para este trabalho não seja necessário mencionar o material particulado em suspensão na atmosfera com diâmetro aerodinâmico inferior a 2.5 $\mu g/m^3$ ($PM_{2.5}$) é relevante referir estas partículas, uma vez que possuem um tamanho mais reduzido, apresentam uma maior probabilidade de penetrar o sistema respiratório (APA, 2021e). Para além disso, as PM podem ligar-se a hidrocarbonetos ou metais pesados e serem absorvidos, por meio do processo respiratório, na circulação sanguínea (APA, 2021e).

1.2.3 Controlo das emissões de poluentes

O setor industrial e energético, as atividades de exploração agropecuárias/agrícolas e as suas atividades conexas como, a compostagem, produção de biogás, atividades de incineração e aplicação de fertilizantes são os maiores contribuintes da poluição atmosférica, existindo, assim, vários regimes legais em matérias de ambiente (APA, 2021f).

Através da regulamentação da União Europeia (UE) tem ocorrido o decréscimo das emissões de óxido de azoto, partículas e ainda, de hidrocarbonetos e monóxido de carbono provenientes das fontes móveis, contribuindo para a proteção da qualidade do ar e a redução das emissões de gases com efeito de estufa (APA, 2021f).

Entre os regimes legais aplicáveis destacam-se o Regime de Emissões Industriais (REI) que integra os requisitos de controlo de emissões em diversas atividades e o Regime de Emissões para o Ar (REAR) que determina que, para a execução das atividades, as infraestruturas emissoras de poluentes provenientes da queima de combustíveis fósseis têm de obter uma licença - Título de Emissões para o Ar (TEAR), onde são estabelecidas as condições para o acompanhamento e verificação do cumprimento dos requisitos impostos para a proteção do ambiente e os VL de emissão (APA, 2021f,g).

Relativamente à avaliação da qualidade do ar, a APA é responsável por promover e implementar a política de avaliação e gestão da qualidade do ar ambiente estabelecida no Decreto-Lei n.º 102/2010, de 23 de setembro (APA, 2021c). Esta legislação estipula os objetivos da qualidade do ar que incluem os VL, os valores-alvo e os níveis críticos consoante os poluentes. Estes valores e os métodos e critérios usados para a avaliação são comuns para todos os países da UE, possibilitando a comparação dos resultados de vários locais e regiões (APA, 2021c).

Torna-se, assim, relevante diferenciar cinco conceitos que se encontram definidos no Decreto-Lei 102/2010, de 23 de setembro, artigo 2.º, alíneas l, m, x, gg e hh:

- **Limiar de alerta**: "um nível acima do qual uma exposição de curta duração apresenta riscos para a saúde humana da população em geral e a partir do qual devem ser adotadas medidas imediatas";
- **Limiar de informação**: "um nível acima do qual uma exposição de curta duração apresenta riscos para a saúde humana de grupos particularmente sensíveis da população, a partir do qual é necessário a divulgação imediata de informações adequadas";
- **Nível crítico**: "um nível fixado com base em conhecimentos científicos, acima do qual podem verificar-se efeitos nocivos diretos em recetores como árvores, outras plantas ou ecossistemas naturais, mas não em seres humanos";
- **Valor alvo**: "um nível fixado com o intuito de evitar, prevenir ou reduzir os efeitos nocivos na saúde humana e ou no ambiente, a atingir, na medida do possível, durante um determinado período de tempo";
- **Valor limite**: "um nível fixado com base em conhecimentos científicos com o intuito de evitar, prevenir ou reduzir os efeitos nocivos na saúde humana e ou no ambiente, a atingir num prazo determinado e que, quando atingido, não deve ser excedido".

Em Portugal é obrigatório efetuar a avaliação de conformidade legal e comunicar a respetiva informação à Comissão Europeia. Para além dos valores referidos, estão ainda estabelecidos, limiares de informação e alerta para determinados poluentes cuja exposição a elevadas concentrações, em períodos de curta duração, é mais alarmante (APA, 2021h).

Segundo o artigo 24.º, alínea 2 da legislação supramencionada, sempre que níveis medidos excedam os VL, as CCDR e as DRA elaboram planos de qualidade do ar e adotam medidas,

para garantir que as concentrações dos poluentes cumprem com os objetivos estipulados e evitar que a população esteja exposta a níveis que representem riscos significativos para a saúde (APA, 2021h). Adicionalmente, quando esta situação ocorre, compete à APA disponibilizar ao público, através do site na *Internet*, a informação transmitida à Comissão Europeia e sensibilizar a população para esta problemática pois, os cidadãos devem contribuir para melhorar a qualidade do ar através da adoção de boas práticas (APA, 2021c).

Para complementar este trabalho, os valores obtidos nas redes de monitorização foram comparados com os VL de NO_2 e PM_{10} estabelecidos no regime legal. Estes definem-se como níveis fixados com o intuito de evitar, prevenir ou reduzir os efeitos nocivos na saúde humana e/ou no ambiente e quando atingidos, não devem ser excedidos (APA, 2021h).

Segundo a CCDR (2022) e o artigo 18.º da legislação supramencionada, no anexo XII encontram-se os valores legislados para os poluentes abordados neste trabalho (Tabela 1.1).

Poluente	**Valor**	**Período de referência**	**Concentração**	**Nº de excedências permitidas**
NO_x	Nível crítico	Ano civil	30 $\mu g/m^3$	
NO_2	Limiar alerta	Uma hora	400 $\mu g/m^3$	
	Valor limite	Uma hora	200 $\mu g/m^3$	Não exceder mais de 18 vezes por ano civil
		Ano civil	40 $\mu g/m^3$	
PM_{10}	Valor limite	Diário	50 $\mu g/m^3$	Não exceder mais de 35 vezes por ano civil
		Ano civil	40 $\mu g/m^3$	

Tabela 1.1: Valores legislados para NO_x, NO_2 e PM_{10}

No que diz respeito ao NO_2, o indicador utilizado é a média anual, ou seja, o valor agregado com base nas concentrações horárias medidas em cada estação, sendo posteriormente comparado com o respetivo VL (40 $\mu g/m^3$). A análise da qualidade do ar das zonas e aglomerações expostas a este poluente efetua-se considerando o pior valor obtido nas estações pertencentes a cada uma dessas unidades territoriais (APA, 2023a).

Para controlar a exposição diária às PM_{10}, sobretudo nas cidades, foi estabelecido o VL diário de 50 $\mu g/m^3$ (que não deve ser excedido mais de 35 dias por ano civil) e VL anual de 40 $\mu g/m^3$. São utilizados dois indicadores para avaliar a exposição a estas partículas: o número de excedências ao VL diário e o VL da média anual. A tendência de evolução de exposição da população às partículas inaláveis realiza-se com base na agregação nacional dos valores médios anuais, associados à pior situação registada em cada zona/aglomeração (APA, 2023b).

1.3 Estrutura e objetivos do trabalho

A presente dissertação tem por objetivo identificar as zonas envolventes às Estações de Monitorização (EM) com comportamentos semelhantes de poluição de ar, contribuindo para a

identificação de possíveis EM redundantes. Para tal, foram escolhidas sete estações na zona de Lisboa e obtidas as concentrações de NO_2 e PM_{10} no sistema de informação da QualAR. Posteriormente, através de técnicas de estatística multivariada, verificaram-se as estações que manifestavam comportamentos similares em termos de concentração de poluentes, aferindo as que, futuramente, poderão vir a ser encerradas ou deslocadas para um novo local. Contudo, devido à falta de observações, e tal como será explicado no capítulo seguinte, foram somente identificadas as possíveis EM redundantes para NO_2.

Este trabalho encontra-se dividido em quatro capítulos e referências.

- No primeiro capítulo apresenta-se uma breve introdução da dissertação, para além de se indicar o objetivo em estudo.
- No segundo capítulo apresenta-se a metodologia utilizada bem como, um enquadramento de cada um dos métodos estatísticos utilizados.
- O terceiro capítulo trata dos resultados obtidos neste estudo, concluindo-se quais as EMQAr que parecem ser redundantes.
- No quarto e último capítulo são apresentadas as conclusões que se retiraram desta dissertação.
- As referências são expostas no fim do estudo em questão.

Esta página foi intencionalmente deixada em branco.

Capítulo 2

Metodologia

As informações contidas neste capítulo foram retiradas de livros (Härdle & Hlávka., 2015; Alkarkhi & Alqaraghuli , 2020; Johnson & Wichern. , 2014; Zelterman. , 2015) e apontamentos das disciplinas de Complementos de Estatística para a Engenharia do Mestrado de Engenharia da Qualidade e Ambiente, de Estatística Multivariada da Licenciatura em Matemática Aplicada à Tecnologia e à Empresa e de Técnicas de Estatística Multivariada das Licenciaturas em Engenharia Química e Biológica e em Engenharia Informática e de Computadores (Fernandes & Ramos, 2022a,b,c,d,e,f,g,h), do Instituto Superior de Engenharia de Lisboa.

2.1 Estações de Monitorização da Qualidade do Ar

A RMQAr é constituída por várias EMQAr, que são geridas pelas CCDR da região onde operam, sob responsabilidade da APA. Entre 2001 e 2019 existiram mais de 25 EMQAr na Área Metropolitana de Lisboa. Ao longo destes 19 anos, a RMQAr sofreu várias modificações, nomeadamente a desativação e/ou ativação de EM e ainda algumas alterações no tipo de poluentes que monitorizam.

Atualmente, em Portugal existem 69 EM das quais, 16 encontram-se em zonas rurais, 11 em áreas suburbanas e 42 em zonas urbanas. Desta vasta gama de estações urbanas, para este trabalho, foram selecionadas, 6 do concelho de Lisboa (Avenida da Liberdade ou Av. Liberdade, Beato, Santa-Cruz de Benfica ou Benfica, Entrecampos, Olivais e Restelo) e uma do concelho da Amadora (Alfragide).

As estações de tráfego localizadas em Entrecampos, Benfica e Avenida da Liberdade medem as concentrações de poluentes resultantes das emissões rodoviárias. Contrariamente, as estações de fundo em Olivais, Beato, Restelo e Alfragide avaliam a exposição média da população a concentrações de fundo.

A Tabela 2.1 apresenta as principais caraterísticas das sete EM.

A Figura 2.1, obtida através do *Google Earth Pro*, mostra a localização geográfica das EM.

Ao analisar detalhadamente os dados disponíveis no site da QualAr, foi possível perceber que, para os diversos poluentes monitorizados nas sete EM, encontravam-se em falta muitas observações. Devido a isto, a escolha do poluente a usar no estudo só foi realizada após avaliação da informação disponível, tendo-se decidido utilizar as medições de NO_2. Este era o único poluente atmosférico que possuía mais observações e que, poderia assim, conduzir a resultados mais fiáveis.

Para este estudo foram recolhidos dados de 2016 a 2019, perfazendo um total de $n = 35064$ observações. O agrupamento dos dados dos quatro anos consecutivos deveu-se, à forte probabilidade dos resultados e conclusões anuais serem semelhantes entre si, pois as

Tabela 2.1: EMQAr

Concelho	**Estação**	**Tipo de estação**	**Tipo de área**	**Coordenadas**	
				Latitude	Longitude
Amadora	Alfragide	Urbana	Fundo	38.738889	-9.207500
Lisboa	AvLiberdade	Urbana	Tráfego	38.721149	-9.146152
	Beato	Urbana	Fundo	38.733686	-9.114497
	Benfica	Urbana	Tráfego	38.748787	-9.201764
	Entrecampos	Urbana	Tráfego	38.748567	-9.149012
	Olivais	Urbana	Fundo	38.769783	-9.107292
	Restelo	Urbana	Tráfego	38.705738	-9.209461

Figura 2.1: Localização das sete EMQAr.

concentrações do poluente não diferem, significativamente, de ano para ano.

Porém, só foram consideradas as observações quando cada concentração de massa média horária estava disponível para todas as EM ao mesmo tempo. Desta forma, quando pelo menos uma das EM não apresentava a concentração do poluente em um determinado horário do ano, foram eliminados todos os respetivos dados das outras EM. Assim, o estudo foi somente realizado com $n = 18380$ observações e $p = 7$ variáveis.

As medições foram efetuadas de forma contínua, registando-se a concentração mássica média horária, em $\mu g/m^3$.

2.2 *Software* estatístico utilizado na análise de dados

R é uma linguagem de programação orientada para o manuseamento, análise e visualização de dados. O *software* foi criado por John Chambers e fornece uma ampla variedade de técnicas estatísticas e gráficas (R, 2023).

O conjunto integrado de recursos do *software* estatístico R inclui:

- Um tratamento eficaz de dados e o seu respetivo armazenamento;
- Um conjunto de operadores para o cálculo de matrizes;
- Uma coleção ampla, coerente e integrada de ferramentas para análise de dados;
- Facilidades na produção gráfica para análise de dados;
- Uma linguagem de programação bem desenvolvida, simples e eficaz.

Um dos pontos fortes do R é a facilidade na produção dos gráficos, incluindo, quando necessário, símbolos matemáticos e fórmulas. Contudo, deve-se ter cuidado perante a escolha do *design* para gráficos de pequenas dimensões, embora o usuário mantenha o controlo total na produção dos mesmos.

Ao contrário de outros *softwares* de análise de dados, o R contém ferramentas flexíveis e podem ser adicionadas, pelos usuários, novas funcionalidades através de novas funções e pacotes. São fornecidos com a distribuição de base cerca de oito pacotes no entanto, estão disponíveis muitos mais em vários repositórios.

O R tem um formato de documentação semelhante ao LaTeX, que é usado para fornecer documentação, tanto *online* quanto em cópia impressa.

Este *software* está disponível para as plataformas UNIX e sistemas similares (incluindo FreeBSD e Linux), Windows e MacOS.

Toda a análise estatística da dissertação foi realizada utilizando a versão R 4.2.3.

Inicialmente, foi obtido um resumo estatístico para um conjunto de variáveis contínuas, usando a função *basicStats* incluída no pacote *fBasics*.

Relativamente à análise de componentes principais (ACP), a adequação dos dados à utilização desta técnica foi validada pelo teste de Kaiser-Meyer-Olkin, com a função *KMO*, e pelo teste de esfericidade de *Bartlett*, com a função *cortest.bartlett*, ambos incluídos no pacote *psych*. Para esta técnica foi utilizada a função *principal* e o método *varimax*, também inclusos no pacote *psych*. Já, os *scree plots* de *Cattell* foram obtidos usando a função *screeplot*.

Para a análise de *clusters* (AC), a matriz de dissimilaridade (matriz de distância) foi calculada utilizando a função *dist* e a abordagem hierárquica aglomerativa, com o método de ligação média, foi implementada utilizando a função *hclust*, ambas incluídas no pacote *stats*. A atribuição de EM a cada *cluster*, usando a distância euclidiana e o método hierárquico aglomerativo, com ligação média, foi obtida por meio da função *cutree*, que também está incluída no pacote *stats*. Os dendrogramas foram obtidos usando a função *plot*. Os coeficientes de correlação cofenética de Pearson foram obtidos por meio da função *cor*, juntamente com a função *cophenetic*. Ainda nesta técnica de estatística multivariada, para o método de partição não hierárquico *K-means*, a atribuição de EM em cada *cluster* foi obtida usando a função *kmeans*. A representação gráfica dos *clusters* foi obtida através da função *fviz_ cluster*, incluída no pacote *factoextra*, juntamente com o pacote *ggplot2*. O índice para avaliar a qualidade dos *clusters* foi obtido por meio da função *eclust*, também incluída no pacote *factoextra*.

Na análise de regressão linear, as estimativas dos coeficientes da reta de regressão e os valores de R^2 e R^2_{ajust} foram obtidas com as funções *summary* e *lm*. A regressão *stepwise* e o valor de *Akaike Information Criterion* (AIC) foi realizada com a função *step*.

2.3 Conceitos e métodos estatísticos

Dada a importância do tema desta dissertação, vários investigadores têm-se dedicado ao estudo da qualidade do ar através do uso de diversos métodos e técnicas. A estatística descritiva, a correlação linear, a regressão linear simples e múltipla, a análise de componentes principais e a análise de *clusters* são algumas das técnicas estatísticas mais utilizadas para prosseguir com este tipo de trabalho.

Todas as ciências possuem uma linguagem própria, e o mesmo acontece em Estatística. Antes de se efetuar o enquadramento da metodologia é necessário conhecer o significado de certos termos como, população, amostra e variável estatística.

A população é um conjunto de elementos com uma ou mais caraterísticas em comum e com interesse para o estudo. As populações, consoante o número de elementos que as compõem, podem ser finitas ou infinitas. Contudo, para realizar um estudo estatístico é muitas vezes usada somente uma parte da população, isto é, uma amostra.

Uma amostra corresponde a um subconjunto finito da população obtida através de uma amostragem aleatória e deve ser representativa da população, ou seja, deve conter em proporção tudo o que a população possui. Para além disso, a sua formação deve ser efetuada com alguma precaução uma vez que, a dimensão tem de ser suficiente grande para que as caraterísticas se aproximem, o mais possível das caraterísticas da população e que, todos os elementos da população possam ter a mesma oportunidade de fazer parte da amostra.

Com o intuito de complementar esta informação a Tabela 2.2 diferencia alguns conceitos.

Tabela 2.2: Parâmetro, estimadores e estimativas

Parâmetros	**Estatísticas/Estimadores**	**Estimativas**
Caraterística da população. É um valor fixo, usualmente desconhecido.	Caraterística da amostra aleatória. É uma variável aleatória.	Valor assumido pelo estimador para uma amostra em particular.
Média populacional (μ)	Média amostral aleatória $\left(\overline{X}\right)$	Média amostral observada $(\overline{x})$
Variância populacional $\left(\sigma^2\right)$	Variância amostral aleatória $\left(S^2\right)$	Variância amostral observada $\left(s^2\right)$
Desvio-padrão populacional (σ)	Desvio-padrão amostral aleatório (S)	Desvio-padrão amostral observado (s)

As variáveis estatísticas são caraterísticas comuns aos elementos da população em estudo e podem ser classificadas como qualitativas, se estiverem relacionadas com uma qualidade, expressas em categorias, ou como quantitativas, se forem expressas em valores numéricos e atribuídas unidades de medida. Estas últimas podem, ainda, dividir-se em variáveis estatísticas discretas, assumindo valores isolados ou variáveis estatísticas contínuas, podendo tomar qualquer valor dentro de um intervalo ou reunião de intervalos de números reais.

Para realizar um estudo estatístico descritivo seguem-se os seguintes passos:

1. Definição do problema a ser estudado (variável);

2. Planificação do processo para resolver o problema, decidindo-se a utilização da população ou da amostra;
3. Recolha de dados através de questionários, observações, experimentação e/ou pesquisa bibliográfica;
4. Organização dos dados em tabelas ou gráficos;
5. Análise e interpretação dos dados.

A estatística incorpora a estatística descritiva e a inferência estatística, sendo que esta última é composta pela estimação pontual (por exemplo, da média, variância e desvio padrão amostral), pela estimação por intervalos de confiança para os parâmetros e pelos testes de hipóteses. Para efetuar um estudo estatístico é necessário recorrer a distribuições amostrais:

- Distribuição normal reduzida ($Z \sim N(0,1)$);
- Distribuição Qui-quadrado $\left(X \sim \chi^2(n)\right)$;
- Distribuição *t-Student* ($T \sim t(n)$);
- Distribuição *F-Snedecor* ($X \sim F(n_1, n_2)$).

Qualquer uma destas distribuições, depende do tipo de população, da dimensão da amostra e do conhecimento ou não da variância populacional (no caso de se realizar intervalos de confiança e/ou testes de hipóteses com o valor médio populacional).

2.3.1 Estatística descritiva de dados multivariados

A estatística descritiva descreve as propriedades relativas a um conjunto de dados. Os métodos descritivos como, as medidas de localização de tendência central e de tendência não central, as medidas de dispersão, as medidas de assimetria e as medidas de achatamento/curtose permitem sintetizar a diversidade das informações contidas nesses mesmos dados.

Neste estudo, algumas medidas não serão usadas nomeadamente, a mediana, a moda e as medidas de assimetria e curtose.

A análise descritiva será efetuada com dados multivariados existindo um conjunto de n observações em p variáveis observadas. O conjunto das $n \times p$ observações dá origem a uma matriz de dados observados.

	Variável 1	Variável 2	$\cdots$	Variável j	$\cdots$	Variável p
Indivíduo 1	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1p}
Indivíduo 2	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Indivíduo i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Indivíduo n	x_{n1}	x_{n2}	$\cdots$	x_{nj}	$\cdots$	x_{np}

Esta matriz tem dimensão $n \times p$, sendo X_1, X_2, ..., X_p as variáveis iniciais, e cada uma das n unidades representam os indivíduos, tratamentos ou observações. A notação x_{ij} designa o valor da variável j para o indivíduo i, com $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, p$.

Assim, a matriz X com n linhas e p colunas será:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}.$$

As medidas de localização de tendência central são a média, a mediana e a moda. A posição relativa destas medidas fornece, em geral, informações sobre a curva da distribuição, podendo esta ser simétrica (quando os valores da média, mediana e moda são iguais), assimétrica positiva (quando o valor da média é superior ao valor da mediana) ou assimétrica negativa (quando a média tem um valor inferior ao da mediana) em relação a um eixo.

A média amostral da variável j ($\overline{x}_j$) corresponde à soma de todos os valores da variável (x_{ij}) a dividir pelo número total de observações (n) e vem expressa na mesma unidade de medida da variável.

$$\overline{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n},$$

com $j = 1, 2, \ldots p$.

As medidas de localização de tendência não central são os quantis e dividem, por exemplo, os dados em quatro (quartis), em dez (decis) ou em cem partes iguais (percentis). Nestes casos existem no total, três quartis $\left(\frac{1}{4}, \frac{2}{4}, \frac{3}{4}\right)$, nove decis $\left(\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}\right)$ e noventa e nove percentis $\left(\frac{1}{100}, \frac{2}{100}, \ldots, \frac{99}{100}\right)$, podendo ser determinados através da seguinte expressão:

$$Q_p = (1 - k)x_{(i)} + kx_{(i+1)},$$

com

$$i = \lfloor np + 1 - p \rfloor$$

e

$$k = np + 1 - p - i.$$

Relativamente aos quartis, o primeiro $\left(Q_{\frac{1}{4}}\right)$ fica à esquerda da mediana e limita 25% das observações de menor valor. O segundo quartil $\left(Q_{\frac{2}{4}} = Q_{\frac{1}{2}}\right)$ equivale à mediana logo, limita 50% das observações de menor valor e, à direita da mediana encontra-se o terceiro quartil $\left(Q_{\frac{3}{4}}\right)$, limitando 75% das observações de menor valor. A amostra tem que estar ordenada para obter os quartis.

Com base nos quartis, é possível obter os valores que apresentam um grande afastamento das restantes observações, designados de *outliers* moderados ou severos/extremos. Após a identificação dos *outliers* deve ser investigada a origem do seu afastamento. Para obtê-los

calcula-se, primeiramente, o intervalo-interquartis (IQ), utilizando os valores do primeiro e terceiro quartis:

$$IQ = Q_{\frac{3}{4}} - Q_{\frac{1}{4}}.$$

São considerados *outliers* moderados, os valores observados que pertencem a um destes intervalos:

$$Q_{\frac{1}{4}} - 3IQ \leq x_i < Q_{\frac{1}{4}} - 1,5IQ$$

ou

$$Q_{\frac{3}{4}} + 1,5IQ < x_i \leq Q_{\frac{3}{4}} + 3IQ$$

e *outliers* severos, os valores observados que pertencem a um destes intervalos

$$x_i < Q_{\frac{1}{4}} - 3IQ$$

ou

$$x_i > Q_{\frac{3}{4}} + 3IQ.$$

O diagrama de extremos e quartis, também denominado de *boxplot*, representa os quartis, indicando a presença ou não de *outliers* e a simetria relativamente à distribuição dos valores observados.

Na Figura 2.2 apresenta-se um *boxplot* onde, m é o menor dos valores observados que não é considerado *outlier*, M é o maior dos valores observados que não é considerado *outlier*, $\circ$ e $*$ representam, respetivamente, *outliers* moderados e *outliers* extremos ou severos.

Figura 2.2: Diagrama de extremos e quartis.

Embora as medidas de localização forneçam informações importantes, não permitem compreender a variabilidade dentro de um conjunto de dados. Assim sendo, as medidas descritivas de dispersão ou de variabilidade descrevem a forma como os diversos valores de uma variável estatística se distribuem em redor dos valores centrais.

A variância $\left(s_j^2\right)$ é uma medida de dispersão, sendo expressa na unidade de medida da variável j ao quadrado. Por sua vez, o desvio-padrão (s_j), expressa-se na unidade de medida da variável j e indica a variabilidade média dos dados relativamente à média. Em conjuntos de dados com pequena dispersão, os valores encontram-se mais próximos da média, resultando num valor de variância e desvio-padrão mais reduzido. Contrariamente, dados com maior dispersão possuem uma maior variância e desvio-padrão.

A variância é dada por:

$$s_j^2 = s_{jj} = \frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}{n-1} = \frac{\sum_{i=1}^{n} x_{ij}^2 - n\overline{x}_j^2}{n-1}$$

e o desvio-padrão é dado por:

$$s_j = \sqrt{s_j^2} = \sqrt{s_{jj}},$$

com $j = 1, 2, \ldots, p$.

Uma outra medida de dispersão de natureza relativa é o coeficiente de variação (cv_j). Este mede o grau de concentração dos valores em torno da média, sendo que quanto menor o seu valor, mais representativa dos dados se torna a média. Se o valor deste coeficiente for inferior ou igual a 15%, os dados apresentam uma fraca variabilidade. Se o valor varia entre 15% e 30% os dados apresentam uma variabilidade média. Mas, caso seja igual ou superior a 30% os dados exibem uma variabilidade mais elevada. Este coeficiente é dado por:

$$cv_j = \frac{s_j}{\mid \overline{x}_j \mid} \times 100\%,$$

com $j = 1, 2, \ldots, p$.

Contudo, quando se verifica a existência de *outliers* torna-se mais adequado determinar o coeficiente de variação resistente. A sua interpretação é exatamente a mesma que é feita para o coeficiente de variação. Este coeficiente é determinado por:

$$cv_r = \frac{IQ}{Q_{\frac{1}{2}}} \times 100\%$$

Esta é também uma medida de dispersão de natureza relativa, ou seja, independente das unidades de medida dos dados e é particularmente útil quando pretendemos comparar a dispersão de duas distribuições em que as variáveis não estão expressas na mesma unidade e/ou que têm localizações centrais muito diferentes.

Uma medida da associação linear entre as observações das variáveis j e k é a covariância:

$$s_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)}{n-1} = \frac{\sum_{i=1}^{n} x_{ij}x_{ik} - n \times \overline{x}_j \times \overline{x}_k}{n-1},$$

com $j = 1, 2, \ldots, p$ e $k = 1, 2, \ldots, p$ e $j \neq k$.

Se a covariância entre as variáveis é positiva, as variáveis relacionam-se de forma direta, ou seja, aumentam ou diminuem conjuntamente. Se a covariância entre as variáveis é negativa, as variáveis relacionam-se de forma inversa, ou seja, quando uma aumenta a outra diminui. Caso não exista nenhuma associação entre os valores das duas variáveis, a covariância é zero. Para além disso, observa-se que, quando $j = k$, a covariância corresponde à variância e ainda, $s_{jk} = s_{kj}$ para todo o j e k. O valor da covariância depende das unidades de medida das variáveis.

É possível averiguar o grau de associação linear entre as variáveis através do coeficiente de correlação (r_{jk}) que é determinado pela seguinte expressão:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2}},$$

para $j = 1, 2, \ldots, p$ e $k = 1, 2, \ldots, p$ e $j \neq k$.

Observa-se assim que, quando $j = k$ o coeficiente assume o valor 1. O sinal da covariância e do coeficiente de correlação amostral é o mesmo. Ao contrário da medida descritiva anterior, o valor deste coeficiente não depende das unidades de medida das variáveis.

Este coeficiente pode variar entre (-1) e (+1), inclusive, sendo que, caso o coeficiente de correlação tenha um valor próximo destes valores está-se perante uma correlação muito forte. Caso assuma um dos valores mencionados, a correlação é, respetivamente, linear positiva perfeita ou linear negativa perfeita. Se r_{jk} for negativo existe uma associação linear negativa, ou seja, as variáveis tendem a variar em sentidos opostos, enquanto que se r_{jk} for positivo existe uma associação linear positiva, ou seja, as variáveis tendem a variar no mesmo sentido. Por fim, se, o valor for estiver próximo de zero indica que a associação linear é fraca, sendo que, se for nulo não existe correlação linear. Todavia, uma correlação só é considerada aceitável se $\mid r_{jk} \mid \geq 0.7$.

A Tabela 2.3 indica os diferentes graus de associação linear entre as variáveis.

Tabela 2.3: Grau de associação linear entre as variáveis

Intervalo de valores	**Grau de associação linear**
$\mid r_{jk} \mid \in [0.7; 0.85[$	Razoável
$\mid r_{jk} \mid \in [0.85; 0.95[$	Forte
$\mid r_{jk} \mid \in [0.95; 0.9999]$	Muito Forte
$\mid r_{jk} \mid = 1$	Perfeita

Algumas das medidas supramencionadas podem ser escritas em forma de matriz:

- Vetor das médias:

$$\overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix};$$

- Matriz de variâncias covariâncias:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix};$$

- Matriz de correlações:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}.$$

2.3.2 Regressão linear

2.3.2.1 Regressão linear simples

A regressão linear simples permite estabelecer relações estatísticas entre a variável dependente (Y) e a variável independente (x). Esta metodologia é usada para prever o comportamento da variável Y através da utilização de um modelo matemático que se ajuste aos dados amostrais das variáveis – reta dos mínimos quadrados.

Para esta regressão são apenas usadas duas variáveis, pelo que a correlação entre elas é considerada simples.

Diz-se que existe correlação entre as variáveis se entre elas existir uma relação estatística e se a intensidade de um fenómeno associado à primeira variável é acompanhada, tendencialmente, pela intensidade do outro.

Após a recolha e análise dos dados amostrais das variáveis, marcam-se os pontos (x, y) num diagrama de dispersão. Este diagrama serve para avaliar a linearidade e a correlação entre as variáveis, podendo esta ser positiva ou negativa.

A correlação procura determinar o grau de relação entre as variáveis enquanto que, a regressão linear procura determinar uma expressão matemática que descreva a relação entre essas variáveis.

Tal como se fez na subsecção anterior, na regressão linear também é usada a covariância linear amostral ($cov[x, y]$) e o coeficiente de correlação linear amostral (r). A interpretação de ambas as medidas efetua-se da mesma maneira, ou seja, se o valor da covariância for positivo existe uma relação linear positiva e, se o valor for inferior a zero significa que, existe uma relação linear negativa entre as variáveis. Relativamente à correlação, os valores podem variar entre (-1) e (+1), inclusive, sendo que, caso o coeficiente de correlação linear amostral tenha um valor próximo destes valores está-se perante uma correlação linear muito forte.

A covariância linear amostral é dada por:

$$cov[x, y] = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n \times \overline{x} \times \overline{y}}{n-1} = \frac{s_{xy}}{n-1},$$

designando-se

$$s_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y}$$

por soma cruzada.

O coeficiente de correlação linear amostral é dado por:

$$r = \frac{cov(x, y)}{s_x \times s_y} = \frac{s_{xy}}{\sqrt{s_{xx} \times s_{yy}}},$$

onde s_x e s_y são os desvios-padrão das variáveis, e

$$s_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = (n-1)\, s_x^2$$

e

$$s_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = (n-1)\, s_y^2$$

representam as somas de quadrados de x e de y, respetivamente.

A regressão linear simples pode ser descrita através da seguinte equação:

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

sendo Y a variável explicada ou dependente, x a variável explicativa ou independente, β_0 e β_1 são os parâmetros da reta a ajustar, em que β_0 representa a interseção da reta ao eixo vertical e β_1 o declive da reta, e ε representa uma variável do tipo residual, que inclui outros fatores explicativos de Y e ainda possíveis erros de medição.

Esta expressão descreve uma reta que, quando ajustada aos dados do diagrama usando o método dos mínimos quadrados, se designa por reta de regressão ou reta ajustada, passando a ter seguinte expressão:

$$\hat{y} = a + bx,$$

onde a e b correspondem às estimativas de β_0 e β_1, respetivamente.

De forma a medir a qualidade do ajustamento feito pelo método dos mínimos quadrados, utiliza-se o coeficiente de determinação $\left(r^2\right)$ que, determina a percentagem de variabilidade da variável Y que é explicada à custa da variável x.

Este coeficiente tem um valor a variar entre 0 ou 1. Quando tem o valor 1, todas as observações estão sob a reta de regressão e quando tem valor 0, o modelo não tem utilidade na explicação da variabilidade da variável Y.

Este pode ser calculado através do quadrado do coeficiente de correlação ou através da seguinte expressão:

$$r^2 = \frac{s_{xy}^2}{s_{xx} \times s_{yy}}.$$

2.3.2.2 Regressão linear múltipla

Na regressão linear múltipla são utilizadas mais de duas variáveis. As k variáveis independentes são utilizadas para explicar a variação de Y, variável dependente.

As condições da regressão linear múltipla são análogas às da regressão linear simples pois, para cada conjunto de valores x_j há uma subpopulação de valores de Y estatisticamente independentes com, distribuição normal e variâncias iguais.

Esta regressão pode ser escrita com a equação:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

sendo o modelo ajustado escrito da seguinte forma:

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik},$$

com $i = 1, 2, \ldots, n$, a variável dependente ou resposta, Y, e as variáveis independentes ou preditoras $x_1, x_2, \ldots, x_k$.

À semelhança da regressão linear simples, o coeficiente de determinação $\left(R^2\right)$ também varia entre 0 e 1. Contudo, na regressão linear múltipla, um elevado valor deste coeficiente não implica necessariamente que o modelo seja um bom ajustamento, pois a adição de uma variável, quer seja significativa ou não, aumenta sempre esse valor. Assim, determinados modelos podem ser pouco fiáveis embora tenham um valor elevado de R^2. Por esse motivo, e de forma a obter uma melhor ideia da qualidade do modelo, utiliza-se o coeficiente de determinação ajustado $\left(R^2_{ajust}\right)$, que tem em conta também o número de variáveis do modelo e só aumenta se houver vantagem na adição de uma nova variável. Este coeficiente é dado por:

$$R^2_{ajust} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2),$$

onde $p = k + 1$ e k corresponde ao número de variáveis independentes.

Para se ter uma ideia da percentagem de contribuição de cada variável x recorre-se aos coeficientes de determinação parciais. Inicialmente, é efetuada a regressão linear simples entre a variável dependente e as diferentes variáveis independentes. De seguida, os valores de R^2 e

R^2_{ajust} são calculados e comparados, podendo assim tirar ilações sobre a possível exclusão de determinada variável independente da equação.

A seleção de variáveis independentes para serem usadas no modelo de regressão é extremamente importante, uma vez que nem todas são necessárias para modelar a variável dependente. Por um lado, o modelo deve conter número suficiente de variáveis independentes, devendo excluir-se as variáveis inter-correlacionadas entre si mas, por outro lado, o modelo torna-se mais acessível se possuir menos variáveis. As técnicas de seleção de variáveis mais usuais são:

- A técnica para escolher o "melhor" modelo de regressão através do R^2_{ajust} que consiste em verificar para que variáveis o R^2_{ajust} apresenta um maior valor. Muitas vezes, o valor do coeficiente começa a diminuir à medida que o número de variáveis independentes no modelo aumenta.
- A segunda técnica designada de regressão *stepwise*, consiste num procedimento que constrói uma sequência de modelos de regressão, adicionando ou removendo variáveis em cada passo. O *AIC* penaliza o modelo por ter demasiadas variáveis. Caso o *AIC* aumente, o ajustamento piora e, se diminuir, o ajustamento melhora. O *AIC* é dado por:

  $$AIC = n \ln\left(\frac{SQE}{n}\right) + 2k,$$

  sendo a soma de quadrados do erros dada por:

  $$SQE = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{Y}.$$

  As matrizes desta última fórmula são dadas por:

  $$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

  $$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

  e

  $$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}.$$

  A regressão *stepwise* pode ser feita em 3 direções: *Forward, Backward ou Both*. Após análise, escolhe-se, a direção onde se encontra o menor valor de AIC e escreve-se a equação de regressão com os dados presentes no *output*:

  – Na direção *Forward*, o modelo inicial contém apenas a constante $\hat{\beta}_0$. O procedimento procura e retêm as variáveis que diminuem o valor AIC sendo que, o processo só pára quando nenhuma variável independente conseguir diminuir mais esse valor.

– Na direção *Backward*, o modelo contém todas as variáveis independentes a que está associado com um determinado valor AIC. À medida que se retira uma variável independente verifica-se se o valor AIC decresce.
– A direção *Both* conjuga ambas as direções precedentes pois, cada vez que uma variável independente é adicionada ao modelo, é realizado um teste de remoção da variável independente menos útil.

2.3.3 Testes de hipóteses paramétricos e não paramétricos

Os testes de hipóteses são compostos pela hipótese nula (H_0) e pela hipótese alternativa (H_1) que é, normalmente, a negação de H_0.

Para retirar uma conclusão destes testes recorre-se a regra de decisão estatística, ou seja, rejeitar ou não rejeitar H_0. Existem testes de hipóteses não paramétricos e testes de hipóteses paramétricos. Os **testes de hipóteses não paramétricos** podem ser usados para testar a hipótese de que uma distribuição de frequências observadas se ajusta (ou adere) a uma determinada distribuição amostral. O testes de aderência ou de qualidade de ajuste que se decidiu usar neste trabalho são os testes de *Kolmogorov-Smirnov* com correção de *Lilliefors* (KSL) e de *Shapiro-Wilk*. É importante mencionar que este último teste deve ser utilizado quando a dimensão da amostra é reduzida inclusive, o *software R* só permite a utilização quando existem 3000 a 5000 valores observados. Os **testes de hipóteses paramétricos** permitem tomar decisões acerca dos parâmetros da população a partir dos valores observados em amostras. Estes incluem testes de homocedasticidade (teste de Levene e teste de Bartlett) e testes para parâmetros da população.

Quando se pretende efetuar testes de hipóteses, é importante verificar os pressupostos do modelo, nomeadamente, os erros ou os resíduos serem independentes e normalmente distribuídos, com valor esperado nulo e variância constante, quando usados na regressão linear simples ou múltipla.

Esta verificação é feita através da análise de resíduos ou erros observados que são obtidos calculando as diferenças entre os valores observados (y_i) e os valores estimados pela reta ajustada ($\hat{y}_i$):

$$e_i = y_i - \hat{y}_i.$$

O estudo da normalidade dos resíduos pode ser feito com recurso, ao **teste de *Kolmogorov-Smirnov*** com correção de ***Lilliefors*** ou ao **teste de *Shapiro-Wilk***.

Nestes testes o primeiro passo é formular as hipóteses e estabelecer nível de significância (α) e o tamanho da amostra (n). As hipóteses são:

- H_0: A amostra é proveniente de uma população com distribuição normal;
- H_1: A amostra é proveniente de uma população com distribuição diferente da distribuição normal.

Antes de validar a normalidade dos resíduos com o teste de ajustamento pode-se observar esta suspeita no gráfico *Q-Q Plot* ou gráfico quantil-quantil. Este gráfico consiste apenas em uma avaliação meramente visual, pelo que se torna numa avaliação subjetiva, não servindo como prova. Assim, um *Q-Q Plot* é um gráfico de dispersão obtido pela representação de pares ordenados dos quantis de duas distribuições, sendo que se os dois conjuntos de quantis forem provenientes da mesma distribuição, o conjunto de pontos obtido formará uma linha que é aproximadamente reta.

De seguida, determina-se o valor da estatística de teste e o valor crítico tabelado, permitindo assim tomar uma decisão. Para tal, no *software* R pode-se encontrar informações como, o

valor de $D_{observado}$ no caso do teste de *Lilliefors-Kolmogorov-Smirnov*, $W_{observado}$ no caso do teste de *Shapiro-Wilk*, e $p - value$.

Se ocorrer alguma das seguintes situações deve-se rejeitar H_0:

- Se $D_{observado} \geq D_{critico,\alpha}$ (valor tabelado);
- Se $W_{observado} \leq W_{critico,\alpha}$ (valor tabelado).

Também se deve rejeitar H_0 se $\alpha \geq p - value$. Neste caso, concluí-se, para um dado α, que os resíduos são provenientes de uma população com distribuição diferente da distribuição normal.

Nos **testes paramétricos** a hipótese nula é uma afirmação sobre o valor de um parâmetro populacional e contém uma das seguintes condição de igualdade:

$$H_0 : \theta = \theta_0 \qquad \text{ou} \qquad H_0 : \theta \leq \theta_0 \qquad \text{ou} \qquad H_0 : \theta \geq \theta_0.$$

Para além disso, a hipótese alternativa condiciona o tipo de teste a realizar, podendo este pode ser bilateral, unilateral à esquerda ou unilateral à direita. Esta hipótese comporta uma destas três formas:

$$H_1 : \theta \neq \theta_0 \qquad \text{ou} \qquad H_1 : \theta < \theta_0 \qquad \text{ou} \qquad H_1 : \theta > \theta_0.$$

Para testar um dado valor de um parâmetro, as etapas são ligeiramente diferentes. Os primeiros passos passam por identificar o parâmetro a testar, especificar as hipóteses, identificar o tipo de população e estabelecer α e n. Logo de seguida, identifica-se a estatística de teste e a respetiva distribuição amostral através do quadro resumo. Por fim, para se tomar uma decisão, determinam-se as regiões crítica e de aceitação e o valor da estatística de teste. Desta forma, deve-se rejeitar H_0 quando o valor da estatística de teste pertence à região crítica ou de rejeição, e não se deve rejeitar H_0 quando o valor da estatística de teste pertence à região de aceitação.

Pode-se **testar a significância do modelo de regressão linear simples**, ou seja, se efetivamente existe uma relação linear entre as variáveis em estudo, seguindo-se as etapas de um teste paramétrico com as seguintes hipóteses:

- H_0: $\beta_1 = 0$
- H_1: $\beta_1 \neq 0$

Usa-se a estatística de teste:

$$T_0 = \frac{B - (\beta_1)_0}{\frac{S_{y.x}}{\sqrt{S_{xx}}}} \sim t_{n-2}$$

e a distribuição amostral associada é a *t-Student*.

Quando H_0 não é rejeitada diz-se que não existe uma relação linear entre as variáveis x e Y, ou seja, o modelo de regressão ajustado não é significativo e não deve ser utilizado.

Tal como mencionado no início desta subseção, na regressão linear múltipla, os testes requerem que os ε_i sejam independentes e normalmente distribuídos com uma média próxima ou igual a zero e desvio-padrão σ. A inferência estatística relativa à normalidade dos resíduos e aos parâmetros é realizada de forma análoga na regressão linear simples.

Com um dado nível de significância α, as hipóteses para testar a significância do modelo de regressão linear múltipla são:

- H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

- H_1: $\beta_j \neq 0$ para algum j com $j = 1, 2, \ldots, k$

A estatística de teste é dada por:

$$F_0 = \frac{\frac{SQR}{k}}{\frac{SQE}{n-p}} = \frac{QMR}{QME} \sim F(k; n-p),$$

com k o número de variáveis independentes e $p = k + 1$, a soma de quadrados da regressão dada por:

$$SQR = \widehat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{Y} - n\overline{y}^2,$$

e a soma de quadrados do erros dada por:

$$SQE = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{Y}$$

sendo a distribuição amostral associada a distribuição *F-Snedecor*. Tem-se, ainda, a soma de quadrados total dada por:

$$SQT = \mathbf{Y}'\mathbf{Y} - n\overline{y}^2,$$

com $SQT = SQR + SQE$.

Para que, pelo menos, uma variável x contribua para explicar a variação da variável Y e o modelo seja considerado significativo, deve-se rejeitar H_0 e isso acontece se:

- $F_0 \geq F\left(k, n-p, 1-\alpha\right)$ ou
- $\alpha \geq p - value$.

A partir deste teste não é possível ter conhecimento da importância de cada variável x_j no modelo e, por esse motivo, pode ser testada a significância dos vários coeficientes de regressão.

- H_0: $\beta_j = 0$
- H_1: $\beta_j \neq 0$ para um j, com $j = 1, 2, \ldots, k$

Usa-se a estatística de teste:

$$T_0 = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \sim t_{n-p},$$

sendo $se(\widehat{\beta}_j) = \sqrt{\hat{\sigma}^2 c_{jj}}$ o erro padrão estimado de $\widehat{\beta}_j$, com c_{jj} o j-ésimo elemento da diagonal principal da matriz

$$\mathbf{C} = \left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

correspondente a $\widehat{\beta}_j$. A distribuição amostral associada é a *t-Student*.

O parâmetro a testar diz-se significativo, ou seja, a variável x_j contribuí para explicar a variável Y se se rejeitar H_0. Isso ocorre se se verificarem as seguintes condições:

- $T_0 \leq -t_{n-p;1-\frac{\alpha}{2}}$ ou $T_0 \geq t_{n-p;1-\frac{\alpha}{2}}$ ou
- $\alpha \geq p - value$.

O **teste de homocedasticidade ou de homogeneidade das variâncias** serve para testar a hipótese de que a dispersão dos dados em torno da média é igual para todos os grupos, ou seja, se as variâncias populacionais são iguais ou se existem pelo menos duas que são diferentes.

Como primeira etapa deve-se especificar as hipóteses e estabelecer o nível de significância α e a dimensão da amostra n. As hipóteses são:

- H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$
- H_1: $\exists i, j : \sigma_i^2 \neq \sigma_j^2$

com $i \neq j$, $i = 1, 2, \ldots, k$ e $j = 1, 2, \ldots, k$.

Posteriormente, para ser possível tomar uma decisão, determina-se o valor da estatística de teste e o valor crítico tabelado. Tal como anteriormente, no *software* R encontram-se informações como, o valor de $F_{observado}$ no caso do teste de Levene, $Q_{observado}$ no caso do teste de Bartlett, e $p - value$.

Se ocorrer algumas das seguintes situações deve-se rejeitar H_0:

- Se $F_{observado} \geq F_{k-1;n-k;1-\alpha}$, sendo k o número de amostras e $n = n_1 + n_2 + ... + n_k$;
- Se $Q_{observado} \geq \chi^2_{k-1;1-\alpha}$, sendo k o número de amostras.

Também se deve rejeitar H_0 se $\alpha \geq p - value$. Assim, neste caso, concluí-se, que com um determinado nível de significância α, as variâncias populacionais das variáveis não são idênticas.

2.3.4 Análise de componentes principais

A ACP permite reduzir a dimensão dos dados e passar de um número elevado de variáveis descritivas p para um conjunto menor de variáveis k, com perda mínima de informação.

Deste modo, a ACP é um método estatístico multivariado que transforma um conjunto de variáveis iniciais correlacionadas entre si, e portanto de alguma forma redundantes, num outro conjunto de variáveis mutuamente ortogonais (não correlacionadas), designadas de componentes principais.

Todas as k variáveis (PC) juntas devem explicar a maioria da variação das p variáveis iniciais. O novo conjunto de variáveis $Y_1, Y_2, ..., Y_p$ é obtido por ordem decrescente de importância, ou seja, a primeira explica o máximo possível da variância dos dados originais, a segunda o máximo possível da variância ainda não explicada e, por assim adiante:

$$Var(Y_1) \geq Var(Y_2) \geq ... \geq Var(Y_p)$$

As componentes principais são extraídas da matriz de variâncias covariâncias (matriz S) ou da matriz de correlações (matriz R). Para determinar estas componentes é preciso calcular uma dessas matrizes, encontrar os vetores próprios que estão associados a valores próprios e, por fim, escrever as combinações lineares, que serão as novas variáveis.

Por norma, quando as unidades de medida são idênticas para todas as p variáveis originais e as variâncias não têm valores muito diferentes, recorre-se à matriz de variâncias covariâncias. Contudo, quando há uma elevada discrepância das variâncias entre as variáveis originais ou as unidades de medida destas variáveis são diferentes, usa-se a matriz de correlações. Esta matriz coincide com a matriz S para variáveis estandardizadas, em que os seus elementos são obtidos a partir das variáveis iniciais, subtraindo a média $(\overline{x}_j)$ e dividindo pelo desvio-padrão $\left(s_{x_j}\right)$:

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_{x_j}}.$$

Contudo, o procedimento matemático para a obtenção das componentes principais é realizado exatamente da mesma maneira como com a matriz de variâncias covariâncias $\mathbf{S}$ e as componentes passam a ser obtidas através de vetores próprios associados a valores próprios da matriz R.

Como mencionado, cada nova variável Y_j é uma combinação linear das p variáveis iniciais $X_1, X_2, \ldots, X_p$:

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p = \boldsymbol{a}_j'\boldsymbol{X}$$

ou quando estandardizadas

$$Y_j = a_{1j}Z_1 + a_{2j}Z_2 + ... + a_{pj}Z_p = \boldsymbol{a}_j'\boldsymbol{Z},$$

sendo

$$\boldsymbol{a}_j' = \begin{bmatrix} a_{1j} & a_{2j} & \cdots & a_{pj} \end{bmatrix}$$

um vetor de componentes tais que:

$$\boldsymbol{a}_j'\boldsymbol{a}_j = \sum_{i=1}^{p} a_{ij}^2 = 1$$

e

$$\boldsymbol{a}_j'\boldsymbol{a}_r = 0,$$

para $j \neq r$, $j = 1, 2, \ldots, p$ e $r = 1, 2, \ldots, p$.

Encontra-se, a primeira componente principal (Y_1), escolhendo o vetor de constantes ($\boldsymbol{a}_1$) de modo a maximizar a variância dessa componente.

A obtenção dos vetores próprios ($\boldsymbol{a}_j$) é feita com a determinação dos valores próprios (λ) da matriz de variâncias covariâncias ou da matriz de correlações. Como $Var[Y_j] = \lambda_1$ escolhe-se para a primeira componente o maior valor próprio obtido, isto é, λ_1. Assim, a primeira componente principal ($Y_1 = \boldsymbol{a}_1'X$), terá coeficientes correspondentes ao vetor próprio ($\boldsymbol{a}_1'$) associado ao valor próprio mais elevado (λ_1).

Seguindo este processo, encontrar-se-ão as restantes componentes, todas não correlacionadas entre si e com variância decrescente.

Torna-se ainda relevante mencionar que, a soma das variâncias das variáveis originais X_i é igual à soma das variâncias das componentes Y_i e que, por sua vez, é igual à soma de todos os valores próprios. Desta forma, concluí-se que a j-ésima componente explica:

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

da variância total original, e ainda que as primeiras k componentes explicam:

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

da variância total, sendo $u = 1, 2, \ldots, p$.

Assim, c_j representa a percentagem da variância total explicada pela componente principal Y_j:

$$c_j = \frac{\lambda_j}{\sum_{j=1}^{p}\lambda_j} \times 100\%.$$

Caso as componentes sejam calculadas a partir de variáveis estandardizadas, a soma dos valores próprios passa a ser p e assim, a variância explicada pela componente Y_j é calculada através da seguinte expressão:

$$c_j = \frac{\lambda_j}{p} \times 100\%.$$

De maneira a facilitar a aplicação desta análise segue-se a seguinte metodologia:

1. Os primeiros passos correspondem à estimação da matriz de correlações e verificação da existência de variáveis correlacionadas em número elevado. Para **averiguar a adequação dos dados à aplicação da técnica** utiliza-se o Teste de Esfericidade de Bartlett ou a Estatística de Kaiser-Meyer-Olkin (KMO).

   O teste de esfericidade de Bartlett, testa a hipótese da matriz de correlações ser uma matriz identidade e o seu determinante ser igual a 1, logo, de as variáveis não estarem correlacionadas entre si.

   A aplicação da ACP pressupõe que se rejeite H_0, existindo assim, correlações significativas entre as variáveis.

   - $H_0 : \rho = I$
   - $H_1 : \rho \neq I$

   A estatística de teste é dada por:

$$\chi_0^2 = -\left[n - 1 - \frac{1}{6}(2p+5)\right] \ln \mid \mathbf{R} \mid \sim \chi_{\frac{1}{2}p(p-1)}$$

   ou

$$\chi_0^2 = -\left[n - 1 - \frac{1}{6}(2p+5)\right] \sum_{i=1}^{p} \ln(\lambda_i) \sim \chi_{\frac{1}{2}p(p-1)}$$

   e o $p - value$ é dado por:

$$p - value = P\left[\chi^2_{\frac{1}{2}p(p-1)} \geq \chi_0^2\right].$$

   Tal como mencionado na subsecção 2.3.3, deve-se rejeitar H_0 se o valor da estatística de teste verificar a regra de decisão estatística:

$$\chi_0^2 \geq \chi^2_{\frac{p(p-1)}{2};1-\alpha}$$

   ou se

$$\alpha \geq p - value.$$

   Contudo, o outro método torna-se mais acessível uma vez que, basta determinar, com a expressão matemática abaixo, a estatística de KMO e interpretar o seu valor consoante a tabela 2.4, sendo que um valor mais próximo de 1 indica a existência de uma correlação forte entre as variáveis:

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j a_{ij}^2},$$

   sendo r_{ij} o coeficiente de correlação observado entre as variáveis i e j e os a_{ij} correspondem aos coeficientes de correlação parcial que devem de ser próximos de zero, uma vez que as componentes são ortogonais entre si.

2. Após a aplicação da ACP obtêm-se tantas componentes como variáveis. No entanto, o objetivo é selecionar apenas algumas sendo necessário, **determinar quais as componentes a excluir da análise**. Usualmente, para decidir quantas componentes se devem reter são aplicados, pelo menos, dois dos critérios apresentados abaixo:

   - Representar num gráfico *scree plot* a percentagem de variância explicada por cada componente. Quando esta percentagem diminui, a curva passa a ser praticamente paralela ao eixo das abcissas. Assim, devem-se selecionar todas as componentes até que a linha que as une comece a ficar horizontal. Para compreender este critério, a Figura 2.3 exemplifica este tipo de gráfico.

Tabela 2.4: KMO

KMO	**Adequação da ACP**
$[0.90; 1.00]$	Muito boa
$[0.80; 0.90[$	Boa
$[0.70; 0.80[$	Média
$[0.60; 0.70[$	Razoável
$[0.50; 0.60[$	Má
$[0.00; 0.50[$	Inaceitável

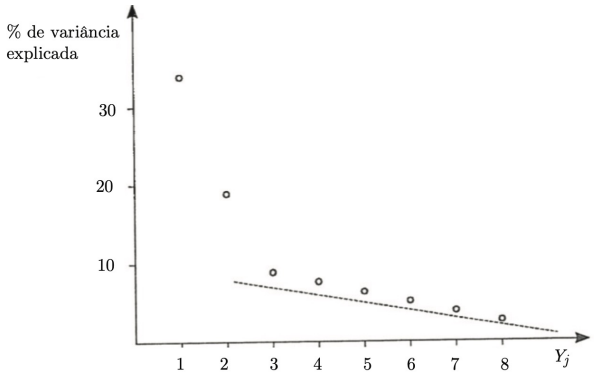

Figura 2.3: Exemplo de um *Scree plot*.

- Incluir componentes suficientes para explicar, no mínimo, 70% da variância explicada acumulada. Tanto este critério como o anterior podem ser utilizados quando a análise é realizada com a matriz S ou com a matriz R.
- Quando aplicada uma matriz R, pode-se usar o critério de Kaiser que sugere excluir as componentes cujos valores próprios são inferiores à média, ou seja, inferiores a 1.
- O critério de Bartlett somente se aplica, quando as componentes principais são derivadas de uma matriz de variâncias covariâncias. Assim sendo, devem-se reter as componentes cuja variância é significativamente diferente de 0. Com essa finalidade, testa-se a seguinte hipótese:
  - $H_0 = \lambda_{p-k} = \lambda_{p-k+1} = \cdots = \lambda_p$

    A estatística de teste é dada por:

$$\chi_0^2 = M \left[-\ln \mid \mathbf{S} \mid + \sum_{j=1}^{k} \ln \lambda_j + (p-k) \ln l \right] \sim \chi^2_{\frac{1}{2}(p-k+1)(p-k+2)},$$

com

$$M = n - k - \frac{1}{6} \left[2\,(p-k) + 1 + \frac{2}{p-k} \right]$$

e

$$l = \frac{1}{p-k}\left[\text{tr}\,(\mathbf{S}) - \sum_{j=1}^{k}\lambda_j\right]$$

e o $p - value$ é dado por:

$$p - value = P\left[\chi^2_{\frac{1}{2}(p-k+1)(p-k+2)} \geq \chi^2_0\right].$$

Deve-se rejeitar H_0 se:

$$\chi^2_{\text{observado}} \geq \chi^2_{\frac{1}{2}(p-k+1)(p-k+2)}$$

ou se

$$\alpha \geq p - value.$$

Não rejeitar H_0 significa que não existirá grande interesse em reter mais do que as k primeiras componentes principais.

3. Posteriormente, **verificam-se os pesos e as correlações entre as variáveis iniciais e as componentes principais**.

   Desta forma, é possível perceber, em cada uma das componentes, quais as variáveis iniciais que têm um maior peso, considerando o valor absoluto do coeficiente de cada variável. Assim sendo, quanto maior for o valor absoluto do coeficiente da variável na componente principal, maior o peso da variável para a componente.

   As fórmulas seguintes permitem calcular, respetivamente, as correlações ou *loadings* entre as variáveis originais e cada uma das componentes principais permitindo saber quais as variáveis originais (*standardizadas* ou não) que têm maior correlação com cada uma das componentes. A correlação entre a componente principal e a variável original é dada por:

$$r_{Y_j,X_i} = a_{ij}\frac{\sqrt{\lambda_j}}{\sqrt{s_{ii}}} = a_{ij}\frac{\sqrt{\lambda_j}}{\sqrt{s_i^2}}$$

   ou

$$r_{Y_j,Zi} = a_{ij}\sqrt{\lambda_j},$$

   no caso da matriz ser estandardizada.

4. A próxima etapa passa pela **construção da matriz de *scores*** individuais que, equivale a substituir a matriz de dados originais de dimensão $n \times p$ por uma matriz $n \times k$.

   Após a redução das variáveis, as k componentes principais serão os novos indivíduos e toda a análise será feita com base nos valores dessas componentes.

   As Tabelas 2.5 e 2.6 ilustram, respetivamente, a substituição da matriz de dados originais e estandardizados, por uma nova matriz que contém os valores das componentes principais.

   Para obter essas componentes, escrevem-se as combinações lineares das variáveis originais através do procedimento descrito abaixo. Este processo é repetido até serem geradas todas as componentes para análise.

Tabela 2.5: Exemplo tabela matriz.

Objetos (indivíduos)	Variáveis				Valores das componentes			
	X_1	X_2	$\cdots$	X_p	Y_1	Y_2	$\cdots$	Y_k
1	x_{11}	x_{12}	$\cdots$	x_{1p}	y_{11}	y_{12}	$\cdots$	y_{1k}
2	x_{21}	x_{22}	$\cdots$	x_{2p}	y_{21}	y_{22}	$\cdots$	y_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{np}	y_{n1}	y_{n2}	$\cdots$	y_{nk}

Tabela 2.6: Exemplo tabela matriz *standard*.

Objetos (indivíduos)	Variáveis				Valores das componentes			
	Z_1	Z_2	$\cdots$	Z_p	Y_1	Y_2	$\cdots$	Y_k
1	z_{11}	z_{12}	$\cdots$	z_{1p}	y_{11}	y_{12}	$\cdots$	y_{1k}
2	z_{21}	z_{22}	$\cdots$	z_{2p}	y_{21}	y_{22}	$\cdots$	y_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	z_{n1}	z_{n2}	$\cdots$	z_{np}	y_{n1}	y_{n2}	$\cdots$	y_{nk}

$$y_{11} = a_{11}x_{11} + a_{21}x_{12} + \cdots + a_{p1}x_{1p}$$

$$y_{21} = a_{11}x_{21} + a_{21}x_{22} + \cdots + a_{p1}x_{2p}$$

$$\vdots$$

$$y_{n1} = a_{11}x_{n1} + a_{21}x_{n2} + \cdots + a_{p1}x_{np}.$$

ou

$$y_{11} = a_{11}z_{11} + a_{21}z_{12} + \cdots + a_{p1}z_{1p}$$

$$y_{21} = a_{11}z_{21} + a_{21}z_{22} + \cdots + a_{p1}z_{2p}$$

$$\vdots$$

$$y_{n1} = a_{11}z_{n1} + a_{21}z_{n2} + \cdots + a_{p1}z_{np}.$$

Depois calculam-se as componentes centradas na média, subtraindo a cada *score* a média dos *scores* da respetiva componente. Assim todas as componentes têm a média dos seus *scores* nula.

5. Como último passo, procede-se à **rotação das componentes principais que não foram excluídas**, por forma a facilitar a sua interpretação.

O método mais utilizado, designado de método de rotação ortogonal de VARIMAX, foi proposto por Kaiser e pretende que, para cada componente existam apenas alguns pesos significativos e todos os outros próximos de 0. Isto tem como intuito, maximizar a variação entre os pesos de cada componente principal.

A proporção da variância explicada mantém-se constante, apenas se distribuí de forma diferente para que sejam maximizadas as diferenças entre as contribuições das variáveis: aumentando as que mais contribuem e diminuindo os pesos das que menos contribuem.

Depois de efetuada a rotação, torna-se mais simples identificar e interpretar cada componente principal a partir dos pesos das variáveis que a compõem. Assim, quanto mais perto de 1 estiver esse peso, mais forte é a associação entre essa variável e a componente.

Em geral, são considerados como significativos os pesos iguais ou superiores a 0.5 em valor absoluto.

No *software* R, os valores dos pesos entre (-0.1) e $(+0.1)$ não aparecem por defeito sendo assim, considerados aproximadamente 0.

2.3.5 Análise de *clusters*

A AC permite dividir, em grupos, um conjunto de dados relativos a p variáveis medidas em vários objetos ou indivíduos. Estes grupos, denominados de *clusters*, agrupam os objetos ou indivíduos semelhantes entre si.

A vasta diversidade de métodos para a definição dos grupos conduz a agrupamentos muito distintos, tanto em número, como em conteúdo. Porém, pretende-se que exista, sempre, uma coesão interna dos *clusters*, ou seja, os indivíduos que pertencem ao mesmo *cluster* devem ser mais semelhantes entre si, do que indivíduos de *clusters* diferentes, e um isolamento externo, ou seja, que os *clusters* estejam separados entre si.

Este processo é composto por **cinco fases**, todavia pode-se recorrer a uma etapa preliminar que consiste na visualização gráfica, através do diagrama de perfil ou do *scree plot*, do número de *clusters* a formar. O diagrama de perfil é adequado quando se tem um número moderado de variáveis, sendo que antes da análise de *clusters*, nomeadamente da construção deste diagrama, se as variáveis não possuírem a mesma unidade de medida, os dados devem ser estandardizados. No diagrama, as variáveis encontram-se no eixo das abcissas (eixo x), a escala de valores no eixo das ordenadas (eixo y) e cada ponto do gráfico representa o valor da variável correspondente. Contudo, a efetividade desta técnica é afetada quando o número de observações é muito elevado, pois a imagem fica confundida.

O *scree plot* consiste numa representação gráfica da soma de quadrados dos desvios dentro dos grupos *versus* número de *clusters*. Neste gráfico deve-se somente considerar o número de *clusters* a partir do qual a curva passa a ser quase paralela aos eixo x.

Fases para a obtenção dos *clusters*:

1. **Seleção dos indivíduos ou objetos**;
2. **Seleção das variáveis** que melhor caraterizam cada indivíduo, permitindo a sua classificação num dado grupo. Se for necessário, transformar as variáveis;
3. **Construção da medida de dissemelhança/semelhança**;

   A análise de *clusters* opera sobre dois tipos de estruturas de dados, identificados por dois formatos de matrizes. Uma matriz de dados, $X = [x_{ij}]$, que inclui variáveis qualitativas e variáveis quantitativas contínuas ou discretas e uma matriz de dissemelhanças,

$D = [d_{ij}]$, ou de semelhanças, $S = [s_{ij}]$, onde, d_{ij} representa o valor da dissemelhança e s_{ij} o valor da semelhança entre o objeto i e o objeto j.

As ideias subjacentes ao processo de construção de *clusters* são a semelhança e a dissemelhança. Intuitivamente, a dissemelhança reflete o grau de diferença, afastamento ou divergência e a semelhança mede o grau de parecença ou proximidade entre dois indivíduos.

Torna-se, assim, importante conhecer uma forma de definir dissemelhanças e semelhanças entre objetos ou entre variáveis, com base nas observações efetuadas em cada um dos objetos.

Em geral, os algoritmos utilizam matrizes de dissemelhanças, objetos por objetos ou variáveis por variáveis, e isso faz com que uma primeira tarefa da análise de *clusters* seja a construção dessa matriz, a partir de uma matriz de dados (objetos por variáveis).

A medida de dissemelhança mais utilizada para variáveis quantitativas é a distância euclidiana. Assumindo que a matriz dos dados é representada por $X = [x_{ij}]$, com $i = 1, \ldots, n$ e $j = 1, \ldots, p$, a distância euclidiana entre o indivíduo i e o indivíduo j é definida por:

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

No entanto, esta dissemelhança não é a mais adequada quando as variáveis têm unidades de medida diferentes, variâncias muito diferentes ou são correlacionadas, uma vez que nestes casos, as variáveis intervêm com pesos diferentes na determinação das dissemelhanças. Além disso, a distância euclidiana é sensível a mudanças de escala, porque isso pode implicar a mudança das distâncias e a sua ordem, afetando, consequentemente, o resultado da análise de *clusters*. Para ultrapassar este inconveniente utiliza-se a distância euclidiana estandardizada:

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

onde

$$z_{rk} = \frac{x_{rk} - \overline{x}_{\cdot k}}{s_k},$$

com $r = 1, \ldots, n$. As estimativas do valor médio da variável k são obtidas fazendo:

$$\overline{x}_{\cdot k} = \frac{\sum_{r=1}^{n} x_{rk}}{n}$$

e a estimativa do desvio padrão é dado por:

$$s_k = \left[\frac{\sum_{r=1}^{n} (x_{rk} - \overline{x}_{\cdot k})^2}{n-1} \right]^{\frac{1}{2}}.$$

Substituindo z_{ik} e z_{jk} pelas novas expressões obtém-se:

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

4. **Escolha do método a aplicar aos dados**;

Nesta etapa utilizam-se métodos hierárquicos e/ou não hierárquicos.

Nos **modelos hierárquicos**, os grupos formam uma hierarquia caraterizada pelo facto de dois grupos serem disjuntos ou um deles estar contido no outro. Para tal recorre-se a:

- **Algoritmos aglomerativos ou ascendentes**: atuam a partir dos n objetos iniciais, encarados como grupos com um só objeto, formando novos grupos por aglutinação sucessiva de grupos. No final forma-se um grupo que incluirá todos os n indivíduos.
- **Algoritmos divisivos ou descendentes**: contrariamente, este procedimento atua a partir de um grupo inicial com todos os n indivíduos, formando novos grupos por divisão sucessiva até chegar a n grupos singulares de um só objeto.

Qualquer que seja o algoritmo adotado, a estrutura hierárquica corresponde a um esquema de uma árvore em posição invertida, com a raiz voltada para cima e os ramos voltados para baixo, chamado de dendrograma. Os nós internos deste gráfico representam os *clusters* e a altura dos troncos indica a distância a que os *clusters* se fundem.

Dada uma matriz de dissemelhanças, existem muitas formas de definir a distância entre quaisquer dois *clusters* ou objetos. Após a escolha da definição apropriada, determina-se a distância crítica (d^*) que, corresponde à distância dada pelo nível mínimo, crítico ou de fusão a que os objetos se ligam para formar um novo *cluster*.

A grande vantagem desta representação é mostrar como os sucessivos grupos se vão formando ao longo do processo hierárquico, quer subindo quer descendo a árvore. Por esse motivo, o dendrograma também é usado para determinar a quantidade de grupos a considerar, inicialmente, na AC. No entanto, como desvantagem, não revela a informação relativa às dissemelhanças iniciais.

Dos dois algoritmos apresentados, usualmente, opta-se por escolher o do tipo aglomerativo com o argumento, de que os algoritmos divisivos são mais exigentes a nível computacional.

O **procedimento aglomerativo** é descrito nos seguintes passos:

(a) Considerar os n objetos iniciais como n grupos singulares. A dissemelhança entre os grupos coincide com a matriz de dissemelhanças entre os objetos, $D = [d_{ij}]$;

(b) Identificar o elemento mais pequeno da matriz. Isto equivale a identificar os dois grupos mais semelhantes (grupo A e B), com n_A e n_B, e a sua dissemelhança (d_{AB}), ou seja, a distância entre esses grupos. É importante mencionar que, definir a distância significa determinar um método hierárquico aglomerativo que lhe fica associado;

(c) Unir os grupos à distância crítica (d_{AB}). Atualizar a matriz D eliminando as linhas e as colunas correspondentes aos grupos A e B e, introduzir uma nova linha e coluna com as dissemelhanças calculadas entre o novo grupo (AB) e cada um dos restantes grupos;

(d) Repetir os passos das duas últimas alíneas (b e c), até obter um único grupo que incluirá todos os objetos.

Assim sendo, **os métodos hierárquicos aglomerativos mais comuns** são:

- **Ligação simples ou método do vizinho mais próximo (*single linkage*)**:

  Neste método, a dissemelhança entre os dois grupos é determinada pelos objetos mais próximos, ou seja, a menor das $n_A n_B$ dissemelhanças entre cada elemento de

A e de cada elemento de B:

$$d_{AB} = \min\{d_{ij} : i\epsilon A, j\epsilon B\}.$$

Cada vez que um objeto é adicionado a um grupo, as distâncias do novo grupo aos restantes tornam-se menores ou ficam inalteradas. Assim, vão se formando grupos maiores, deixando os objetos isolados na sua posição. Isto pode ser considerado um ponto positivo para este método, uma vez que permite detetar *outliers*.

- **Ligação completa ou método do vizinho mais afastado (*complete linkage*)**:
  De modo antagónico, a dissemelhança entre dois grupos é a maior das $n_A n_B$ dissemelhanças entre cada elemento de A e de B, servindo-se, assim, dos dois elementos mais afastados:

$$d_{AB} = \max\{d_{ij} : i\epsilon A, j\epsilon B\}.$$

  Quando um objeto é acrescentado a um grupo, a distância do novo grupo aos restantes aumenta ou fica inalterada.

  Este método demonstra ser mais vantajoso que o anterior, pois permite formar grupos pequenos que depois são aglutinados para formar grupos maiores, indo assim, ao encontro do principal objetivo da análise de *clusters*.

- **Ligação média (*unweighted pair-group method using the average approach* - UPGMA)**:
  O método da ligação média considera toda a informação dos grupos, uma vez que a dissemelhança entre dois grupos é a média das dissemelhanças entre todos os pares de objetos, formados com um objeto de cada grupo:

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

  Um grupo é formado pelo conjunto de indivíduos, em que cada um deles tem mais semelhanças, em média, com todos os indivíduos do mesmo grupo do que com todos os indivíduos de qualquer outro grupo.

Os **modelos não hierárquicos** operam sobre uma matriz de dados e exigem que o número de grupos seja fixado à partida.

A partição inicial deve ser definida com base no conhecimento do problema em estudo, ser escolhida ao acaso ou obtida com base no resultado da aplicação prévia de outro método de análise (por exemplo, um método hierárquico).

Há muitos critérios para a formação de *clusters* pertencentes a uma partição inicial.

Com o argumento de que, os grupos devem satisfazer propriedades básicas de coesão interna e isolamento externo, um dos critérios usado assenta na análise de uma matriz de dados do tipo contínuo $X_{n\times p}$ e na equação:

$$T = W + B,$$

onde T, W e B correspondem a matrizes associadas à variação total dos dados, à variação dentro dos grupos e entre grupos, respetivamente:

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}) (\mathbf{x}_{ij} - \overline{\mathbf{x}})',$$

onde $\mathbf{x}_{ij}$ é o vetor de observações do objeto j no grupo i,

$$\overline{\mathbf{x}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbf{x}_{ij}}{\sum_{i=1}^{k} n_i},$$

$$\sum_{i=1}^{k} n_i = n$$

é o vetor das médias de cada uma das p variáveis nos n objetos,

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}) (\mathbf{x}_{ij} - \overline{\mathbf{x}})',$$

e

$$B = \sum_{i=1}^{k} n_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}}) (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})'.$$

Para $p = 1$, o número ideal de *clusters* ou por outras palavras, a melhor partição é aquela em que W é mínimo ou B é máximo, isto é, quanto maior for a homogeneidade interna e maior a separação entre grupos.

No caso multivariado minimizar a soma de quadrados dentro dos grupos equivale a minimizar o traço da matriz W, $\mathrm{tr}\,(W)$. Isto equivale a minimizar a soma de quadrados das distâncias euclidianas entre os objetos e as médias dos respetivos grupos:

$$\mathrm{tr}(W) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)' (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d_{ij,i}^2,$$

onde $d_{ij,i}$ é a distância euclidiana do objeto j do grupo i à média do grupo i.

Este algoritmo é conhecido por **algoritmo das k-médias** ou **método de particionamento das k-médias** e a sequência de passos é:

(a) Particionar os indivíduos em k *clusters* (fixado previamente);

(b) Calcular os centróides dos *clusters*;

(c) Determinar as distâncias entre cada indivíduo e os centróides dos vários *clusters*;

(d) Deslocar os objetos de forma a que cada um fique colocado no grupo de partição que tem o centróide mais próximo;

(e) Recalcular os centróides dos novos grupos formados;

(f) Repetir os dois passos anteriores, até não ser possível efetuar mais deslocações.

Desta forma, todos os indivíduos estão num *cluster* cujo centróide é o mais próximo deles, obtendo-se os *clusters* finais.

5. **Discussão e apresentação dos resultados**.

Nesta última fase devem ser validados os resultados obtidos e deve ser feita uma descrição e interpretação dos *clusters* finais.

A validação dos resultados obtidos pode ser feita com base em dois coeficientes. A qualidade do ajuste é julgada pelo modo como as distâncias originais entre os indíviduos foram preservadas na matriz de proximidades.

- **Coeficiente de correlação cofenética**:

  É somente aplicado quando se utilizam métodos de aglomeração hierárquicos:

$$C = \frac{\sum_{i<j} \left(d_{ij} - \overline{d}_{ij}\right) \left(d^*_{ij} - \overline{d}^*_{ij}\right)}{\sqrt{\left[\sum_{i<j} \left(d_{ij} - \overline{d}_{ij}\right)^2\right] \left[\sum_{i<j} \left(d^*_{ij} - \overline{d}^*_{ij}\right)^2\right]}},$$

  onde, d_{ij} e d^*_{ij} representam, respetivamente, a distância original e a distância na matriz de proximidades entre as observações i e j.

  Caso, os valores do coeficente sejam próximos ou iguais a 1, a solução é de boa qualidade. Quando se obtêm valores inferiores a 0.8 deve-se ponderar se existe ou não uma estrutura hierárquica nas observações e, talvez, utilizar outros métodos de aglomeração.

- **Coeficiente de *silhouette***:

  Permite medir o quão bem uma observação está associada a um *cluster* e estima a distância média entre os *clusters*. O gráfico obtido permite perceber o quão próximo cada ponto de um *cluster* se encontra dos pontos nos *clusters* vizinhos. O coeficiente de *silhouette* da observação i é dado por:

$$Si = \frac{b_i - a_i}{\max\left(a_i, b_i\right)},$$

  onde a_i corresponde à dissemelhança média entre i e todos os outros pontos pertencentes ao *cluster* e $b_i = \min_C d\left(i, C\right)$ representa a dissemelhança entre i e o seu *cluster* "vizinho", ou seja, o *cluster* mais próximo ao qual i não pertence.

  Caso os valores de S_i sejam próximo de 1 significa que as observações estão bem agrupadas, ou melhor dizendo, o objeto i é semelhante aos outros objetos do seu grupo. Se o coeficiente assumir um valor em torno de 0 indica que a observação está entre dois *clusters*. Mas, se as observações tomarem um valor negativo, estão provavelmente colocadas no *cluster* errado.

Tal como referido no início deste subcapítulo, a dificuldade na escolha dos métodos e algoritmos a adotar é elevada. Os métodos hierárquicos não requerem um conhecimento prévio do número de *clusters*, contudo são desvantajosos pelo facto de que, quando um objeto é atribuído a um *cluster* não pode mais sair do mesmo. Para além disso, os algoritmos disponíveis são numerosos, pois dependem das dissemelhanças entre os objetos e do critério de agregação dos *clusters* usado. A melhor forma de superar este incoveniente, é operar com vários métodos, comparar os resultados e verificar se são consistentes.

Esta página foi intencionalmente deixada em branco.

Capítulo 3

Resultados

Nesta dissertação existem 7 variáveis em estudo, sendo que cada uma delas contém 18380 observações, que correspondem às medições de NO_2 de 2016 a 2019.

As variáveis são:

- **Alfragide**: Concentração de NO_2 na EM de Alfragide, em $\mu g/m^3$;
- **AvLiberdade**: Concentração de NO_2 na EM da Avenida da Liberdade, em $\mu g/m^3$;
- **Beato**: Concentração de NO_2 na EM do Beato, em $\mu g/m^3$;
- **Benfica**: Concentração de NO_2 na EM de Benfica, em $\mu g/m^3$;
- **Entrecampos**: Concentração de NO_2 na EM de Entrecampos, em $\mu g/m^3$;
- **Olivais**: Concentração de NO_2 na EM de Olivais, em $\mu g/m^3$;
- **Restelo**: Concentração de NO_2 na EM do Restelo, em $\mu g/m^3$.

3.1 Estatística descritiva de dados multivariados

Neste estudo foi usado o *software* R. A partir da amostra foram definidas as variáveis quantitativas por ano, mês e estação do ano.

O diagrama de extremos e quartis ou *boxplot* dá uma ideia da variação dos dados observados das variáveis através dos quartis. Este diagrama identifica os quartis ($1^{\underline{o}}$ e 3^{o}Q), a mediana (2^{o}Q), os valores externos (mínimo e máximo) e os *outliers*.

Nos *boxplots* da Figura 3.1 é possível verificar que todas as variáveis apresentam muitos *outliers*, o que poderá afetar tanto este trabalho como futuros estudos. Contudo, devido ao elevado número de observações, não foram determinados os que são *outliers* moderados e os que são *outliers* severos. Para além disso, verifica-se que os diagramas das EM são semelhantes, excepto o da EM da AvLiberdade, destacando-se negativamente por ter maiores valores observados do poluente em causa.

Após a análise dos gráficos de extremos e quartis para cada EM, por ano, mês e estação do ano realça-se o facto, do ano 2017 apresentar maiores valores do poluente em estudo em todas as EM e uma diminuição dos valores do poluente no verão e nos meses de julho e agosto, também em todas as EM. Como exemplo, apresentam-se os diagramas para a EM de Entrecampos (Figura 3.2).

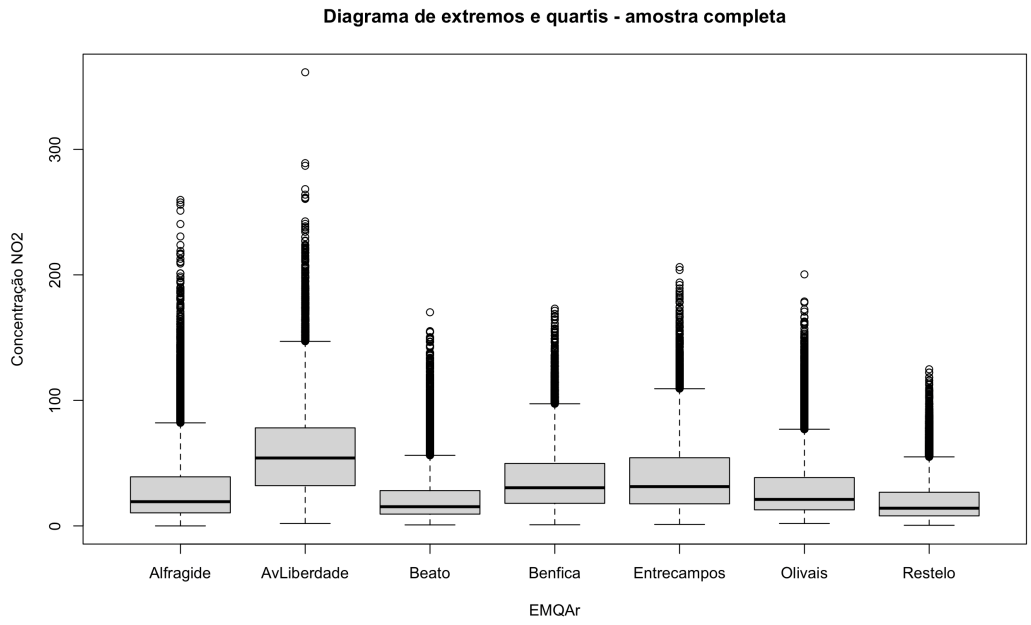

Figura 3.1: Diagramas de extremos e quartis das variáveis durante os 4 anos consecutivos.

De seguida, obtiveram-se as medidas estatísticas para cada estação (variável) (Figura 3.3).

Através dos resultados apresentados na Figura 3.3, e através do exemplo da variável Alfragide, concluí-se que:

Em Alfragide, a concentração de NO_2 tem um valor mínimo e máximo de, respetivamente, 0 $\mu g/m^3$ e 259.700 $\mu g/m^3$. O 1º quartil é 10.400 $\mu g/m^3$, ou seja, pelo menos 25% das observações relativas à concentração de NO_2 estão abaixo desse valor. O 2º quartil é de 19.300 $\mu g/m^3$, ou seja, 50% das concentrações do poluente na estação de Alfragide são menores ou iguais a 19.300 $\mu g/m^3$. O 3º quartil é 39.100 $\mu g/m^3$, sendo que 75% das observações são inferiores ou iguais a esse valor.

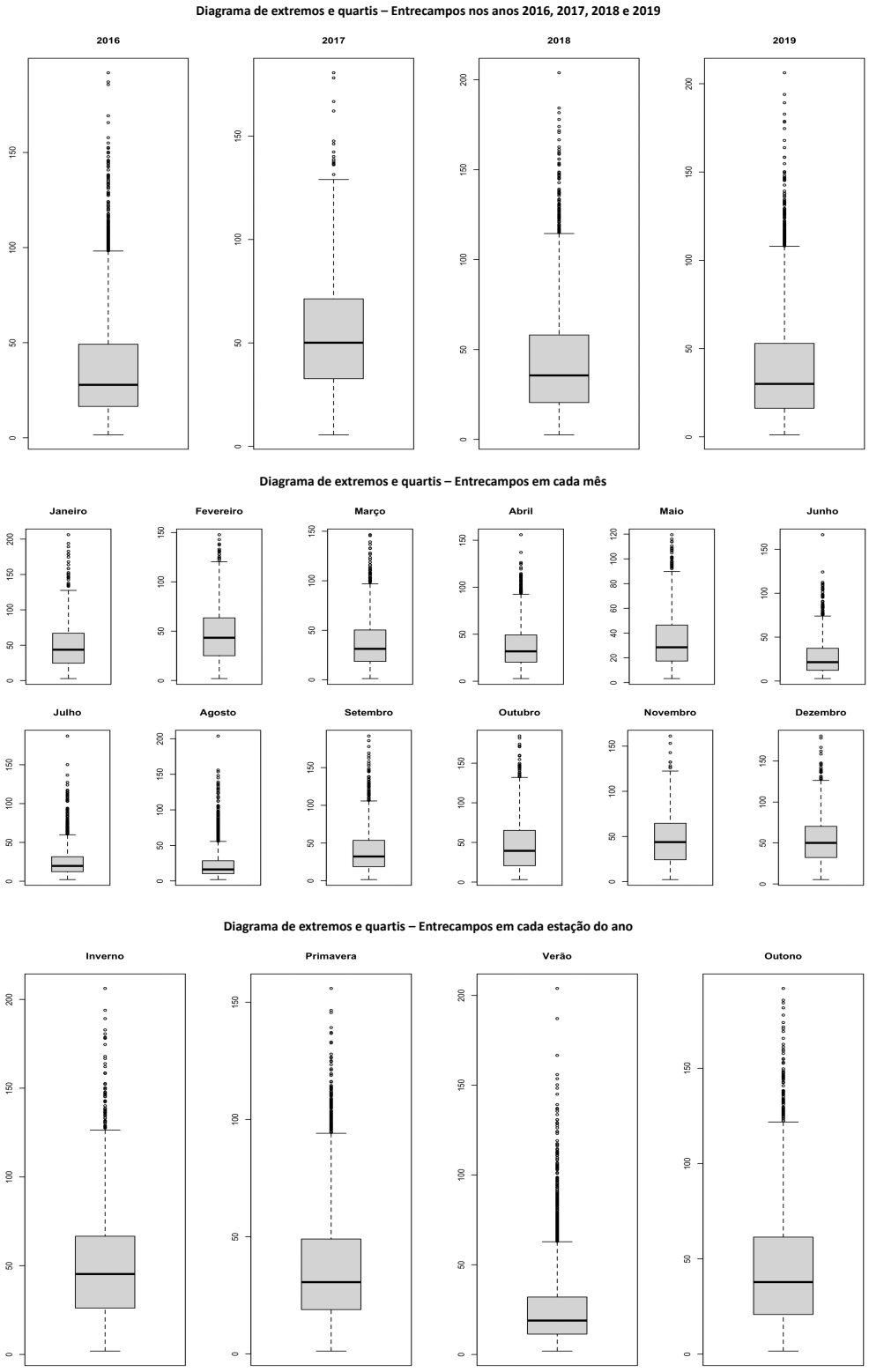

Figura 3.2: Diagramas de extremos e quartis da EM de Entrecampos em cada ano, mês e estação do ano.

	Alfragide	AvLiberdade	Beato	Benfica	Entrecampos	Olivais	Restelo
nobs	18380.000	18380.000	18380.000	18380.000	18380.000	18380.000	18380.000
Minimum	0.000	1.900	0.800	0.900	1.200	1.900	0.500
Maximum	259.700	361.400	170.200	173.100	206.200	200.400	124.800
1. Quartile	10.400	32.100	9.400	18.000	17.600	12.800	8.000
3. Quartile	39.100	78.100	28.125	49.725	54.300	38.500	26.800
Mean	29.417	58.706	22.283	36.517	38.812	29.510	20.325
Median	19.300	54.100	15.300	30.400	31.300	21.100	14.100
Variance	803.525	1213.976	387.077	633.726	761.998	585.846	315.625
Stdev	28.347	34.842	19.674	25.174	27.604	24.204	17.766

Figura 3.3: *Output* com resumo alargado por variável.

Relativamente às medidas de localização de tendência central, foram obtidas para cada uma das estações, as concentrações médias de NO_2 (Figura 3.4).

Alfragide	AvLiberdade	Beato	Benfica	Entrecampos	Olivais	Restelo
29.41659	58.70603	22.28254	36.51736	38.81199	29.51031	20.32487

Figura 3.4: Médias das variáveis.

Verificam-se que os valores são relativamente semelhantes, com exceção da AvLiberdade (58.706 $\mu g/m^3$). A concentração média do poluente para as estações de Alfragide, Beato e Benfica foi, respetivamente, 29.417 $\mu g/m^3$, 22.283 $\mu g/m^3$ e 36.517 $\mu g/m^3$. Enquanto que para as estações de Entrecampos, Olivais e Restelo foi 38.812 $\mu g/m^3$, 29.510 $\mu g/m^3$ e 20.325 $\mu g/m^3$.

Dado que existem muitos *outliers*, a medida de localização mais adequada para caraterizar as variáveis é a mediana. À semelhança do que acontece nas médias, verifica-se que o valor da mediana para a AvLiberdade é distinto dos restantes valores.

Alfragide	AvLiberdade	Beato	Benfica	Entrecampos	Olivais	Restelo
29.41659	58.70603	22.28254	36.51736	38.81199	29.51031	20.32487

Figura 3.5: Medianas das variáveis.

As medidas descritivas de dispersão nomeadamente, a variância e o desvio-padrão da amostra, foram obtidas para cada uma das estações (Figuras 3.6 e 3.7).

Alfragide	AvLiberdade	Beato	Benfica	Entrecampos	Olivais	Restelo
803.5250	1213.9756	387.0766	633.7259	761.9979	585.8458	315.6251

Figura 3.6: Variâncias das variáveis.

Alfragide	AvLiberdade	Beato	Benfica	Entrecampos	Olivais	Restelo
28.34652	34.84215	19.67426	25.17391	27.60431	24.20425	17.76584

Figura 3.7: Desvios-padrão das variáveis.

Os valores elevados dos desvios-padrão indicam que os dados apresentam uma grande variabilidade média.

Embora as variáveis tenham a mesma unidade de medida, as médias e as variâncias não são semelhantes e por esse motivo, opta-se por estudar a variabilidade de cada variável através do coeficiente de variação.

Alfragide	AvLiberdade	Beato	Benfica	Entrecampos	Olivais	Restelo
96.36235	59.35021	88.29452	68.93683	71.12316	82.01965	87.40937

Figura 3.8: Coeficiente de variação das variáveis.

Os dados da amostra apresentam uma variabilidade elevada, uma vez que em todas as variáveis o $cv_j > 30\%$, concluindo-se que a média é pouco representativa dos dados. Em comparação com as outras variáveis, Alfragide é a variável com maior variabilidade, isto é, a que apresenta maior dispersão de valores relativamente à concentração de NO_2 (Figura 3.8).

Devido à existência de muitos *outliers* determinou-se o coeficiente de variação resistente, comprovando que Alfragide é a variável com maior variabilidade (Figura 3.9).

Alfragide	AvLiberdade	Beato	Benfica	Entrecampos	Olivais	Restelo
148.70466	85.02773	122.38562	104.35855	117.25240	121.80095	133.33333

Figura 3.9: Coeficiente de variação resistente das variáveis.

No diagrama da Figura 3.10 visualiza-se o padrão de associação entre as variáveis, ou seja, se existe uma relação positiva ou negativa. Se a curva de ajustamento aos dados for uma reta, significa que a relação entre as variáveis é linear. Caso contrário, pode existir uma relação não linear ou não existir, de todo, relação entre as variáveis. A título de exemplo, o segundo quadrado da primeira coluna representa o gráfico de dispersão entre Alfragide (eixo x) e a AvLiberdade (eixo y). Este mesmo gráfico é replicado no segundo quadrado da primeira linha, mas com as variáveis nos eixos opostos. Assim, as caixas acima da diagonal são representações em espelho dos gráficos abaixo.

Ao analisar o diagrama é possível aferir que existe uma relação positiva entre todas as variáveis, pois à medida que uma variável aumenta, a outra variável também aumenta. Para além disso, deduz-se que parece existir uma relação linear positiva entre as variáveis.

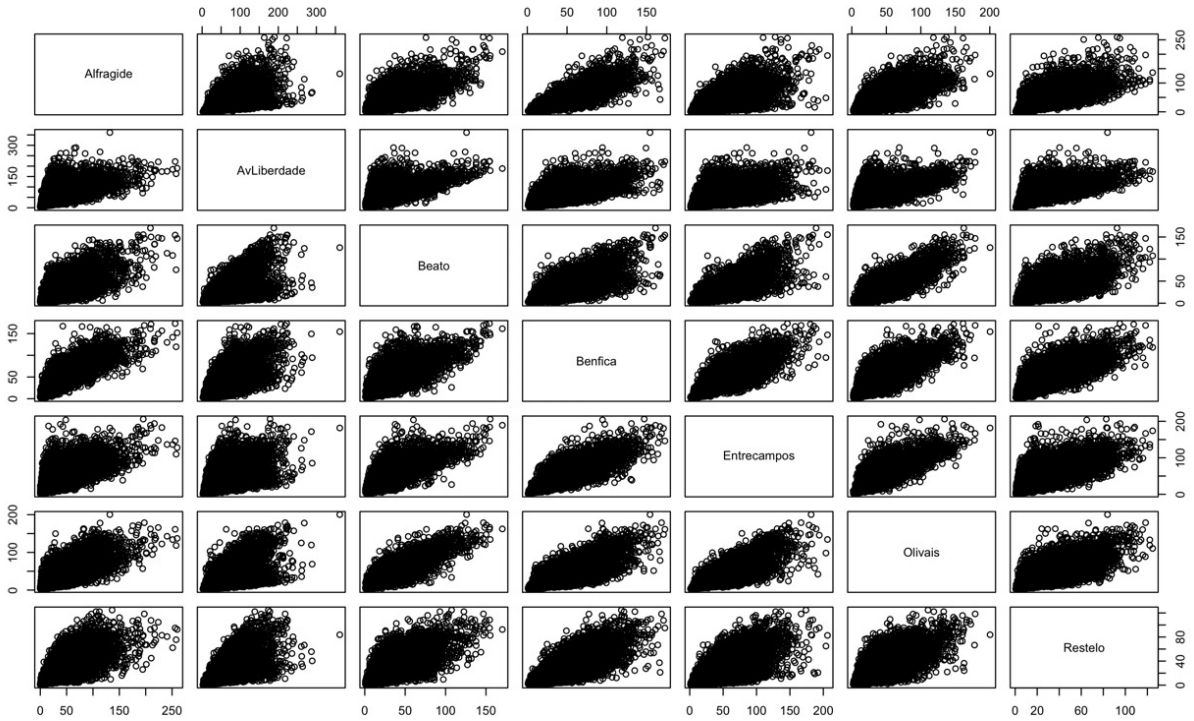

Figura 3.10: Diagrama de dispersão entre as variáveis.

A partir da matriz de variâncias-covariâncias da Figura 3.11 verifica-se que as covariâncias entre as variáveis são positivas, ou seja, todos os pares de variáveis tendem a aumentar ou diminuir conjuntamente. Por exemplo, a covariância entre Alfragide e Beato corresponde ao valor de 417.397 $\mu g/m^3$ o que significa, que quanto maior a concentração de NO_2 na estação de Alfragide maior é a concentração desse poluente na estação do Beato.

	Alfragide	AvLiberdade	Beato	Benfica	Entrecampos	Olivais	Restelo
Alfragide	803.5250	626.6057	417.3968	611.6776	543.1358	531.2971	397.1135
AvLiberdade	626.6057	1213.9756	427.6977	565.1160	476.1809	498.9080	423.2481
Beato	417.3968	427.6977	387.0766	380.4255	425.8425	416.9162	267.6886
Benfica	611.6776	565.1160	380.4255	633.7259	580.2716	500.7852	362.5992
Entrecampos	543.1358	476.1809	425.8425	580.2716	761.9979	553.0992	342.5944
Olivais	531.2971	498.9080	416.9162	500.7852	553.0992	585.8458	327.4923
Restelo	397.1135	423.2481	267.6886	362.5992	342.5944	327.4923	315.6251

Figura 3.11: Matriz de variâncias-covariâncias.

Na Figura 3.12 encontra-se a matriz de correlações.

Para ser mais percetível a análise de correlações foi obtido um correlograma (Figura 3.13).

A correlação linear amostral é positiva em todos os casos, ou seja, quando a concentração do poluente aumenta em uma EM, a concentração também aumenta na EM correlacionada. Existe uma correlação fraca entre a AvLiberdade e todas as outras variáveis (Entrecampos, Olivais, Beato, Alfragide, Benfica, Restelo) e entre Alfragide e Entrecampos, uma vez que os seus coeficientes de correlação estão compreendidos entre os valores 0.5 e 0.7. As variáveis mais fortemente correlacionadas são Beato e Olivais com $r = 0.88$. Assim, quanto maior a

concentração do poluente na estação de Olivais, maior a concentração na estação de Beato. Os restantes pares de variáveis têm uma correlação elevada, logo existe uma relação linear positiva forte entre elas.

É de notar que as conclusões retiradas do diagrama de dispersão são as mesmas do correlograma.

	Alfragide	AvLiberdade	Beato	Benfica	Entrecampos	Olivais	Restelo
Alfragide	1.0000000	0.6344387	0.7484295	0.8571803	0.6941156	0.7743656	0.7885499
AvLiberdade	0.6344387	1.0000000	0.6239265	0.6442908	0.4950968	0.5915941	0.6837609
Beato	0.7484295	0.6239265	1.0000000	0.7681046	0.7841040	0.8755050	0.7658533
Benfica	0.8571803	0.6442908	0.7681046	1.0000000	0.8350331	0.8218812	0.8107563
Entrecampos	0.6941156	0.4950968	0.7841040	0.8350331	1.0000000	0.8278173	0.6985824
Olivais	0.7743656	0.5915941	0.8755050	0.8218812	0.8278173	1.0000000	0.7615943
Restelo	0.7885499	0.6837609	0.7658533	0.8107563	0.6985824	0.7615943	1.0000000

Figura 3.12: Matriz de correlações.

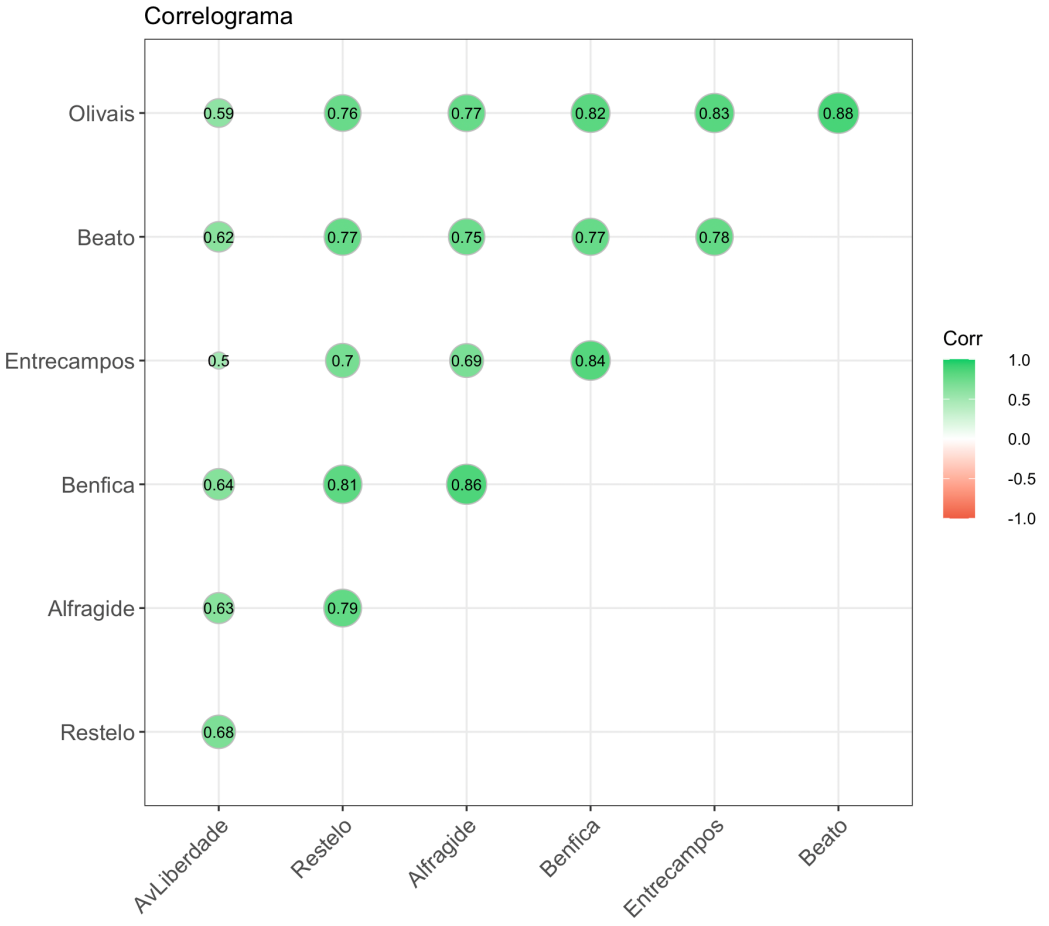

Figura 3.13: Correlograma.

3.2 Análise de Componentes Principais

Para esta técnica estatística existe a necessidade de estandardizar os dados, uma vez que as variâncias das variáveis têm um valor muito diferente, por exemplo, a variância de Olivais assume um valor de 585.846 $(\mu g/m^3)^2$ enquanto que a variância de AvLiberdade é igual a 1213.976 $(\mu g/m^3)^2$. Assim, a matriz utilizada na ACP corresponde à matriz de correlações.

Dada a forte correlação entre as variáveis, verificada na análise descritiva de dados, decidiu-se aplicar o método de ACP. Contudo, realizaram-se dois testes para averiguar adequação dos dados à aplicação desta técnica estatística.

Primeiramente, realizou-se o teste de esfericidade de Bartlett. Após formular as hipóteses e estabelecer um nível de significância de 5%, verificou-se, no *output* da Figura 3.14, o valor da estatística de teste (139045.6) e o valor de *p-value* (0).

```
$chisq          $p.value     $df
[1] 139045.6    [1] 0        [1] 21
```

Figura 3.14: Teste de esfericidade de Bartlett.

Como $\alpha = 0.05 \geq p - value = 0$, deve-se rejeitar H_0, com o nível de significância de 5%, concluindo que existem correlações estatisticamente significativas entre as variáveis. Deste modo, a técnica ACP é adequada aos dados.

O segundo teste corresponde à estatística KMO, que tal como o teste de esfericidade de Bartlett, compara as correlações entre as variáveis.

De seguida, obteve-se o *output* da Figura 3.15. Dado que $KMO = 0.90$ verifica-se, com base na Tabela 2.4, que a adequação dos dados é muito boa.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor(amostra))
Overall MSA =  0.9
MSA for each item =
  Alfragide AvLiberdade       Beato     Benfica Entrecampos     Olivais     Restelo
       0.90        0.93        0.89        0.85        0.87        0.91        0.95
```

Figura 3.15: Estatística KMO.

Assim, concluí-se que a técnica de ACP é adequada aos dados.

Posteriormente, executaram-se dois testes de hipóteses para averiguar a homogeneidade das variâncias, o teste de Levene e o teste de Bartlett.

Através dos *outputs* da Figura 3.16, os valores das estatísticas de teste correspondem a 1463.4, no caso do teste de Levene, e 11101, no caso do teste de Bartlett. Relativamente ao *p-value*, ambos os testes apresentam o mesmo valor.

Assim, verifica-se que se deve rejeitar a hipótese nula, pois $\alpha = 0.05 \geq p - value = 2.2 \times 10^{-16}$, comprovando que as variâncias das variáveis são estatisticamente diferentes.

```
Levene's Test for Homogeneity of Variance (center = mean)
          Df F value    Pr(>F)
group      6  1463.4 < 2.2e-16 ***
      128653


    Bartlett test of homogeneity of variances

data:  list(Alfragide, AvLiberdade, Beato, Benfica, Entrecampos, Olivais, Restelo)
Bartlett's K-squared = 11101, df = 6, p-value < 2.2e-16
```

Figura 3.16: Testes de homocedasticidade de Levene e Bartlett.

Para obtenção das componentes principais foi necessário calcular os valores e vetores próprios, com a matriz de correlações. Ao usar a matriz de correlações estão a usar-se os dados estandardizados.

Utilizando o *software* R obteve-se os desvios-padrão de cada componente:

```
Standard deviations (1, .., p=7):
[1] 2.3336325 0.7595798 0.5684954 0.4924949 0.4571785 0.3394726 0.2953062
```

Figura 3.17: Desvios-padrão de cada componente.

O quadrado destes correspondem às variâncias das componentes que são numericamente iguais aos valores próprios associados a cada componente (Figura 3.18). Os valores próprios são:

```
[1] 5.446 0.577 0.323 0.243 0.209 0.115 0.087
```

Figura 3.18: Valores próprios de cada componente.

Onde correspondem os seguintes vetores próprios, sendo que a cada coluna corresponde um valor próprio:

```
Rotation (n x k) = (7 x 7):
                PC1         PC2        PC3        PC4         PC5         PC6         PC7
Alfragide   0.3827844  0.07861921 -0.5984875  0.2004213 -0.51965798  0.18314743 -0.38125030
AvLiberdade 0.3198451  0.80538612  0.3511740 -0.3236001 -0.11259215 -0.01822048 -0.08946014
Beato       0.3877407 -0.14712904  0.4734175  0.4651872 -0.13290024  0.53093360  0.29653660
Benfica     0.3998529 -0.07819712 -0.3584077 -0.3802234 -0.02619105 -0.08300456  0.74390939
Entrecampos 0.3724818 -0.46305296  0.1604892 -0.5756348  0.20215152  0.25197715 -0.43054541
Olivais     0.3944572 -0.25602517  0.2817378  0.1988372 -0.20101039 -0.78125312 -0.09581540
Restelo     0.3828980  0.19350333 -0.2465285  0.3510750  0.78591332 -0.05912584 -0.10373052
```

Figura 3.19: Vetores próprios.

Uma vez calculados os vetores próprios, de seguida escrevem-se as 7 componentes principais, que são combinações lineares das variáveis iniciais estandardizadas.

Assim obtêm-se:

$Y_1 = 0.383Z_1 + 0.320Z_2 + 0.388Z_3 + 0.400Z_4 + 0.372Z_5 + 0.394Z_6 + 0.383Z_7$

$Y_2 = 0.079Z_1 + 0.805Z_2 - 0.147Z_3 - 0.078Z_4 - 0.463Z_5 - 0.256Z_6 + 0.194Z_7$
$Y_3 = -0.599Z_1 + 0.351Z_2 + 0.473Z_3 - 0.358Z_4 + 0.160Z_5 + 0.282Z_6 - 0.247Z_7$
$Y_4 = 0.200Z_1 - 0.324Z_2 + 0.465Z_3 - 0.380Z_4 - 0.576Z_5 + 0.199Z_6 + 0.351Z_7$
$Y_5 = -0.520Z_1 - 0.113Z_2 - 0.133Z_3 - 0.026Z_4 + 0.202Z_5 - 0.201Z_6 + 0.786Z_7$
$Y_6 = 0.183Z_1 - 0.018Z_2 + 0.531Z_3 - 0.083Z_4 + 0.252Z_5 - 0.781Z_6 - 0.059Z_7$
$Y_7 = -0.381Z_1 - 0.089Z_2 + 0.297Z_3 + 0.744Z_4 - 0.431Z_5 - 0.096Z_6 - 0.104Z_7$,

onde z_i, com $i = 1, 2, ..., 7$, corresponde, às variáveis estandardizadas de Alfragide (Z_1), AvLiberdade (Z_2), Beato (Z_3), Benfica (Z_4), Entrecampos (Z_5), Olivais (Z_6) e Restelo (Z_7).

Analisando cada uma das componentes principais sabe-se que:

Na 1ª componente, as variáveis Z_3, Z_4 e Z_6 são as que têm maior peso, sendo que todas estas têm uma contribuição positiva. No entanto, todas as variáveis têm um peso muito idêntico.

Em relação à 2ª componente, a variável Z_2 apresenta maior peso e contribui positivamente para esta componente. Embora a variável Z_5 possua um valor inferior, têm um elevado peso e uma contribuição negativa.

Na 3ª componente, as variáveis Z_1 e Z_3 são as que têm maior peso e contribuem, respetivamente, negativa e positivamente para esta componente.

Relativamente à 4ª componente, as variáveis Z_3 e Z_5 são as que têm mais peso e contribuem, respetivamente, positiva e negativamente para essa componente.

Na 5ª componente, a variável Z_7 é a que tem maior peso e contribuí positivamente para esta componente. Contudo, em comparação com as restantes variáveis, a variável Z_1 também apresenta um elevado peso na componente, contribuindo negativamente para esta.

Na 6ª componente, as variáveis Z_3 e Z_6 são as que apresentam maior peso e contribuem, positiva e negativamente para esta componente.

Por fim, a variável Z_4 é a que apresenta maior peso na 7ª componente com uma contribuição positiva. Porém, a variável Z_5 também têm um elevado peso para esta componente, contribuindo negativamente para esta.

Como se verá mais à frente, são somente escolhidas duas a três componentes. Para comprovar o que foi mencionado anteriormente, a Figura 3.20 mostra, graficamente, os pesos das variáveis em cada uma das três primeiras componentes.

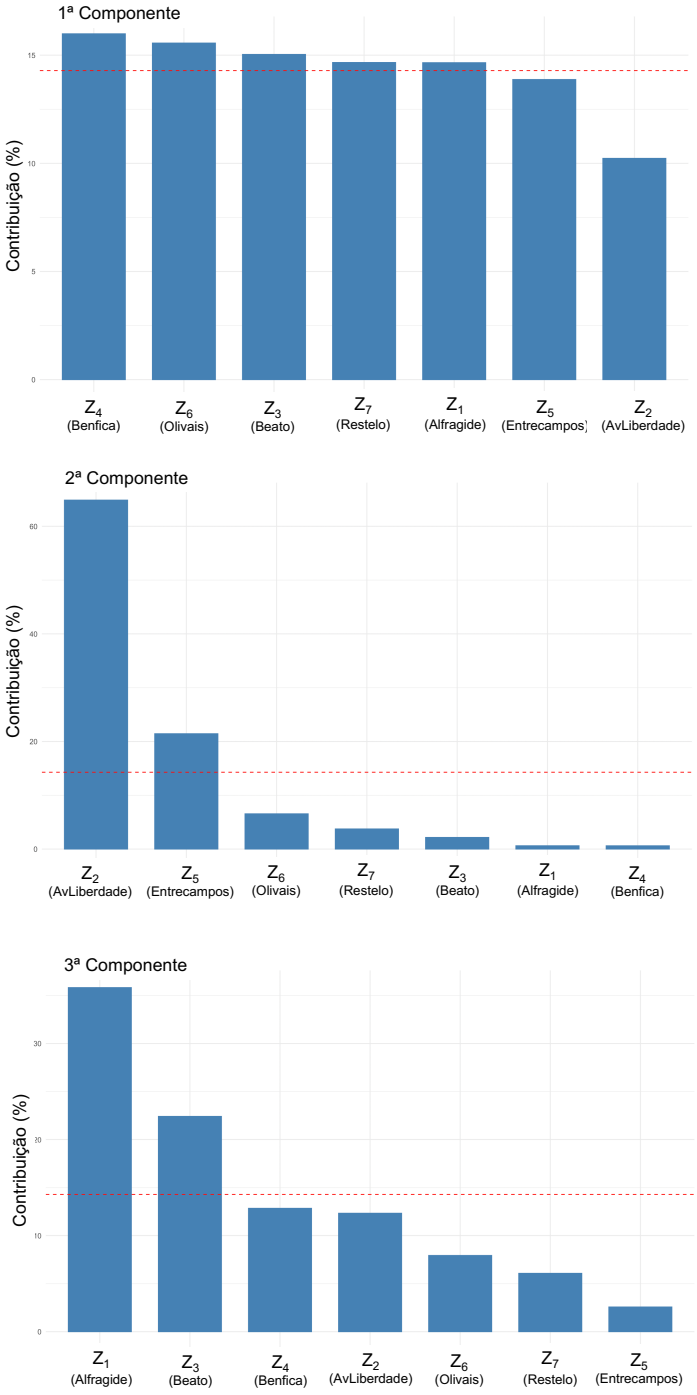

Figura 3.20: Contribuição das variáveis para a 1ª, 2ª e 3ª componente.

Dado que se obtêm tantas componentes como variáveis, a escolha do número de componentes principais a reter foi realizada com a aplicação de três critérios:

- **Variância explicada acumulada**:
  Primeiro determinou-se a contribuição de cada componente principal para a explicação da variância total.

Este critério diz que se deve incluir componentes suficientes para explicar mais de 70% da variância total. Através da Figura 3.21 concluí-se que só uma componente já explica 77.8%. Contudo, para alcançar o objetivo desta dissertação, devem-se considerar duas ou três componentes, uma vez que explicam, respetivamente, 86.0% e 90.7% da variância total explicada.

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.334 0.75958 0.56850 0.49249 0.45718 0.33947 0.29531
Proportion of Variance 0.778 0.08242 0.04617 0.03465 0.02986 0.01646 0.01246
Cumulative Proportion  0.778 0.86040 0.90657 0.94122 0.97108 0.98754 1.00000
```

Figura 3.21: Variância total explicada das componentes principais.

- ***Scree Plot***:

  O *scree plot* representa graficamente a percentagem da variância explicada por cada componente.

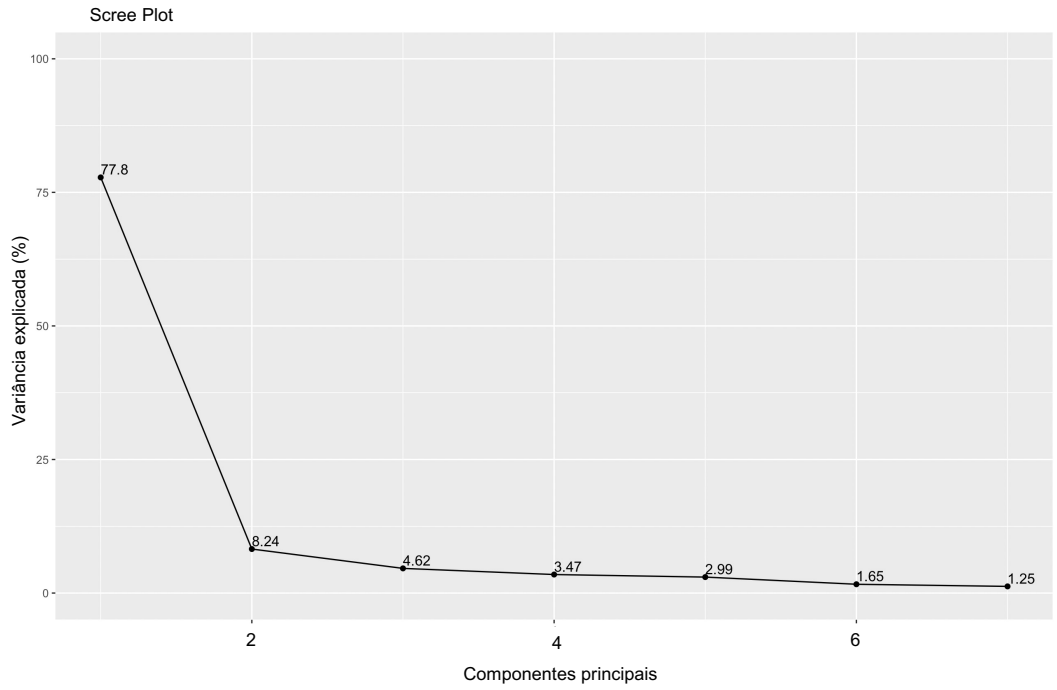

Figura 3.22: *Scree Plot.*

No gráfico da Figura 3.22 observa-se que a reta fica quase paralela com o eixo das abcissas, ou seja, com declive reduzido a partir da 4ª componente, devendo-se assim selecionar duas a três componentes principais.

- **Critério de Kaiser**:

  Segundo este critério devem-se incluir as componentes cujos valores próprios são superiores a 1. Desta forma, com base na Figura 3.18, apenas a 1ª componente principal deveria ser selecionada.

Tendo em conta estes critérios e os objetivos do trabalho, optou-se por realizar o estudo tanto para duas como para três componentes.

a) ACP com 2 componentes

Os valores das correlações entre as variáveis iniciais estandardizadas (Z_i) e as duas primeiras componentes principais (Y_j) selecionadas são apresentadas na Tabela 3.1. Para além disso, foi obtido o gráfico com as correlações entre as variáveis e as componentes (Figura 3.23).

Tabela 3.1: Correlações entre as variáveis iniciais estandardizadas e as duas componentes selecionadas

PC1	PC2
$r_{Y_1,Z_1} = 0.893$	$r_{Y_2,Z_1} = 0.060$
$r_{Y_1,Z_2} = 0.746$	$r_{Y_2,Z_2} = 0.612$
$r_{Y_1,Z_3} = 0.905$	$r_{Y_2,Z_3} = -0.112$
$r_{Y_1,Z_4} = 0.933$	$r_{Y_2,Z_4} = -0.059$
$r_{Y_1,Z_5} = 0.869$	$r_{Y_2,Z_5} = -0.352$
$r_{Y_1,Z_6} = 0.921$	$r_{Y_2,Z_6} = -0.194$
$r_{Y_1,Z_7} = 0.894$	$r_{Y_2,Z_7} = 0.147$

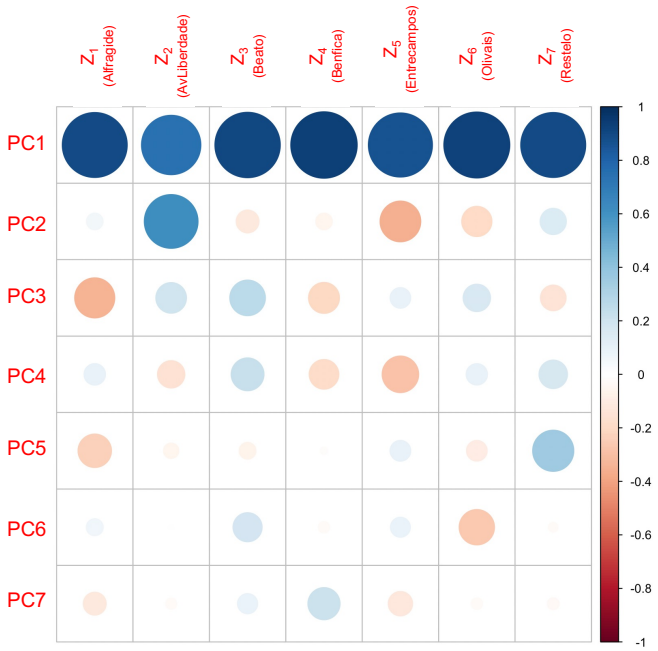

Figura 3.23: Gráfico com correlações entre as variáveis iniciais estandardizadas e as componentes.

Com base nesses valores e no gráfico concluí-se que todas as variáveis originais estandardizadas têm uma correlação positiva forte com a PC1. Porém, a variável Z_2 (AvLiberdade) apresenta uma correlação ligeiramente inferior em comparação com as restantes.

Observa-se que na PC2 todas as variáveis iniciais têm uma correlação baixa, com exceção da variável Z_2 (AvLiberdade).

Então, pode-se concluir que estas componentes se complementam.

De seguida determinaram-se valores (*scores*) das duas componentes principais centradas na média. Inicialmente existia uma matriz de dimensão 18380×7 que foi substituída por uma matriz de dimensão 18380×2, sendo que para esta nova matriz as entradas correspondem aos *scores* (Tabela 3.2).

Tabela 3.2: Scores da PC1 e PC2

Scores da PC1	**Scores da PC2**
$Y_{11} = -1.253$	$Y_{12} = -0.771$
$Y_{21} = -0.249$	$Y_{22} = -1.146$
...	...
$Y_{142851} = -0.641$	$Y_{142851} = -0.900$

A partir dos gráficos da Figura 3.24 verifica-se que na PC1 existem *scores* positivos e negativos. No entanto, os *scores* negativos têm todos um valor muito semelhante. Enquanto que os positivos apresentam um maior leque de valores, ou seja, os *scores* têm uma maior variabilidade e, contrariamente ao que era esperado, não existe uma diferença entre os anos, meses ou estações do ano. Na PC2, a quantidade de valores positivos e negativos é praticamente semelhante e a variabilidade é idêntica para os positivos e negativos. Também se verifica que não há uma grande diferença entre anos, meses ou estações do ano.

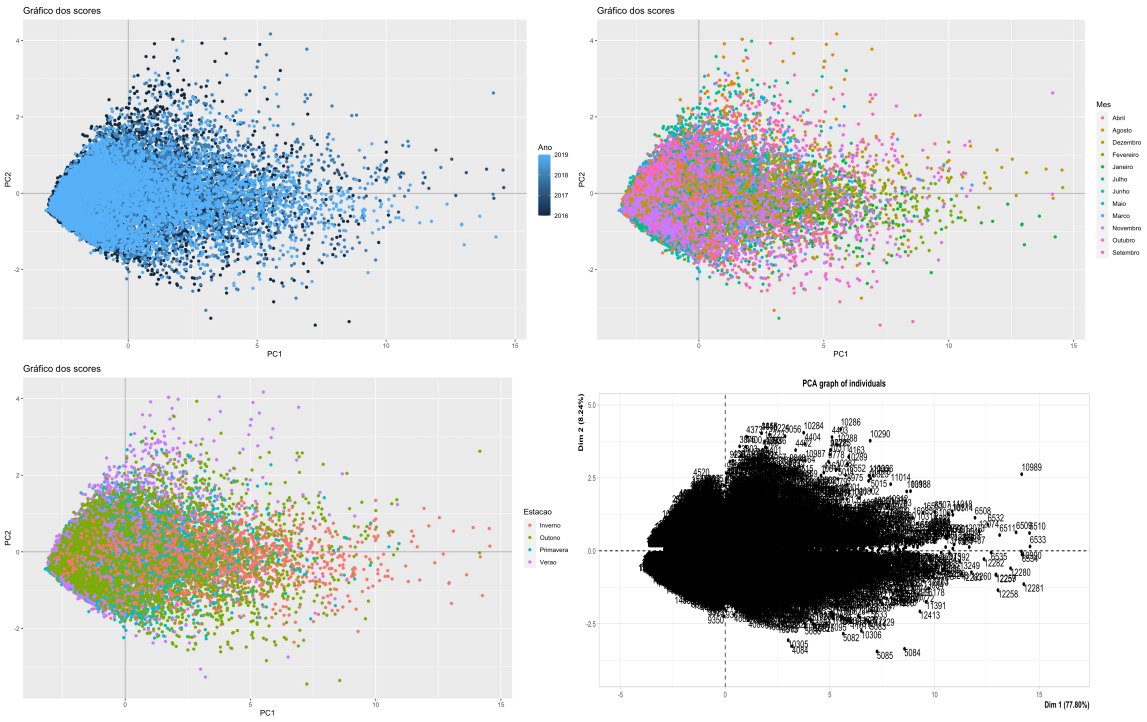

Figura 3.24: Gráfico de *scores* da PC1 e PC2 por ano, mês, estação do ano e observações.

A aplicação do método de rotação ortogonal VARIMAX permite obter um novo conjunto de pesos das variáveis estandardizadas para cada componente selecionada. Deste modo, aumentam os valores dos pesos que mais contribuem para a formação da componente e diminuem os pesos das que menos contribuem. Quanto mais próximo de 1, em valor absoluto, estiver esse peso, mais forte é a associação entre essa variável e a componente, enquanto que um peso próximo de 0 permite concluir que essa variável pouco contribui para a formação da componente.

O *output* da Figura 3.25 obtém-se quando se aplica o método de rotação ortogonal VARIMAX às componentes selecionadas.

```
Loadings:
            RC1   RC2
Alfragide   0.702 0.556
AvLiberdade 0.268 0.927
Beato       0.809 0.421
Benfica     0.802 0.480
Entrecampos 0.915 0.203
Olivais     0.868 0.362
Restelo     0.653 0.628

                 RC1   RC2
SS loadings    3.880 2.142
Proportion Var 0.554 0.306
Cumulative Var 0.554 0.860
```

Figura 3.25: Rotação ortogonal VARIMAX com 2 componentes.

Assim, as componentes principais passam a ser:

$Y_1 = 0.702Z_1 + 0.268Z_2 + 0.809Z_3 + 0.802Z_4 + 0.915Z_5 + 0.868Z_6 + 0.653Z_7$

$Y_2 = 0.556Z_1 + 0.927Z_2 + 0.421Z_3 + 0.480Z_4 + 0.203Z_5 + 0.362Z_6 + 0.628Z_7$

Relativamente à 1ª componente, anteriormente as variáveis que tinham maior peso eram Z_3 (Beato), Z_4 (Benfica) e Z_6 (Olivais). Depois do método de rotação, todas as variáveis têm um peso significativo (>0.5) na obtenção da componente, exceto a variável Z_2 (AvLiberdade). Contudo, verifica-se que as variáveis que têm um maior peso são Z_5 (Entrecampos), Z_6 (Olivais) e Z_3 (Beato) e Z_4 (Benfica).

Na 2ª componente, a variável que tinha uma maior contribuição era a Z_2 (AvLiberdade). Com a rotação, concluí-se que a variável com maior peso continua a ser Z_2 (AvLiberdade), porém as variáveis Z_7 (Restelo) e Z_1 (Alfragide) apresentam valores superiores a 0.5, logo são considerados significativos.

A distribuição da proporção da variância total explicada por cada componente é agora mais equilibrada após a aplicação do método de rotação de VARIMAX.

A Figura 3.26 permite verificar que os valores dos *scores* obtidos após rotação não variam muito consoante os quatro anos consecutivos, os meses e as estações do ano, tal como acontecia antes.

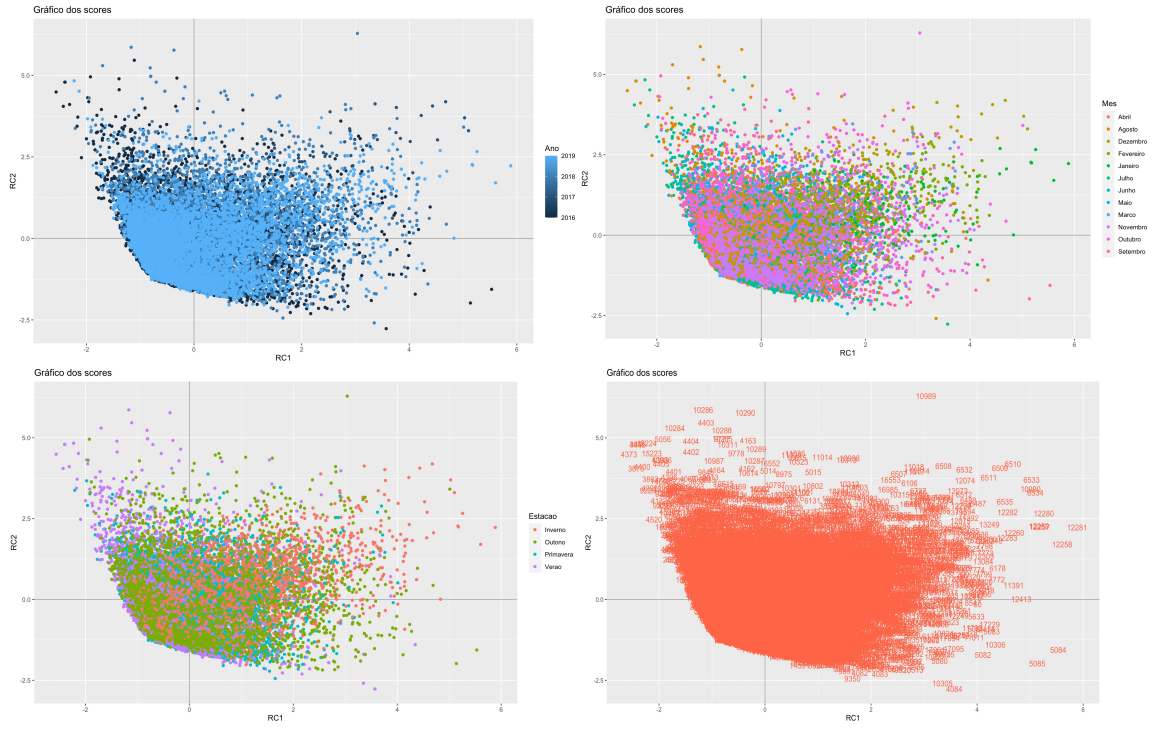

Figura 3.26: Gráficos de *scores* da RC1 e RC2 após rotação por, ano, mês, estação do ano e observações.

b) ACP com 3 componentes

Os valores das correlações entre as variáveis iniciais estandardizadas (Z_i) e as três primeiras componentes principais (Y_j) encontram-se na Tabela 3.3.

Tabela 3.3: Correlações entre as variáveis iniciais e as três componentes selecionadas

PC1	**PC2**	**PC3**
$r_{Y_1,Z_1} = 0.893$	$r_{Y_2,Z_1} = 0.060$	$r_{Y_3,Z_1} = -0.340$
$r_{Y_1,Z_2} = 0.746$	$r_{Y_2,Z_2} = 0.612$	$r_{Y_3,Z_2} = 0.200$
$r_{Y_1,Z_3} = 0.905$	$r_{Y_2,Z_3} = -0.112$	$r_{Y_3,Z_3} = 0.269$
$r_{Y_1,Z_4} = 0.933$	$r_{Y_2,Z_4} = -0.059$	$r_{Y_3,Z_4} = -0.204$
$r_{Y_1,Z_5} = 0.869$	$r_{Y_2,Z_5} = -0.352$	$r_{Y_3,Z_5} = 0.091$
$r_{Y_1,Z_6} = 0.921$	$r_{Y_2,Z_6} = -0.194$	$r_{Y_3,Z_6} = 0.160$
$r_{Y_1,Z_7} = 0.894$	$r_{Y_2,Z_7} = 0.147$	$r_{Y_3,Z_7} = -0.140$

As conclusões retiradas entre a PC1 e PC2 são as mesmas do ponto anterior.

Através dos valores da Tabela 3.3 e do gráfico da Figura 3.23 verifica-se que a variável Z_1 (Alfragide) é a que tem uma correlação mais forte com a PC3, mas ainda assim é uma correlação mais fraca do que as anteriores.

De seguida, determinaram-se os *scores* das três primeiras componentes principais. A matriz original de dimensão 18380×7 foi substituída por uma matriz de dimensão 18380×3 com os valores apresentados na Tabela 3.4.

Tabela 3.4: Scores da PC1, PC2 e PC3

Scores da PC1	**Scores da PC2**	**Scores da PC3**
$Y_{11} = -1.253$	$Y_{12} = -0.771$	$Y_{13} = 0.041$
$Y_{21} = -0.249$	$Y_{22} = -1.146$	$Y_{23} = 0.442$
$\ldots$	$\ldots$	$\ldots$
$Y_{142851} = -0.641$	$Y_{142851} = -0.900$	$Y_{142853} = 0.078$

Em relação aos gráficos de *scores* entre a PC1 e PC2, as conclusões são iguais às do tópico anterior.

Nos gráficos das Figuras 3.27 e 3.28 verifica-se que na PC3 a quantidade de valores positivos e negativos é praticamente semelhante e, tal como se visualizou anteriormente para duas componentes principais, não ocorre variação entre os anos, meses e estações do ano.

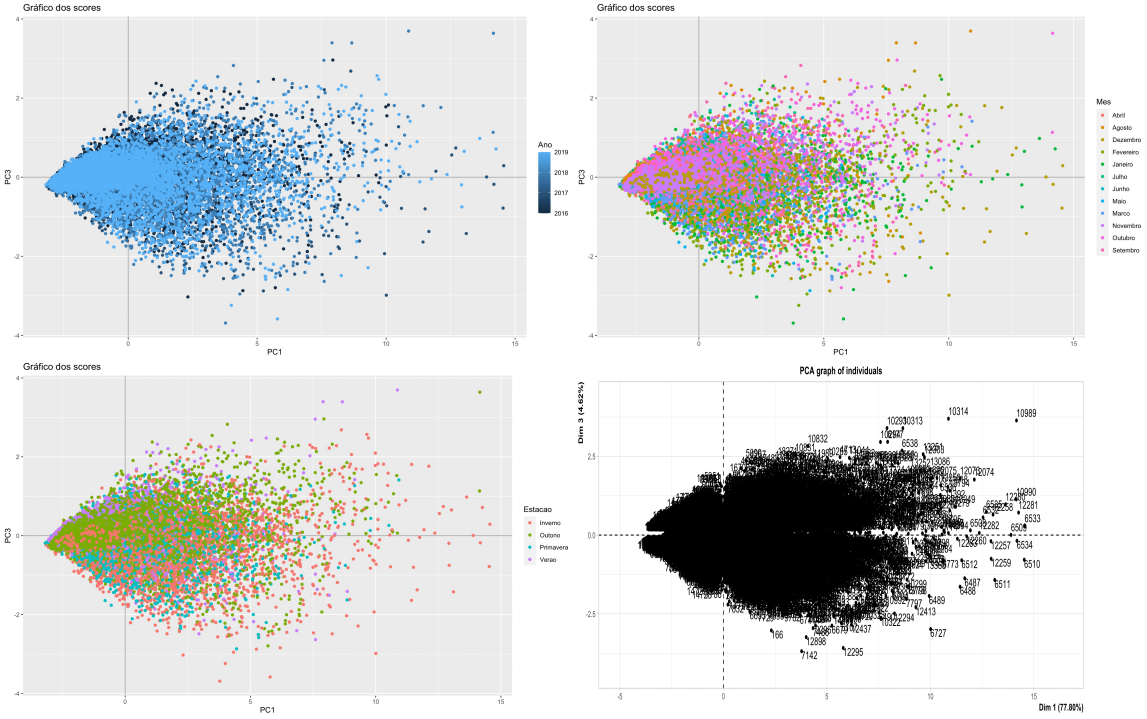

Figura 3.27: Gráfico de *scores* da PC1 e PC3 por ano, mês, estação do ano e observações.

Figura 3.28: Gráfico de *scores* da PC2 e PC3 por ano, mês, estações do ano e observações.

```
Loadings:
            RC1   RC3   RC2
Alfragide   0.405 0.807 0.320
AvLiberdade 0.249 0.304 0.904
Beato       0.799 0.319 0.404
Benfica     0.564 0.717 0.289
Entrecampos 0.827 0.433 0.127
Olivais     0.805 0.413 0.303
Restelo     0.450 0.649 0.464

                 RC1   RC3   RC2
SS loadings    2.719 2.139 1.489
Proportion Var 0.388 0.306 0.213
Cumulative Var 0.388 0.694 0.907
```

Figura 3.29: Rotação ortogonal VARIMAX com 3 componentes.

Com base no *output* da Figura 3.29, as componentes principais após a aplicação do método de rotação de VARIMAX são:

$Y_1 = 0.405Z_1 + 0.249Z_2 + 0.799Z_3 + 0.564Z_4 + 0.827Z_5 + 0.805Z_6 + 0.450Z_7$

$Y_3 = 0.807Z_1 + 0.304Z_2 + 0.319Z_3 + 0.717Z_4 + 0.433Z_5 + 0.413Z_6 + 0.649Z_7$

$Y_2 = 0.320Z_1 + 0.904Z_2 + 0.404Z_3 + 0.289Z_4 + 0.127Z_5 + 0.303Z_6 + 0.464Z_7$

Verifica-se que, com rotação, as variáveis na $1^{\underline{a}}$ componente que apresentam um maior peso são Z_3 (Beato), Z_4 (Benfica), Z_5 (Entrecampos) e Z_6 (Olivais).

Na $3^{\underline{a}}$ componente, a variável que tinha uma maior contribuição era Z_1 (Alfragide). Depois de se efetuar a rotação, verifica-se que a variável Z_1 (Alfragide) continua com um maior peso, contudo as variáveis Z_4 (Benfica) e Z_7 (Restelo) também apresentam pesos estatisticamente significativos.

Para a $2^{\underline{a}}$ componente, só a variável Z_2 (AvLiberdade) é que apresenta um maior peso para a componente.

Por fim, concluí-se que a $3^{\underline{a}}$ componente passa a explicar mais da variância total do que a $2^{\underline{a}}$ componente.

No entanto a proporção da variância total explicada por cada componente fica distribuída de forma mais equilibrada.

As conclusões a retirar das Figuras 3.30, 3.31 e 3.32 mantêm-se iguais às verificadas anteriormente, embora, em alguns casos, os valores estejam representados em diferentes quadrantes.

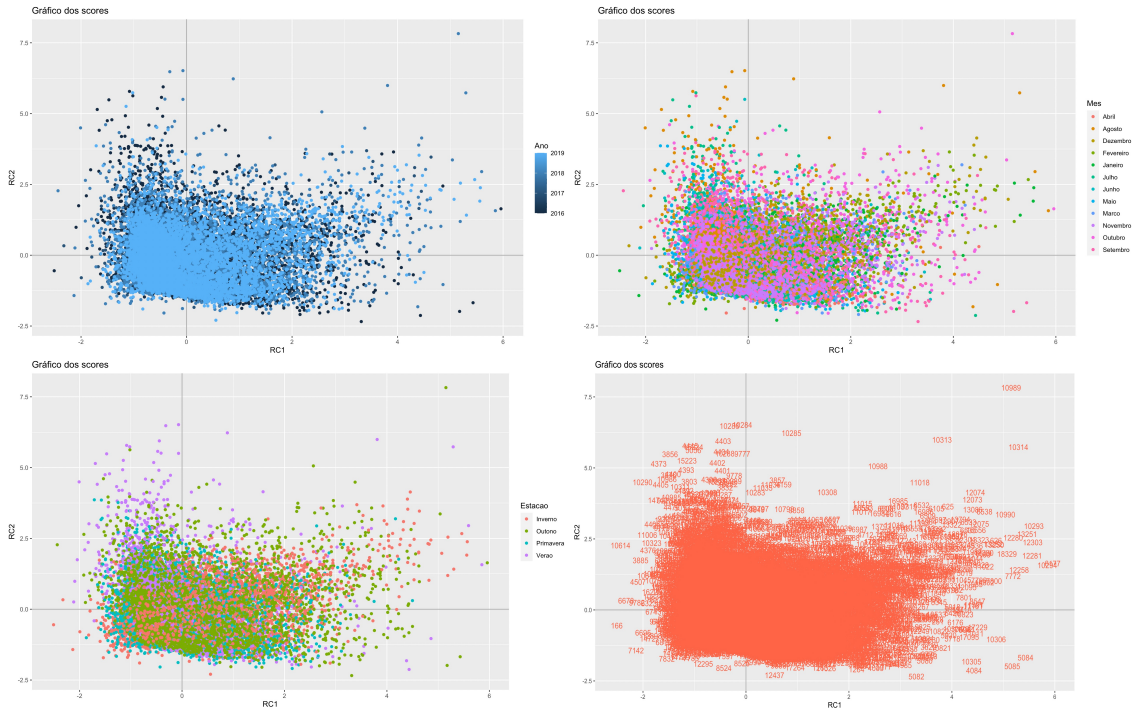

Figura 3.30: Gráficos de *scores* da RC1 e RC2 após rotação por, ano, mês, estação do ano e observações.

Figura 3.31: Gráficos de *scores* da RC1 e RC3 após rotação por, ano, mês, estação do ano e observações.

Figura 3.32: Gráficos de *scores* da RC2 e RC3 após rotação por, ano, mês, estação do ano e observações.

3.3 Análise de *Clusters*

A partir de um conjunto de dados, a AC constrói grupos, de forma a que os objetos do mesmo grupo (*cluster*) sejam mais semelhantes do que os objetos localizados em *clusters* diferentes. Por esse motivo, esta ferramenta é importante para agrupar EM com concentrações idênticas de poluentes atmosféricos.

Antes de prosseguir com a análise selecionaram-se os objetos/indivíduos, as variáveis quantitativas e estandardizaram-se os dados no *software* estatístico R.

Na análise de *clusters* as observações estão em linha e as variáveis em coluna, obtendo-se os *clusters* em função das observações. Uma vez que o pretendido é obter os *clusters* em função das localizações é necessário transpor a matriz Z.

A determinação do número adequado de *clusters* é uma etapa importante nesta análise. Como etapa preliminar representou-se o *scree plot* para auxiliar na escolha do número de *clusters*.

Analisando o *scree plot* da Figura 3.33 não é evidente o número de *clusters* a partir do qual o declive começa a ser reduzido, não sendo possível ter uma ideia precisa de quantos *clusters* se podem considerar na análise. Mais à frente tomaremos a decisão quanto ao número de *clusters* finais.

Para esta análise, é necessário calcular as distâncias entre as observações, tendo sido usada a distância euclidiana estandardizada.

Segundo o abordado na subseção 2.3.5, podem-se usar vários métodos de análise de *clusters* e, no final, comparar os resultados para determinar qual o melhor. Estes métodos subdividem-se em hierárquicos e métodos não hierárquicos. No que diz respeito aos métodos hiérarquicos só se utilizaram algoritmos aglomerativos.

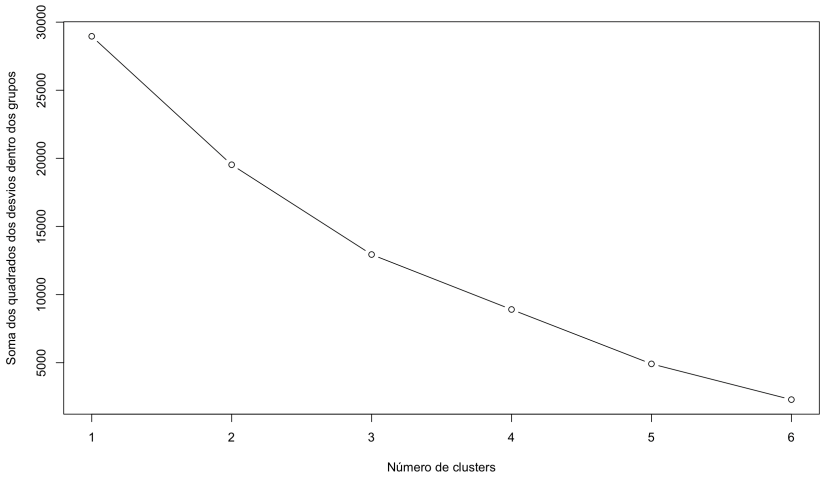

Figura 3.33: *Scree plot* para determinar o número de *clusters*.

- **Método hierárquico aglomerativo**

  Para realizar um algoritmo aglomerativo existem vários métodos como, o método da ligação simples, da ligação completa e da ligação média. Cada um destes pode fornecer resultados diferentes. Ainda é importante referir que dentro de cada método, usou-se a matriz de distância euclidiana estandardizada.

  – **Distância euclidiana e o método da ligação simples:**

  Neste método a dissemelhança entre dois grupos é a menor das dissemelhanças entre cada elemento de um grupo para o outro, ou seja, é determinada pelos indivíduos mais próximos. Assim, há uma tendência para que os grupos se juntem, formando outros maiores.

  Dada a incerteza relativa ao número adequado de *clusters* realizou-se, tal como na ACP, o estudo para $k = 2$ e $k = 3$.

  a) 2 *clusters*

  Foi obtido o dendrograma, a representação gráfica da análise de *clusters* e o gráfico de *silhouette*.

  Com base na Figura 3.34, os *clusters* são:

  ***Cluster* 1** - Alfragide, Beato, Benfica, Entrecampos, Olivais e Restelo;

  ***Cluster* 2** - AvLiberdade.

  Ao analisar a representação gráfica verifica-se que o *cluster* 1 apresenta uma fraca coesão interna, pois os indivíduos que o constituem estão afastados.

  Para validar a solução encontrada determinou-se o coeficiente de *silhouette*, que mede o quão bem uma observação está associada a um *cluster* e estima a distância média entre estes. Apesar do gráfico de *silhouette* (Figura 3.34) indicar diversos valores, o mais importante para esta análise corresponde ao coeficiente *silhouette* médio geral. Caso este coeficiente assuma um valor médio geral elevado significa que foi feito um bom agrupamento.

  No gráfico de *silhouette* observa-se que as EM parecem estar nos *clusters* corretos, dado que nenhuma tem valores negativos. Os dois *clusters* apresentam valores de coeficiente muito distintos, sendo que no *cluster* 2 é onde se verifica o valor zero. Isto acontece, pois esse *cluster* só contêm uma EM. Para além disso,

duas estações pertencentes ao *cluster* 1 apresentam um coeficiente inferior a 0.25, havendo assim a possibilidade de poderem pertencer a outro grupo, isto é, de estarem entre dois *clusters*. O coeficiente médio geral é muito baixo (0.22) logo esta solução não é de boa qualidade.

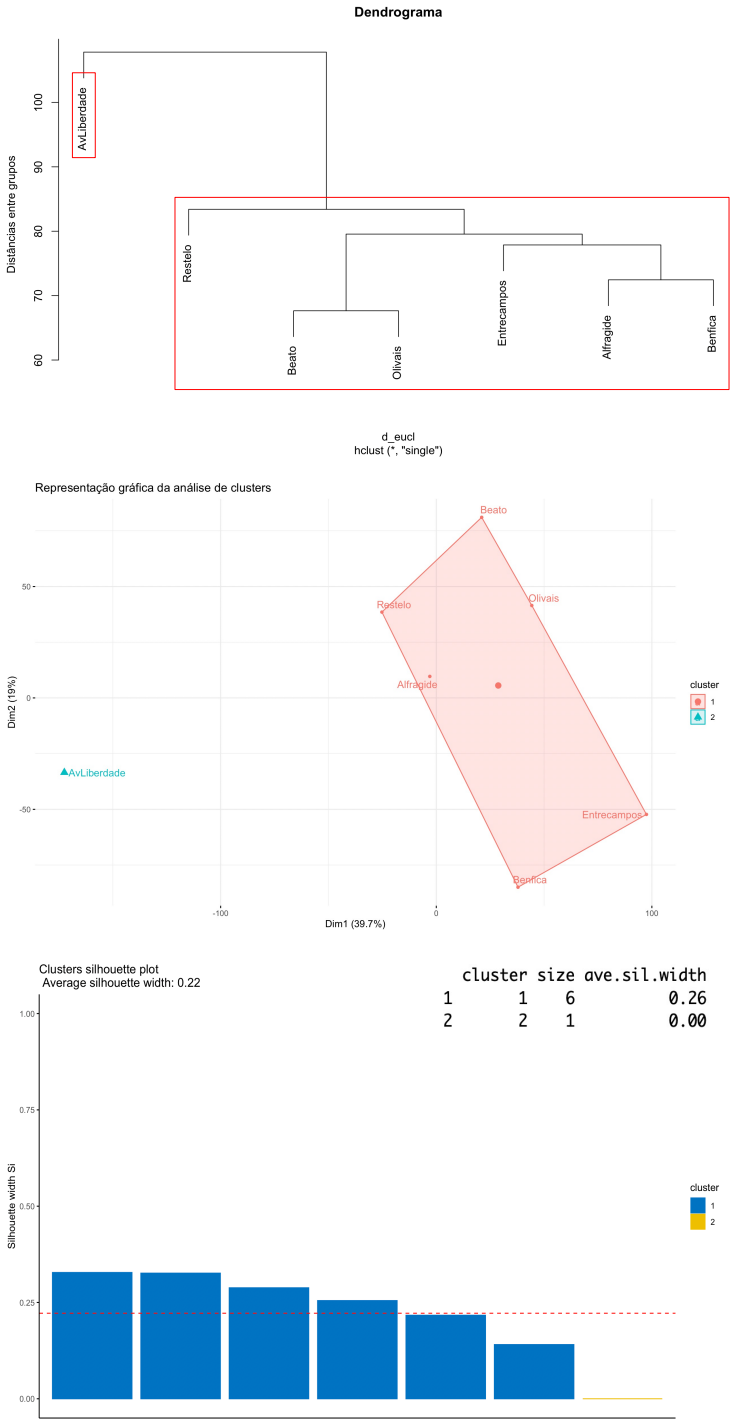

Figura 3.34: Dendrograma, Representação de *clusters* e Gráfico de *silhouette* para $k = 2$ com método da ligação simples.

b) AC com 3 *clusters*

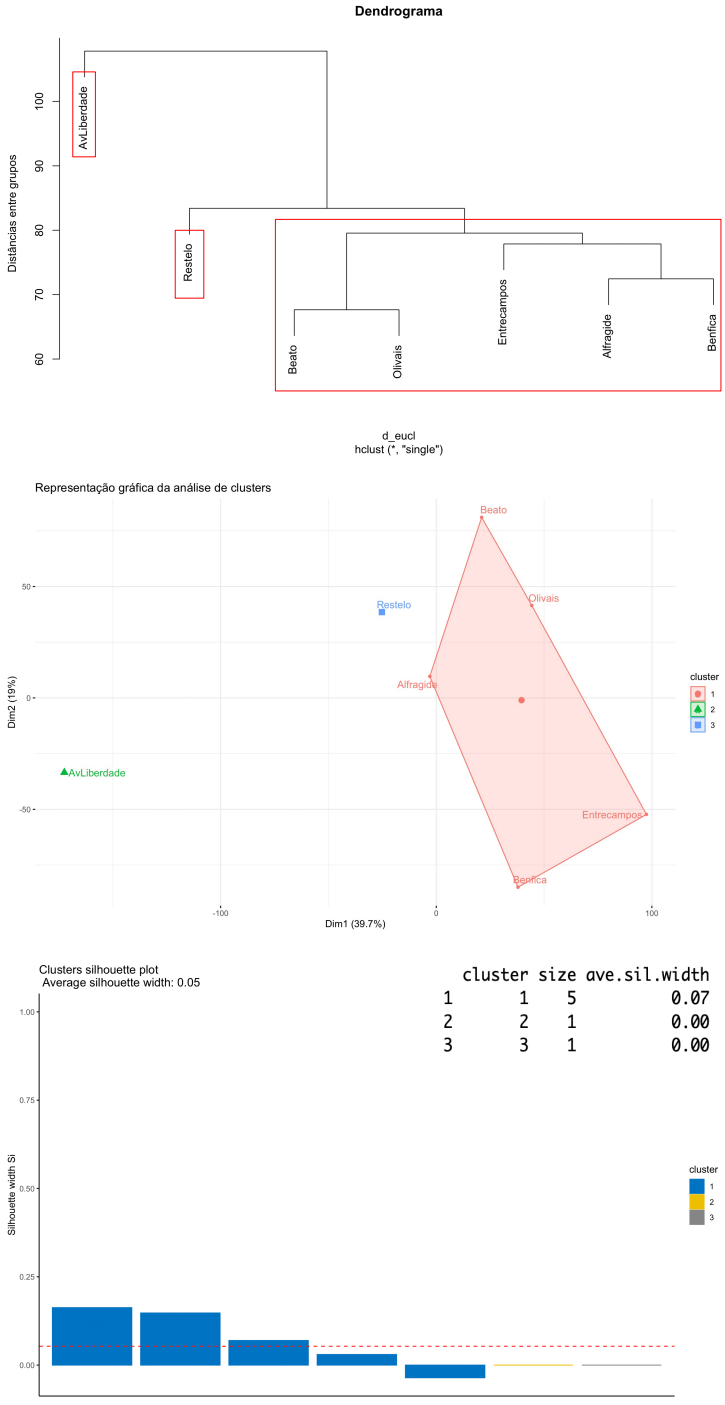

Figura 3.35: Dendrograma, Representação de *clusters* e Gráfico de *silhouette* para $k = 3$ com método da ligação simples.

Através do dendrograma e da representação de *clusters* da Figura 3.35 concluí-se que a única diferença entre estes e os anteriores foi a separação da EM do Restelo para outro grupo diferente. Pode-se verificar que o *cluster* 2 é o que se encontra mais afastado dos restantes. Todavia, continua a existir

uma reduzida coesão interna no maior grupo devido à elevada distância entre as diversas EM.

Os *clusters* são:

***Cluster* 1** - Alfragide, Beato, Olivais, Entrecampos e Benfica;

***Cluster* 2** - AvLiberdade;

***Cluster* 3** - Restelo.

Com o gráfico de *silhouette* (Figura 3.35) verifica-se que no *cluster* 1 encontram-se duas EM provavelmente mal classificadas, pois uma apresenta um coeficiente negativo e a outra está abaixo da média dos coeficientes de *silhouette* (linha vermelha a tracejado). Os *clusters* 2 e 3 têm coeficiente com valor zero, porque contêm somente uma EM. Ainda se verifica que todos os grupos apresentam um coeficiente inferior a 0.25 e que o coeficiente de *silhouette* médio piorou (0.05).

O coeficiente de correlação cofenética consiste em outro método para validar as soluções encontradas e é igual independentemente do número de *clusters* usado, pois avalia a solução obtida pelo método aplicado e não pelo número de *clusters* que se decide formar. Este coeficiente assume um valor de 0.872 o que significa, que este método de ligação simples é adequado, pois, apesar das conclusões anteriores, a solução preserva as distâncias originais na solução encontrada.

Através da análise dos diversos gráficos podem surgir dúvidas relativamente à colocação de algumas EM em determinados *clusters*, contudo, apesar do coeficiente de *silhouette* ser menor parece que a solução $k = 3$ faz mais sentido no contexto do estudo e que os *clusters* obtidos são mais coesos.

– **Distância euclidiana e o método da ligação completa:**

Este método serve-se dos dois elementos mais afastados para derivar a medida de proximidade entre os grupos. Há assim a tendência para grupos grandes não crescerem mais e os pequenos formarem grupos maiores.

a) 2 *clusters*

A constituição dos *clusters* e o gráfico de *silhouette* é igual para quando se considerou $k = 2$ no método da ligação simples, logo as conclusões são as mesmas (Figura 3.36). Assim, verifica-se que a solução não é de boa qualidade e onde:

***Cluster* 1** - Alfragide, Beato, Benfica, Entrecampos, Olivais e Restelo;

***Cluster* 2** - AvLiberdade.

a) 3 *clusters*

No dendrograma e na representação de *clusters* da Figura 3.37 observa-se que o agrupamento dos *clusters* é totalmente diferente do que foi previamente analisado. Os *clusters* são:

***Cluster* 1** - Restelo, Alfragide e Benfica;

***Cluster* 2** - AvLiberdade;

***Cluster* 3** - Beato, Olivais e Entrecampos.

Apesar de continuar a existir um grande afastamento dos indivíduos contidos nos *clusters* 1 e 2, estes demonstram ser mais coesos do que no método da ligação simples.

Com o gráfico de *silhouette* (Figura 3.37) concluí-se que, como não existem valores negativos, todas as estações parecem estar no *cluster* correto. Contudo, no *cluster* 1, uma EM poderá estar mal classificada, pois encontra-se abaixo da média dos coeficientes de *silhouette* (linha vermelha a tracejado). O *cluster* 2

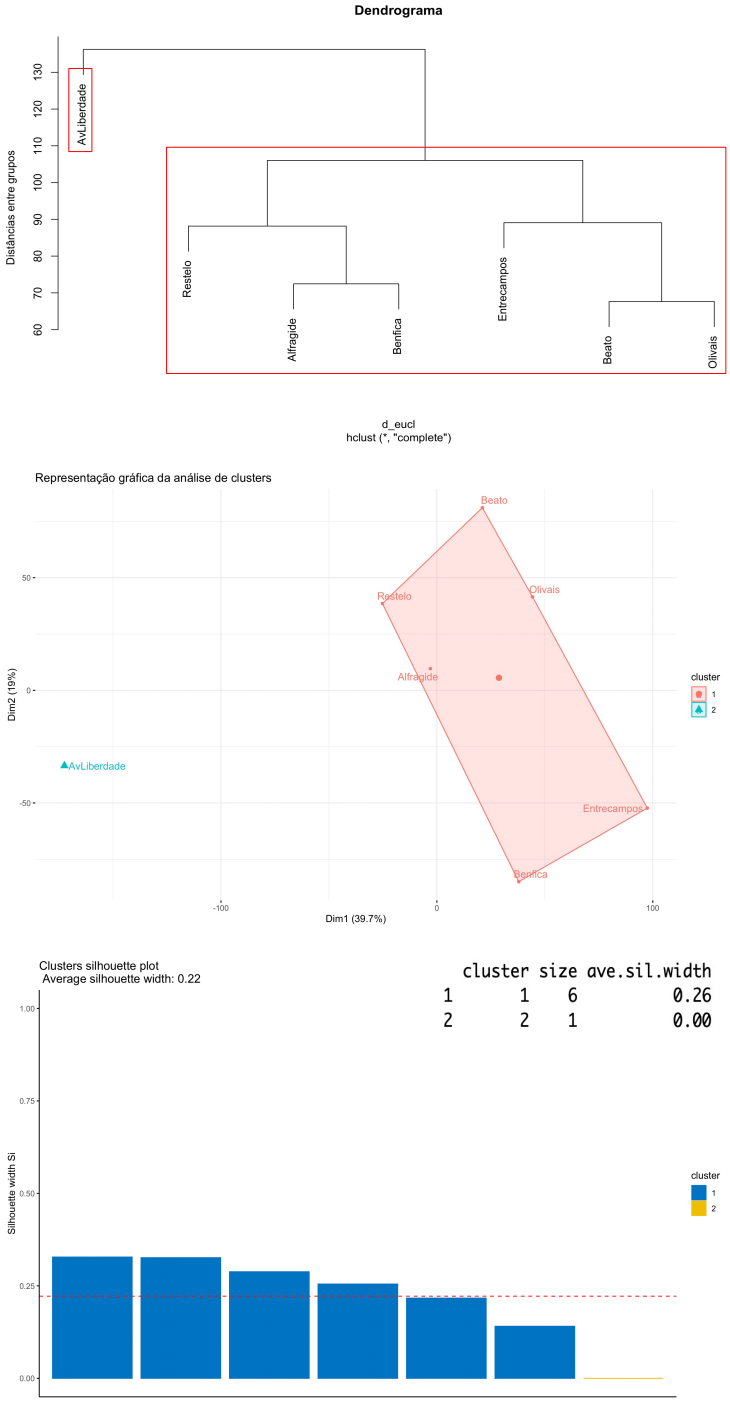

Figura 3.36: Dendrograma, Representação de *clusters* e Gráfico de *silhouette* para $k = 2$ com método da ligação completa.

têm um coeficiente igual a zero. Ainda é possível destacar que todos os grupos apresentam um coeficiente inferior a 0.25. O coeficiente de *silhouette* médio corresponde ao valor de 0.12, concluindo que a solução não é de boa qualidade.

Com o auxílio do *software* R, obteve-se um valor de 0.889 para o coeficiente de

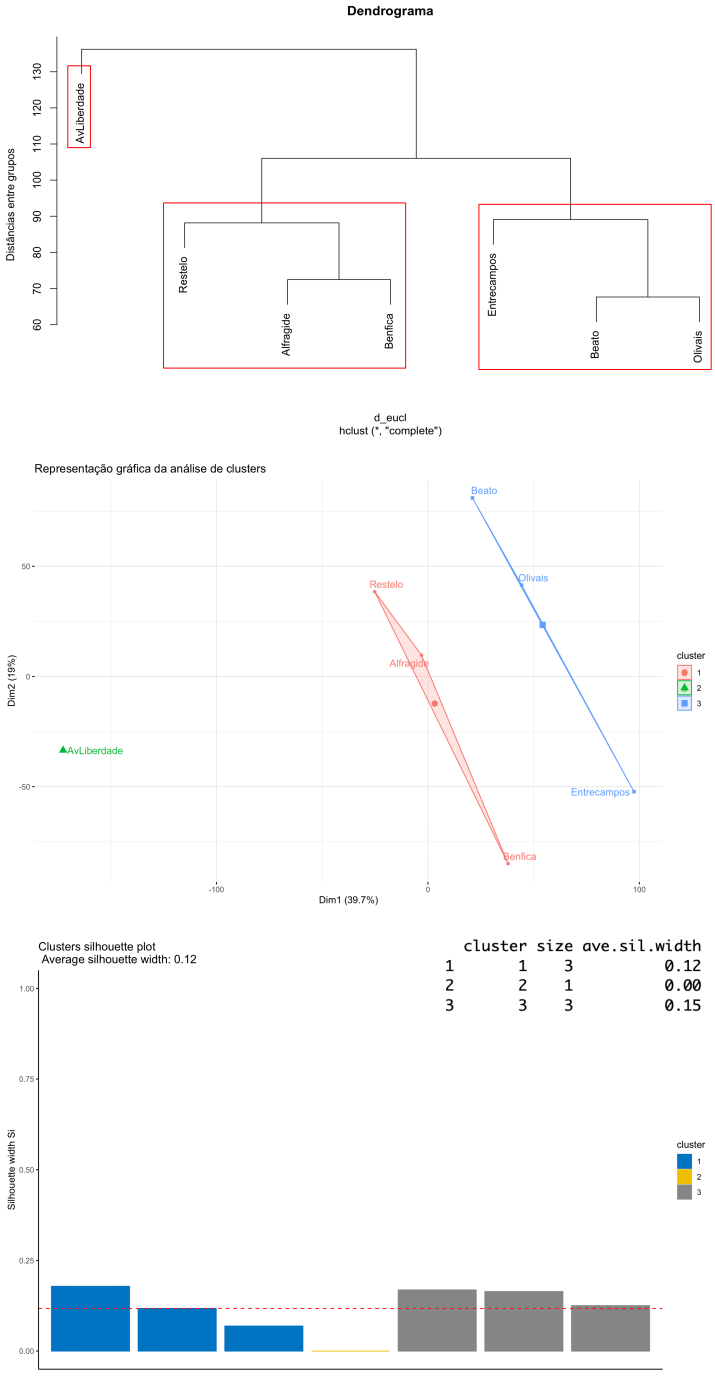

Figura 3.37: Dendrograma, Representação de *clusters* e Gráfico de *silhouette* para $k = 3$ com método da ligação completa.

cofenética. Sendo maior do que o coeficiente do método anterior, pode-se concluir que dos dois métodos, este permitiu obter um melhor resultado.

Através da análise dos diversos gráficos podem surgir dúvidas relativamente à colocação de alguns EM em determinados *clusters*, contudo, apesar do coeficiente de *silhouette* ser menor parece que a solução $k = 3$ faz mais sentido no contexto do

estudo e que os *clusters* obtidos são mais coesos.

– **Distância euclidiana e o método da ligação média:**

No método da ligação média a dissemelhança entre dois grupos é a média das distâncias entre todos os pares de observações.

a) 2 *clusters*

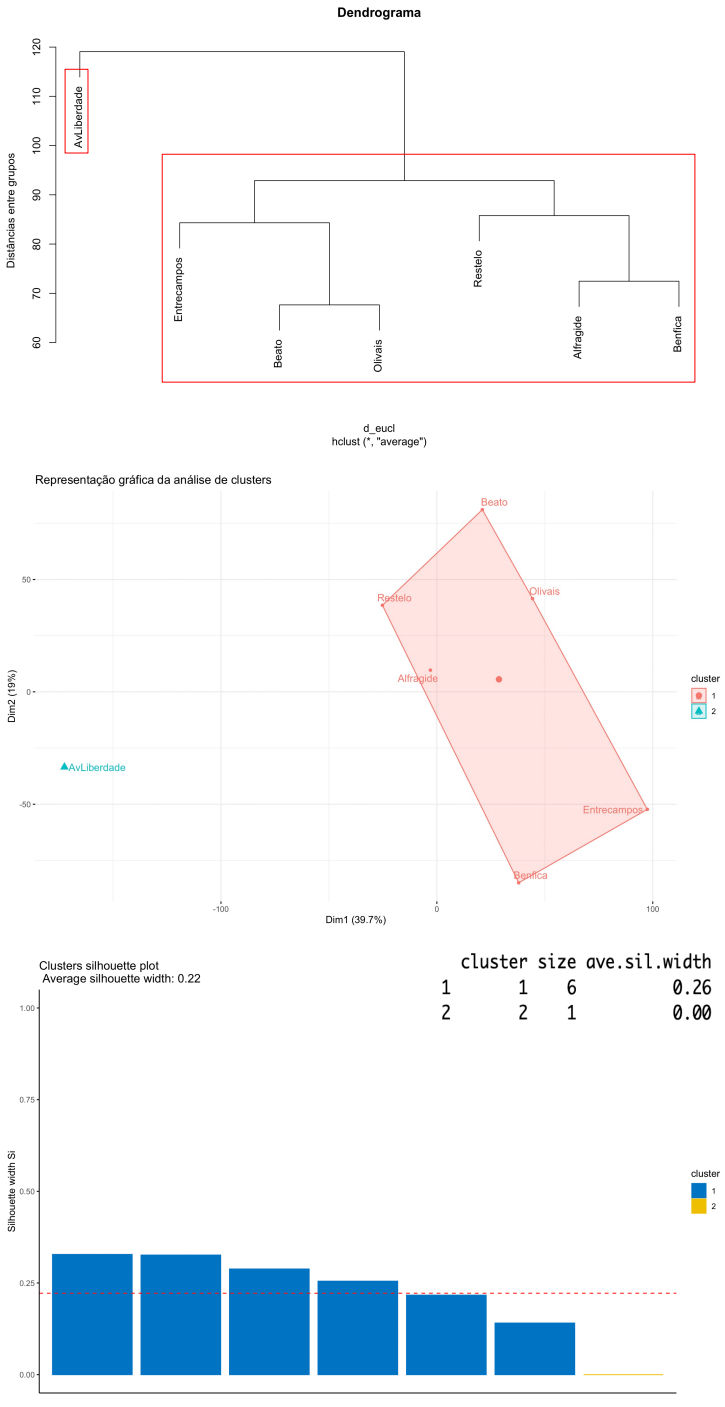

Figura 3.38: Dendrograma, Representação de *clusters* e Gráfico de *silhouette* para $k = 2$ com método da ligação média.

No método de ligação média, os dois *clusters* são:
***Cluster* 1** - Alfragide, Beato, Benfica, Entrecampos, Olivais e Restelo;
***Cluster* 2** - AvLiberdade.

Visto que os *clusters* são os mesmos que nos método da ligação simples e completa para $k = 2$, as conclusões são precisamente as mesmas (Figura 3.38). Para além disso, as explicações acerca do gráfico de *silhouette* são as mesmas, verificando-se que a solução não é de boa qualidade.

a) 3 *clusters*

O dendrograma e a representação de *clusters* da Figura 3.39 são idênticos aos apresentados no método da ligação completa para $k = 3$.

Os *clusters* são:
***Cluster* 1** - Restelo, Alfragide e Benfica;
***Cluster* 2** - AvLiberdade;
***Cluster* 3** - Beato, Olivais e Entrecampos.

Desta forma, as interpretações encontradas nos gráficos do método da ligação completa são precisamente as mesmas para os do método da ligação média. Ou seja, mesmo com uma divisão mais clara dos grupos pode-se ainda observar a fraca coesão entre os *clusters* 1 e 2, a possível classificação incorrecta de uma EM no *cluster* 1, a dúvida sobre onde deve ser incluído o *cluster* 2 e o valor reduzido do coeficiente de *silhouette* médio (0.12).

De todos os métodos realizados, este foi o que apresentou um maior valor de coeficiente de cofenética (0.897), permitindo assim obter um melhor resultado.

Apesar de ter um coeficiente de *silhouette* médio menor, analisando os outros dois gráficos, concluí-se que a solução $k = 3$ faz mais sentido no contexto do estudo, para além dos *clusters* obtidos parecerem mais coesos.

- **AC usando um método não hierárquico:**

Os métodos não hierárquicos distinguem-se por terem diferentes princípios e cujos resultados não constituem hierarquias.

O método de partição *k-means* exige que o número de grupos seja fixado à partida e analisa algumas partições, extraindo a melhor. Os resultados provenientes deste método podem ser afetados pela seleção inicial dos grupos e pela presença de *outliers*.

Tendo como referência a análise realizada com o método hierárquico, fixaram-se previamente dois e três *clusters*.

a) 2 *clusters*

A partir da Figura 3.40 podem-se escrever os seguintes *clusters*:
***Cluster* 1** - Alfragide, Beato, Benfica, Entrecampos, Olivais e Restelo;
***Cluster* 2** - AvLiberdade.

Averigua-se que os grupos formados são iguais aos vistos anteriormente no método da ligação simples, completa e média para $k = 2$, concluindo que o *cluster* 1 não apresenta uma boa coesão interna devido ao afastamento das observações. Com o diagrama de *silhouette* (Figura 3.40) podem-se retirar precisamente as mesmas conclusões dos métodos supramencionados, ou seja, as EM estão classificadas no *cluster* correto já que nenhum apresenta um valor de *silhouette* negativo. Porém, permanece a dúvida relativamente ao *cluster* 2 pois este só contêm uma EM.

A soma de quadrados entre os grupos é de 32.6%.

a) 3 *clusters*

Averigua-se na Figura 3.41 que os *clusters* são:

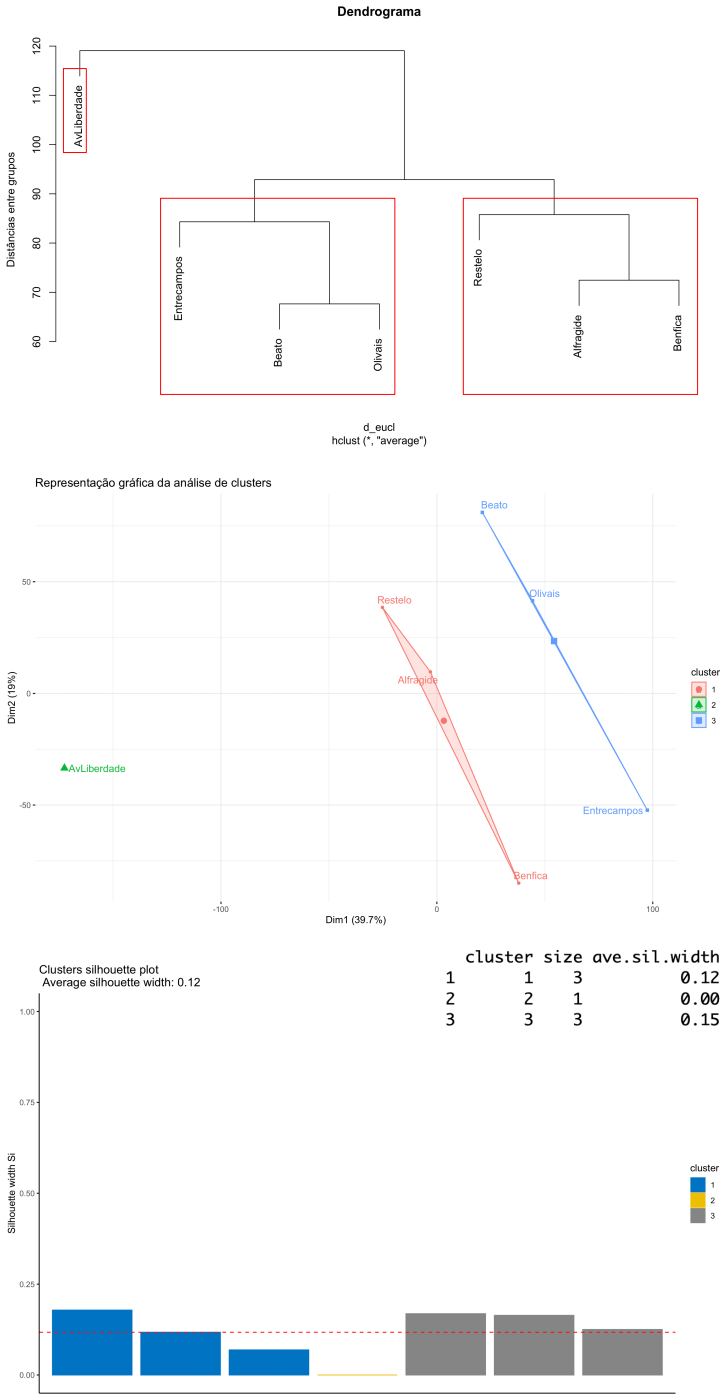

Figura 3.39: Dendrograma, Representação de *clusters* e Gráfico de *silhouette* para $k = 3$ com método da ligação média.

***Cluster* 1** - Beato, Olivais e Entrecampos;
***Cluster* 2** - AvLiberdade;
***Cluster* 3** - Restelo, Alfragide e Benfica.

É possível constatar que os *clusters* são os mesmos que no método da ligação completa e média para $k = 3$ logo as soluções retiradas são idênticas. Assim,

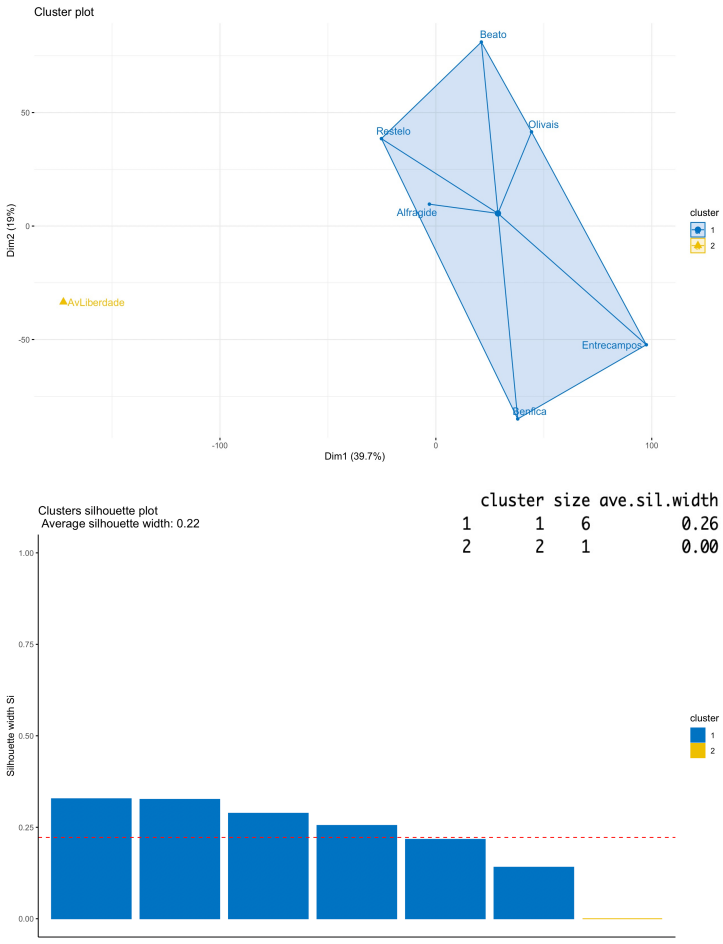

Figura 3.40: Representação de *clusters* e Gráfico de *silhouette* para $k = 2$ com método *k-means*.

concluí-se que em comparação com $k = 2$, os grupos encontram-se bem separados. Contudo, o *cluster* 1 contêm somente uma EM. Para além disso, uma das estações pertencentes ao *cluster* 3, poderá estar mal classificada ou entre dois *clusters*, pois encontram-se abaixo da média do valor de *silhouette*.

A soma de quadrados entre os grupos é de 55.3%.

- **Conclusão dos resultados da AC**

  Para ter conhecimento de qual o melhor tipo de método (hierárquico ou não hierárquico) compararam-se os resultados.

  Os resultados demonstram ser consistentes, uma vez que a composição dos grupos e o coeficiente de *silhouette* na ligação completa, média e *k-means* são muito parecidos. Contudo, no método da ligação simples, os três grupos são diferentes dos apresentados nos outros métodos e com um coeficiente de *silhouette* menor.

  Com base nos valores do coeficiente de cofenética, os melhores resultados (com um valor de 0.898) são obtidos com o método da ligação média com recurso à distância euclidiana quando se considera $k = 3$.

  O método de particionamento das $k - means$, analogamente aos métodos hierárquicos, apresenta resultados semelhantes, mas melhores para $k = 3$, que tem um valor estatisticamente mais significativo na soma de quadrados entre grupos (55.3%) quando

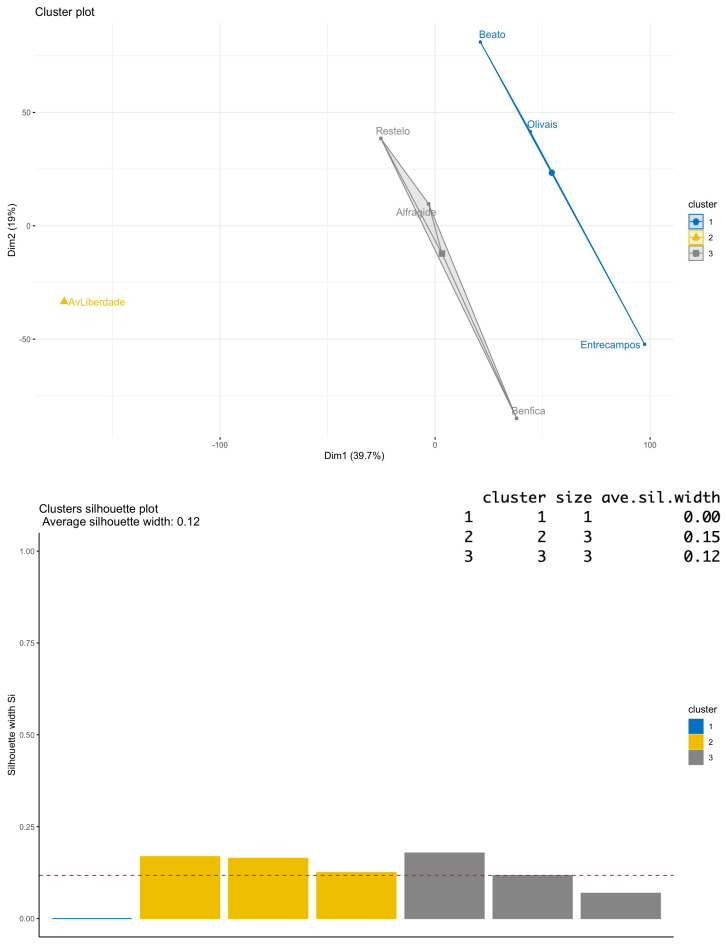

Figura 3.41: Representação de *clusters* e Gráfico de *silhouette* para $k = 3$ com método *k-means*.

comparado com $k = 2$ *clusters* (32.6%).

Observa-se que a única diferença entre o método da ligação média e o método *k-means* é a ordem dos *clusters*. Como solução final deste capítulo escolheu-se a ligação média para $k = 3$ e concluiu-se que as EM agrupadas em cada um dos *clusters* correspondem às EM com maior peso em cada componente principal obtida após rotação (RC).

Assim, verifica-se que o *cluster* 1 contém as estações do Restelo, Alfragide e Benfica, o *cluster* 2 a estação da AvLiberdade e o *cluster* 3 as estações de Beato, Olivais e Entrecampos.

Conclui-se que estes resultados são semelhantes aos obtidos para ACP, pois as estações com maior peso na RC1 corresponde ao Cl.3, na RC2 ao Cl.2 e na RC3 ao Cl.1.

3.4 Regressão Linear

Pretende-se usar a regressão linear para obter estimativas da concentração de NO_2, para observações que se encontram em falta, usando EM redundantes. Para além disso, é possível prever o comportamento de alguma EM redundante que futuramente seja encerrada.

Os resultados obtidos nas subseções anteriores apontam para a existência de dois grupos de EM redundantes: Restelo, Alfragide, Benfica; Beato, Olivais e Entrecampos. Com estes

dois grupos realizou-se o estudo de regressão linear múltipla.

Como se pode verificar nos dendrogramas das Figuras da subseção 3.3 em cada um dos grupos existem EM que parecem ter maiores semelhanças entre si nomeadamente, as estações de Alfragide e Benfica; Beato e Olivais. Com estas estações realizou-se o estudo da regressão linear simples.

Primeiramente foram obtidos os coeficientes do modelo de regressão ajustado da variável Y sobre a x e escritas as equações matemáticas da Tabela 3.5. As expressões da variável Y sobre a x_1 e a x_2 são apresentadas na Tabela 3.6.

Estas expressões matemáticas seguem um dos seguintes modelos:

$$\hat{y}_i = a + bx_i$$

e

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}.$$

De seguida, determinaram-se os valores dos coeficientes de determinação.

Tabela 3.5: Expressões matemáticas dos modelos de regressão simples e os respetivos coeficientes de determinação (r^2)

Modelos de regressão simples	r^2
Alfragide $= -5.830 + 0.965\times$Benfica	0.735
Benfica $= 14.124 + 0.761\times$Alfragide	0.735
Beato $= 1.282 + 0.712\times$Olivais	0.766
Olivais $= 5.510 + 1.077\times$Beato	0.766

Através destas retas preveem-se os valores de Y com base nos valores de X. Contudo não é possível usar a mesma equação da reta para estimar o comportamento de X a partir dos valores de Y quando não existe uma correlação perfeita entre as variáveis ($r \neq \pm 1$), pois nesse caso as duas retas não coincidem, sendo concorrentes.

Recorreu-se à técnica de *stepwise* para se selecionar as variáveis independentes a serem usadas nos modelos de regressão linear múltipla. A Figura 3.42 permite exemplificar os *outputs* obtidos no R, e onde se encontra sinalizado o menor valor de AIC.

Após análise, escolheu-se a direção onde se encontrava o menor valor de AIC e escreveram-se as equações de regressão com os dados presentes nesses *outputs*.

Verificou-se que em todas as direções, o menor AIC era obtido quando se consideravam as duas variáveis independentes em estudo. Assim sendo, os resultados são consistentes em todas as direções e os melhores modelos são os apresentados na Tabela 3.6.

Posteriormente, analisaram-se os pressupostos associados aos resíduos. Primeiramente, através do gráfico da Figura 3.43, verificou-se que os erros são provenientes de uma população com distribuição diferente da distribuição normal dado que, o conjunto de pontos obtidos não forma uma reta.

Continuamente, e de forma a comprovar o resultado do gráfico anterior, realizou-se o teste de ajustamento KSL. Tanto na regressão simples como na regressão múltipla verificou-se, com um nível de significância de 5%, que se deve rejeitar H_0, uma vez que $D_{observado} \geq$

Tabela 3.6: Expressões matemáticas dos modelos de regressão múltipla e os respetivos coeficientes de determinação ajustados $\left(r^2_{ajust}\right)$

Modelos de regressão múltipla	$r^2_{ajustado}$
Entrecampos $= 10.494 + 0.357\times$ Beato $+0.690\times$ Olivais	0.700
Beato $= -0.190 + 0.134\times$ Entrecampos $+0.585\times$ Olivais	0.778
Olivais $= 0.910 + 0.723\times$Beato $+0.322\times$ Entrecampos	0.818
Restelo $= 0.720 + 0.221\times$Alfragide $+0.359\times$ Benfica	0.690
Alfragide $= -5.582 + 0.436\times$Restelo $+0.716\times$ Benfica	0.760
Benfica $= 11.201 + 0.512\times$Alfragide $+0.505\times$ Restelo	0.783

```
> step(lm(Entrecampos~1), direction="forward", scope=~Beato+Olivais)
Start:  AIC=121969.6
Entrecampos ~ 1

          Df Sum of Sq      RSS    AIC
+ Olivais  1   9597202  4407558 100723
+ Beato    1   8610394  5394365 104436
<none>                 14004760 121970

Step:  AIC=100723
Entrecampos ~ Olivais

        Df Sum of Sq     RSS    AIC
+ Beato  1    211245 4196313  99822
<none>               4407558 100723

Step:  AIC=99822.24
Entrecampos ~ Olivais + Beato


Call:
lm(formula = Entrecampos ~ Olivais + Beato)

Coefficients:
(Intercept)      Olivais        Beato
    10.4942       0.6903       0.3566
```

```
> step(lm(Entrecampos~Beato+Olivais), direction="backward")
Start:  AIC=99822.24
Entrecampos ~ Beato + Olivais

          Df Sum of Sq     RSS    AIC
<none>                 4196313  99822
- Beato    1    211245 4407558 100723
- Olivais  1   1198052 5394365 104436

Call:
lm(formula = Entrecampos ~ Beato + Olivais)

Coefficients:
(Intercept)        Beato      Olivais
    10.4942       0.3566       0.6903


> step(lm(Entrecampos~Beato+Olivais), direction="both")
Start:  AIC=99822.24
Entrecampos ~ Beato + Olivais

          Df Sum of Sq     RSS    AIC
<none>                 4196313  99822
- Beato    1    211245 4407558 100723
- Olivais  1   1198052 5394365 104436

Call:
lm(formula = Entrecampos ~ Beato + Olivais)

Coefficients:
(Intercept)        Beato      Olivais
    10.4942       0.3566       0.6903
```

Figura 3.42: *Output* da técnica stepwise, nas 3 direções, relativo ao modelo Entrecampos $\sim$ Beato + Olivais.

$D_{crtico,\alpha} = 0.00654$ $\left(\alpha = 0.05 \geq p - value = 2.2 \times 10^{-16}\right)$. Assim, concluí-se que os erros não são normalmente distribuídos (Figuras 3.43, 3.44, e 3.45).

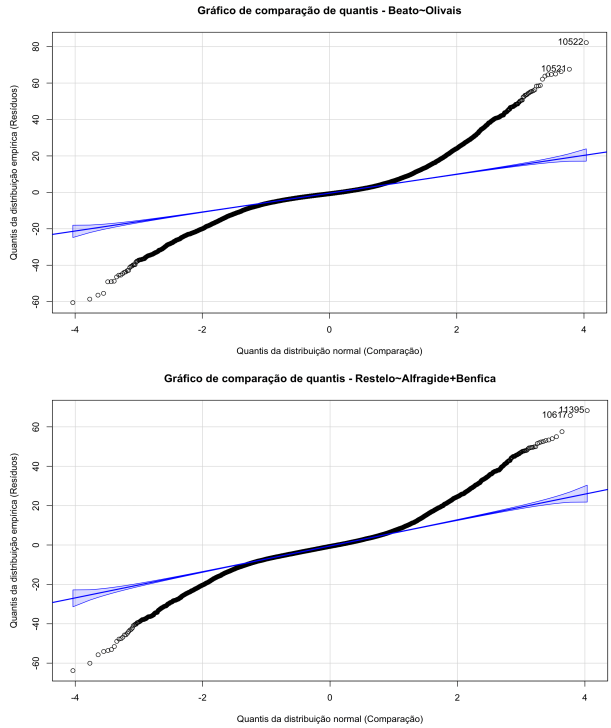

Figura 3.43: Gráficos da normalidade dos modelos Beato ∼ Olivais e Restelo ∼ Alfragide + Benfica.

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(lm(Alfragide ~ Benfica))
D = 0.11044, p-value < 2.2e-16

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(lm(Benfica ~ Alfragide))
D = 0.063509, p-value < 2.2e-16

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(lm(Beato ~ Olivais))
D = 0.1242, p-value < 2.2e-16

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(lm(Olivais ~ Beato))
D = 0.11928, p-value < 2.2e-16
```

Figura 3.44: *Outputs* do teste KSL para a regressão simples.

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(lm(Entrecampos ~ Beato + Olivais))
D = 0.11424, p-value < 2.2e-16

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(lm(Beato ~ Entrecampos + Olivais))
D = 0.11667, p-value < 2.2e-16

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(lm(Olivais ~ Beato + Entrecampos))
D = 0.12031, p-value < 2.2e-16

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(lm(Restelo ~ Alfragide + Benfica))
D = 0.095541, p-value < 2.2e-16

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(lm(Alfragide ~ Restelo + Benfica))
D = 0.13068, p-value < 2.2e-16

    Lilliefors (Kolmogorov-Smirnov) normality test

data:  resid(lm(Benfica ~ Alfragide + Restelo))
D = 0.057934, p-value < 2.2e-16
```

Figura 3.45: *Outputs* do teste KSL para a regressão múltipla.

A verificação dos pressupostos é importante, visto que a inferência estatística no modelo de regressão linear se baseia nesses pressupostos. Assim, se houver violação dos mesmos, como é o caso, a utilização do modelo deve ser vista com algum cuidado, pois poderá conduzir a conclusões que não sejam totalmente fiáveis. Poderá ser interessante considerar outros modelos ou confirmar os resultados obtidos com os modelos de regressão usando novas leituras do poluente nas respetivas EM e verificar se o valor da previsão obtida através do modelo é próximo do valor real.

Apesar da violação deste pressuposto verificou-se em todos os modelos uma média dos erros aproximadamente igual a zero e um gráfico de resíduos (e_i) versus *fitted values* ($\hat{y}_i$) com uma mancha de pontos dispersos aleatoriamente numa faixa horizontal centrada no eixo das abcissas, ou seja, torno da reta $y = 0$, sem nenhum comportamento ou tendência. Como exemplo apresentam-se os gráficos para os modelos Beato $\sim$ Olivais e Restelo $\sim$ Alfragide + Benfica (Figura 3.46).

Através da análise destes gráficos existem indícios de que a variância dos erros é constante, ou seja, existe homocedasticidade dos erros.

Tendo em conta os resultados obtidos anteriormente decidiu-se manter estes modelos e testar a sua significância.

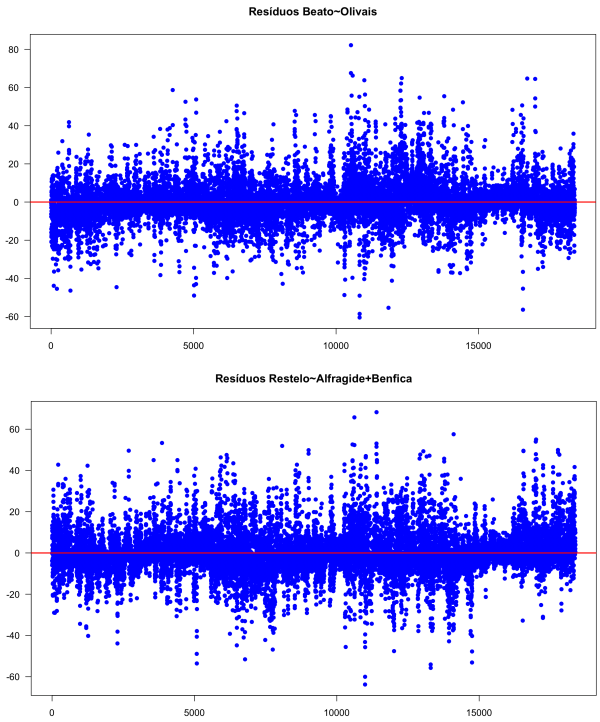

Figura 3.46: Gráficos de resíduos dos modelos Beato $\sim$ Olivais e Restelo $\sim$ Alfragide + Benfica.

Para a regressão linear simples foi testada a significância do modelo, verificando-se, através dos *outputs* da Figura 3.47, que existe relação entre as variáveis, ou seja, o modelo é significativo, dado que o valor das estatísticas de teste dos testes de significância do modelo pertencem à região crítica $\left(\alpha = 0.05 \geq p - value = 2.2 \times 10^{-16}\right)$.

```
Call:
lm(formula = Alfragide ~ Benfica)

Residuals:
     Min       1Q   Median       3Q      Max
-112.656   -7.654   -0.419    5.728  148.091

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.830279   0.189734  -30.73   <2e-16 ***
Benfica      0.965208   0.004278  225.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.6 on 18378 degrees of freedom
Multiple R-squared:  0.7348,  Adjusted R-squared:  0.7347
F-statistic: 5.091e+04 on 1 and 18378 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Beato ~ Olivais)

Residuals:
    Min      1Q  Median      3Q     Max
-60.520  -3.951  -0.747   3.055  82.207

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.281577   0.110579   11.59   <2e-16 ***
Olivais     0.711648   0.002897  245.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.507 on 18378 degrees of freedom
Multiple R-squared:  0.7665,  Adjusted R-squared:  0.7665
F-statistic: 6.033e+04 on 1 and 18378 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Benfica ~ Alfragide)

Residuals:
    Min      1Q  Median      3Q     Max
-93.434  -8.602  -1.351   7.003 120.395

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.124195   0.137825   102.5   <2e-16 ***
Alfragide    0.761243   0.003374   225.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.97 on 18378 degrees of freedom
Multiple R-squared:  0.7348,  Adjusted R-squared:  0.7347
F-statistic: 5.091e+04 on 1 and 18378 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Olivais ~ Beato)

Residuals:
    Min      1Q  Median      3Q     Max
-87.589  -5.680  -1.831   4.096  88.522

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.510017   0.130346   42.27   <2e-16 ***
Beato       1.077090   0.004385  245.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.7 on 18378 degrees of freedom
Multiple R-squared:  0.7665,  Adjusted R-squared:  0.7665
F-statistic: 6.033e+04 on 1 and 18378 DF,  p-value: < 2.2e-16
```

Figura 3.47: *Outputs* dos testes paramétricos para a significância dos modelos de regressão linear simples.

Com recurso aos *outputs* da Figura 3.48 observa-se, que $F_{observado} \geq F(2, 18377, 0.95) = 2.9966$ $\left(\alpha = 0.05 \geq p - value = 2.2 \times 10^{-16}\right)$ em qualquer um dos modelos, logo deve-se rejeitar H_0 do teste à significância dos modelos, concluindo-se que os modelos são globalmente significativos. Desta forma, sabe-se que, pelo menos uma variável x_i contribuí para explicar a variação da variável Y.

Usando o mesmo *output* concluí-se que se deve rejeitar H_0 nos testes à significância de cada um dos parâmetros do modelo, pois o valor da estatística de teste associada a cada teste pertence à região crítica $(\alpha = 0.05 \geq p - value)$ Assim, os parâmetros (β_1, β_2) são significativos e as variáveis independentes contribuem para explicar a variação de Y. Ou seja, por exemplo, as concentrações do poluente nas estações do Beato e Olivais contribuem para explicar a variação das concentrações do poluente na EM de Entrecampos.

```
Call:
lm(formula = Entrecampos ~ Beato + Olivais)

Residuals:
    Min      1Q  Median      3Q     Max
-60.858  -9.480  -3.766   6.766 138.935

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.49416    0.17640   59.49   <2e-16 ***
Beato        0.35661    0.01172   30.42   <2e-16 ***
Olivais      0.69032    0.00953   72.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.11 on 18377 degrees of freedom
Multiple R-squared:  0.7004,  Adjusted R-squared:  0.7003
F-statistic: 2.148e+04 on 2 and 18377 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Beato ~ Entrecampos + Olivais)

Residuals:
    Min      1Q  Median      3Q     Max
-58.625  -4.297  -0.453   3.218  84.354

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.190231   0.118254  -1.609    0.108
Entrecampos  0.134397   0.004419  30.416   <2e-16 ***
Olivais      0.584764   0.005039 116.038   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.277 on 18377 degrees of freedom
Multiple R-squared:  0.7777,  Adjusted R-squared:  0.7777
F-statistic: 3.215e+04 on 2 and 18377 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Olivais ~ Beato + Entrecampos)

Residuals:
    Min      1Q  Median      3Q     Max
-74.132  -4.898  -0.833   3.425  75.506

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.910042   0.131341   6.929 4.38e-12 ***
Beato       0.723143   0.006232 116.038  < 2e-16 ***
Entrecampos 0.321725   0.004442  72.434  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.32 on 18377 degrees of freedom
Multiple R-squared:  0.8184,  Adjusted R-squared:  0.8183
F-statistic: 4.14e+04 on 2 and 18377 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Restelo ~ Alfragide + Benfica)

Residuals:
    Min      1Q  Median      3Q     Max
-63.791  -4.945  -0.703   3.946  68.251

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.719982   0.131747   5.465 4.69e-08 ***
Alfragide   0.221133   0.004995  44.268  < 2e-16 ***
Benfica     0.358731   0.005625  63.776  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.887 on 18377 degrees of freedom
Multiple R-squared:  0.6903,  Adjusted R-squared:  0.6903
F-statistic: 2.048e+04 on 2 and 18377 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Alfragide ~ Restelo + Benfica)

Residuals:
    Min      1Q  Median      3Q     Max
-82.854  -6.907  -0.521   4.601 144.763

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.582211   0.180454  -30.93   <2e-16 ***
Restelo      0.435754   0.009844   44.27   <2e-16 ***
Benfica      0.715883   0.006947  103.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.88 on 18377 degrees of freedom
Multiple R-squared:  0.7603,  Adjusted R-squared:  0.7603
F-statistic: 2.915e+04 on 2 and 18377 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Benfica ~ Alfragide + Restelo)

Residuals:
    Min      1Q  Median      3Q     Max
-68.451  -8.038  -1.188   6.341 123.294

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.200924   0.132873   84.30   <2e-16 ***
Alfragide    0.511583   0.004964  103.05   <2e-16 ***
Restelo      0.505164   0.007921   63.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.73 on 18377 degrees of freedom
Multiple R-squared:  0.7828,  Adjusted R-squared:  0.7828
F-statistic: 3.312e+04 on 2 and 18377 DF,  p-value: < 2.2e-16
```

Figura 3.48: *Outputs* dos testes paramétricos para a significância dos modelos de regressão linear múltipla.

3.5 Verificação do cumprimento do VL anual

Com base na Tabela 1.1 e na análise da Figura 3.49 verifica-se que:

- Em 2016, somente a EM da AvLiberdade é que excede o VL anual estabelecido na legislação;
- Em 2017, as EM de Alfragide, da AvLiberdade, de Benfica, de Entrecampos, e de Olivais apresentam valores de concentração superiores a 40 $\mu g/m^3$;
- Em 2018, as EM da AvLiberdade e de Entrecampos apresentam valores de concentração superiores a 40 $\mu g/m^3$;

- Em 2019, ocorre exatamente o mesmo que em 2016, dado que a EM da AvLiberdade é a única que apresenta uma concentração anual de NO_2 superior ao VL anual estabelecido na legislação.

É de notar que tal como se analisou nos resultados, o ano 2017 apresenta os maiores valores do poluente em estudo em todas as EM. Para além disto, a EM da AvLiberdade apresenta um comportamento muito diferente das restantes estações, sendo que esta apresenta sempre elevados níveis de poluição ao longo dos anos. Esta EM é considerada uma estação de tráfego e por isso, para reduzir as emissões de NO_2 devem-se, utilizar transportes menos poluentes e adotar estratégias para controlar as emissões nas fontes móveis como, a aplicação dos filtros de partículas e modificação dos veículos para ocorrer a recirculação dos gases de escape (Silva, 2019). Ainda se aconselha o uso combustível líquido ou gasoso produzido a partir de biomassa designados de biocombustíveis. Estes são atualmente fontes renováveis de energia incorporadas nos combustíveis rodoviários convencionais e em Portugal, os mais procurados são o biodisel e o bioetanol (ERSE, s.d.).

```
2016
Alfragide AvLiberdade       Beato     Benfica Entrecampos     Olivais     Restelo
 23.01452    57.62855    18.99853    33.42097    35.50048    26.65540    17.70653
2017
Alfragide AvLiberdade       Beato     Benfica Entrecampos     Olivais     Restelo
 70.31284    73.63146    35.99729    55.80127    55.15316    47.12007    32.94213
2018
Alfragide AvLiberdade       Beato     Benfica Entrecampos     Olivais     Restelo
 34.48181    62.54173    24.21888    39.34115    42.60262    32.10019    21.91832
2019
Alfragide AvLiberdade       Beato     Benfica Entrecampos     Olivais     Restelo
 28.15141    55.22093    22.85378    35.63091    37.59230    28.72003    20.58620
```

Figura 3.49: *Outputs* das concentrações médias anuais de NO_2 em cada EM.

Esta página foi intencionalmente deixada em branco.

Capítulo 4

Conclusões e trabalho futuro

O NO_2 é um poluente que resulta da queima de combustíveis, a altas temperaturas, nos motores dos veículos automóveis. A concentração deste poluente é consideravelmente mais elevada perto de estradas e em áreas urbanas, pois as emissões do transporte rodoviário estão próximas do solo e estão distribuídas por áreas densamente povoadas.

A poluição do ar no concelho de Lisboa está acima do recomendado pela OMS, provocando, no ser humano, efeitos nefastos nos sistemas pulmonar e cardiovascular.

Algumas EM não continham medições, o que dificultou a escolha das amostras a serem usadas. Entre os dois poluentes abordados neste trabalho decidiu-se utilizar somente as amostras relativas à concentração de NO_2, dado que era o poluente com menos observações em falta e, simultaneamente, um dos principais poluentes atmosféricos. Com o objetivo de identificar as EM da cidade de Lisboa com comportamentos semelhantes de poluição foi realizado um estudo, de longo prazo, com diversas ferramentas estatísticas.

Concluiu-se que existem dois conjuntos de EM que parecem ser redundantes para NO_2: Restelo, Alfragide e Benfica; Entrecampos, Beato e Olivais. Todavia, nestes grupos existem estações que parecem ter maiores semelhanças entre si, como é o caso da estação de Alfragide e Benfica; Beato e Olivais. Para além disso, verificou-se que a estação da AvLiberdade não é redundante e por isso não pode ser removida.

A identificação de EM redundantes permitiu verificar a existência de uma gestão ineficaz da RMQAr. Para possibilitar a otimização desta rede, as estações redundantes poderiam ser deslocadas para novas localizações, ampliando a área monitorizada sem aumentar os custos de manutenção.

Torna-se pertinente realçar, que no mesmo cluster, alguma EM pode encontrar-se mais afastada das restantes. Facto este pode ser explicado pelo vento e a sua contribuição na dispersão do poluente atmosférico.

A regressão linear permitiu estimar as concentrações de NO_2 em EM removidas, utilizando as concentrações das restantes estações em funcionamento. Os modelos mostraram que as EM selecionadas foram suficientes para inferir as concentrações do poluente atmosférico.

Ainda foi demonstrado que as maiores concentrações anuais de NO_2 foram observadas na maioria das estações de tráfego, tendo-se verificado uma diferença entre o ano de 2017 e os restantes anos.

Este trabalho demonstrou ser vantajoso para próximos estudos, pois só é possível atuar a nível da qualidade do ar com um conhecimento prévio sobre as EM existentes.

Esta página foi intencionalmente deixada em branco.

Bibliografia

Alkarkhi, A. F., & Alqaraghuli, W. A. (2020). *Applied statistics for environmental science with R*. Elsevier.

Agência Portuguesa do Ambiente. (2021a). *Redes de medição*. Ministério do Ambiente e da Ação Climática. https://apambiente.pt/ar-e-ruido/redes-de-medicao

Agência Portuguesa do Ambiente. (2021b). *Delimitação zonas e aglomerações*. Ministério do Ambiente e da Ação Climática. https://apambiente.pt/ar-e-ruido/delimitacao-zonas-e-aglomeracoes

Agência Portuguesa do Ambiente. (2021c). *Qualidade do ar*. Ministério do Ambiente e da Ação Climática. https://www.apambiente.pt/ar-e-ruido/qualidade-do-ar

Agência Portuguesa do Ambiente. (2021d). *Óxidos de Azoto NO_x*. Ministério do Ambiente e da Ação Climática. https://apambiente.pt/ar-e-ruido/oxidos-de-azoto-nox

Agência Portuguesa do Ambiente. (2021e). *Partículas em suspensão PM*. Ministério do Ambiente e da Ação Climática. https://apambiente.pt/ar-e-ruido/particulas-em-suspensao-pm

Agência Portuguesa do Ambiente. (2021f). *Controlo de emissões*. Ministério do Ambiente e da Ação Climática. https://www.apambiente.pt/ar-e-ruido/controlo-de-emissoes

Agência Portuguesa do Ambiente. (2021g). *Regime de Emissões para o Ar (REAR)*. Ministério do Ambiente e da Ação Climática. https://apambiente.pt/ar-e-ruido/regime-de-emissoes-para-o-ar-rear

Agência Portuguesa do Ambiente. (2021h). *Objetivos de qualidade do ar*. Ministério do Ambiente e da Ação Climática. https://apambiente.pt/ar-e-ruido/objetivos-de-qualidade-do-ar

Agência Portuguesa do Ambiente. (2023a). *Poluição Atmosférica por Dióxido de Azoto*. Ministério do Ambiente e da Ação Climática. https://rea.apambiente.pt/content/poluição-atmosférica-por-dióxido-de-azoto

Agência Portuguesa do Ambiente. (2023b). *Poluição por Partículas Inaláveis*. Ministério do Ambiente e da Ação Climática. https://rea.apambiente.pt/content/poluição-por-part%C3%ADculas-inaláveis

Castro, A., Araújo, R., & Silva, G. (2013). Qualidade do ar - Parâmetros de controle e efeitos na saúde humana: uma breve revisão. *Holos*, *5*, 107-121. http://www.redalyc.org/articulo.oa?id=481548607010

Comissão de Coordenação e Desenvolvimento Regional de Lisboa e Vale do Tejo. (2022). *Regulamentação - Qualidade do Ar*. Ministério da

Coesão Territorial e Ministério do Ambiente e da Ação Climática. https://www.ccdr-lvt.pt/ambiente/qualidade-do-ar/regulamentacao-qualidade-do-ar/

Decreto-Lei n.º 102/2010, de 23 de setembro. Diário da República n.º 186/2010, Série I, 4177 - 4205. https://files.diariodarepublica.pt/1s/2010/09/18600/0417704205.pdf

European Environmental Agency. (2020a). *Air pollution*. https://www.eea.europa.eu/themes/air/intro

European Environmental Agency. (2020b). *Air quality in Europe - 2020 report* (EEA Report No. 09-2020). European Union. https://www.eea.europa.eu/publications/air-quality-in-europe-2020-report

Entidade Reguladora dos Serviços Energéticos. (s.d.). *Biocombustíveis*. Entidade Reguladora dos Serviços Energéticos. https://www.erse.pt/combustiveis-e-gpl/funcionamento/biocombustiveis/

Fernandes, R., & Ramos, P. (2022a). *Estatística descritiva e análise exploratória de dados*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2022b). *Análise descritiva de dados multivariados*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2022c). *Correlação e regressão linear*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2022d). *Regressão Linear Múltipla*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2022e). *Análise de componentes principais*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2022f). *Análise de clusters*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2022g). *Inferência estatística*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2022h). *Testes de hipóteses*. Instituto Superior de Engenharia de Lisboa.

Gokce, H. B., Arıoglu, E., Copty, N., Onay, T., & Gun, B. (2020). Exterior air quality monitoring for the Eurasia Tunnel in Istanbul. *Science of the Total Environment*, *699*, 134312. https://doi.org/10.1016/j.scitotenv.2019.134312

Gomes, J. (2022). *Modelização da Qualidade do Ar - 2*. Instituto Superior de Engenharia de Lisboa.

Härdle, W., & Hlávka, Z. (2015). *Multivariate statistics* (Second Edition). Springer.

Johnson, R., & Wichern, D. (2014). *Applied multivariate statistical analysis* (6th ed.). Pearson.

Mobilizar. (s.d.). *Monitorização da Qualidade do Ar*. ZERO - Associação Sistema Terrestre Sustentável. https://mobilizar.pt/ar/monitorizacao-da-qualidade-do-ar/

Pires, J., Pereira, M., Alvim-Ferraz, M., & Martins, F. (2009). Identification of redundant air quality measurements through the use of principal component analysis. *Atmospheric Environment*, *43*, 3837-3842. https://doi.org/10.1016/j.atmosenv.2009.05.013

Pires, J., Sousa, S., Pereira, M., Alvim-Ferraz, M., & Martins, F. (2008a). Management of air quality monitoring using principal component and cluster analysis - Part I: SO_2 and PM_{10}. *Atmospheric Environment*, *42*, 1249-1260. https://doi.org/10.1016/j.atmosenv.2007.10.044

Pires, J., Sousa, S., Pereira, M., Alvim-Ferraz, M., & Martins, F. (2008b). Management of air quality monitoring using principal component and cluster analysis - Part II: CO, NO_2 and O_3. *Atmospheric Environment*, *42*, 1261-1274. https://doi.org/10.1016/j.atmosenv.2007.10.041

QualAR. (2020). *A Rede de Medição*. Ministério da Coesão Territorial e Ministério do Ambiente e da Ação Climática. https://qualar.apambiente.pt/node/qualar-network

R. (2023). *What is R?*. The R Project for Statistical Computing. https://www.r-project.org/about.html

Rani, B., Singh, U., Sharma, D., Chuhan, A., & Maheshwari, R. (2011). Photochemical Smog Pollution and Its Mitigation Measures. *Journal of Advanced Scientific Research*, *2*(4), 28-33. https://sciensage.info/index.php/JASR/article/view/56

Roser, M. (2021). Data review: how many people die from air pollution?. *Our World in Data*. https://ourworldindata.org/data-review-air-pollution-deaths

Silva, J. M. (2019). *Monitorização e Tratamento de Poluentes Atmosféricos - NO_x*. Instituto Superior de Engenharia de Lisboa.

World Health Organization. (2021). *WHO global air quality guidelines. Particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization. https://iris.who.int/bitstream/handle/10665/345329/9789240034228-eng.pdf?sequence=1

Zelterman D. (2015). *Applied multivariate statistical with R*. Springer.