Universidade do Minho
Escola de Ciências

Ana Cristina Gonçalves Pedra

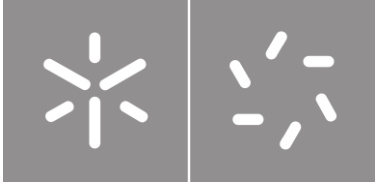
Séries Temporais e Análise de Risco na Avaliação da Qualidade de Água de uma Bacia Hidrográfica

outubro de 2023

Séries Temporais e Análise de Risco na Avaliação da Qualidade de Água de uma Bacia Hidrográfica

Ana Pedra

UMinho | 2023



Universidade do Minho
Escola de Ciências

Ana Cristina Gonçalves Pedra

Séries Temporais e Análise de Risco na Avaliação da Qualidade de Água de uma Bacia Hidrográfica

Dissertação de Mestrado
Mestrado em Estatística para Ciência de Dados

Trabalho efetuado sob a orientação da
Professora Doutora Arminda Manuela Andrade Pereira Gonçalves
e da
Professora Doutora Irene Vitória Ribeiro Brito

outubro de 2023

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Atribuição
CC BY

https://creativecommons.org/licenses/by/4.0/

Agradecimentos

Em primeiro lugar, gostaria de agradecer à Professora A. Manuela Gonçalves e à Professora Irene Brito, as minhas orientadoras, pela imensa partilha de conhecimento, constante disponibilidade e acompanhamento. Por tudo isto, obrigada.

Agradeço ao Engenheiro José Vitorino, da Agência Portuguesa do Ambiente (APA) pela sua disponibilidade e pela partilha de informação fundamental para a realização do presente trabalho.

Agradeço à minha família por todo o apoio e por me fazerem acreditar que com trabalho, foco e determinação todos os objetivos são alcançados.

Agradeço ainda aos meus amigos, pelo companheirismo e amizade e ainda a todos aqueles que de uma forma ou de outra contribuíram para o meu desenvolvimento pessoal e profissional.

Por fim e não menos importante, agradeço à Fundação para a Ciência e Tecnologia (FCT) e, por conseguinte, ao Centro de Matemática da Universidade do Minho (CMAT) pela bolsa de investigação UIDB/00013/2020.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Resumo

A água é um recurso natural limitado, insubstituível e indispensável. Neste sentido, monitorizar a sua qualidade é essencial para evitar problemas ambientais e de saúde pública. Os métodos estatísticos são ferramentas importantes para controlar e prever mudanças no processo de monitorização da qualidade da água.

O principal objetivo deste trabalho é desenvolver novas metodologias que combinem conceitos da teoria do risco e abordagens de séries temporais, para prever e analisar a qualidade da água. As metodologias são ilustradas a partir de um conjunto de dados da Bacia Hidrográfica do Rio Douro (em Portugal) em termos de variáveis ambientais de qualidade da água, medidas mensalmente em 18 estações de monitorização e registadas no período de janeiro de 2002 a dezembro de 2013. Em particular, este estudo foca-se nas variáveis de qualidade da água: Azoto, Carência Bioquímica de Oxigénio a 5 dias, Carência Química de Oxigénio, Condutividade, Nitratos, Oxigênio Dissolvido, pH e Temperatura da Água.

Foi realizada uma análise de *clusters* para agrupar estações homogéneas em termos de qualidade de água. Como a qualidade da água depende da quantidade de chuva, esta abordagem diferencia e estuda separadamente o período de maio a setembro (período seco) e o período de outubro a abril (período húmido), após considerar a totalidade dos dados. Várias medidas de risco, como a entropia, o valor em risco, entre outras, foram calculadas para os *clusters* considerados, com o objetivo de avaliar o risco de poluição da água de cada um. As mesmas medidas foram aplicadas aos dados mensais para identificar os meses mais poluídos do ano. Posteriomente, foram aplicadas abordagens de séries temporais como os modelos SARIMA e os métodos de alisamento exponencial aos *clusters* para obter previsões para os últimos 12 meses e comparar os resultados com os resultados da classificação através das medidas de risco.

Assim, identificam-se *clusters* (estações de amostragem) com maior risco e conclui-se que os níveis de poluição são superiores no período seco. A classificação dos meses sugere que o risco é maior nos meses de junho a outubro, o que aponta uma divisão alternativa de análise nos períodos seco e húmido. A análise de séries temporais dos *clusters* revela uma sazonalidade evidente e sugere que a abordagem mais apropriada para prever valores futuros depende da variável em estudo.

Palavras-chave: Agrupamentos, Bacia Hidrográfica, Medidas de Risco, Qualidade da Água, Rio Douro, Séries temporais.

Abstract

Water is a limited, irreplaceable and indispensable natural resource. In this sense, monitoring its quality is essential to avoid environmental and public health problems. Statistical methods are important tools for the control and for forecasting changes in the management process of water quality.

The main objective of this work is to develop new methodologies, combining concepts from risk theory and times series approaches, to predict and analyze water quality. The methodologies are illustrated using a data set of the Douro River basin (in Portugal) in terms of environmental water quality variables, measured monthly in 18 monitoring stations and recorded in the period from January 2002 to December 2013. In particular, this study focuses on the water quality variables: Ammonical Nitrogen, 5-day Biological Oxygen Demand, Chemical Oxygen Demand, Conductivity, Nitrate, Dissolved Oxygen, pH and Water Temperature.

A cluster analysis was carried out to group homogeneous stations in terms of water quality. Since water quality depends on the flow variation, this approach differentiates and study separately the time horizon from May to September (dry period) and the time horizon from October to April (wet period), after considering all the data together. Several risk measures, such as value at risk, probability of excess, among others were determined for the considered clusters in order to assess the risk of water pollution for each one. The same measures were applied to the monthly data to identify the most critical months for water pollution in the year. After that, time series modeling approaches like SARIMA models and exponential smoothing methods were applied to the clusters to obtain predictions for the last 12 months and compare the results with the results of the classification obtained through the risk measures.

This identifies clusters (sampling stations) with the highest risk and concludes that the pollution levels are higher in the dry period. The classification of months suggest that the risk is higher in the months from June to October, which points out an alternative division of analysis into the dry and wet periods. The cluster time series analysis reveals a clear seasonality pattern and suggest that the most appropriate approach to forecast values depend on the variable under study.

Keywords: Clusters, Watershed, Risk Measures, Water Quality, Douro River, Time Series.

Conteúdo

1 Introdução **1**

1.1 Estado da Arte 2

1.2 Objetivos 5

1.3 Estrutura da Dissertação 5

2 Medidas de Risco **7**

2.1 Medidas de Risco Simples 8

2.2 Medida de Risco Valor Esperado - Variância 16

2.3 Distribuição de Frequências 17

3 Séries Temporais **19**

3.1 Definição de Série Temporal 19

3.1.1 Componentes e Decomposição 20

3.2 Processos Estocásticos 21

3.2.1 Processos Estocásticos Estacionários 22

3.2.2 Processos Estocásticos Não Estacionários 25

3.3 Valores em Falta 27

4 Modelos e Métodos de Previsão **29**

4.1 Modelos SARIMA e Metodologia Box-Jenkins 29

4.1.1 Modelos de Processos Estacionários 30

4.1.2 Modelos de Processos Não Estacionários 33

4.1.3 Metodologia Box-Jenkins 34
4.1.4 Análise de Resíduos 36
4.1.5 Seleção do Melhor Modelo 38
4.1.6 Previsões Pontuais e Intervalares 39
4.2 Métodos de Alisamento Exponencial 40
4.2.1 Método de Alisamento Exponencial Simples 40
4.2.2 Método de Holt 41
4.2.3 Método de Holt-Winters 42
4.2.4 Previsões Pontuais e Intervalares 44
4.3 Medidas de Avaliação 44

5 Análise de Agrupamento (*Clustering*) 47

5.1 Estrutura dos Dados 47
5.2 Seleção dos Indivíduos e Variáveis 48
5.3 Medidas de Proximidade 48
5.4 Métodos de Análise de Agrupamento 50

6 Aplicação à Bacia Hidrográfica do Rio Douro 53

6.1 Descrição da Base de Dados 54
6.2 Análise Exploratória de Dados 60
6.3 Análise de Agrupamentos (*Clustering*) 60
6.4 Classificação dos *Clusters* através de Medidas de Risco 84
6.5 Classificação dos Meses através de Medidas de Risco 93
6.6 Análise das Séries Temporais dos *Clusters* 99
6.6.1 Análise Exploratória das Séries 102
6.6.2 Condutividade 105
6.6.3 Oxigénio Dissolvido 113
6.7 *Clusters*: Medidas de Risco *vs* Séries Temporais 120

7 Conclusão **125**

7.1 Trabalho Futuro 127

A Medidas Descritivas da Análise de *Clusters* utilizando o Conjunto das Variáveis de Qualidade **134**

B Representações Gráficas das Séries Temporais dos *Clusters* **136**

C Previsões Pontuais e Intervalares do Conjunto de Métodos Aplicados **141**

Lista de Tabelas

3.1 Transformações usuais de Box-Cox. 26

6.1 Descrição e localização das estacões de monitorização. 55

6.2 Fronteiras de qualidade dos parâmetros físico-químicos aplicáveis em rios e albufeiras. 58

6.3 Descrição das variáveis e unidades de medida. 59

6.4 Medidas descritivas da variável AZOTO, para os *clusters* obtidos. . . 70

6.5 Medidas descritivas da variável CBO_5, para os *clusters* obtidos. . . . 72

6.6 Medidas descritivas da variável CQO, para os *clusters* obtidos. 74

6.7 Medidas descritivas da variável COND, para os *clusters* obtidos. . . . 76

6.8 Medidas descritivas da variável NI, para os *clusters* obtidos. 78

6.9 Medidas descritivas da variável OD, para os *clusters* obtidos. 80

6.10 Medidas descritivas da variável pH, para os *clusters* obtidos. 82

6.11 Medidas descritivas da variável TEMP, para os *clusters* obtidos. . . . 84

6.12 Distribuição de frequências do OD no *Cluster* 1_T. 85

6.13 Medidas de risco por *cluster* para a variável CBO_5. 87

6.14 Ordens dos valores de cada medida de risco para a variável CBO_5 e para cada *cluster* no período observado, nos meses secos e nos meses húmidos. 87

6.15 Ordens dos valores de cada medida de risco para os *clusters* da variável CBO_5. 87

6.16 Medidas de risco por *cluster* para a variável COND. 88
6.17 Ordens dos valores de cada medida de risco para a variável COND e para cada *cluster* no período observado, nos meses secos e nos meses húmidos. 89
6.18 Ordens dos valores de cada medida de risco para os *clusters* da variável COND. 89
6.19 Medidas de risco por *cluster* para a variável OD. 90
6.20 Ordens dos valores de cada medida de risco para a variável OD e para cada *cluster* no período observado, nos meses secos e nos meses húmidos. 90
6.21 Ordens dos valores de cada medida de risco para os *clusters* da variável OD. 90
6.22 Medidas de risco por *cluster* para a variável TEMP. 91
6.23 Ordens dos valores de cada medida de risco para a variável TEMP e para cada *cluster* no período observado, nos meses secos e nos meses húmidos. 91
6.24 Ordens dos valores de cada medida de risco para os *clusters* da variável TEMP. 92
6.25 Medidas de risco por mês para a variável CBO_5. 94
6.26 Ordens dos valores de cada medida de risco para a variável CBO_5. . . 94
6.27 Medidas de risco por mês para a variável COND. 95
6.28 Ordens dos valores de cada medida de risco para a variável COND. . 95
6.29 Medidas de risco por mês para a variável OD. 96
6.30 Ordens dos valores de cada medida de risco para a variável OD. . . . 97
6.31 Medidas de risco por mês para a variável TEMP. 98
6.32 Ordens dos valores de cada medida de risco para a variável TEMP. . 98
6.33 Medidas descritivas das séries de COND, dos *Clusters* 1_T e 2_T. 102

6.34 Medidas descritivas das séries de OD, dos *Clusters* 1_T e 2_T. 104
6.35 Resultados do ajustamento do modelo SARIMA à série dos logarítmos da COND do *Cluster* 1_T. 105
6.36 Resultados do ajustamento do método aditivo de Holt-Winters à série de COND do *Cluster* 1_T. 108
6.37 Resultados do ajustamento do modelo SARIMA à série dos logarítmos de COND do *Cluster* 2_T. 109
6.38 Resultados do ajustamento do método aditivo de Holt-Winters à série de COND do *Cluster* 2_T. 111
6.39 Medidas de avaliação de cada um dos modelos aplicados a cada um dos *clusters* de COND. 112
6.40 Resultados do ajustamento do modelo SARIMA à série dos logarítmos do OD do *Cluster* 1_T. 113
6.41 Resultados do ajustamento do método aditivo de Holt-Winters à série de OD do *Cluster* 1_T. 115
6.42 Resultados do ajustamento do modelo SARIMA à série dos logarítmos do OD do *Cluster* 2_T. 117
6.43 Resultados do ajustamento do método aditivo de Holt-Winters à série de OD do *Cluster* 2_T. 119
6.44 Medidas de avaliação de cada um dos modelos aplicados a cada um dos *clusters* de OD. 120
6.45 Ordens dos *clusters* de COND calculadas através de séries temporais e medidas de risco. 122
6.46 Ordens dos *clusters* de OD calculadas através de séries temporais e medidas de risco. 123

A.1 Medidas descritivas das variáveis de qualidade, para os *clusters* obtidos considerando o período observado. 134

A.2 Medidas descritivas das variáveis de qualidade, para os *clusters* obtidos considerando os meses secos. 135
A.3 Medidas descritivas das variáveis de qualidade, para os *clusters* obtidos considerando os meses húmidos. 135

C.1 Valores observados e previsões pontuais e intervalares resultantes dos processos de modelação SARIMA e Holt-Winters para o *Cluster* 1_T de COND. 141
C.2 Valores observados e previsões pontuais e intervalares resultantes dos processos de modelação SARIMA e Holt-Winters para o *Cluster* 2_T de COND. 141
C.3 Valores observados e previsões pontuais e intervalares resultantes dos processos de modelação SARIMA e Holt-Winters para o *Cluster* 1_T de OD. 142
C.4 Valores observados e previsões pontuais e intervalares resultantes dos processos de modelação SARIMA e Holt-Winters para o *Cluster* 2_T de OD. 142

Lista de Figuras

2.1 Função entropia de Shannon da variável aleatória X com duas realizações possíveis. 11

6.1 Região hidrográfica do rio Douro e respetivas sub-bacias. Fonte: Adaptado de Zucco (2016). 54

6.2 Representação da Região Hidrográfica do Rio Douro e localização geográfica das estações de monitorização em estudo. 55

6.3 Dendrogramas das variáveis de qualidade da água, para o período observado (em cima), para os meses secos (à esquerda) e para os meses húmidos (à direita). 62

6.4 *Boxplots* da variável AZOTO (à esquerda) e da variável CBO_5 (à direita), dos *clusters* do conjunto das variáveis de qualidade da água para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 63

6.5 *Boxplots* da variável CQO (à esquerda) e da variável COND (à direita), dos *clusters* do conjunto das variáveis de qualidade da água para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 64

6.6 *Boxplots* da variável NI (à esquerda) e da variável OD (à direita), dos *clusters* do conjunto das variáveis de qualidade da água para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 65

6.7 *Boxplots* da variável pH (à direita) e da variável TEMP (à esquerda), dos *clusters* do conjunto das variáveis de qualidade da água para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 66

6.8 Dendrogramas da variável AZOTO (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 69

6.9 Dendrogramas da variável CBO_5 (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 71

6.10 Dendrogramas da variável CQO (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 73

6.11 Dendrogramas da variável COND (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 75

6.12 Dendrogramas da variável NI (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 77

6.13 Dendrogramas da variável OD (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 79

6.14 Dendrogramas da variável pH (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 81

6.15 Dendrogramas da variável TEMP (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo). 83

6.16 Histograma e curva de densidade da variável OD no *Cluster* 1_T. . . . 85

6.17 Série temporal de COND do *Cluster* 1_T. 100

6.18 Série temporal de COND do *Cluster* 2_T. 101

6.19 Série temporal de OD do *Cluster* 1_T. 101

6.20 Série temporal de OD do *Cluster* 2_T. 102

6.21 *Boxplots* das série de COND dos *Clusters* 1_T e 2_T (à esquerda) e *boxplots* dos valores de COND dos *Clusters* 1_T e 2_T divididos por meses (à direita). 103

6.22 *Boxplots* das série de OD dos *Clusters* 1_T e 2_T (à esquerda) e *boxplots* dos valores de OD dos *Clusters* 1_T e 2_T divididos por meses (à direita).104

6.23 Série dos resíduos, respetivo histograma, curva de densidade, FAC e FACP após a aplicação do modelo SARIMA à série de COND do *Cluster* 1_T. 106

6.24 Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do modelo SARIMA para a COND do *Cluster* 1_T. 107

6.25 Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do método de Holt-Winters para a COND do *Cluster* 1_T. 108

6.26 Série dos resíduos, respetivo histograma, curva de densidade, FAC e FACP após a aplicação do modelo SARIMA à série de COND do *Cluster* 2_T. 110

6.27 Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do modelo SARIMA para a COND do *Cluster* 2_T. 110

6.28 Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através o método de Holt-Winters para a COND do *Cluster* 2_T. 112

6.29 Série dos resíduos, respetivo histograma, curva de densidade, FAC e FACP após a aplicação do modelo SARIMA à série de OD do *Cluster* 1_T. 114

6.30 Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do modelo SARIMA para o OD do *Cluster* 1_T. 115

6.31 Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do método de Holt-Winters para o OD do *Cluster* 1_T. 116

6.32 Série dos resíduos, respetivo histograma, curva de densidade, FAC e FACP após a aplicação do modelo SARIMA à série de OD do *Cluster* 2_T. 117

6.33 Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do modelo SARIMA para o OD do *cluster* 2_T. 118

6.34 Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através o método de Holt-Winters para o OD do *Cluster* 2_T. 119

6.35 Previsões pontuais calculadas através dos modelos SARIMA para os *Clusters* 1_T e 2_T de COND. 121
6.36 Previsões pontuais calculadas através do método de Holt-Winters para os *Clusters* 1_T e 2_T de OD. 123

B.1 Série temporal de COND do *Cluster* 1_T. 136
B.2 Série temporal de COND do *Cluster* 2_T. 137
B.3 Decomposição aditiva da série de COND do *Cluster* 1_T. 137
B.4 Estimativas e previsões pontuais e intervalares obtidas para a COND do *Cluster* 1_T através do modelo SARIMA. 138
B.5 Estimativas e previsões pontuais e intervalares obtidas para a COND do *Cluster* 1_T através do método de Holt-Winters. 138
B.6 Decomposição aditiva da série de COND do *Cluster* 2_T. 139
B.7 Decomposição aditiva da série de OD do *Cluster* 1_T. 139
B.8 Decomposição aditiva da série de OD do *Cluster* 2_T. 140

Lista de Abreviaturas, Siglas e Acrónimos

ADF - *Augmented Dickey-Fuller* (em português, Dickey-Fuller Aumentado)

ALV - Albufeira Alvão - V. Real

ANO - Ano de registo da observação

APA - Agência Portuguesa do Ambiente

AR - *Autoregressive* (em português, Autorregressivo)

ARIMA - *Autoregressive Integrated Moving Average* (em português, Autorregressivo Integrado de Médias Móveis)

ARMA - *Autoregressive Moving Average* (em português, Autorregressivo de Médias Móveis)

AUR - Praia Aurora

AZOTO - Azoto Amoniacal

BAS - Albufeira Bastelo

BAT - Ponte Bateira

CAM - Albufeira Camba

CBO_5 - Carência Bioquímica de Oxigénio a 5 dias

CMAT - Centro de Matemática da Universidade do Minho

COND - Condutividade

COR - Foz Corgo

CQO - Carência Química de Oxigénio

DA - Desvio Absoluto

DF - Dickey-Fuller

DNGRH - Direção Nacional de Gestão de Recursos Hídricos

DP - Desvio Padrão

DQA - Diretiva Quadro da Água

E - Entropia

EPAM - Erro Percentual Absoluto Médio

EQM - Erro Quadrático Médio

EST - Estação de Monitorização

FAC - Função de Autocorrelação

FACP - Função de Autocorrelação Parcial

FCT - Fundação para a Ciência e Tecnologia

FEM - Função de Excesso Médio

FEMI - Função Excesso Médio Inferior

FEMS - Função Excesso Médio Superior

IQA - Índice de Qualidade da Água

KPSS - Kwiatkowski-Phillips-Schmidt-Shin

LON - Albufeira Fonte Longa

M - Média

MA - *Moving Average* (em português, Médias Móveis)

MAR - Quinta Maravilha

MEL - Melres

MES - Mês de registo da observação

MOL - Moledo

NI - Nitratos

O - Ordem

OD - Oxigénio Dissolvido

OF - Ordem Final

P - Probabilidade

PEN - Penude

PER - Período (Seco ou Húmido)

PGRH - Plano de Gestão da Região Hidrográfica

PP - Probabilidade de Perda

PPI - Probabilidade de Perda Inferior

PPS - Probabilidade de Perda Superior

PPT - Phillips-Perron Teste

RAN - Albufeira Ranhados

REQM - Raiz do Erro Quadrático Médio

RH3 - Região Hidrográfica do rio Douro

SAR - *Seasonal Autoregressive* (em português, Autorregressivo Sazonal)

SARIMA - *Seasonal Autoregressive Integrated Moving Average* (em português, Autorregressivo Integrado de Médias Móveis Sazonal)

SARMA - *Seasonal Autoregressive Moving Average* (em português, Autorregressivo e de Médias Móveis Sazonal)

SEM - Semealho (Albufeira Torrão)

SER - Albufeira Serra Serrada

SI - Semivariância Inferior

SMA - *Seasonal Moving Average* (em português, Médias Móveis Sazonal)

SNIRH - Sistema Nacional de Informação de Recursos Hídricos

SO - Soma das Ordens

SOR - Albufeira Sordo

SS - Semivariância Superior

STL - *Seasonal-Trend Decomposition Procedure Based on Loess* (em português, Processo de decomposição tendêncio-sazonal baseado em loess)

TAM - Foz Tâmega (Albufeira Crestuma)

TEL - Ponte Vale Telhas

TEMP - Temperatura da Água

TVaR - *Tail Value at Risk* (em português, Valor da Cauda em Risco)

V - Variância

VaR - *Value at Risk* (em português, Valor em Risco)

VAU - Vau

Capítulo 1

Introdução

A água é um recurso natural limitado, insubstituível e de valor inestimável. Do total de água disponível no nosso planeta, 97% é água salgada (encontrando-se nos mares e oceanos) e apenas 3% é água doce. Desses 3% um pouco mais de 2% está nos glaciares e calotes polares sobrando menos de 1%, presente nos rios, lagos e águas subterrâneas, para as diversas atividades humanas e equilíbrio dos ecossistemas, pelo *World Wide Fund for Nature* (WWF).

Retomando ao tempo das primeiras civilizações, da Mesopotâmia, do Egito e da Índia vemos que todas elas se instalaram de forma natural em torno de rios , tal facto deve-se à água, às cheias sazonais que aumentavam a fertilidade dos solos permitindo criar mais alimento, satisfazer as necessidades de mais habitantes e desenvolver as populações. Na época, o principal risco que as populações enfrentavam era o de ocorrência de catástrofes naturais, tais como inundações, seca, sismos, entre outros.

Mais tarde, a partir do século XVIII, a revolução industrial provocou alterações profundas no modo de vida das populações, o desenvolvimento e inovação trouxeram consigo doenças e níveis de poluição nunca antes vistos.

Desde então o desenvolvimento económico, o crescimento populacional e as consequentes alterações climáticas têm vindo a prejudicar de forma significativa o meio ambiente, afetando a qualidade das águas dos rios e mares em todo o mundo. Assim, a noção de risco que outrora se referia apenas aos riscos ambientais foi sendo diversificada e adequada às diversas áreas. Hoje, os indivíduos debatem-se não só com riscos naturais e biológicos, como também com riscos financeiros, sociais, de contrair doenças, etc.

Deste modo, não só os níveis de poluição da água foram aumentando (ou equivalentemente, qualidade da água diminuindo) em todo o planeta, como o conceito de risco se foi expandindo.

Até à atualidade diversos esforços foram e continuam a ser feitos para travar esta poluição, mas, mesmo assim estima-se que em Portugal 56% das massas de água, superficiais e subterrâneas, estão classificadas como tendo estado inferior a *Bom* segundo a Associação Sistema Terrestre Sustentável (ZERO).

Neste trabalho, realizado no âmbito de uma Bolsa de Investigação (BI) sobre o tema "Análise de Risco na Avaliação da Qualidade de Água de uma Bacia Hidrográfica", com a referência I&D UIDB/00013/2020, financiado pelo Centro de Matemática (CMAT), pretende-se aliar a área de teoria de risco à avaliação da qualidade da água. Em particular, ir-se-á analisar e comparar a qualidade da água de superfície de diversas estações de monitorização pertencentes à bacia hidrográfica do rio Douro.

Todos os resultados, desde gráficos, funções, entre outros, apresentadas neste trabalho foram realizados ou implementados utilizando a linguagem R, em particular usando o programa *RStudio* (versão 4.2.0).

1.1 Estado da Arte

Os primeiros estudos sobre teoria de risco remontam ao início do século XX. De um modo geral, esta área estuda a forma como as pessoas detetam, avaliam e respondem ao risco e incerteza de diversas situações.

Um dos primeiros trabalhos realizados neste âmbito foi apresentado por Frank Knight, em 1921 e distingue as noções de risco e incerteza, que é resumida pela afirmação: "if you don't know for sure what will happen, but you know the odds, that's risk, and if you don't even know the odds, that's uncertainty".

Naturalmente que, desde o passado até ao presente a humanidade se debate com o risco de diversas situações. Ora, o risco de uma ação específica é determinado pela perceção individual de risco, indivíduos diferentes têm diferentes perceções (Brachinger e Weber, 1997), logo o risco é uma característica subjetiva de cada ação. Assim sendo, como é que podemos descrever o risco? MacKenzie (2014) refere que uma das principais formas de o fazer é através de índices de risco, que resumem o risco utilizando números ou categorias, como palavras, letras ou cores.

Com o avanço dos estudos bastantes foram os métodos apresentados para construir índices de risco. Alguns deles fracassados, como as matrizes de risco (Cox, 2008) e outros vitoriosos que são atualmente utilizados nas mais diversas áreas. Em particular, Mano (1989) desenvolveu um índice de qualidade da água (IQA) com base em vetores de qualidade tais como pH, oxigénio dissolvido, azoto amoniacal,

entre outros e avaliou a qualidade da água superficial das bacias hidrográficas do rio Minho e do rio Ave.

A presente dissertação foca-se no estudo de medidas de risco que são medidas que permitem resumir o risco através de um número real e, por isso, são entendidas como um subconjunto dos índices de risco.

Ao longo dos anos várias foram as medidas apresentadas por diversos autores. Shannon (1948), Markowitz (1959), Kaas *et al.* (2009) e Klugman (2019) são autores que definiram medidas de risco que quantificam a perceção de risco de uma forma objetiva, tais como variância, probabilidade de perda, entropia, função de excesso médio, valor em risco, entre outras. Já Pollatsek e Tversky (1970) foram os responsáveis por desenvolver uma medida de risco que quantifica a perceção de risco de uma forma objetiva e subjetiva, designada medida de risco valor esperado-variância.

Os primeiros estudos sobre séries temporais remontam ao final do século XIX e início do século XX. Esta área, que se dedica ao estudo de variáveis que variam ao longo tempo tem por objetivo fazer previsões para o futuro, estimar observações passadas que por algum motivo não puderam ser observadas, entre outros (Alpuim, 1997).

Um dos primeiros estudos nesta área foi apresentado por Galton (1877) e embora fosse focado na hereditariedade, introduziu contribuições relevantes para a compreensão dos padrões das séries temporais.

Mais tarde, em 1927, Yule apresentou a noção de séries estocásticas, destacando que cada série temporal pode ser vista como a realização de um processo estocástico (Gooijer e Hydman, 2006). Baseadas nesta ideia, várias foram as propostas apresentadas para modelar e prever séries temporais ao longo dos anos, desde propostas clássicas de aplicação de médias móveis, a modelos SARIMA, métodos de alisamento exponencial, modelos de regressão, redes neurais, processos de espaço de estados, entre outros. A presente dissertação destaca os métodos baseados em modelos SARIMA e os métodos de alisamento exponencial.

Os modelos SARIMA, introduzidos por Box e Jenkins (1970), descrevem uma série como função das suas observações passadas e combinação linear de choques aleatórios e decompõem-se em vários modelos mais simples (autorregressivo, de médias móveis, autorregressivo e de médias móveis). Já os métodos de alisamento exponencial, introduzidos no final da década de 1950 (Brown, 1959, Holt, 1957 e Winters, 1960), são métodos que descrevem uma série através de combinações ponderadas das suas observações passadas.

Alguns destes métodos perdem a sua eficácia e conduzem a erros quando existe

um número significativo de observações em falta, algo comum quando as séries são referentes a dados reais. Hyndman e Athanasopoulos (2018) são autores que referem duas alternativas para lidar com o problema. A primeira passa por considerar apenas os dados posteriores ao último valor ausente, mas tal só é possível se os valores considerados forem suficientes para gerar previsões significativas e a segunda passa por substituir os valores em falta por estimativas.

Dada a existência de abordagens distintas, há a necessidade de determinar entre elas qual a mais eficaz. Esta questão vê-se respondida através da aplicação de medidas de avaliação que determinam a precisão dos modelos e métodos, no entanto, nem todas as medidas são adequadas para todas as aplicações. Por exemplo, medidas como o erro quadrático médio, EQM e a raiz do erro quadrático médio, REQM são úteis para comparar diferentes métodos aplicados ao mesmo conjunto de dados, no entanto não devem ser utilizadas para comparar dados com diferentes escalas. Nestes casos, medidas como, o erro percentual absoluto médio, EPAM, são preferíveis, embora não possam ser aplicadas a séries com valores nulos.

A análise de séries temporais pode ser aplicada a diferentes áreas, como a área de economia, medicina, entre outras. No presente trabalho, o seu intuito é avaliar a qualidade da água de uma bacia hidrográfica. Atualmente, já existem alguns estudos neste âmbito.

Worrall e Burt (1999) e Arya e Zhang (2015) são autores que utilizaram modelos autorregressivos e de médias móveis (ARMA) para prever, respetivamente, os níveis de nitratos em águas de bacias hidrográficas da Inglaterra e a temperatura e quantidade de oxigénio dissolvido da água no rio Stillaguamish, em Washington, obtendo em ambos os casos boas previsões. Em particular, para o primeiro caso obteve erros percentuais médios inferiores a 10%.

Já Costa e Monteiro (2015) aplicaram um modelo misto de espaço de estados para diferenciar a qualidade de água de diferentes estações de monitorização do Rio Vouga através da variável de qualidade oxigénio dissolvido, OD. Posteriormente, em 2018, os mesmos autores compararam a precisão de modelos de espaço de estados com modelos de regressão com erros correlacionados, utilizando para o efeito a mesma base de dados.

Encontram-se ainda na literatura estudos que agregam a análise de séries temporais a outras abordagens, como os estudos de Costa e Gonçalves (2011) e Gonçalves e Alpuim (2011), que tem por objetivo a monitorização da qualidade de água, mas aplicam previamente uma análise de *clusters*. O intuito desta análise é distinguir os níveis de qualidade de água de diferentes locais, já que a qualidade da água num

determinado local reflete as condições predominantes na área envolvente, para além de ser afetada pela quantidade de precipitação.

1.2 Objetivos

De uma forma muito sucinta os principais objetivos deste trabalho são:

- Estudar medidas de risco e de incerteza;
- Estudar a modelação e previsão de séries temporais;
- Estabelecer metodologias no contexto de teoria de risco e de séries temporais;
- Aplicar as metodologias estudadas para analisar e avaliar a qualidade da água de superfície de uma bacia hidrográfica;
- Comparar os processos obtidos com os processos usuais de classificação dos níveis de poluição de águas de superfície utilizados pela Direção Nacional de Gestão de Recursos Hídricos (DNGRH).

1.3 Estrutura da Dissertação

Face aos objetivos propostos a presente dissertação encontra-se dividida em diversos capítulos.

No presente capítulo (Capítulo 1) é feita uma breve introdução ao tema em questão e são definidos os objetivos a atingir, para além de se apresentar alguns trabalhos desenvolvidos na área (Estado da Arte).

No Capítulo 2 são apresentadas e definidas várias medidas de risco. Estas medidas, normalmente utilizadas no contexto financeiro serão aplicadas de forma inovadora aos nossos dados, com o objetivo de avaliar a qualidade da água de superfície das diversas estações de monitorização.

O Capítulo 3 é dedicado à introdução e exposição das noções teóricas relacionadas com a análise de séries temporais e o Capítulo 4 à apresentação de vários modelos e métodos de previsão das mesmas.

No Capítulo 5 são apresentadas as noções teóricas que suportam a análise de *clusters* através de métodos hierárquicos aglomerativos, posteriormente aplicados com o intuito de identificar grupos de estações (de monitorização) com níveis de qualidade de água semelhantes.

No Capítulo 6 são inicialmente exploradas as diversas leis que permitem avaliar a qualidade das águas de superfície, é feito o enquadramento da zona de estudo e uma análise preliminar da base de dados. Posteriormente, são aplicadas as diversas abordagens que permitem alcançar o principal objetivo da dissertação. Inicialmente aplicam-se metodologias de análise de *clusters*, de seguida apresentam-se abordagens baseadas em medidas de risco (classificação de *clusters* e meses) e por fim efetuam-se análises de séries temporais (estimação e previsão).

Por fim, no Capítulo 7 são apresentadas as conclusões e expostas as principais sugestões de trabalho futuro.

Capítulo 2

Medidas de Risco

O conceito de **risco** está presente nas mais diversas áreas de conhecimento. Em todas elas, o risco refere-se ao risco de uma decisão alternativa, por outras palavras, é uma característica negativa de uma decisão alternativa. No entanto, o risco não é somente uma característica objetiva já que depende da perceção de cada indivíduo. Deste modo, o risco pode ser medido de forma objetiva e subjetiva.

No seguimento desta ideia surge o conceito de **preferência de risco** que se refere à preferência de um dado indivíduo por uma alternativa sob condições de risco. Geralmente, um indivíduo prefere uma alternativa a outra baseando-se não só no risco de cada uma delas como também nos restantes aspetos positivos ou negativos que o mesmo lhes atribui, onde um aspeto positivo pode simplesmente ser o conhecimento do indivíduo da alternativa.

Questionando um indivíduo sobre a sua preferência (de risco) entre alternativas, obtém-se uma ordenação de preferência, intitulada ordenação de risco.

Definição 2.0.1 *Seja $\mathcal{A}$ um espaço de alternativas e $A, B \in \mathcal{A}$ duas quaisquer alternativas. A **ordenação de risco** é dada pela relação binária $\succsim$, onde $A \succsim B$ afirma que a alternativa A possui um risco igual ou superior ao risco da alternativa B ou, equivalentemente, que B é preferível a A.*
Simplificando, $A \succ B$ significa que o risco de A é superior ao risco de B e $A \sim B$ que A e B possuem o mesmo risco.

Ora, a ordenação de risco pode ser representada numericamente por funções $R : \mathcal{A} \to \mathbb{R}$, onde

$$A \succsim B \;\Leftrightarrow\; R(A) \geq R(B). \tag{2.1}$$

Estas funções, que permitem quantificar o risco, designam-se **funções de medição de risco** ou simplesmente **medidas de risco**. Assim, tem-se que A possui um risco igual ou superior a B se e só se a medida de risco de A é superior à medida de risco de B.

Nas seguintes secções serão apresentadas e definidas algumas das medidas de risco referidas por Brachinger e Weber (1997), Kaas *et al.* (2009) e Klugman (2019) que permitem modelar a preferência dos indivíduos.

De acordo com Mackenzie (2014), estas medidas podem ser classificadas em quatro categorias: momentos de uma distribuição, quantis de uma distribuição, funções de desutilidade e fatores combinados. Medidas consideradas funções de desutilidade e fatores combinados não serão estudadas nesta dissertação.

2.1 Medidas de Risco Simples

As medidas de risco simples, apresentadas nesta secção, são medidas que permitem quantificar o risco somente de forma objetiva. Estas têm um papel fulcral neste estudo, pois apesar de existirem medidas de risco que permitem quantificar o risco de forma objetiva e subjetiva, a subjetividade não interfere na qualidade da água, por outras palavras, a perceção de cada indivíduo não altera os níveis de poluição da água.

Variância e Desvio Padrão

Definição 2.1.1 *Seja X uma variável aleatória, f_X a sua função densidade de probabilidade e F_X a sua função de distribuição. Então, a **variância** e o **desvio padrão** são, respetivamente, dados por*

$$\sigma^2 = Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 dF_X(x) \;\; e \;\; \sigma = \left[\int_{-\infty}^{+\infty} (x-\mu)^2 dF_X(x)\right]^{1/2}, \quad (2.2)$$

onde $\mu = \mathbb{E}(X) = \int_{-\infty}^{+\infty} x dF_X(x)$.

Note-se que, no caso de X ser discreta $dF_X(x) = F_X(x) - F_X(x-dx)$ e no caso de X ser contínua $dF_X(x) = f_X(x)dx$.

Geralmente, um aumento na medida de dispersão variância (consequentemente, desvio padrão) traduz-se numa situação de maior risco.

Nas seguintes medidas e sempre que se mencionar, $dF_X(x)$ representa o diferencial do Integral de Riemann-Stieltjes.

Desvio Absoluto

Definição 2.1.2 *Seja X uma variável aleatória, f_X a sua função densidade de probabilidade e F_X a sua função de distribuição. Então, o* ***desvio absoluto*** *é dado por*

$$\int_{-\infty}^{+\infty} |x-\mu| dF_X(x), \tag{2.3}$$

onde μ é o valor médio de X.

Semivariância Inferior e Semivariância Superior

No cálculo da variância considera-se todos os valores que X pode tomar, mas, em diversos casos apenas interessa considerar os valores de X inferiores a μ, ou contrariamente, os valores de X superiores a μ. Assim sendo, da variância derivaram as duas medidas seguintes.

Definição 2.1.3 *Seja X uma variável aleatória, f_X a sua função densidade de probabilidade e F_X a sua função de distribuição. Então, a* ***semivariância inferior*** *e a* ***semivariância superior*** *são dadas, respetivamente, por*

$$\int_{-\infty}^{\mu} (x-\mu)^2 dF_X(x) \;\; e \;\; \int_{\mu}^{+\infty} (x-\mu)^2 dF_X(x), \tag{2.4}$$

onde μ é o valor médio de X.

Probabilidade de Perda

Definição 2.1.4 *Seja X uma variável aleatória, f_X a sua função densidade de probabilidade e F_X a sua função de distribuição. Então, a* ***probabilidade de perda*** *ou* ***probabilidade de ruína*** *é dada por*

$$P(X \leq r) = \int_{-\infty}^{r} dF_X(x), \tag{2.5}$$

onde o valor conhecido r representa o quantil (de probabilidade p de X, calculada a partir de r) a partir do qual um indivíduo tem uma perda.

Entropia

A medida de risco entropia difere das medidas anteriormente apresentadas na medida em que é independente do tipo de dados, apenas depende das probabilidades,

o que permite que a mesma seja calculada tanto para variáveis quantitativas como para variáveis qualitativas.

Definição 2.1.5 *Seja X uma variável aleatória discreta com n realizações possíveis $x_i, i = 1, \ldots, n$ e $p_i, i = 1, \ldots, n$, a probabilidade de cada realização. Então a* ***entropia****, $H(X)$, é dada por*

$$H(X) = H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \ln(p_i). \tag{2.6}$$

Originalmente, Shannon (1948) definiu e apresentou a entropia através do logarítmo na base dois, ou seja, como sendo dada por

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i). \tag{2.7}$$

Considere-se a entropia de Shannon e uma variável aleatória X com duas realizações possíveis

$$X = \begin{cases} x_1, & p_1 \\ x_2, & p_2 = 1 - p_1. \end{cases}$$

Consequentemente,

$$\begin{aligned} H(X) &= -\sum_{i=1}^{2} p_i \log_2(p_i) \\ &= -p_1 \log_2(p_1) - p_2 \log_2(p_2) \\ &= -p_1 \log_2(p_1) - (1 - p_1) \log_2(1 - p_1). \end{aligned}$$

Uma vez que $p_1 \in [0, 1]$, a representação de $H(X)$ é dada pelo gráfico na Figura 2.1.

Através da análise deste conclui-se que quando as realizações possíveis são equiprováveis, ou seja $p_1 = p_2 = \frac{1}{2}$, a entropia é máxima e toma o valor 1. De facto,

$$H(X) = -\left(\tfrac{1}{2} \times \log_2(\tfrac{1}{2}) + \tfrac{1}{2} \times \log_2(\tfrac{1}{2})\right) = -\log_2(\tfrac{1}{2}) = 1.$$

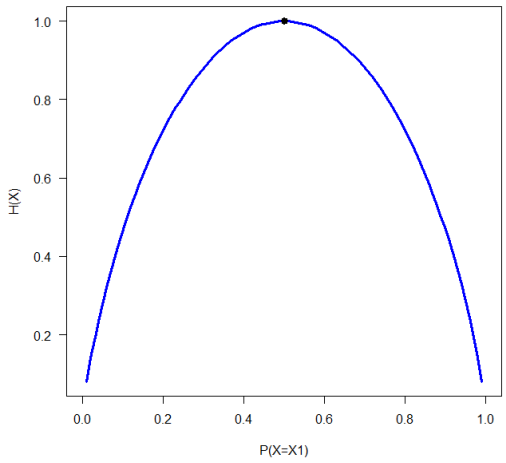


Figura 2.1: Função entropia de Shannon da variável aleatória X com duas realizações possíveis.

De uma forma mais prática, a entropia é uma medida que representa o grau de incerteza entre decisões, veja-se o exemplo 2.1.6.

Exemplo 2.1.6 *Sejam $X_i, i = 1, 2$, variáveis aleatórias independentes que representam a probabilidade de ao lançar uma moeda, i, sair cara ou coroa. Suponha-se que*

$$X_1 = \begin{cases} 1, & p_{11} = 0,90 \\ 0, & p_{12} = 0,10 \end{cases} \qquad X_2 = \begin{cases} 1, & p_{21} = 0,60 \\ 0, & p_{22} = 0,40, \end{cases}$$

onde 1 representa o evento sair cara e 0 representa o evento sair coroa.

Ambas as moedas são viciadas, mas possuem diferentes probabilidades de sair cara, 90% da moeda 1 contra 60% da moeda 2. Lançando a moeda 1 tem-se quase a certeza de que a face voltada para cima será cara, ou seja, o grau de incerteza é reduzido, já lançando a moeda 2 a certeza de obter a face cara é mais reduzida, ou seja, o grau de incerteza quanto ao resultado é mais elevado. Assim, diz-se que a entropia de X_2 é superior à entropia de X_1. De facto,

$$H(X_2) = -0,60 \times \ln(0,60) - (1 - 0,40) \times \ln(1 - 0,40) = 0,613$$
$$H(X_1) = -0,90 \times \ln(0,90) - (1 - 0,90) \times \ln(1 - 0,90) = 0,325.$$

Função de Excesso Médio

Definição 2.1.7 *Seja X uma variável aleatória que representa o risco (perda). Para um dado valor d tal que $P(X > d)$ existe e é não nula, a variável $Y^P = X - d$, onde $X > d$, é a* ***variável de excesso*** *e o valor esperado de $X - d$, dado que $X > d$, é a* ***função de excesso médio*** *dada por*

$$e(d) = E(Y^P) = E(X - d|X > d). \tag{2.8}$$

Valor em Risco

Segundo Klugman (2019), o valor em risco é a medida de risco mais utilizada. No contexto financeiro, a mesma representa o montante necessário para assegurar, com um certo grau de confiança, que uma dada empresa não se torna insolvente.

Definição 2.1.8 *Seja X uma variável aleatória que representa o risco (perda). O* ***valor em risco*** *de X com um nível de confiança p, $VaR_p(X)$ ou π_p, é definido como sendo o percentil 100p da distribuição da variável X.*
Em particular, se a variável X possui uma distribuição contínua, pode-se afirmar que o valor em risco de X é o valor π_p que satisfaz $P(X < \pi_p) = p$. Contrariamente, se a variável X possui uma distribuição discreta, diz-se que o valor em risco de X é o menor valor π_p que satisfaz $P(X \leq \pi_p) \geq p$.

Note-se que, no caso contínuo, para todo p, $\pi_p(X) = Q_p(X)$. Por outras palavras, o valor em risco de probabilidade p de X é o quantil de probabilidade p de X. Além disso, da definição 2.1.4 tem-se que $Q_p(X) = r$, logo $\pi_p(X) = r$.

Exemplo 2.1.9 *Sejam Z_1 e Z_2 variáveis aleatórias contínuas que representam perdas, tais que*

$$F_{Z_1}(90) = 0,95$$
$$F_{Z_2}(80) = 0,95.$$

Suponha-se que entre as duas se pretende escolher a que traduz uma situação de maior segurança, ou seja, de menor perda.

Tem-se que,

$$F_{Z_1}(90) = 0,95 \Leftrightarrow P(Z_1 \leq 90) = 0,95 \Leftrightarrow P(Z_1 \geq 90) = 0,05 \Leftrightarrow \pi_{0,95}(Z_1) = 90$$
$$F_{Z_2}(80) = 0,95 \Leftrightarrow P(Z_2 \leq 80) = 0,95 \Leftrightarrow P(Z_2 \geq 80) = 0,05 \Leftrightarrow \pi_{0,95}(Z_2) = 80.$$

Logo, o valor em risco de Z_1 é superior ao valor em risco de Z_2, com p=0,95, ou seja, a variável que representa um menor risco, ou equivalentemente, maior segurança é a variável Z_2.

Valor da Cauda em Risco

Definição 2.1.10 *Seja X uma variável aleatória que representa perdas. O* ***valor da cauda em risco*** *de X com um nível de confiança p, $TVaR_p(X)$, é a perda esperada dado que a perda excede o percentil $100p$ da distribuição da variável X, ou seja, é definida por*

$$TVaR_p(X) = E(X|X > \pi_p). \tag{2.9}$$

Teorema 2.1.11 *(Relação entre TVaR e VaR) Note-se que, se X é uma variável contínua, então para um nível de confiança p, $TVaR_p(X)$ é a média de todos os valores VaR superiores a p, ou seja*

$$TVaR_p(X) = \frac{\int_p^1 VaR_u(X)du}{1-p}. \tag{2.10}$$

Demonstracão 2.1.12 *Da definição de $TVaR$, tem-se que*

$$TVaR_p(X) = E(X|X > \pi_p) = \frac{\int_{\pi_p}^{+\infty} xdF(x)}{P(X > \pi_p)} = \frac{\int_{\pi_p}^{+\infty} xdF(x)}{1-p}. \tag{2.11}$$

Logo,

$$(1-p)\, TVaR_p(X) = \int_{\pi_p}^{+\infty} xdF(x).$$

Considere-se a seguinte mudança de variável, $u = F(x)$. Então $x = Q_u(X)$. Ao mesmo tempo, $du = dF(x)$. Ora se x tende para $+\infty$, então $u = \lim_{x\to+\infty} F(x) = 1$ e se $x = \pi_p$, então $u = F(\pi_p) = p$.

Deste modo,

$$\begin{aligned}\int_{\pi_p}^{+\infty} xdF(x) &= \int_p^1 Q_u(x)du \\ &= \int_p^1 \pi_u(x)du \\ &= \int_p^1 VaR_u(x)du.\end{aligned}$$

Em suma,

$$TVaR_p(X) = \frac{\int_p^1 VaR_u(x)du}{1-p}. \tag{2.12}$$

Teorema 2.1.13 *(Relação entre TVaR, VaR e a função de excesso médio) Se X é uma variável contínua então $TVaR_p(X)$ pode ser reescrito pela soma*

$$TVaR_p(X) = VaR_p(X) + e(\pi_p). \tag{2.13}$$

Demonstracão 2.1.14 *Da definição de função de excesso médio, tem-se que*

$$e(d) = E(X - d|X > d) = \frac{\int_d^{+\infty}(x-d)dF(x)}{P(X > d)} \tag{2.14}$$

e substituindo d por π_p, obtém-se

$$\begin{aligned}
e(\pi_p) = E(X - \pi_p|X > \pi_p) &= \frac{\int_{\pi_p}^{+\infty}(x-\pi_p)dF(x)}{P(X > \pi_p)} \\
&= \frac{\int_{\pi_p}^{+\infty}(x-\pi_p)dF(x)}{1-p} \\
&= \frac{\int_{\pi_p}^{+\infty} xdF(x) - \int_{\pi_p}^{+\infty}\pi_p dF(x)}{1-p} \\
&= \frac{\int_{\pi_p}^{+\infty} xdF(x)}{1-p} - \frac{\int_{\pi_p}^{+\infty}\pi_p dF(x)}{1-p} \\
&= TVaR_p(X) - \frac{\pi_p\int_{\pi_p}^{+\infty} dF(x)}{1-p} \\
&= TVaR_p(X) - \frac{\pi_p[F(x)]_{\pi_p}^{+\infty}}{1-p} \\
&= TVaR_p(X) - \frac{\pi_p(1-p)}{1-p} \\
&= TVaR_p(X) - \pi_p \\
&= TVaR_p(X) - VaR_p(X),
\end{aligned}$$

ou seja,

$$TVaR_p(X) = VaR_p(X) + e(\pi_p).$$

Exemplo 2.1.15 *Seja X uma variável aleatória tal que $X \sim N(\mu, \sigma)$. A função densidade de probabilidade de X é dada por*

$$f(x) = \tfrac{1}{\sqrt{2\pi}\sigma}\ exp\big[-\tfrac{1}{2}(\tfrac{x-\mu}{\sigma})^2\big].$$

Considere-se que ϕ representa a função densidade de probabilidade de uma distribuição Normal padrão, $\phi(x) = \frac{1}{\sqrt{2\pi}}\, exp[-\frac{1}{2}x^2]$, e Φ representa a correspondente função de distribuição, $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, exp[-\frac{1}{2}t^2]dt$.
Então,

$$VaR_p(X) = \mu + \sigma\Phi^{-1}(p)$$

e consequentemente

$$TVaR_p(X) = \mu + \sigma\frac{\phi[\Phi^{-1}(p)]}{1-p}.$$

Da definição de VaR_p tem-se que $P(X > \pi_p) = 1 - p$. Ora,

$$\begin{aligned} P(X > \pi_p) = 1 - p &\Leftrightarrow P(X \leq \pi_p) = p \\ &\Leftrightarrow P\left(\frac{X-\mu}{\sigma} \leq \frac{\pi_p - \mu}{\sigma}\right) = p \\ &\Leftrightarrow P\left(Z \leq \frac{\pi_p - \mu}{\sigma}\right) = p \\ &\Leftrightarrow \Phi\left(\frac{\pi_p - \mu}{\sigma}\right) = p \\ &\Leftrightarrow \frac{\pi_p - \mu}{\sigma} = \Phi^{-1}(p) \\ &\Leftrightarrow \pi_p = \sigma\Phi^{-1}(p) + \mu \\ &\Leftrightarrow VaR_p(X) = \mu + \sigma\Phi^{-1}(p). \end{aligned}$$

Como X é uma variável contínua, com distribuição Normal, da definição de $TVaR$ tem-se que

$$(1-p)\, TVaR_p(X) = \int_{\pi_p}^{+\infty} x dF(x) = \int_{\pi_p}^{+\infty} x\frac{1}{\sqrt{2\pi}\sigma} exp\big[-\tfrac{1}{2}\big(\tfrac{x-\mu}{\sigma}\big)^2\big]dx.$$

Considere-se a seguinte mudança de variável, $y = \frac{x-\mu}{\sigma}$. Então $x = \sigma y + \mu$. Ao mesmo tempo, $dy = \frac{1}{\sigma}dx$, ou seja, $dx = \sigma dy$. Ora se $x = \pi_p$, então $y = \frac{\pi_p - \mu}{\sigma} = \Phi^{-1}(p)$ e se x tende para $+\infty$, então $y = \lim_{x\to+\infty} \frac{x-\mu}{\sigma} = +\infty$.

Assim,

$$\begin{aligned}
(1-p)\,TVaR_p(X) &= \int_{\Phi^{-1}(p)}^{+\infty} (\sigma y+\mu)\frac{1}{\sqrt{2\pi}\sigma} exp\big[-\frac{1}{2}y^2\big]\sigma dy \\
&= \int_{\Phi^{-1}(p)}^{+\infty} (\sigma y+\mu)\frac{1}{\sqrt{2\pi}} exp\big[-\frac{1}{2}y^2\big] dy \\
&= \frac{\sigma}{\sqrt{2\pi}}\int_{\Phi^{-1}(p)}^{+\infty} y\; exp\big[-\frac{1}{2}y^2\big]dy + \mu\int_{\Phi^{-1}(p)}^{+\infty}\frac{1}{\sqrt{2\pi}}\; exp\big[-\frac{1}{2}y^2\big]dy \\
&= -\frac{\sigma}{\sqrt{2\pi}}\int_{\Phi^{-1}(p)}^{+\infty} -y\; exp\big[-\frac{1}{2}y^2\big]dy + \mu\Big(1-\int_{-\infty}^{\Phi^{-1}(p)}\frac{1}{\sqrt{2\pi}}\; exp\big[-\frac{1}{2}y^2\big]dy\Big) \\
&= -\frac{\sigma}{\sqrt{2\pi}}\big[exp\{-\frac{1}{2}y^2\}\big]_{\Phi^{-1}(p)}^{+\infty} + \mu\big(1-\Phi(\Phi^{-1}(p))\big) \\
&= \frac{\sigma}{\sqrt{2\pi}} exp\{-\frac{1}{2}\Phi^{-1}(p)^2\} + \mu\big(1-p\big) \\
&= \sigma\phi(\Phi^{-1}(p)) + \mu(1-p),
\end{aligned}$$

ou seja,

$$TVaR_p(X) = \mu + \sigma\tfrac{\Phi^{-1}(p)}{1-p}.$$

2.2 Medida de Risco Valor Esperado - Variância

Pollatsek e Tversky (Brachinger and Weber, 1997) abordaram a Teoria de Risco de um ponto de vista diferente, tendo em conta o contexto financeiro. Estes interpretaram as variáveis aleatórias como sendo lotarias/valores monetários e provaram que satisfeitas algumas propriedades é possível criar uma medida de risco única que combina medidas simples anteriormente vistas, a medida de risco valor esperado - variância.

Contrariamente às medidas vistas até ao momento, esta medida quantifica o risco de forma não só objetiva, como também subjetiva.

Definição 2.2.1 *Seja X uma variável aleatória que representa perdas ou ganhos, E(X) o seu valor médio e Var(X) a sua variância. Então,* ***a medida de risco valor esperado-variância*** *é dada por*

$$R(X) = \theta\, Var(X) - (1-\theta)\, E(X), \tag{2.15}$$

onde θ *é um número real único tal que* $0 < \theta \leq 1$.

O papel de θ é especificar a contribuição relativa de $E(X)$ e $Var(X)$ para o risco, assim θ é um valor conhecido e específico para cada indivíduo. Portanto, pode-se dizer que o valor esperado e variância determinam a quantificação objetiva e o parâmetro θ determina a quantificação subjetiva. Se θ toma valores muito próximos de zero, então o indivíduo atribui uma importância quase nula à variância, ou equivalentemente, atibui a maioria da importância ao valor esperado. Já se $\theta = \frac{1}{2}$ então o indivíduo considera que as duas medidas afetam de igual forma o risco. No caso de $\theta = 1$ então o indivíduo atribui a totalidade da importância à variância, considerando que o valor esperado de X não interfere com o risco. Para além do referido, para $\theta \in]0, 1]$ o risco varia conforme a variância varia (aumenta/diminui se a variância aumenta/diminui) e varia de forma inversa à variação do valor médio (aumenta se o valor médio diminui ou vice-versa).

Estudos empíricos posteriores de Coombs e Bowen (Brachinger and Weber, 1997) vieram provar que a medida de risco de Pollatsek e Tversky era insuficiente para determinar o risco, uma vez que fixada a variância e o valor médio, o risco variava, contrariamente ao espectável.

2.3 Distribuição de Frequências

A **distribuição de frequências** de uma variável descreve a distribuição dos seus valores ao longo de um conjunto de diferentes classes. Uma forma natural de observar esta distribuição é através de um histograma, no entanto, no caso da variável ser de natureza contínua, é necessário primeiramente proceder à divisão do conjunto de observações em intervalos de classes.

Seja S um conjunto com I subconjuntos, isto é, $S = \{S_1, \ldots, S_I\}$ e sejam $r_{i1}, r_{i2}, ..., r_{iT}$ as observações de cada subconjunto S_i registadas ao longo de T tempos. Definindo,

$$r_{min} = min_{1 \leq i \leq I}\{r_{i1}, r_{i2}, \ldots, r_{iT}\} \tag{2.16}$$

$$r_{max} = max_{1 \leq i \leq I}\{r_{i1}, r_{i2}, \ldots, r_{iT}\}, \tag{2.17}$$

o intervalo $[r_{min}, r_{max}]$ contém todas as observações de S (Brito, 2023). Dividir este intervalo em N subintervalos de igual dimensão não é uma tarefa fácil, uma vez que diferentes valores de N podem evidenciar diferentes características dos dados. Neste sentido recorre-se à Regra de Rice, que dita que

$$N = \lceil 2\sqrt[3]{T} \rceil. \tag{2.18}$$

Por sua vez, a amplitude de cada subintervalo é dada por

$$\Delta = \frac{r_{max} - r_{min}}{N}. \tag{2.19}$$

Deste modo, obtém-se as classes $[r_0, r_1[, [r_1, r_2[, \ldots, [r_{N-1}, r_N]$, onde $r_0 = r_{min}$, $r_N = r_{max}$ e $r_n = r_{n-1} + \Delta$, com $r = 1, \ldots, N$.

A distribuição de frequências é definida pelo conjunto de representantes das classes e pelas respetivas frequências relativas, onde o representante de cada classe é o valor médio das observações da mesma e a frequência relativa é dada pela divisão do número de observações que a ela pertencem e o número total de observações de cada S_i, respetivamente, T.

Capítulo 3

Séries Temporais

A análise de séries temporais designa um conjunto de procedimentos estatísticos utilizados para estudar observações recolhidas ao longo do tempo. Segundo Alpuim (1997), esta tem-se vindo a tornar fundamental para o estudo das mais variadas áreas de conhecimento. De facto, desde a economia, a medicina, várias são as aplicações possíveis. Por exemplo, esta análise pode ser utilizada para estudar o mercado de vendas, prever riscos de investimento, monitorizar surtos epidemiológicos ou até mesmo avaliar a concentração de um determinado composto na água. Deste modo, neste capítulo serão apresentadas as principais noções teóricas de séries temporais.

3.1 Definição de Série Temporal

Uma **série temporal** é um conjunto de observações medidas sequencialmente no tempo. Segundo Chatfield (2004), estas são designadas como contínuas, caso as medições sejam feitas de forma contínua no tempo, ou como discretas, caso as medições sejam feitas em momentos específicos. Geralmente, nas séries temporais discretas, as observações (medições) são igualmente espaçadas no tempo, por exemplo são registadas diariamente, semanalmente, mensalmente, etc. Além disso, as séries podem ser ainda designadas como univariadas, se são compostas por observações de uma única variável ou como multivariadas, se são compostas por observações de duas ou mais variáveis.

Na prática, as séries temporais contínuas são transformadas em séries temporais discretas através de uma discretização em intervalos de tempo iguais. Este processo não resulta numa perda significativa de informação desde que o intervalo de amostragem seja significativamente pequeno. Ao longo desta dissertação, serão estudadas apenas séries temporais discretas.

3.1.1 Componentes e Decomposição

Pearsons (1919) refere que a variação de uma série temporal pode ser decomposta em quatro componentes principais: a componente tendência (T), a componente sazonal (S), a componente cíclica (C) e a componente aleatória (E).

A **tendência** (T) é a inclinação que a série possui ao longo do tempo, podendo ser crescente ou decrescente, linear ou não linear. Esta pode ser consequência do facto das observações dependerem de uma componente determinística, embora alguns autores defendam que a mesma pode possuir uma natureza estocástica. Se uma série temporal não exibir uma variação crescente ou decrescente diz-se que a série é estacionária em média.

A **sazonalidade** (S) corresponde a um padrão, de aumento ou diminuição, que se repete em períodos específicos da série. Este padrão é originado por variações periódicas de duração fixa, que são explicadas por diversos fatores, como por exemplo, a estação do ano, picos de vendas são geralmente registados no verão, época de férias e ainda no natal, época de festividades. Esta componente é classificada como aditiva, caso não dependa do nível da série, ou como multiplicativa, caso seja proporcional ao mesmo. Por vezes esta componente é responsável por omitir outras componentes, quando tal acontece é natural removê-la através do processo ajustamento sazonal.

A **componente cíclica** (C) é um padrão de flutuação, que não apresenta qualquer periocidade definida, nem efeito associado à sazonalidade. Muitas vezes é difícil distinguir a componente cíclica da componente tendência, assim, quando tal sucede é natural aglomerar as duas componentes e designar esta junção por componente de tendência-cíclica.

Por fim, a **componente aleatória** (E), tipicamente designada por **aleatoriedade** é a componente que detém toda a variação não explicada pelas componentes anteriores e representa o ruído aleatório. Quando este ruído é modelado por um processo estocástico de variáveis aleatórias independentes e identicamente distribuídas é intitulado ruído branco.

A maioria dos métodos tradicionais para a análise de séries temporais baseia-se na decomposição da série nas suas componentes e descreve o valor da série num determinado tempo através das mesmas.

Considere-se um tempo t, sejam Y_t o valor da série no tempo t, T_t a tendência (tendência-cíclica, caso a série possua uma componente cíclica) no tempo t, S_t a sazonalidade no tempo t e E_t a aleatoriedade no tempo t.

O **modelo de decomposição aditivo** obtém cada valor da série como a soma das suas componentes, ou seja, como

$$Y_t = T_t + S_t + E_t. \tag{3.1}$$

Este modelo é apropriado quando a magnitude das oscilações sazonais não varia com o nível da série. Caso estas aumentem ou diminuam de forma proporcional à tendência, o modelo de decomposição multiplicativo é o mais apropriado.

O **modelo de decomposição multiplicativo** obtém cada valor da série como o produto das suas componentes, ou seja, como

$$Y_t = T_t \times S_t \times E_t. \tag{3.2}$$

Geralmente, estes modelos são convertidos em modelos de decomposição aditivos através da aplicação de uma transformação logarítmica, obtendo-se:

$$log(Y_t) = log(T_t) + log(S_t) + log(E_t). \tag{3.3}$$

Existem ainda modelos que combinam os modelos anteriormente apresentados, como o **modelo de decomposição multiplicativo com erros aditivos** onde cada valor da série é dado por

$$Y_t = T_t \times S_t + E_t. \tag{3.4}$$

3.2 Processos Estocásticos

Uma série temporal pode ser encarada como um conjunto de observações de um **processo estocástico**.

Definição 3.2.1 *Um processo estocástico é qualquer família ou coleção de variáveis aleatórias* $\{Y_t,\, t \in \mathcal{T}\}$*, onde* $\mathcal{T}$ *é um conjunto de índices que representa o tempo.*

Ao conjunto de variáveis aleatórias Y_t dá-se o nome de espaço de estados e ao conjunto de índices $\mathcal{T}$ dá-se o nome de espaço de parâmetros. Dependendo da natureza de $\mathcal{T}$ estes processos são classificados como discretos ($\mathcal{T} = \mathbb{Z}$ ou $\mathcal{T} = \mathbb{N}$) ou contínuos ($\mathcal{T} = \mathbb{R}$ ou, como é mais frequente, $\mathcal{T} = \mathbb{R}^+$).

Deste modo, dado um processo estocástico $\{Y_t,\, t \in \mathcal{T}\}$, uma série temporal é um conjunto de observações do processo estocástico nos instantes $t_1, \ldots, t_n$. Regra geral, estas observações são igualmente espaçadas no tempo e por isso considera-se o tempo inteiro, isto é, $t = 0, \pm 1, \pm 2, \ldots$.

Note-se que, o conjunto de valores observados que compõem a série é uma realização única de um processo estocástico, por outras palavras, um mesmo processo estocástico gera infinitas séries.

Os processos estocásticos classificam-se em **estacionários** e **não estacionários**.

3.2.1 Processos Estocásticos Estacionários

Segundo Cordeiro (2011), um processo estocástico é estacionário se possui propriedades probabilísticas estáveis ou invariantes ao longo do tempo.

Definição 3.2.2 *Um processo estocástico* $\{Y_t,\, t \in \mathcal{T}\}$ *diz-se* ***estritamente (ou fortemente) estacionário*** *se e só se as distribuições conjuntas de* $(Y_{t_1}, \ldots, Y_{t_n})$ *e de* $(Y_{t_1+\delta}, \ldots, Y_{t_n+\delta})$ *são iguais,* $\forall \delta \in \mathbb{R}$ *e* $\forall$ *n-úplo* $(t_1, \ldots, t_n)$*, isto é,*

$$F_{(Y_{t_1},\ldots,Y_{t_n})}(y_1, \ldots, y_n) = F_{(Y_{t_1+\delta},\ldots,Y_{t_n+\delta})}(y_1, \ldots, y_n) \tag{3.5}$$

em todos os pontos $(y_1, \ldots, y_n)$*.*

Deste modo, num processo estritamente estacionário a distribuição de qualquer conjunto de margens não se altera quando se efetuam translações no tempo.

A propriedade de estacionariedade é de difícil verificação, já que implica o conhecimento de todas as distribuições marginais (Cordeiro, 2011). Assim, na prática esta propriedade é substituída pela estacionariedade de 2ª ordem, que embora seja uma propriedade mais fraca descreve, no geral, o mesmo tipo de comportamento.

Definição 3.2.3 *Um processo estocástico* $\{Y_t,\, t \in \mathcal{T}\}$ *diz-se* ***estacionário de 2ª ordem*** *ou* ***fracamente estacionário*** *se e só se todos os momentos até à 2ª ordem das distribuições* $(Y_{t_1}, \ldots, Y_{t_n})$ *e* $(Y_{t_1+\delta}, \ldots, Y_{t_n+\delta})$ *existem e são iguais. Deste modo, num processo estacionário de 2ª ordem:*

1. *O valor médio não depende de t, isto é,* $E[Y_t] = \mu(t) = \mu$*;*
2. *A variância não depende de t, isto é,* $Var[Y_t] = \sigma^2(t) = \sigma^2$*;*
3. *A covariância de* Y_{t_1} *e* Y_{t_2} *depende apenas da diferença* $t_2 - t_1$*, isto é,* $\forall t_1, t_2, Cov(Y_{t_1}, Y_{t_2}) = \gamma(|t_2 - t_1|)$*.*

Note-se que se um processo é estritamente estacionário e os seus momentos até à 2ª ordem são finitos, então é também estacionário de 2ª ordem.

De agora em diante, um processo estacionário de 2ª ordem será designado simplesmente por processo estacionário.

Funções de autocovariância, autocorrelação e autocorrelação parcial

Definição 3.2.4 *A **função de autocovariância** de um processo estacionário mede a intensidade com que covariam pares de valores do processo separados por um intervalo de amplitude k e é definida por*

$$\gamma_k = Cov(Y_t, Y_{t+k}) = E[(Y_t - \mu)(Y_{t+k} - \mu)]. \tag{3.6}$$

Esta função caracteriza-se pelas seguintes propriedades:

1. $\gamma_0 = Cov(Y_t, Y_t) = Var(Y_t) = \sigma^2$;
2. $\gamma_k = \gamma_{-k}$, isto é, a função é par;
3. $|\gamma_k| \leq \gamma_0$, como consequência da Desigualdade de Cauchy-Schwarz que refere que $|E(XY)| \leq \sqrt{E[X^2]E[Y^2]}), \forall X, Y$;
4. $\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\gamma(|t_i - t_j|) \geq 0, \forall \alpha_1, \ldots, \alpha_n \in \mathbb{R}$ e $\forall t_1, \ldots, t_n$, ou seja, a função é semidefinida positiva.

Definição 3.2.5 *A **função de autocorrelação** (FAC) de um processo estacionário mede a correlação entre pares de valores do processo separados por um intervalo de amplitude k e é definida por*

$$\rho_k = Cor(Y_t, Y_{t+k}) = \frac{Cov(Y_t, Y_{t+k})}{\sqrt{Var(Y_t)Var(Y_{t+k})}} = \frac{Cov(Y_t, Y_{t+k})}{Var(Y_t)} = \frac{\gamma_k}{\gamma_0}. \tag{3.7}$$

A representação gráfica de ρ_k em função de k é denominada por correlograma. O comportamento deste permite identificar certas características da série e, portanto, auxilia de forma significativa, a escolha do modelo que melhor se adequa à série.

Na generalidade dos casos, à medida que k aumenta, ρ_k e consequentemente γ_k diminuem. Segundo Cordeiro (2011), é expectável que a capacidade de memória da maioria dos processos seja limitada e, portanto, à medida que o desfasamento k aumenta, o que sucedeu no momento t é pouco refletido no momento $t + k$. Deste modo, pode-se dizer que à medida que $k \to \infty$, $\rho_k \to 0$.

Esta função possui as seguintes propriedades:

1. $\rho_0 = Cor(Y_t, Y_t) = 1$;
2. $\rho_k = \rho_{-k}$, isto é, a função é par;
3. $|\rho_k| \leq 1$, como consequência da Desigualdade de Cauchy-Schwarz;

4. $\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\rho(|t_i-t_j|)\geq 0, \forall\alpha_1,\ldots,\alpha_n\in\mathbb{R}$ e $\forall t_1,\ldots,t_n$, ou seja, a função é semidefinida positiva.

Por vezes, pretende-se não só estudar a correlação de modo global como também a correlação parcial entre duas observações com um desfasamento k, isto é, a correlação entre Y_t e Y_{t+k} depois de eliminar o efeito que as variáveis intermédias, $Y_{t+1}, Y_{t+2}, \ldots, Y_{t+k-1}$, exercem sobre elas. Deste modo, surge a função de correlação parcial (FACP).

Definição 3.2.6 *O conjunto de* ***autocorrelações parciais*** *de intervalo k é dado por $\{\phi_{kk} : k = 1, 2, \ldots\}$, onde*

$$\phi_{kk} = Cor[Y_t, Y_{t+k}|Y_{t+1}, Y_{t+2}, \ldots, Y_{t+k-1}] = \frac{|P_k^*|}{|P_k|}, \tag{3.8}$$

sendo P_k^* a matriz de autocorrelações $k \times k$ em que a última coluna é substituída por $[\rho_1, \rho_2, \ldots, \rho_k]^T$ e P_k a matriz de autocorrelações $k \times k$ dada por

$$P_k = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \ldots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \ldots & \rho_{k-2} \\ \rho_2 & \rho_1 & 1 & \ldots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \ldots & \ldots & 1 \end{bmatrix}. \tag{3.9}$$

Sabe-se que:

1. $\phi_{11} = \rho_1$;
2. $\phi_{22} = \frac{\rho_2-\rho_1^2}{1-\rho_1^2}$;
3. $\phi_{33} = \frac{\rho_3(1-rho_1^2)+\rho_1(\rho_1^2+\rho_2^2-2\rho_2)}{(1-\rho_2)(1+\rho_2-2\rho_1^2)}$.

Ruído Branco

Um processo estocástico diz-se um **processo puramente aleatório** ou **processo de ruído branco** quando é formado por uma sucessão de variáveis aleatórias independentes e identicamente distribuídas, com média e variância constantes, isto é, um processo estocástico diz-se um ruído branco se satisfaz as seguintes propriedades:

1. $E[\epsilon_t] = \mu$;
2. $Var(\epsilon_t) = \sigma^2$;

3. $\forall k, Cov(\epsilon_t, \epsilon_{t+k}) = \gamma_k = 0$.

Assim, um ruído branco é um processo estacionário cujas funções de autocorrelação (FAC) e autocorrelação parcial (FACP) são nulas, para todo $k \neq 0$.

Um ruído branco raramente é observado em séries de dados reais, no entanto este desempenha um papel fulcral na escolha de modelos de previsão, já que, um bom modelo de previsão é aquele que produz erros de previsão com um comportamento semelhante a um ruído branco.

3.2.2 Processos Estocásticos Não Estacionários

As séries associadas a fenómenos ambientais são regra geral não estacionárias. O processo que descreve uma série pode ser não estacionário se a média e/ou variância não forem constantes, ou seja, dependerem do tempo. Quando tal sucede utilizam-se transformações que estabilizam a média e/ou a variância, tornando a série estacionária, já que este pressuposto é fundamental quando se pretende obter previsões com base em observações passadas.

Transformações para a Estacionariedade

Para **estabilizar a média** utilizam-se processos de diferenciação (regular), que resultam da aplicação do operador ∇.

A diferenciação de primeira ordem de uma série Y_t é dada por

$$\nabla Y_t = Y_t - Y_{t-1}, t = 2, 3, \ldots, n. \tag{3.10}$$

Se esta não for suficiente para transformar a série numa série estacionária deve-se aplicar uma segunda diferenciação (diferenciação de segunda ordem), isto é,

$$\nabla^2 Y_t = \nabla(\nabla Y_t) = \nabla(Y_t - Y_{t-1}) = Y_t - 2Y_{t-1} + Y_{t-2}, t = 3, 4, \ldots, n. \tag{3.11}$$

Seja d a ordem do operador de diferenciação que permite diferenciar uma série d vezes. Então,

$$\nabla^d Y_t = \nabla(\nabla^{d-1} Y_t), para\ d \geq 1\ e\ t = d+1, \ldots, n. \tag{3.12}$$

No entanto, deve-se evitar a diferenciação excessiva de uma série, já que a variância aumenta à medida que a ordem de diferenciação aumenta. O ideal é determinar a série estacionária obtida pela menor diferenciação possível.

Para **estabilizar a variância** utilizam-se transformações paramétricas, baseadas na transformação de Box-Cox, que é dada por

$$Z_t = T(Y_t) = \begin{cases} \frac{Y_t^{\lambda}}{\lambda}, \lambda \neq 0, \\ log(Y_t), \lambda = 0 \end{cases} , \lambda \in [-1, 1]. \tag{3.13}$$

A Tabela 3.1 apresenta as transformações mais usuais que derivam da transformação de Box-Cox.

Tabela 3.1: Transformações usuais de Box-Cox.

λ	Transformação
-1	$Z_t = \frac{1}{Y_t}$
$-0,5$	$Z_t = \frac{1}{\sqrt{Y_t}}$
0	$Z_t = \log(Y_t)$
$0,5$	$Z_t = \sqrt{Y_t}$
1	$Z_t = Y_t$

Note-se que algumas destas transformações apenas estão definidas para séries temporais de valores positivos, nomeadamente a transformação referente ao valor nulo de λ. Para ultrapassar esta dificuldade, nas séries com valores negativos, deve-se encontrar uma contante c, de tal modo que $Y_t + c > 0$ e posteriormente recorrer-se à transformação referida.

Análise de Estacionariedade

Uma forma de analisar a estacionariedade de uma série é através da representação gráfica da mesma ao longo do tempo. Este tipo de análise é subjetivo e portanto carece da confirmação através de testes estatísticos formais. O teste de Dickey-Fuller (DF), o teste Dickey-Fuller Aumentado (ADF), o teste Phillips-Perron (PPT) e o teste Kwiatkowski-Phillips-Schmidt-Shin (KPSS) são alguns dos testes mais utilizados para o efeito, mas nem todos possuem as mesmas hipóteses. Nos três primeiros, a hipótese nula é a da presença de uma raiz unitária, ou seja, pressupõem a não estacionariedade e quando esta é rejeitada fornecem o grau de diferenciação (regular) necessário para tornar a série estacionária. Já o teste KPSS a hipótese nula é que a série é estacionária. Em todos os casos, a hipóteses nula é rejeitada a um nível de significância α, ou seja quando o valor de prova é igual ou inferior a α. Para mais informações consultar os trabalhos de Dickey e Fuller (1979), Said e Dickey (1984), Phillips e Perron (1988) e Kwiatkowski *et al.* (1992).

3.3 Valores em Falta

Na análise de séries temporais a existência de observações em falta ou omissas pode ser problemática. Esta ausência que pode ocorrer por diversos motivos, tais como falta de registo dos dados devido à indisponibilidade humana ou avaria do aparelho de registo, entre outros, pode provocar estimativas e previsões imprecisas. Embora alguns métodos possam ser implementados com a sua existência e sem qualquer problema, existem outros que não a suportam.

Uma forma de lidar com o problema passa por substituir os valores em falta por estimativas. No entanto, apesar de já existirem bastantes *packages* disponíveis no *software* R que dispõem de funções de imputação, no que toca a imputação de séries temporais univariadas a escolha ainda é escassa, já que as mesmas têm de considerar a natureza temporal dos dados.

Moritz *et al.* (2015) são autores que testaram e compararam algumas das funções atualmente disponíveis realçando a função *na.interp()*, do *package forecast*, devido à precisão dos resultados obtidos nos vários cenários.

Esta função utiliza interpolação linear para preencher os valores em falta quando os dados não apresentam sazonalidade. Na prática, ela estima o(s) valor(es) em falta ajustando um segmento de reta entre a observação anterior e posterior ao(s) valor(es) ausente(s), ou seja, assume que a relação entre os diversos valores é aproximadamente linear. Segundo Chapra e Canale (2015), a equação da função de interpolação linear (segmento de reta) é dada por

$$f(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0), \tag{3.14}$$

onde x é a variável independente, x_0 e x_0 são valores conhecidos da variável independente e $f(x)$ é o valor da variável dependente para o valor x da variável independente.

No caso de dados com componente sazonal esta função aplica primeiramente uma decomposição STL (*Seasonal-Trend Decomposition Procedure Based on Loess)*, posteriormente uma interpolação linear aos dados ajustados sazonalmente e, por fim, readiciona a componente sazonal. A decomposição STL, proposta por Cleveland *et al.* (1990) é um método que decompõem uma série temporal nas três componentes tendência, sazonalidade e aleatoriedade, aplicado para o efeito sucessivas suavizações *loess*. Na prática, este processo defende que cada valor da série temporal $Y_i, i = 1, \ldots, n$ pode ser obtido através da expressão

$$Y_i = T_i + S_i + E_i, \tag{3.15}$$

onde T_i é a componente tendência, S_i a componente sazonal e E_t a componente aleatória.

Capítulo 4

Modelos e Métodos de Previsão

O intuito da análise de séries temporais é compreender o comportamento dos dados ao longo do tempo, retirar conclusões e fazer inferências sobre o futuro. Assim, é importante determinar o modelo que melhor descreve o comportamento da série em estudo.

Para o efeito, uma série é tipicamente dividida em duas séries, respetivamente designadas por série de treino e série de teste. A série de treino é utilizada para estimar o modelo e a série de teste é utilizada para avaliar a qualidade preditiva do modelo. Regra geral, a dimensão da série de teste corresponde a 20% da dimensão da série original, no entanto este valor depende do tamanho da amostra e do horizonte temporal de previsão pretendido.

Atualmente, existem diversos modelos de previsão de séries temporais, como os modelos de espaço de estados e os modelos de redes neuronais, entre outros. A presente dissertação foca-se no estudo de duas das abordagens mais utilizadas, nomeadamente, em modelos SARIMA (modelos autorregressivos integrados de médias móveis sazonais) e modelos de alisamento exponencial.

4.1 Modelos SARIMA e Metodologia Box-Jenkins

Os **modelos SARIMA** são modelos que permitem modelar e prever séries temporais estacionárias e não estacionárias, descrevendo a série Y_t como função dos seus valores passados e como combinação de choques aleatórios. Dentro destes modelos, consideram-se os modelos autorregressivos (AR), os modelos de médias móveis (MA) e os modelos autorregressivos de médias móveis (ARMA). Os modelos autorregressivos descrevem o comportamento da série através das suas observações passadas, já os modelos de médias móveis consideram que o comportamento da série é explicado

através de uma sucessão de choques aleatórios e, por fim, os modelos autorregressivos de médias móveis, combinam os dois modelos anteriores. Estes modelos revelam bons resultados quando aplicados a séries estacionárias, no entanto, o mesmo não sucede quando aplicados a séries não estacionárias. Nestes casos, deve-se recorrer a modelos integrados (ARIMA) ou, em caso de sazonalidade a modelos integrados com sazonalidade (SARIMA).

A **metodologia Box-Jenkins** é um processo iterativo de três fases que auxilia a escolha do modelo SARIMA que melhor descreve a série em estudo. Numa primeira fase este procura identificar o modelo mais apropriado, através da análise gráfica dos dados (em particular, pela FAC e FACP empíricas). Posteriomente, procede-se à estimação dos parâmetros do modelo selecionado e, por fim, avalia-se o modelo, através da análise de diagnóstico.

4.1.1 Modelos de Processos Estacionários

Os modelos de processos estacionários, ARMA, são, tal como o nome indica, modelos capazes de modelar séries estacionárias. Regra geral, as séries de fenómenos ambientais, entre outras são não estacionárias, no entanto podem ser transformadas em séries estacionárias através da aplicação de transformações de Box-Cox, tornando possível a modelação destas séries através de modelos ARMA.

Processos autorregressivos de ordem p, AR(p)

Definição 4.1.1 *Um processo Y_t diz-se um* ***processo autorregressivo de ordem p, $AR(p)$****, se satisfaz a equação*

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t, \tag{4.1}$$

onde ϵ_t é um ruído branco de média nula, independente de Y_{t-k} para todo $k \geq 1$.

A equação (4.1) pode ser reeescrita como

$$\Phi_p(B) Y_t = \epsilon_t, \tag{4.2}$$

onde $\Phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ é o polinómio autorregressivo de ordem p, B^k são operadores de atraso tais que $B^k Y_t = Y_{t-k}$ e ϵ_t é um ruído branco de média nula, independente de Y_{t-k} para todo $k \geq 1$.

Se um processo Y_t é um processo AR(p), então a representação gráfica da sua função de autocorrelação, FAC, apresenta um decaimento exponencial ou sinusoidal

amortecido para zero e a sua função de autocorrelação parcial, FACP, é nula para todo o $k > p$, ou seja, apresenta, uma queda brusca para zero a partir do *lag* $p + 1$.

Processo de médias móveis de ordem q, MA(q)

Definição 4.1.2 *Um processo Y_t diz-se um* ***processo de médias móveis de ordem q, MA(q)****, se satisfaz a equação*

$$Y_t = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t, \tag{4.3}$$

ou, alternativamente,

$$Y_t = \Theta_q(B)\epsilon_t, \tag{4.4}$$

onde $\Theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ é o polinómio de médias móveis de ordem q e ϵ_t é um ruído branco de média nula.

Se um processo Y_t é um processo MA(q), então a representação gráfica da sua FAC, apresenta uma queda brusca para zero a partir do *lag* $q + 1$ e a sua FACP apresenta um decaimento exponencial ou sinusoidal amortecido para zero.

Processo autorregressivo e de médias móveis, ARMA(p, q)

A combinação dos dois processos descritos resulta num processo autorregressivo e de médias móveis, ARMA(p, q).

Definição 4.1.3 *Um processo Y_t diz-se um* ***processo autorregressivo e de médias móveis de ordens p e q, ARMA(p, q)****, se satisfaz a equação*

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \tag{4.5}$$

ou, alternativamente,

$$\Phi_p(B) Y_t = \Theta_q(B)\epsilon_t, \tag{4.6}$$

onde $\Phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ e $\Theta_q(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ são, respetivamente, os polinómios autorregressivo e de médias móveis de ordens p e q e ϵ_t é um ruído branco de média nula, independente de Y_{t-k} para todo o $k \geq 1$.

As funções de autocorrelação e autocorrelação parcial de um processo ARMA(p, q) são uma combinação das FAC e FACP de processos AR(p) e MA(q).

Assim, se um processo Y_t é um processo ARMA(p, q), então as representações gráficas da sua FAC e FACP apresentam um decaimento exponencial ou sinusoidal amortecido para zero.

Processo autorregressivo e de médias móveis sazonal, SARMA$(P, Q)_s$

Os processo autorregressivos e de médias móveis sazonais, SARMA$(P, Q)_s$, surgem como uma extensão dos processos ARMA(p, q) e admitem, tal como o nome indica, a presença de sazonalidade na série temporal, ou seja, admitem a existência de correlação significativa entre observações desfasadas em s unidades temporais $(Y_t, Y_{t-s}, Y_{t-2s}, \dots)$.

Definição 4.1.4 *Um processo Y_t diz-se um* ***processo autorregressivo e de médias móveis sazonal de ordens P e Q, SARMA$(P, Q)_s$,*** *se satisfaz a equação*

$$Y_t = \nu_1 Y_{t-s} + \cdots + \nu_P Y_{t-Ps} + \epsilon_t + \eta_1 \epsilon_{t-s} + \cdots + \eta_Q \epsilon_{t-Qs}, \tag{4.7}$$

ou, alternativamente,

$$N_P(B^s) Y_t = H_Q(B^s) \epsilon_t, \tag{4.8}$$

onde $N_P(B^s) = 1 - \nu_1 B^s - \cdots - \nu_P B^{Ps}$ e $H_Q(B^s) = 1 + \eta_1 B^s + \cdots + \eta_Q B^{Qs}$ são, respetivamente, os polinómios autorregressivo e de médias móveis sazonais de ordens P e Q e ϵ_t é um ruído branco de média nula, independente de Y_{t-ks} para todo o $k \geq 1$.

Se um processo Y_t é um processo SARMA$(P, Q)_s$, então as representações gráficas da sua FAC e FACP apresentam um comportamento exponencial ou sinusoidal amortecido sobre os *lags* múltiplos de s e são nulas nos restantes *lags*.

Os processos autorregressivos sazonais de ordem P, SAR$(P)_s$, e os processos de médias móveis sazonais de ordem Q, SMA$(Q)_s$, são casos particulares dos processos SARMA$(P, Q)_s$. A FAC de um processo SAR$(P)_s$ apresenta um decaimento exponencial ou sinusoidal amortecido sobre os *lags* múltiplos de s e é nula nos restantes e a sua FACP apresenta uma queda brusca para zero a partir do P-ésimo *lag* múltiplo de s. A FAC de um processo SMA$(P)_s$ apresenta uma queda brusca para zero a partir do Q-ésimo *lag* múltiplo de s e a sua FACP apresenta uma queda lenta para zero sobre os *lags* múltiplos de s.

4.1.2 Modelos de Processos Não Estacionários

Como foi visto anteriomente, na Secção 3.2.2, uma série temporal pode não ser estacionária. Em tais casos deve-se aplicar uma transformação de Box-Cox e uma ou mais diferenciações de forma a tornar a série estacionária em variância e média, respetivamente. Ao aplicar uma ou mais diferenciações (regulares) obtém-se um modelo "integrado", tipicamente designado por ARIMA.

Processo autorregressivo e de médias móveis integrado, ARIMA(p, d, q)

Definição 4.1.5 *Um processo Y_t diz-se um* ***processo autorregressivo e de médias móveis integrado, ARIMA***(p, d, q)*, se satisfaz a equação*

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d Y_t = (1 + \theta_1 B + \cdots + \theta_q B^q)\epsilon_t, \tag{4.9}$$

ou, alternativamente,

$$\Phi_p(B)\nabla^d Y_t = \Theta_q(B)\epsilon_t, \tag{4.10}$$

onde $\nabla^d Y_t = (1 - B)^d Y_t$, com $d \geq 1$, é a série estacionária após d diferenciações, $\phi_1, \phi_2, \ldots, \phi_p$ são os parâmetros autorregressivos, $\theta_1, \theta_2, \ldots, \theta_q$ são os parâmetros de médias móveis e $\Phi_p(B)$ e $\Theta_q(B)$ são os polinómios autorregressivo e de médias móveis regulares.

A FAC de um processo deste tipo, não estacionário, distingue-se das FAC dos restantes processos, apresentando coeficientes positivos e um decaimento muito lento para zero.

Processo autorregressivo e de médias móveis integrado sazonal, SARIMA$(p, d, q)(P, D, Q)_s$

À semelhança dos modelos ARMA, os modelos ARIMA, podem ser estendidos de forma a incorporar uma componente sazonal. Uma vez que a existência de sazonalidade pressupõem a existência de correlação significativa entre as observações $Y_t, Y_{t-s}, Y_{t-2s}, \ldots$, a diferenciação é alterada dando origem a uma diferenciação sazonal. Uma diferenciação sazonal de primeira ordem é dada por

$$\nabla_s Y_t = Y_t - Y_{t-s} = (1 - B^s)Y_t, \tag{4.11}$$

ou seja, corresponde à diferença entre uma observação no instante t e um observação

registada s momentos antes, instante $t-s$. Esta diferenciação pode ser aplicada a uma série D vezes, definindo-se assim por

$$\nabla_s^D Y_t = (1-B^s)^D Y_t, \tag{4.12}$$

onde $D \geq 1$.

Definição 4.1.6 *Um processo Y_t diz-se um* ***processo autorregressivo e de médias móveis integrado sazonal, SARIMA*** $(p,d,q)(P,D,Q)_s$*, se satisfaz a equação*

$$\Phi_p(B) N_P(B^s) \nabla^d \nabla_s^D Y_t = \Theta_q(B) H_Q(B^s) \epsilon_t, \tag{4.13}$$

onde $\nabla_s^D Y_t = (1-B^s)^d Y_t$, com $D \geq 1$, é a série estacionária após D diferenciações, $\Phi_p(B), N_P(B^s), \Theta_q(B)$ e $H_Q(B^s)$ são os polinómios já referidos, d e D são as ordens de diferenciação das partes regular e sazonal, respetivamente.

4.1.3 Metodologia Box-Jenkins

A metodologia Box-Jenkins é um processo iterativo que auxilia a construção de modelos SARIMA. Esta é composta por três fases, nomeadamente: identificação do modelo, estimação dos parâmetros e validação do modelo (tipicamente designada por análise de diagnóstico).

Identificação do Modelo

A primeira fase da metodologia Box-Jenkins consiste na identificação de um modelo SARIMA$(p,d,q)(P,D,Q)_s$.

Para o efeito deve-se, em primeiro lugar, representar graficamente a série temporal e tentar retirar o máximo de informação possível, identificando-se a existência de variações na média e/ou variância e presença de sazonalidade. Caso a série apresente uma destas características, deve-se procurar retirar o efeito da mesma para que seja possível analisar o comportamento estacionário da série. Recorde-se que caso a série não seja estacionária nem em média nem em variância, deve-se proceder primeiramente à estabilização da variância através de uma transformação de Box-Cox e posteriormente à eliminação da tendência através de uma ou mais diferenciações regulares.

Em segundo lugar, deve-se analisar os gráficos das FAC e FACP da série original e tentar retirar informação extra que a representação gráfica da mesma não tenha

fornecido.

Por fim, deve-se comparar o comportamento das FAC e FACP da série estacionária com o comportamento das FAC e FACP dos modelos teóricos, por forma a identificar o processo que melhor se adequa à modelação da série em questão.

Estimação dos Parâmetros

Depois de identificados os tipos de modelos capazes de descrever o comportamento da série em questão pretende-se estimar os parâmetros dos mesmos. Para o efeito recorre-se a um *software* estatístico já que os métodos de estimação dos mesmos (Método dos Mínimos Quadrados e Método da Máxima Verosimilhança) possuem por vezes um elevado grau de complexidade numérica.

O Método dos Mínimos Quadrados estima os parâmetros do modelo, de tal forma que, o quadrado do valor entre os resíduos do modelo e os dados observados é minimizado. Embora este método seja um dos mais utilizados neste âmbito possui a desvantagem de não gerar estimadores centrados (estimadores cujo valor esperado corresponde ao parâmetro a estimar) no caso de modelos ARMA(p, q).

Já o Método de Máxima Verosimilhança estima os parâmetros do modelo, através de um processo iterativo que maximiza a função de verosimilhança dos parâmetros. A ideia é determinar os valores dos parâmetros de tal forma que os mesmos assumam os valores mais verossímeis de acordo com as observações registadas.

Análise de Diagnóstico

Na última fase da metodologia Box-Jenkins pretende-se avaliar se o modelo identificado (com os respetivos parâmetros) é adequado. Para o efeito, avalia-se a qualidade das estimativas para os parâmetros obtidas e a qualidade de ajustamento do modelo à série em estudo.

A avaliação da qualidade das estimativas dos parâmetros, é tipicamente efetuada através de um teste t-Student, com as seguintes hipóteses

$$H_0 : \beta_i = 0 \text{ vs } H_1 : \beta_i \neq 0,$$

onde β_i são os parâmetros estimados, $i = 1, \ldots k$ e k representa o número de parâmetros estimados. A hipótese nula é rejeitada para um nível de significância α, ou seja quando o valor de prova é igual ou inferior a α, ou equivalentemente, o valor absoluto da estatística de teste, T, é igual ou superior ao valor do quantil $1 - \alpha/2$ da distribuição t-Student com $n - k$ graus de liberdade, onde n representa o número de observações.

A qualidade de ajustamento do modelo é avaliada através dos seus resíduos, que devem cumprir todos os requisitos descritos na Subsecção 4.1.4, ou seja, devem possuir um comportamento semelhante a de um ruído branco. Caso tal não suceda o modelo deve ser rejeitado.

4.1.4 Análise de Resíduos

A presença de aleatoriedade permite, como já foi referido, avaliar a qualidade de ajustamento de um modelo. Os resíduos que correspondem à diferença entre os valores observados e os valores estimados pelo modelo, ou seja, são dados por $e_t = Y_t - \hat{Y}_t$ devem apresentar uma distribuição Normal, de média nula, variância constante e não devem ser correlacionados entre si.

Indepedência

O pressuposto de independência dos resíduos de um modelo deve ser avaliado individual e conjuntamente através da representação gráfica da sua FAC e da realização de testes estatísticos, respetivamente.

A FAC dos resíduos deve ser semelhante à FAC de um ruído branco, isto é, as autocorrelações devem ser significativamente nulas (graficamente pertencerem ao intervalo de limites $\pm 1,96\sqrt{n}$, se o nível de confiança é de 95%).

O teste estatístico mais utilizado para verificar este pressuposto é o teste de Ljung-Box (Ljung e Box, 1978), que avalia se as autocorrelações dos resíduos são nulas. No caso tem-se as hipóteses

$$H_0 : \forall \rho_i : \rho_i = 0,\, i = 1, \ldots, k \quad \text{vs} \quad H_1 : \exists \rho_i : \rho_i \neq 0,\, i = 1, \ldots, k \ .$$

A hipótese nula é rejeitada para um nível de significância α, ou seja, quando o valor de prova é igual ou inferior a α, ou equivalentemente, quando

$$Q = n(n+2) \sum_{i=1}^{k} \frac{\hat{\rho}_i^2}{n-i} > \chi^2_{1-\alpha;k-m},$$

onde n é o número total de observações, $\hat{\rho}_i, i = 1, \ldots, k$ são as estimativas das autocorrelações e $\chi^2_{1-\alpha;k-m}$ é o valor do quantil $1-\alpha$ de uma distribuição Qui-quadrado com $k-m$ graus de liberdade, em que k é o número de autocorrelações a serem testadas e m é o número de parâmetros do modelo estimados.

Na literatura não existe um consenso quanto ao número mais adequado de autocorrelações a serem testadas, k, uma vez que a sua escolha afeta a precisão do teste. Por exemplo, Tsay (2010) sugere que $k \approx ln(n)$, Shumway e Stoffer (2015) referem

a escolha arbitrária de $k = 20$ e Hyndman e Athanasopoulos (2018) propõem que $k = min(10, n/5)$ para séries não sazonais e que $k = min(2s, n/5)$ para séries sazonais. No presente trabalho considerar-se-á a sugestão de Hyndman e Athanasopoulos (2018).

Distribuição Normal

A normalidade dos resíduos de um modelo deve ser avaliada através de representações gráficas e testes estatísticos.

As representações gráficas mais utilizadas neste sentido são: o histograma e o gráfico *QQ-plot*. O histograma permite avaliar se a distribuição dos resíduos se assemelha à função densidade de uma distribuição Normal. Embora esta se caracterize pela sua simetria, se o histograma revelar uma distribuição ligeiramente assimétrica a hipótese de normalidade pode ainda não ser rejeitada pelos testes estatísticos. O gráfico *QQ-plot* representa graficamente os quantis, que devem idealmente formar uma reta que se sobreponha à reta de equação $y = x$. Se tal não suceder, mas se identificar parcialmente uma reta a hipótese de normalidade pode não ser rejeitada.

Os testes estatísticos complementam a análise efetuada pelas representações gráficas. O teste de Shapiro-Wilk e o teste de Kolmogorov-Smirnov são os mais utilizados neste sentido. De notar que o teste de Shapiro-Wilk só deve ser efetuado caso a amostra possua dimensão menor ou igual a 50. Em ambos os casos têm-se as seguintes hipóteses

$$H_0 : \text{Os erros seguem uma distribuição Normal.}$$
$$\text{vs}$$
$$H_1 : \text{Os erros não seguem uma distribuição Normal.}$$

A hipótese nula é rejeitada a um nível de significância α, ou seja, é rejeitada quando o valor de prova é igual ou inferior a α.

Média Nula e Variância Constante

O pressuposto de média nula é avaliado através de um teste t-Student, no entanto este teste só pode ser aplicado se os pressupostos de independência e normalidade se verificarem. No caso, as hipóteses são

$$H_0 : \mu = 0 \text{ vs } H_1 : \mu \neq 0.$$

A hipótese nula é rejeitada para um nível de significância α, ou seja, quando o valor de prova é igual ou inferior a α, ou equivalentemente, quando

$$\frac{\bar{e}\sqrt{n}}{s_e} \geq t_{1-\alpha/2;n-1},$$

onde $\bar{e}$ e s_e são, respetivamente, a média e o desvio padrão dos resíduos, n corresponde ao número total de observações e $t_{1-\alpha/2;n-1}$ o valor do quantil $1-\alpha/2$ da distribuição t-Student com $n-1$ graus de liberdade.

Por fim, o pressuposto de variância constante deve ser avaliado através da representação gráfica dos resíduos do modelo ao longo do tempo.

4.1.5 Seleção do Melhor Modelo

Na análise estatística de dados, em particular, na análise de séries temporais, podem-se encontrar vários modelos que, verificando todos os requisitos da análise de resíduos, demonstram ser adequados para descrever o comportamento da série. Quando tal sucede pretende-se, entre eles, escolher o modelo que produz melhores resultados, isto é, que melhor descreve a série em questão. Para o efeito existem critérios que auxiliam esta escolha, como o Critério de Informação de Akaike (AIC) e o Critério de Informação de Bayesiano (BIC).

Critério de Informação de Akaike

O **Critério de Informação de Akaike** (AIC), apresentado por Akaike (1974), define-se por

$$AIC = -2\log(L) + 2k, \tag{4.14}$$

onde L é a função de verosimilhança e k é o número de parâmetros do modelo SARIMA$(p,d,q)(P,D,Q)_s$ a estimar.

O objetivo é determinados os valores de AIC de dois ou mais modelos escolher o modelo que possui o menor AIC, já que este será o modelo que melhor se ajusta à série em estudo. Note-se que, diferenças de valores de AIC inferiores a duas unidades são insignificantes, logo, quando tal sucede deve-se escolher o modelo mais simples.

Critério de Informação Bayesiano

O **Critério de Informação Bayesiano** (BIC), introduzido por Schwarz (1978), é dado por

$$AIC = -2\log(L) + k \times \log(n), \tag{4.15}$$

onde L é a função de verosimilhança, k é o número de parâmetros do modelo SARIMA$(p,d,q)(P,D,Q)_s$ a estimar e n é o número de observações da série.

Tal como o AIC, o objetivo é selecionar o modelo que possua o menor BIC. A principal diferença entre o AIC e o BIC é a penalidade que os mesmos impõem à complexidade do modelo. Embora o AIC já penalize esta complexidade, através do termo $2k$, a penalização imposta pelo BIC, através do termo $\log(n)$, aumenta logaritmicamente à medida que a dimensão da amostra aumenta. Assim, entre os dois critérios, o BIC incentiva a seleção de um modelo mais simples, isto é, com um menor número de parâmetros.

Para além destes critérios, a medida do erro quadrático médio (EQM), dada pela equação (4.40), Secção 4.3, também pode ser utilizada para comparar modelos.

4.1.6 Previsões Pontuais e Intervalares

As **previsões pontuais** de um modelo podem ser calculadas através da expressão do mesmo, seguindo as etapas:

1. Expandir a equação do modelo e resolvê-la em ordem a Y_t (isto é, obtendo Y_t no lado esquerdo da igualdade e todos os outros termos no lado direito);
2. Reescrever a equação substituindo todas as ocorrências de t por $t+h$;
3. Substituir as observações futuras pelas suas previsões, os erros futuros por zero e os erros passados pelos seus resíduos.

A ideia é começar este processo com $h=1$, ou seja, começar por determinar as previsões a $1-passo$ e replicá-lo para $h=2,\ldots,$ até todas as previsões sejam calculadas.

As **previsões intervalares** são calculadas através da estimativas pontuais e do desvio padrão das mesmas, assumindo que os resíduos cumprem todos os pressupostos descritos na Subsecção 4.1.4. No caso, a previsão intervalar para o instante $t+h$ é dada por

$$\left(\hat{y}_{t+h|t} - z_{1-\alpha/2}\hat{\sigma}_h, \hat{y}_{t+h|t} + z_{1-\alpha/2}\hat{\sigma}_h\right), \tag{4.16}$$

onde $z_{1-\alpha/2}$ é o quantil de probabilidade $1-\alpha/2\%$ da distribuição Normal padrão, $1-\alpha$ é o valor do nível de confiança e $\hat{\sigma}_h$ é a estimativa do desvio padrão para a previsão a $h-passos$. Recorde-se que o nível de confiança mais utilizado é o de 95% e que para este caso se tem $z_{1-\alpha/2}=1,960$.

4.2 Métodos de Alisamento Exponencial

Os **métodos de alisamento exponencial** são métodos não paramétricos que conseguem acomodar mudanças de nível, declive e padrão sazonal. Entre estes métodos, destacam-se: o método de alisamento exponencial simples, desenvolvido por Brown (1959), o método de Holt, desenvolvido por Holt (1957) e o método de Holt-Winters apresentado por Winters (1960) que estende o método de Holt. Cada um destes métodos é adequado para estudar diferentes tipos de séries temporais, respetivamente, séries sem tendência ou sazonalidade, séries unicamente com tendência e séries com tendência e sazonalidade.

À semelhança dos modelos SARIMA, estes métodos descrevem o comportamento da série através das suas observações passadas. No entanto, consideram-se combinações ponderadas das mesmas. A ideia é que, quanto mais recente é a observação, maior é a sua contribuição, ou contrariamente, quanto mais antiga é a observação, menor é a sua contribuição. Assim, estes métodos vão atualizando os pesos a cada passo, podendo-se dizer que a ponderação decresce exponencialmente à medida que a distância temporal das observações aumenta.

Regra geral, estes métodos são dos mais eficientes quando aplicados a séries temporais que apresentam componentes tendência e sazonalidade que variam ao longo do tempo. No entanto, possuem a desvantagem de não permitirem realizar inferências estatísticas diretamente (devido ao seu caráter não paramétrico).

4.2.1 Método de Alisamento Exponencial Simples

O **método de alisamento exponencial simples** é um método de previsão adequado para séries simples, ou seja, séries que não apresentam nem tendência nem sazonalidade. Este método pressupõem que, dada uma observação num tempo t, Y_t, a previsão para o tempo seguinte, $\hat{Y}_{t+1}$, é dada por

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t), \tag{4.17}$$

onde $0 \leq \alpha \leq 1$ é uma constante de alisamento associada ao nível. Se α toma um valor próximo de 1, o erro da observação anterior (dado por $Y_t - \hat{Y}_t$) pesa muito na previsão de $\hat{Y}_{t+1}$, já se α toma um valor próximo de 0, o peso do erro da observação anterior é praticamente nulo.

A equação (4.17) pode ser reescrita como

$$\hat{Y}_{t+1} = \alpha Y_t + (1 - \alpha)\hat{Y}_t, \tag{4.18}$$

logo pode dizer-se que a previsão para o tempo seguinte, $\hat{Y}_{t+1}$, nada mais é do que uma média ponderada entre a observação e previsão anteriores. Para além disso, substituíndo na equação (4.18) $\hat{Y}_t$ por $\alpha Y_{t-1} + (1-\alpha)\hat{Y}_{t-1}$ obtém-se

$$\hat{Y}_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + (1-\alpha)^2\hat{Y}_{t-1}, \tag{4.19}$$

e repetindo esse processo tem-se que

$$\hat{Y}_{t+1} = (1-\alpha)^t\hat{Y}_1 + \alpha\sum_{i=0}^{t-1}(1-\alpha)^i Y_{t-i}. \tag{4.20}$$

Assim, conclui-se que se α toma um valor próximo de 1, as observações mais recentes possuem um maior peso, já se α toma um valor próximo de 0, as observações com um maior peso são as mais antigas.

Este método, é um método recursivo e por isso para que o valor para o instante t seja estimado é necessário o valor do instante anterior, $t-1$. Assim, este método carece da atribuição de uma previsão base para ser iniciado, que tipicamente é a primeira observação, ou seja, considera-se $\hat{Y}_1 = Y_1$.

4.2.2 Método de Holt

O **método de Holt** surge como uma extensão do método de alisamento exponencial simples e permite lidar com séries que apresentam tendência. Enquanto que o método de alisamento exponencial simples apenas carece da estimação do nível cada período de tempo, o método de Holt necessita de estimar o nível e o declive. No caso tem-se

$$\begin{aligned} l_t &= \alpha Y_t + (1-\alpha)(l_{t-1} + b_{t-1}), & (4.21)\\ b_t &= \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}, & (4.22) \end{aligned}$$

onde l_t e b_t representam, respetivamente, o nível e o declive no instante de tempo t e $0 \leq \alpha \leq 1$ e $0 \leq \beta \leq 1$ são as constantes de alisamento associadas ao nível e declive, respetivamente. A previsão a $h - passos$ é dada por

$$\hat{Y}_{t+h} = l_t + hb_t, \tag{4.23}$$

onde $h \geq 1$.

Tipicamente, para iniciar este método, respetivamente, o nível e a tendência,

considera-se $\hat{l}_1 = Y_1$ e $\hat{b}_1 = Y_2 - Y_1$ ou $\hat{b}_1 = (Y_4 - Y_1)/3$.

4.2.3 Método de Holt-Winters

O **método de Holt-Winters** é uma extensão do método de Holt, proposta para os casos em que as séries não só apresentam tendência, como também sazonalidade. Este método distingue o tipo de sazonalidade que pode ser aditiva ou multiplicativa. Assim, pode dizer-se que existem dois modelos distintos. Se a série em estudo apresenta uma sazonalidade independente do nível, diz-se que a sazonalidade é aditiva, já se a amplitude da sazonalidade aumenta à medida que o nível aumenta, diz-se que a sazonalidade é multiplicativa.

Modelo Aditivo

Para o modelo aditivo tem-se que

$$l_t = \alpha(Y_t - s_{t-s}) + (1-\alpha)(l_{t-1} + b_{t-1}), \tag{4.24}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}, \tag{4.25}$$

$$s_t = \gamma(Y_t - l_t) + (1-\gamma)s_{t-s}, \tag{4.26}$$

onde l_t, b_t e s_t representam, respetivamente, o nível, o declive e o fator sazonal no instante de tempo t e $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$ e $0 \leq \gamma \leq 1$ são as constantes de alisamento associadas ao nível, declive e sazonalidade, respetivamente. A previsão a $h - passos$ é dada por

$$\hat{Y}_{t+h} = l_t + hb_t + s_{t-s+h_s^+} \tag{4.27}$$

onde $h \geq 1$ e $h_s^+ = [(h-1) \mod s] + 1$.

Tipicamente, para iniciar este método considera-se que o nível, a tendência e os índices sazonais são dados por

$$\hat{l}_s = \frac{1}{s} \sum_{i=1}^{s} Y_i, \tag{4.28}$$

$$\hat{b}_s = \frac{1}{s^2}\left(\sum_{i=s+1}^{2s} Y_i - \sum_{i=1}^{s} Y_i\right), \tag{4.29}$$

$$\hat{s}_i = Y_i - \hat{l}_s, \tag{4.30}$$

onde $i = 1, \ldots, s$.

Modelo Multiplicativo

Para o modelo multiplicativo tem-se que

$$l_t = \alpha \frac{Y_t}{s_{t-s}} + (1-\alpha)(l_{t-1} + b_{t-1}), \tag{4.31}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}, \tag{4.32}$$

$$s_t = \gamma \frac{Y_t}{l_t} + (1-\gamma)s_{t-s}, \tag{4.33}$$

onde l_t, b_t e s_t representam, respetivamente, o nível, o declive e o fator sazonal no instante de tempo t e $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$ e $0 \leq \gamma \leq 1$ são as constantes de alisamento associadas ao nível, declive e sazonalidade, respetivamente. A previsão a $h - passos$ é dada por

$$\hat{Y}_{t+h} = (l_t + hb_t)s_{t-s+h_s^+}, \tag{4.34}$$

onde $h \geq 1$ e $h_s^+ = [(h-1) \mod s] + 1$.

Para iniciar este método considera-se que o nível é dado pela equação (4.28), que a tendência é dada pela equação (4.29) e que os índices sazonais são dados por

$$\hat{s}_i = \frac{Y_i}{\hat{l}_s}, \tag{4.35}$$

onde $i = 1, \ldots, s$.

Note-se que, para determinar as estimativas iniciais do nível e da sazonalidade são necessárias, respetivamente, s e $2s$ observações, o que determina que estes métodos só podem ser aplicados a séries temporais com pelo menos $2s$ observações.

Para comparar a eficácia de ambos os métodos (aditivo e multiplicativo), ou seja, escolher entre os dois o melhor, deve-se comparar os valores do erro quadrático médio, EQM, que é dado pela equação (4.40) (Subsecção 4.3).

Repare-se que, sendo os métodos de alisamento exponencial simples, de Holt e de Holt-Winters métodos não paramétricos, os seus resíduos não têm de cumprir qualquer tipo de pressuposto.

4.2.4 Previsões Pontuais e Intervalares

Os métodos de alisamento exponencial, sendo métodos não paramétricos são incapazes de produzir intervalos de previsão. No entanto, admitindo que a distribuição dos erros das previsões pontuais não se desvia muito da distribuição Normal, é possível calcular intervalos de previsão empíricos. Estes intervalos são importantes uma vez permitem quantificar a incerteza associada às previsões pontuais.

Assim, segundo Makridakis (1998), o intervalo de previsão empírico calculado a 1-passo, isto é, para um passo à frente é dado pela expressão

$$\left(\hat{Y}_t - z_{1-\alpha/2}\sqrt{EQM}, \hat{Y}_t + z_{1-\alpha/2}\sqrt{EQM}\right), \tag{4.36}$$

onde z representa o quantil da distribuição Normal padrão calculado para um nível de confiança α e EQM (medida do erro quadrático médio) é dada pela equação (4.40) (Subsecção 4.3).

Caso o objetivo seja calcular intervalos de previsão a h-passos, deve-se utilizar a expressão

$$\left(\hat{Y}_{t+h} - z_{1-\alpha/2}\sqrt{EQM_h}, \hat{Y}_{t+h} + z_{1-\alpha/2}\sqrt{EQM_h}\right). \tag{4.37}$$

onde EQM_h é dada pela expressão

$$\text{EQM}_h = \frac{1}{n-h}\sum_{t=h+1}^{n}(Y_t - \hat{Y}_{t|t-h})^2, \tag{4.38}$$

caso a série em estudo não apresente sazonalidade, ou é dada pela expressão

$$\text{EQM}_h = \frac{1}{n-h-s+1}\sum_{t=h+s}^{n}(Y_t - \hat{Y}_{t|t-h})^2, \tag{4.39}$$

caso a série apresente sazonalidade (s). Repare-se que, nestes casos a medida do EQM é determinada somente através da informação da série até ao momento t.

4.3 Medidas de Avaliação

Atualmente, existem diversos métodos de modelação e previsão de séries temporais, uns mais simples e outros mais complexos, no entanto, o aumento de complexidade nem sempre se traduz numa maior qualidade de ajuste ou previsão. Assim, há a necessidade de comparar a eficácia de diferentes métodos, comparação essa que é possível através de medidas de avaliação.

Uma vez que não existe nenhuma medida considerada consensual para avaliar todos os métodos existentes a comparação de dois ou mais métodos deve basear-se em várias medidas de avaliação. Nesta secção ir-se-á apresentar qautro medidas, nomeadamente: o erro quadrático médio, a raiz do erro quadrático médio, o erro percentual absoluto médio e a estatística U de Theil.

Em todas elas e sempre que se mencionar, n representa a dimensão da série, Y_t o valor observado no instante t e $\hat{Y}_t$ a estimativa ou previsão no instante t, de acordo com o conjunto considerado (conjunto de treino ou conjunto de teste, respetivamente).

Erro Quadrático Médio

O erro quadrático médio (EQM), define-se como a média dos quadrados da diferença entre os valores observados e as respetivas estimativas/previsões, ou seja, como

$$\text{EQM} = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2 = \frac{1}{n}\sum_{t=1}^{n} e_t^2, \tag{4.40}$$

onde e_t representa o resíduo do instante t.

Esta medida depende da escala dos dados, por isso, muitas vezes é substituída pela sua raiz (REQM), $REQM = \sqrt{EQM}$, já que esta facilita as interpretações a serem retiradas através da redução da grandeza dos valores.

Dados dois métodos de previsão considera-se melhor, isto é, mais preciso, aquele que possuir um menor EQM, consequentemente, um menor REQM.

Erro Percentual Absoluto Médio

O erro percentual absoluto médio (EPAM) refere-se à percentagem média de erro em relação à escala dos dados e é dado por

$$EPAM = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100. \tag{4.41}$$

Esta medida possui a desvantagem de não admitir observações nulas e ser sensível extremamente sensível a observações muito próximas de zero.

Entre dois métodos de previsão considera-se mais preciso aquele que apresenta um menor EPAM.

Estatística U de Theil

A estatística U de Theil compara o método para o qual é calculada com o método de previsão *naïve* e é dada por

$$U = \sqrt{\frac{\sum_{t=1}^{n-1}\left(\frac{\hat{Y}_{t+1}-Y_{t+1}}{Y_t}\right)^2}{\sum_{t=1}^{n-1}\left(\frac{Y_{t+1}-Y_t}{Y_t}\right)^2}}, \tag{4.42}$$

onde o numerador representa a variação relativa prevista e o denominador representa a variação relativa real. O método de previsão naïve é um método de previsão que assume que $Y_t = Y_{t-1}$ para todo o tempo t.

Se esta medida toma o valor 0, então as previsões do método em estudo são exatas. Já se toma o valor 1 as previsões geradas pelo modelo em estudo são iguais às previsões obtidas através do método *naïve*, ou seja, os dois são igualmente precisos. Se $U > 1$, a precisão do método em estudo é inferior à precisão do método *naïve* e contrariamente, se $U < 1$, a precisão do método em estudo é superior à precisão do método *naïve*.

Dados dois métodos de previsão considera-se mais preciso aquele que apresenta um menor valor de U, supondo que o modelo de referência utilizado em ambos os casos coincide.

Capítulo 5

Análise de Agrupamento (*Clustering*)

A análise de *clusters* designa uma série de procedimentos estatísticos que podem ser usados para classificar objetos ou indivíduos observando apenas as semelhanças ou dissemelhanças entre eles, sem definir previamente critérios de inclusão em qualquer agrupamento (Reis, 1997). O objetivo é construir grupos (*clusters*) de tal forma que objetos dentro do mesmo grupo são mais semelhantes do que objetos situados em grupos diferentes (Branco, 2004).

De forma sucinta, a análise de *clusters* caracteriza-se pelas cinco etapas:

1. Seleção dos indivíduos a serem agrupados;
2. Seleção do conjunto de variáveis a partir das quais será recolhida a informação necessária para o agrupamento dos indivíduos;
3. Definição da medida de semelhança/dissemelhança, entre cada dois indivíduos;
4. Escolha do método de agregação/desagregação dos indivíduos;
5. Validação dos resultados encontrados.

Nas seguintes secções ir-se-á descrever/apresentar os procedimentos necessários a cada uma das etapas descritas.

5.1 Estrutura dos Dados

A análise de *clusters* é aplicada a dois tipos de estrutura de dados, nomeadamente à matriz dos dados ou à matriz de proximidades.

A matriz dos dados, X, é dada por $X = [x_{ij}]$, $i = 1, \ldots, n$ e $j = 1, \ldots, p$, onde x_{ij} representa o valor observado no indivíduo i, para a variável j e consequentemente n representa o número total de indivíduos e p o número total de variáveis.

A matriz de proximidades de dissemelhanças D, é dada por $D = [d_{ij}]$, $i, j = 1, \ldots, n$, onde d_{ij} representa o valor de dissemelhança entre os indivíduos i e j. A matriz de proximidades de semelhanças, S, é definida de forma análoga.

5.2 Seleção dos Indivíduos e Variáveis

O primeiro passo, para realizar uma análise de *clusters* é a seleção dos indivíduos que se pretende agrupar. Para além disso, a escolha das variáveis a partir das quais será recolhida a informação necessária para o agrupamento é fundamental, uma vez que são estas que identificam os indivíduos em estudo, destacando as suas semelhanças ou dissemelhanças.

Naturalmente que, nem sempre as variáveis selecionadas tem a mesma unidade de medida ou estão definidas na mesma escala. Quando tal sucede o valor da medida de semelhança ou dissemelhança selecionada reflete sobretudo as variáveis com maiores valores e maior dispersão. Em alguns casos, esta consequência é favorável, no entanto quando se pretende que a contribuição das diversas variáveis seja semelhante deve-se aplicar um processo de estandardização às variáveis.

Milligan e Cooper (1988) são autores que estudaram e compararam a eficácia de diversos métodos de estandardização, destacando a eficácia da estandardização através da amplitude.

Dada um variável X, a sua estandardização através da amplitude é dada pela expressão

$$Z = \frac{X}{max(X) - min(X)}. \tag{5.1}$$

5.3 Medidas de Proximidade

As medidas de semelhança ou dissemelhança quantificam, respetivamente, a proximidade ou distância dos indivíduos que se pretende agrupar. Assim, quanto mais parecidas forem as características de dois indivíduos, maior é a sua semelhança e quanto mais distintas maior a sua dissemelhança.

Definição 5.3.1 *Uma medida de proximidade diz-se* ***medida de semelhança*** *se satisfaz as propriedades*

1. $s_{ij} \geq 0, \forall i, j$;

2. $s_{ij} = s_{ji}, \forall i, j$;

3. s_{ij} *é tanto maior quanto maior a parecença entre os objetos.*

Tipicamente, $0 \leq s_{ij} \leq 1$, no entanto pode suceder que $-1 \leq s_{ij} \leq 1$ se a semelhança depende do tipo de correlação.

Definição 5.3.2 *Uma medida de proximidade diz-se* ***medida de disemelhança*** *se satisfaz as propriedades*

1. $d_{ij} \geq 0, \forall i, j$;

2. $d_{ij} = d_{ji}, \forall i, j$;

3. $d_{ij} = 0, \forall i$.

Se além das propriedades anteriores se verificam ainda as propriedades

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

5. $d_{ij} = 0 \Leftrightarrow i = j$,

diz-se que a medida de dissemelhança é uma métrica ou uma distância.

A maioria dos métodos de análise de *clusters* utilizam algoritmos baseados em dissemelhanças. Por este motivo, sempre que se possuir uma matriz de semelhanças a mesma deve ser transformada numa matriz de dissemelhanças. Tal é possível através de uma simples transformação, como por exemplo, $d_{ij} = k - s_{ij}$, onde k é uma constante adequada.

As medidas de semelhança ou dissemelhança dependem da natureza das variáveis que podem ser quantitativas ou qualitativas. A presente dissertação apenas considera o estudo de variáveis quantitativas. Para estas, as medidas de distâncias mais utilizadas na comparação de dois indivíduos são: a distância Euclidiana e a distância de Manhattan.

Dada a matriz dos dados, X, a **distância Euclidiana** e a **distância de Manhattan** são, respetivamente, dadas por

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \tag{5.2}$$

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|. \tag{5.3}$$

5.4 Métodos de Análise de Agrupamento

As técnicas de análise de *clusters* podem ser classificadas como hierárquicas e não hierárquicas. Os métodos hierárquicos unem ou dividem os *clusters* de acordo com a proximidade dos dados, não sendo necessária a definição prévia do número de *clusters*. Já os métodos não hierárquicos, que carecem desta exigência, iniciam-se com um grupo definido de *clusters* e iterativamente transferem os elementos até que uma determinada condição seja verificada.

A escolha do tipo de método a utilizar depende da natureza dos dados e do objetivo da análise. Tipicamente, os métodos hierárquicos são preferíveis devido às representações gráficas que facilitam a sua interpretação, no entanto podem ser computacionalmente pesados de executar. Já os métodos não hierárquicos são eficientes mesmo quando aplicados a grandes conjuntos de dados, mas dependem da escolha prévia do número de *clusters* que pode ser difícil de determinar.

Os **métodos hierárquicos** dividem-se em dois tipos, **aglomerativos** e **divisivos**. Os métodos aglomerativos partem de n *clusters*, compostos por um único indivíduo cada e através de uniões sucessivas agrupam todos os indivíduos num único *cluster*. Já os métodos divisivos consideram o processo inverso, ou seja, partem de um único *cluster*, composto por todos os indivíduos e através de separações sucessivas criam n *clusters* compostos por um único indivíduo cada. A presente dissertação foca-se unicamente no estudo de métodos hierárquicos aglomerativos.

O primeiro passo dos métodos hierárquicos aglomerativos consiste em agrupar os dois indivíduos mais próximos, tendo por base a informação fornecida pela matriz de semelhanças/dissemelhanças. De seguida, esta matriz é atualizada de forma a refletir a proximidade entre o *cluster* recém-formado (dois indivíduos mais próximos) e os restantes *clusters* (compostos por um só indivíduo). Posteriormente, agrupam-se os dois *clusters* mais próximos e este processo vai sendo repetido até que todos os indivíduos pertençam a um único *cluster*.

Como referido, no final de cada passo, a matriz de semelhanças/dissemelhanças é atualizada e como pelo menos um *cluster* é composto por mais do que um indivíduo, existe a necessidade de quantificar a proximidade entre um indivíduo e um *cluster* ou entre dois *clusters*. Os métodos mais usuais para este efeito são:

- **Método de Ligação Completa**

  Neste método a distância entre dois *clusters*, A e B, é dada pela distância dos dois indivíduos mais afastados, isto é,

$$d_{AB} = max\{d_{ij} : i \in A, j \in B\}. \tag{5.4}$$

- **Método da Distância Média**

  Neste método a distância entre dois *clusters*, A e B, envolve a totalidade de elementos de cada grupo, n_A e n_B e é dada por

$$d_{AB} = \frac{\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d_{ij}}{n_A n_B}. \tag{5.5}$$

- **Método do Centróide**

  Neste método a distância entre dois *clusters*, A e B, é dada pela distância dos seus respetivos centróides, $\overline{x_A}$ e $\overline{x_B}$, isto é

$$d_{AB} = d(\overline{x_A}\,\overline{x_B}), \tag{5.6}$$

  onde $\overline{x_A} = \frac{\sum_{i \in A} x_i}{n_A}$, $\overline{x_B} = \frac{\sum_{i \in B} x_i}{n_B}$ e x_i é o vetor das p observações do objeto i.

- **Método de Ward**

  Neste método a distância entre dois *clusters*, A e B, é dada pela expressão

$$d_{AB} = \frac{n_A n_B d_{AB}^2}{n_A + n_B}, \tag{5.7}$$

  onde d_{AB}^2 é o quadrado da distância definida pela equação (5.6).

- **Método de Ward.D2**

  Neste método a distância entre dois *clusters*, A e B, é dada pela equação (5.7). De facto, este método é baseado no método de ward, diferindo do mesmo pela elevação ao quadrado das dissemelhanças antes da atualização do agrupamento.

A estrutura decorrente da aplicação destes procedimentos é tipicamente representada graficamente por meio de um **dendrograma**, que nada mais é do que, um esquema que se assemelha a uma árvore, onde os ramos se vão dividindo sucessivamente. Neste, os nós internos representam os *clusters* e a altura dos ramos indica a distância a que dois *clusters* se ligam.

Como referido anteriormente, na aplicação de métodos hierárquicos não é necessário definir previamente o número de *clusters*. De facto, este valor só é determinado

após a construção dos dendrogramas. A ideia é, analisar o dendrograma e determinar onde é que o mesmo deve ser cortado tendo em mente que *clusters* que possuem uma curta distância entre si devem ser agrupados e *clusters* que possuem uma grande distância entre si devem ser separados.

A eficácia de cada método hierárquico pode ser avaliada através do coeficiente de correlação cofenético, r_c, apresentado por Sokal e Rohlf (1962). Esta medida compara a matriz de semelhanças ou dissemelhanças dos dados com a matriz com a matriz de distâncias associada ao dendrograma (tipicamente designada por matriz cofenética). Em particular, o coeficiente de correlação cofenético calcula a correlação de Pearson entre os elementos da matriz de semelhanças ou dissemelhanças e as correspondentes distâncias a que os *clusters* se ligaram representadas no dendrograma pela altura dos ramos.

Assim, quanto maior é o coeficiente de correlação cofenético maior é a concordância entre a matriz de proximidades dos dados e a matriz cofenética, ou seja, mais idênticas são as informações fornecidas por ambas. Deste modo, entre dois ou mais processos deve escolher-se aquele que apresenta o maior coeficiente cofenético, relembrando que este toma valores entre 0 e 1.

Capítulo 6

Aplicação à Bacia Hidrográfica do Rio Douro

A União Europeia estabeleceu, a 23 de outubro de 2000, a Diretiva Quadro da Água (Diretiva 2000/60/CE do Parlamento e Conselho Europeu - DQA) que indica um quadro de ação comunitária para a proteção das águas de superfície interiores, das águas de transição, das águas costeiras e das águas subterrâneas.

Em 2005, a Assembleia da República aprovou a Lei da Água (Lei n.º 58/2005, de 29 de dezembro) transpondo para a ordem jurídica nacional a Diretiva Quadro da Água e estabelecendo ainda medidas para a gestão sustentável das águas.

Atualmente, em Portugal, a entidade responsável por esta gestão é a Agência Portuguesa do Ambiente (APA) a quem cabe entre outros objetivos instituir um sistema de planeamento adaptado às características das águas das diversas regiões hidrográficas. Para tal são recolhidos dados hidrometeorológicos e de qualidade da água nas diversas estações de monitorização que são, posteriormente, divulgados e compilados no repositório do Sistema Nacional de Informação de Recursos Hídricos (SNIRH) e ainda elaborados Planos de Gestão de Região Hidrográfica (PGRH). Ora, a presente dissertação pretende avaliar a qualidade da água da bacia hidrográfica do rio Douro.

A região hidrográfica do rio Douro (RH3), situada no Norte de Portugal, é a região hidrográfica internacional mais extensa da península ibérica. A sua área total é de aproximadamente 97500 km^2, sendo que 18600 km^2 se situam em território Português (19%) e os restantes 78 900 km^2 em território Espanhol (81%).

Segundo o Decreto-Lei n.º 347/2007, de 19 de outubro, esta região engloba a bacia hidrográfica do rio Douro localizada no território Português, as bacias hidrográficas das ribeiras da costa (localizadas entre a bacia hidrográfica do rio Douro

e a bacia hidrográfica do rio Vouga), as respetivas massas de águas de transição e subterrâneas e ainda as massas de água costeiras (tipicamente delimitadas por uma linha de referência localizada a uma distância de uma milha náutica), tal como se pode observar na Figura 6.1.

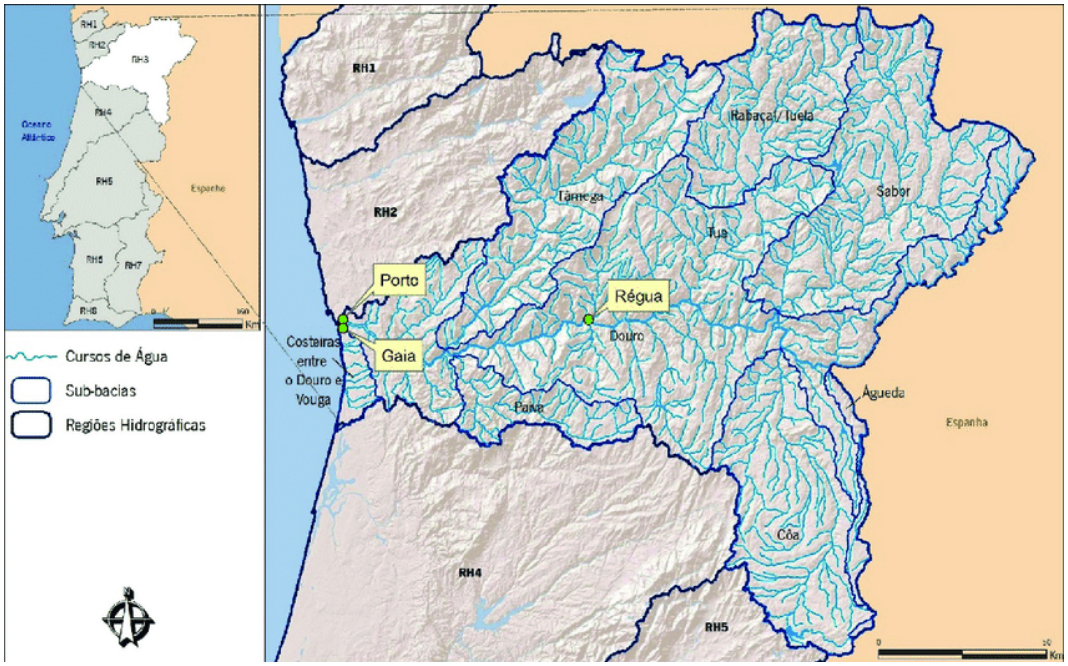


Figura 6.1: Região hidrográfica do rio Douro e respetivas sub-bacias. Fonte: Adaptado de Zucco (2016).

Esta região é composta pelas sub-bacias hidrográficas do Águeda, do Côa, do Douro, do Paiva, costeiras entre Douro e Vouga, do Tâmega, do Tua, do Rabaçal/Tuela e do Sabor e engloba total ou parcialmente os distritos de Aveiro, Braga, Bragança, Guarda, Porto, Vila Real e Viseu, perfazendo um total de 75 concelhos.

A área é composta predominantemente por florestas, terrenos agrícolas e áreas metropolitanas, sendo o fornecimento de água para irrigação (agricultura), o abastecimento das populações e da indústria os principais responsáveis pelo consumo de água da região.

6.1 Descrição da Base de Dados

Na zona da bacia hidrográfica do rio Douro existem diversas estações de monitorização da qualidade de água, no entanto, devido à falta de recursos económicos, algumas delas possuem valores em falta ou foram até mesmo desativadas em determinada altura. Esta falta de dados pode representar um problema do ponto de

vista estatístico, uma vez que pode gerar previsões não significativas. Deste modo, foram selecionadas para o presente estudo 18 estações ativas, segundo o critério de possuírem menos de 5 % de dados omissos. Na Tabela 6.1 são descritas as estações selecionadas e na Figura 6.2 é representada a localização geográfica das mesmas.

Tabela 6.1: Descrição e localização das estacões de monitorização.

Código	Nome	Abreviação	Latitude	Longitude
05K/02	Alb. Alvão - V. Real	ALV	41,350980	-7,797800
05R/01	Alb. Bastelo	BAS	41,394640	-6,660140
05P/01	Alb. Camba	CAM	41,447244	-6,901472
06N/02	Alb. Fonte Longa	LON	41,229627	-7,279693
07M/01	Alb. Ranhados	RAN	41,007689	-7,334808
02Q/02	Alb. Serra Serrada	SER	41,958460	-6,775150
06K/17	Alb. Sordo	SOR	41,270467	-7,789113
07K/01	Foz Corgo	COR	41,154740	-7,773272
07H/06	Foz Tâmega (Alb. Crestuma)	TAM	41,082264	-8,294211
07G/05	Melres	MEL	41,057789	-8,401856
07K/04	Moledo	MOL	41,154423	-7,837018
07J/02	Penude	PEN	41,078119	-7,841385
06I/04	Praia Aurora	AUR	41,274362	-8,072178
07H/04	Ponte Bateira	BAT	41,043360	-8,230788
04N/01	Ponte Vale Telhas	TEL	41,634289	-7,245828
04N/06	Quinta Maravilha	MAR	41,509390	-7,197660
06H/01	Semealho (Alb. Torrão)	SEM	41,195958	-8,160301
04J/09	Vau	VAU	41,490307	-7,966987

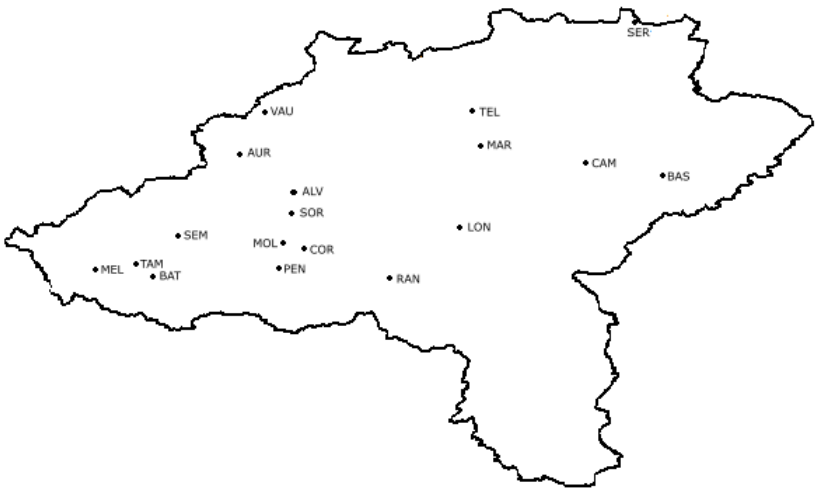


Figura 6.2: Representação da Região Hidrográfica do Rio Douro e localização geográfica das estações de monitorização em estudo.

Os dados em análise, recolhidos nestas estações, são referentes ao período de janeiro de 2002 a dezembro de 2013.

Segundo a Lei n.º 58/2005, de 29 de dezembro (Lei da Água) e nos termos da legislação específica, o bom estado das águas superficiais é o estado global em que se encontra uma massa de águas superficiais quando os seus estados ecológico e químico são considerados pelo menos bons. Por sua vez, entende-se por bom estado ecológico o estado alcançado por uma massa de águas superficiais, classificado como bom e por bom estado químico das águas superficiais o estado químico alcançado por uma massa de águas superficiais em que as concentrações de poluentes cumprem as normas de qualidade ambiental. Tendo em conta as noções descritas, selecionaram-se para este estudo as variáveis de qualidade de água *azoto amoniacal*, *carência bioquímica de oxigénio a 5 dias*, *carência química de oxigénio*, *condutividade*, *nitratos*, *oxigénio dissolvido*, *pH* e *temperatura da água*.

A variável *azoto amoniacal* (AZOTO) refere-se à quantidade de azoto, presente na amostra de água, na forma de amônia (três átomos de hidrogênio ligados a um átomo de nitrogênio - NH_3). Nas águas de superfície, este composto provém, regra geral, do escoamento de fertilizantes agrículas e resíduos de atividades industriais e a sua elevada concentração afeta a fauna e flora da região.

A *carência bioquímica de oxigénio a 5 dias* (CBO_5) mede a quantidade de oxigénio dissolvido necessária para que ocorra a decomposição (por processos biológicos) da matéria orgânica presente na amostra de água durante um período de 5 dias. Valores elevados de CBO_5 indicam uma alta concentração de matéria orgânica que utiliza oxigénio dissolvido para se decompor, logo valores altos de CBO_5 podem-se traduzir em valores baixos de oxigénio dissolvido e consequentemente perda de vida aquática.

A *carência química de oxigénio* (CQO) indica a quantidade de oxigénio necessária para oxidar a matéria orgânica biodegradável, a matéria orgânica não biodegradável e outras substâncias redutoras (substâncias que tem a capacidade de perder ou ganhar elétrons durante uma reação química) presentes na amostra de água. A CQO distingue-se da CBO_5 pois é geralmente medida através de reações químicas mais rápidas, ou seja, que permitem a obtenção de resultados num período de horas, em vez de dias.

A *condutividade* (COND) mede a capacidade que a amostra de água recolhida tem de transmitir corrente elétrica. Valores elevados de condutividade estão associados a uma maior concentração de iões dissolvidos na amostra de água, o que pode ser natural no caso de regiões com solos ricos em sais ou pode indicar o escoamento de

fertilizantes agrícolas ou a descarga de água contaminadas com produtos químicos proveniente de habitações ou indústrias.

A variável *nitratos* (NI) mede a concentração de iões de nitrato na amostra de água. A sua concentração excessiva pode ser originada pelo escoamento de fertilizantes químicos utilizados em excesso ou por descargas domésticas ou industriais.

A variável *oxigénio dissolvido* (OD) refere-se ao nível de oxigénio livre e não composto presente na água, isto é, indica a quantidade de moléculas de oxigénio (O_2) que não estão ligadas a nenhum outro elemento químico naquela amostra de água (*Fondriest Environmental Inc.*, 2023). O oxigénio dissolvido é proveniente do processo de fotossíntese realizado por organismos aquáticos e/ou transferido através do ar pelo processo de difusão, logo a sua concentração varia sazonalmente ou até diariamente. Geralmente, um nível de concentração de oxigénio dissolvido baixo traduz-se num nível elevado de poluição, que tanto pode ser característico da zona em questão, como proveniente de uma zona mais a montante.

O *pH* (pH) é um valor de uma escala numérica adimensional que varia entre 0 e 14 e indica o nível de acidez ou basicidade de uma substância, sendo, mais concretamente, determinado pela concentração de iões de hidrogénio. O valor de pH de uma água pode ser afetado por diversos fatores tais como dissolução de rochas, contaminação por resíduos industriais, absorção de gases atmosféricos, entre outros. Em particular, estas alterações podem colocar em questão ou até mesmo determinar a não sobrevivência de seres aquáticos, que na sua maioria preferem viver em ambientes cujo pH varia entre 6,5 e 9 (*Fondriest Environmental Inc.*, 2023).

A variável *temperatura* (TEMP) refere-se ao valor de temperatura da amostra de água registado, em graus Celsius (^{o}C). Os valores de temperatura variam sazonalmente devido às alterações naturais do clima, no entanto, no caso dos cursos de água também podem variar devido a fatores como a descarga de efluentes com diferentes temperaturas. Para além disso, a temperatura influencia a maioria dos processos físicos, químicos e biológicos que ocorrem na água afetando o valor do pH, do oxigénio dissolvido, entre outros e à semelhança do pH, também as suas variações afetam a sobrevivência e modo de sobrevivência dos organismos aquáticos.

Estas variáveis permitem, individualmente, classificar os níveis de poluição das águas superficiais, como apresentado na Tabela 6.2 (fornecida por um membro da APA). Estes limites são os atualmente considerados pela APA.

Tabela 6.2: Fronteiras de qualidade dos parâmetros físico-químicos aplicáveis em rios e albufeiras.

	Rios			Albufeiras		
Variável \ Estado	*Excelente*	*Bom*	*Razoável*	*Excelente*	*Bom*	*Razoável*
AZOTO	[0,0 ; 0,2]	]0,2 ; 0,4]	]0,4 ; - [	[0,0 ; 0,1]	]0,1 ; 0,2]	]0,2 ; - [
CBO_5	[0,0 ; 3,0]	]3,0 ; 4,0]	]4,0 ; - [	[0,0 ; 3,0]	]3,0 ; 4,0]	]4,0 ; - [
CQO	[0,0 ; 20,0]	]20,0 ; 40,0]	]40,0 ; - [	[0,0 ; 20,0]	]20,0 ; 40,0]	]40,0 ; - [
COND	-	-	]250 ; - [	-	-	]100 ; - [
NI	[0,0 ; 5,0]	]5,0 ; 10,0]	]10,0 ; - [	[0,0 ; 2,0]	]2,0 ; 3,0]	]3,0 ; - [
OD	[8,0 ; 12,0]	[6,0 ; 8,0[	[0,0 ; 6,0]	[8,0 ; 12,0]	[6,0 ; 8,0[	[0,0 ; 6,0]
PH	[6,5 ; 8,5]	[6,0 ; 6,5[∪]8,5 ; 9,0]	[0,0 ; 6,0[∪]9,0 ; - [	[6,5 ; 8,5]	[6,0 ; 6,5[∪]8,5 ; 9,0]	[0,0 ; 6,0[∪]9,0 ; - [
TEMP	-	-	]- ; 6,5[∪]25,5 ; - [	-	-	]- ; 6,5[∪]25,5 ; - [

Tabela elaborada de acordo com a ficha técnica DRH/DEQA 2021 - Critérios para a Classificação das Massas de Água (APA, 2021).

A estas oito variáveis adicionam-se ainda as variáveis *estação*, *mês*, *ano* e *período*, perfazendo um total de 12 variáveis existentes na base de dados.

As variáveis *estação* (EST), *mês* (MES) e *ano* (ANO) identificam, respetivamente, a estação de monitorização, o mês e o ano em que foram registadas as observações.

A variável *período* (PER) identifica a classificação do tipo de mês em que a observação foi recolhida, isto é, distingue os meses de acordo com a sua pluviosidade. Neste caso, os meses de maio a setembro são considerados meses secos e os meses de outubro a abril são considerados meses húmidos. Tal distinção foi efetuada de acordo com a sugestão de um membro da APA.

Na Tabela 6.3 é apresentada uma descrição das variáveis em estudo e são discriminadas as respetivas unidades de medida.

Tabela 6.3: Descrição das variáveis e unidades de medida.

Variável	Descrição	Níveis/Unidades
EST	Identificação da estação de monitorização i	Alb. Alvão, ... i=1,. . . , 18
MES	Identificação do mês	1, ..., 12
ANO	Identificação do ano	2002, ..., 2013
TEMPO	Identificação do tempo em que foi registada a observação na estação de monitorização i	t=1, ..., 144;
PERÍODO	Identificação do período	1 - Meses Secos 2 - Meses Húmidos
AZOTO	Concentração de azoto amoniacal	mg/l
CBO_5	Carência bioquímica de oxigénio a 5 dias	mg/l
CQO	Carência química de oxigénio	mg/l
COND	Capacidade da água de transmitir corrente elétrica	uS/cm
NI	Concentração de nitratos	mg/l
OD	Concentração de oxigénio dissolvido	mg/l
pH	Nível de pH da amostra de água	1, ..., 14
TEMP	Temperatura da amostra de água	^{o}C

6.2 Análise Exploratória de Dados

Uma análise preliminar dos dados revelou que de um modo geral para cada estação existe uma observação por mês, no entanto as mesmas não são temporalmente equidistantes, isto é, não são registadas no mesmo dia do mês. Para além do referido, existem casos em que se registaram várias observações por mês. Como se pretende efetuar uma análise temporal dos dados (séries temporais), para todos os meses em que existem múltiplas observações, as mesmas foram substituídas pela sua média. Assim, cada mês ficou com uma única observação.

Relativamente, aos valores em falta, a sua distribuição aparenta ser aleatória, ou seja, não existe um padrão que nos permita identificar o motivo pelo qual os dados em falta ocorreram exatamente naqueles meses/dias. Como por vezes alguns métodos, em particular de análise de séries temporais, não podem ser aplicados ou conduzem a erros/previsões imprecisas na ausência de observações procurou-se uma forma de evitar o problema. Hyndman e Athanasopoulos (2018) referem que uma forma de o fazer passa por substituir os valores em falta por estimativas, no caso de tal ser possível, isto é, assumindo que a série é suficientemente grande para produzir boas previsões e sugerem a utilização do método de interpolação aplicado no *R* através da função *na.interp()* do package *forecast*.

Este método de interpolação foi aplicado a cada uma das séries temporais de cada estação e de cada variável de qualidade, perfazendo um total de 144 séries univariadas (8 variáveis e 18 estações) em que os seus valores em falta foram substituídos por estimativas.

Em relação às variáveis há ainda a salientar que a variável CBO_5 possui um grande número de valores iguais a dois.

6.3 Análise de Agrupamentos (*Clustering*)

Com o objetivo de agrupar as estações de monitorização de acordo com o seu nível de poluição, foram aplicadas técnicas de agrupamento hierárquicas. Neste estudo exploraram-se duas distâncias, Euclidiana e Manhattan e quatro critérios de agregação, completo, médio, ward.D e ward.D2. A eficácia e comparação dos processos foi determinada através da análise do coeficiente cofenético, escolhendo-se o agrupamento com o valor mais elevado.

Esta análise foi feita para o conjunto das variáveis de qualidade de água e para cada uma das variáveis. Posteriormente, foi aplicada uma nova análise de *clusters*, para os meses secos e para os meses húmidos, considerando novamente o conjunto

das variáveis e cada uma delas.

Uma vez que as diferentes variáveis possuem diferentes unidades de medida, foi necessário recorrer a um método de estandardização. Neste caso os métodos aplicados foram a estandardização padrão, a estandardização através do valor máximo e a estandardização através da amplitude. Para além disso foi também realizada uma análise de *clusters* com os dados originais.

Segundo Milligan e Cooper (1988), os critérios de agregação de ward (ward e ward.D2), utilizando a estandardização através da amplitude são os que produzem resultados mais interpretáveis. Assim, o presente estudo focou-se nestes métodos e considerou a distância Euclidiana, já que esta produziu resultados mais aproximados da realidade conhecida dos níveis de poluição nas estações em estudo.

A estandardização através da amplitude foi também aplicada a cada uma das variáveis, concluindo-se que os *clusters* obtidos são exatamente iguais aos *clusters* obtidos através dos dados das variáveis originais. Deste modo, de forma a ser possível fazer uma comparação, também os *clusters* apresentados para cada uma das variáveis foram obtidos após a aplicação da estandardização através da amplitude.

Conjunto das Variáveis de Qualidade da Água

Na Figura 6.3 apresentam-se os dendrogramas obtidos pelo conjunto das variáveis de qualidade de água, aplicados à totalidade dos dados (em cima) e aos dados separados em períodos, meses secos (à esquerda) e meses húmidos (à direita). Como se pode observar, em todos os casos são obtidos três *clusters*, no entanto os mesmos diferem de acordo com o conjunto de dados considerado. Estas desigualdades sugerem uma possível alteração local dos níveis de poluição ao longo do ano, originada pela variação dos valores de uma ou mais variáveis ao longo do tempo.

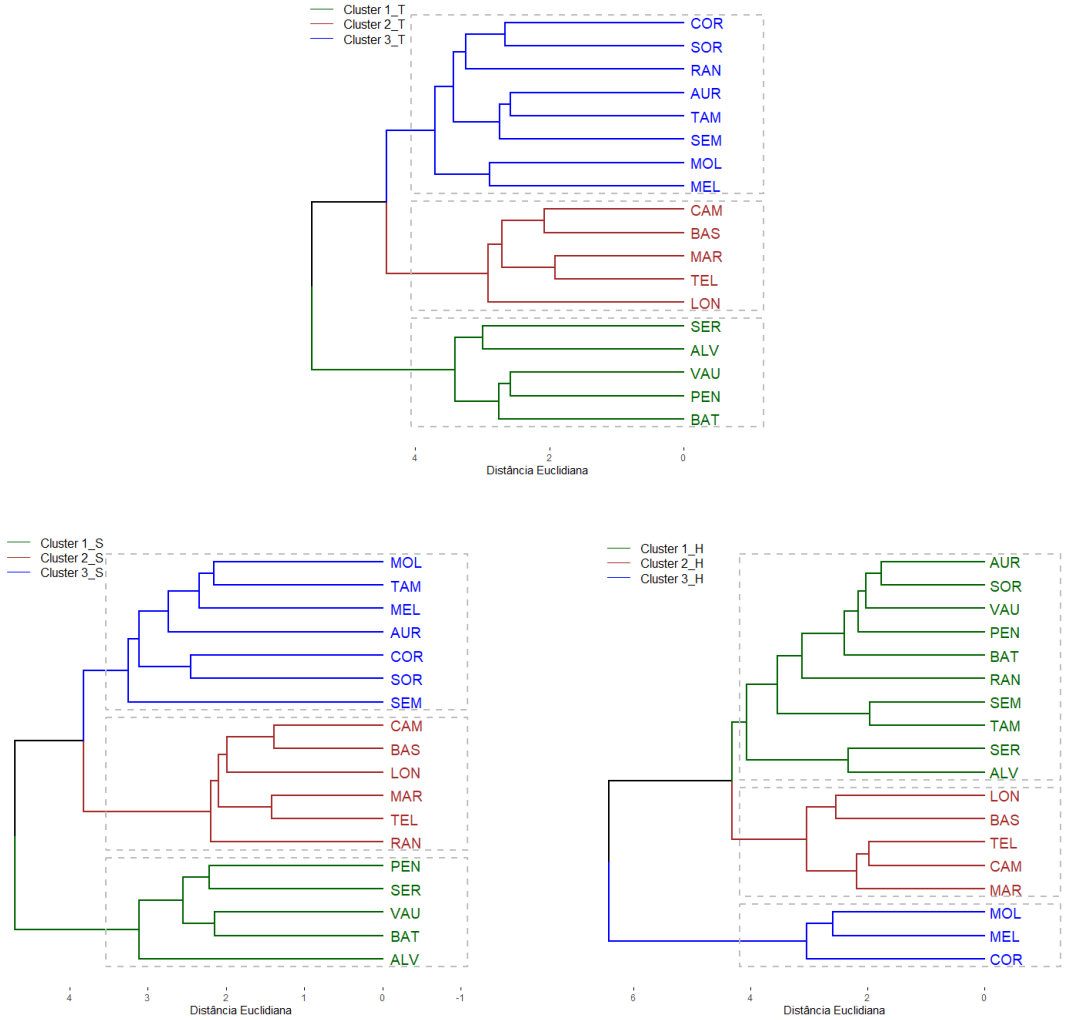


Figura 6.3: Dendrogramas das variáveis de qualidade da água, para o período observado (em cima), para os meses secos (à esquerda) e para os meses húmidos (à direita).

Com o intuito de explorar qual ou quais as diferenças que geram os *clusters*, são apresentadas nas Figuras 6.4, 6.5, 6.6 e 6.7 os *boxplots* (caixas com bigodes) de cada variável e de cada *cluster*, de acordo com o conjunto de dados considerado. As respetivas medidas descritivas são exibidas nas Tabelas A.1, A.2 e A.3 (Apêndice A).

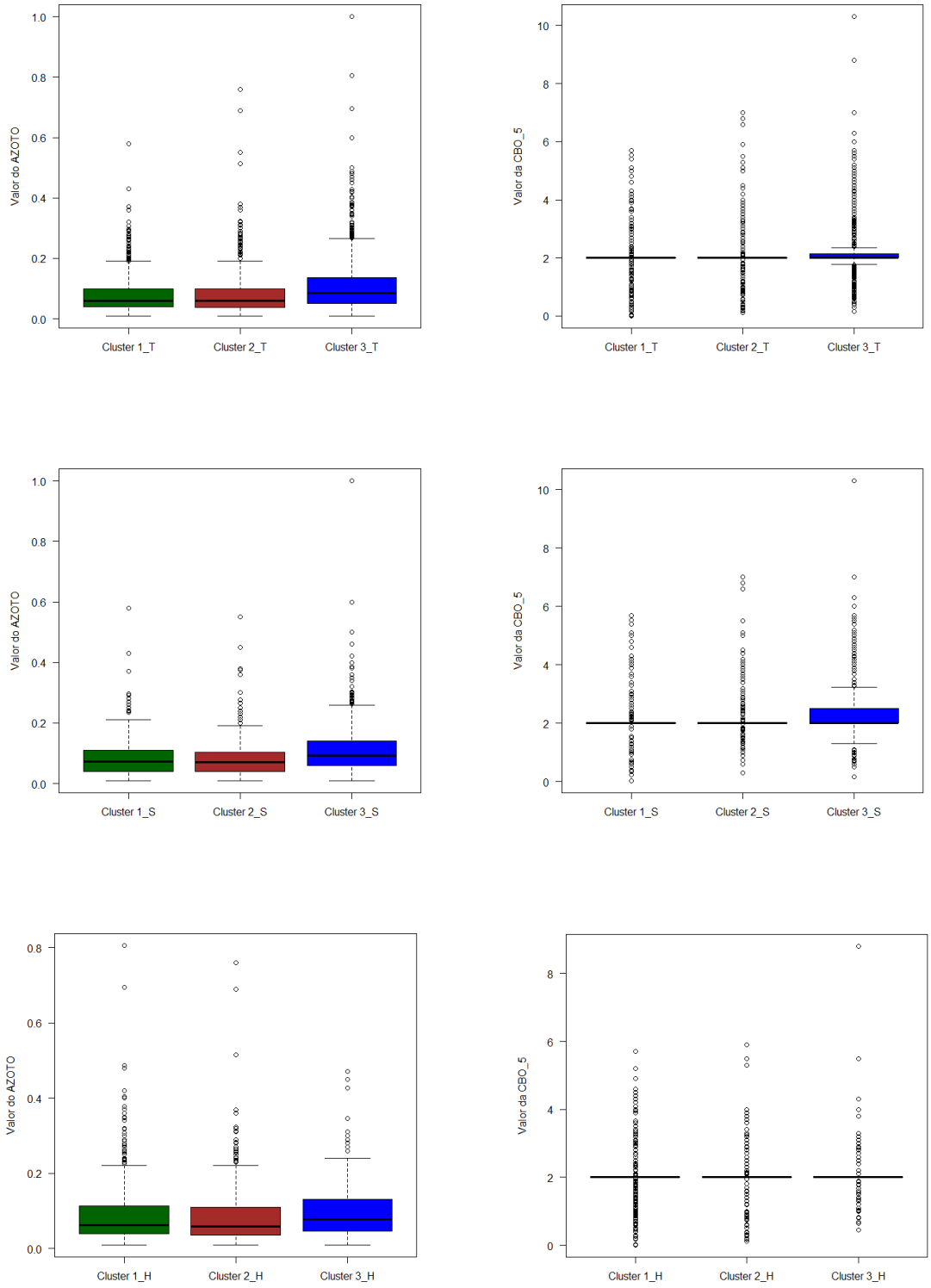


Figura 6.4: *Boxplots* da variável AZOTO (à esquerda) e da variável CBO_5 (à direita), dos *clusters* do conjunto das variáveis de qualidade da água para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

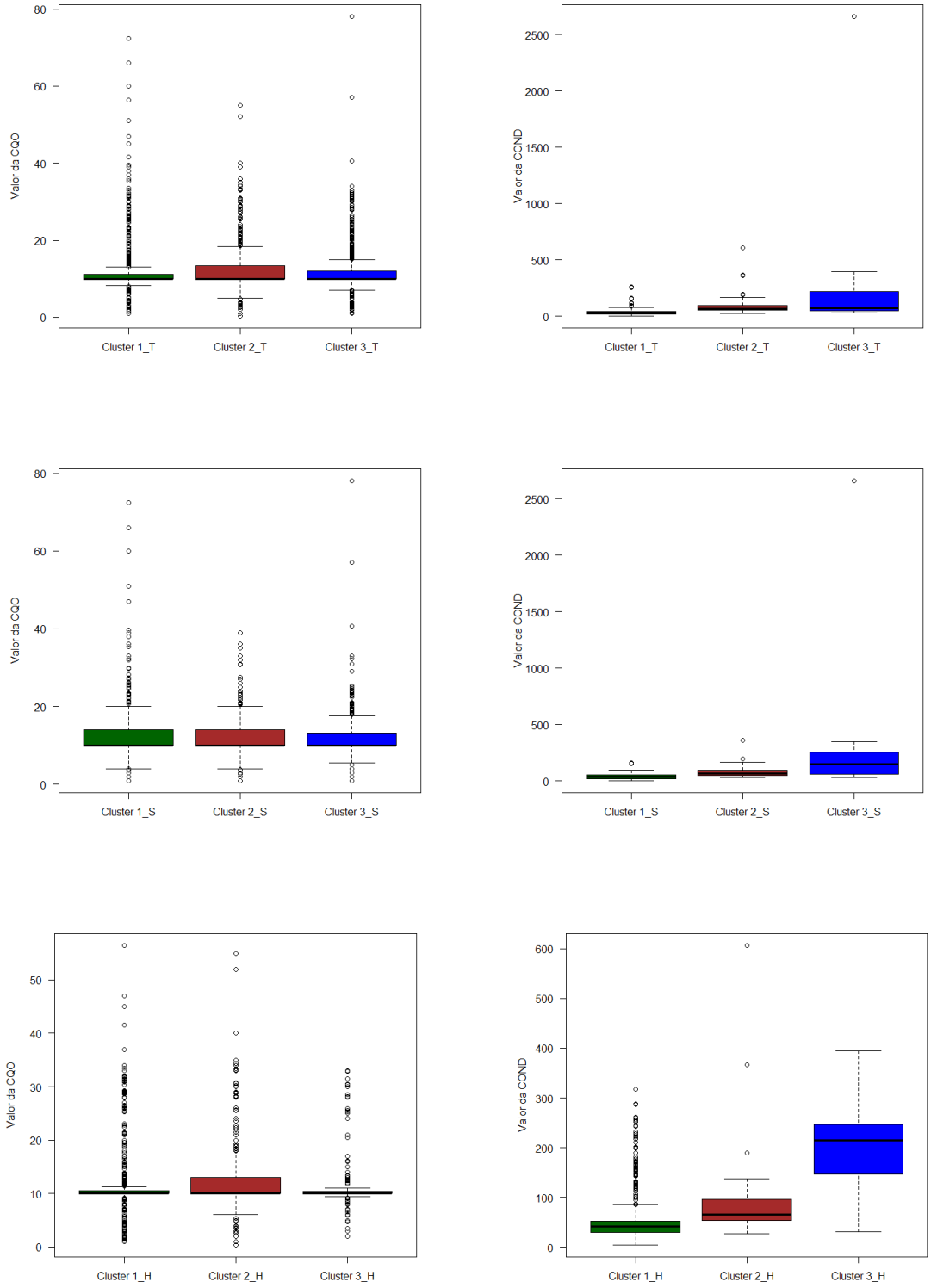


Figura 6.5: *Boxplots* da variável CQO (à esquerda) e da variável COND (à direita), dos *clusters* do conjunto das variáveis de qualidade da água para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

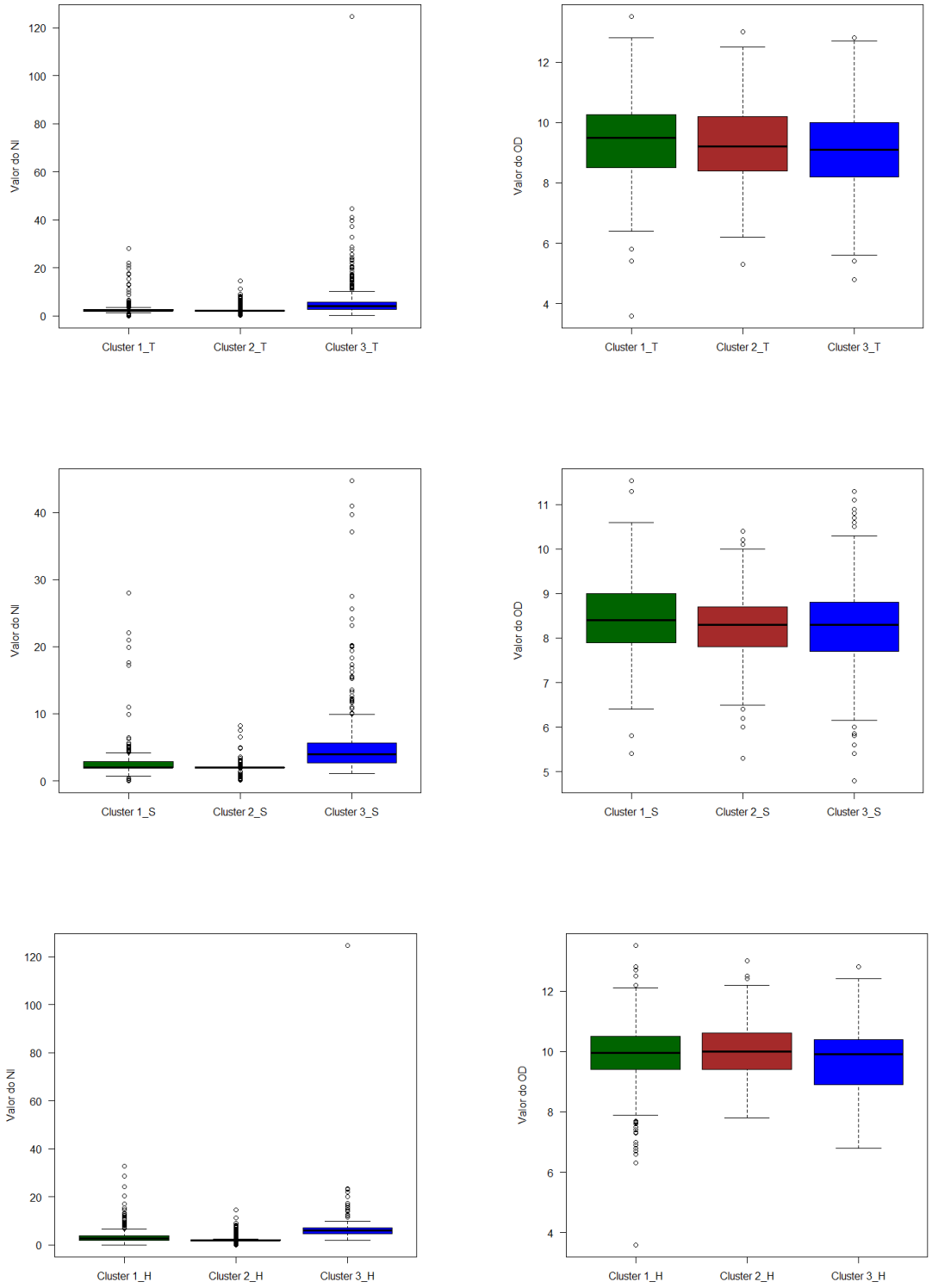


Figura 6.6: *Boxplots* da variável NI (à esquerda) e da variável OD (à direita), dos *clusters* do conjunto das variáveis de qualidade da água para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

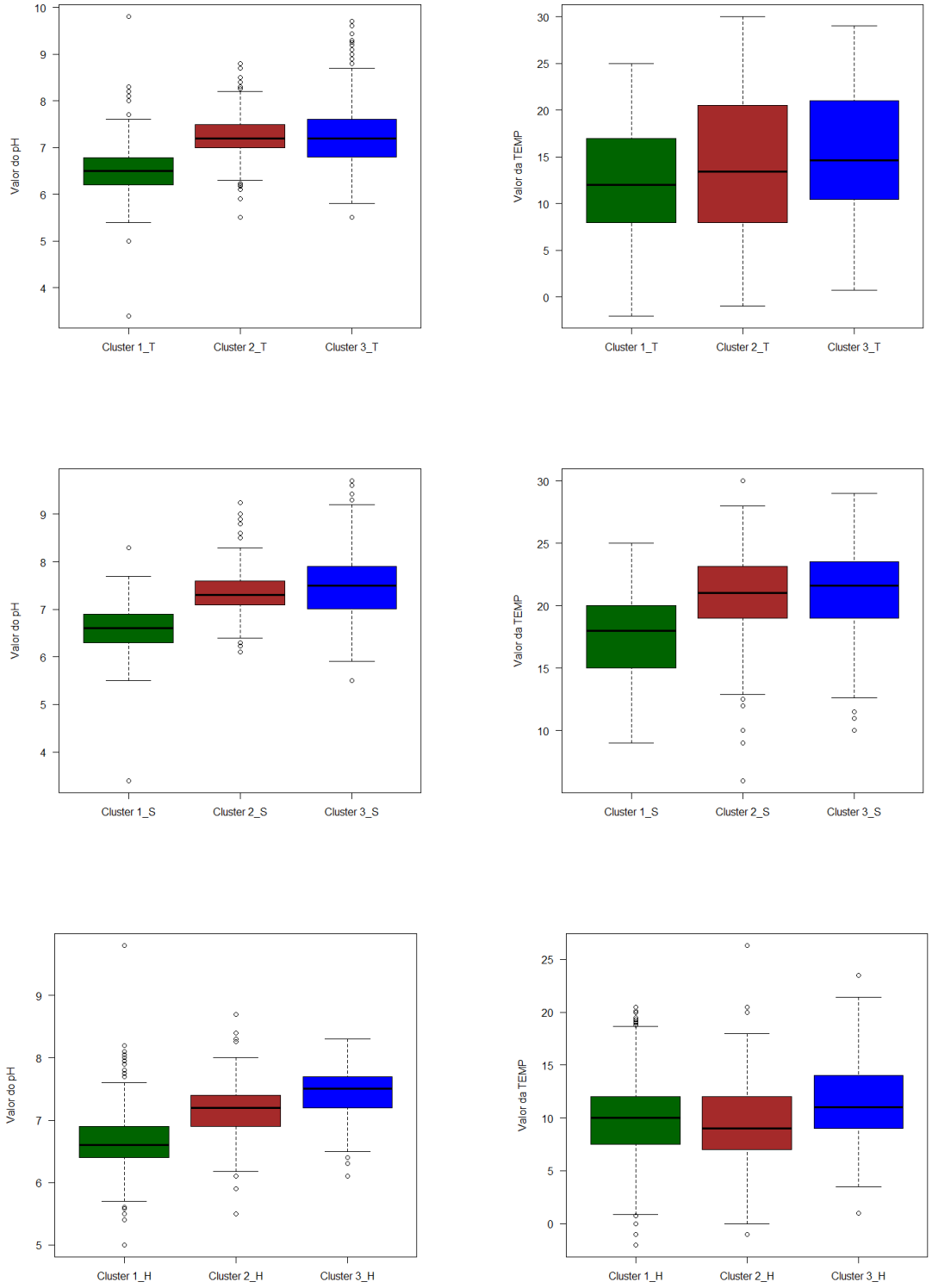


Figura 6.7: *Boxplots* da variável pH (à direita) e da variável TEMP (à esquerda), dos *clusters* do conjunto das variáveis de qualidade da água para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

Os *boxplots* da variável azoto amoniacal presentes na Figura 6.4 destacam, em cada período (total, seco ou húmido), a diferença de valores entre os *Clusters* 1 e 2 e o *Cluster* 3 (1_T, 2_T e 3_T, 1_S, 2_S e 3_S e 1_H, 2_H e 3_H, respetivamente). De facto, para cada conjunto de dados considerado os *Clusters* 3_T, 3_S e 3_H possuem valores mais elevados e variabilidades superiores.

Relativamente à variável CBO_5, é notória a presença de valores atípicos e o destaque dos *Clusters* 3_T e 3_S que contrariamente aos restantes, possuem alguma variabilidade (ver Figura 6.4). Esta diferença que é essencialmente característica das estações de monitorização que constituem o *Cluster* 3_S é observada mesmo quando a totalidade das observações é considerada (ver *Cluster* 3_T).

Na Figura 6.5 observa-se que os valores da variável CQO possuem sempre uma distribuição assimétrica positiva. Para além disso, os *Clusters* 2_T, 2_S e 2_H possuem valores ligeiramente superiores comparativamete aos restantes (*Clusters* 1_T e 3_T, 1_S e 3_S e 1_H e 3_H, respetivamente). No caso do *Cluster* 2_H, este possui ainda uma variabilidade superior comparativamente aos restantes *clusters* dos meses húmidos.

Na mesma Figura (6.5) observa-se que a variável COND possui um valor discrepante que se isola dos demais e que, portanto, poderá ter surgido por um erro de medição ou registo. Além disso, os *Clusters* 3_T, 3_S e 3_H possuem, respetivamente em cada período uma maior variabilidade e valores superiores. Em oposição, os *Clusters* 1_T, 1_S e 1_H são sempre os detentores da menor variabilidade e menores registos.

A análise da Figura 6.6 indica que o *Cluster* 3_T da variável NI possui uma distribuição assimétrica positiva, para além de uma maior variabilidade e valores superiores comparativamente aos restantes *clusters* do mesmo período (*Clusters* 1_T e 2_T). O mesmo sucede quando considerados os dados dos meses secos, ou seja, o *Cluster* 3_S destaca-se em relação aos *Clusters* 1_S e 2_S pelas mesmas características. Ainda em relação a esta variável observa-se que os *Clusters* 3_T e 3_H possuem um *outlier* severo.

Numa primeira análise dos *boxplots* referentes à variável oxigénio dissolvido, presentes na Figura 6.6, não se detetam diferenças significativas entre os *clusters* gerados em cada um dos casos, de facto, para cada período considerado (total, seco e húmido) os *clusters* gerados possuem médias e medianas idênticas. No entanto, numa análise mais atenta repara-se que as concentrações de oxigénio dissolvido são inferiores nos meses secos comparativamente aos meses húmidos.

Na Figura 6.7 apresentam-se os *boxplots* que distinguem os níveis de pH. Observando os mesmos repara-se que os *Clusters* 1_T, 1_S e 1_H possuem valores mais reduzidos comparativamente aos restantes *clusters* registados em cada período (2_T

e 3_T, 2_S e 3_S e 2_H e 3_H, respetivamente).

Por fim, a análise dos *boxplots* da variável TEMP, Figura 6.7, evidencia a diferença notória e expectável de temperatura entre os meses secos e os meses húmidos. Para além disso, no caso dos meses secos, o *cluster* 1_S distingue-se por registar temperaturas de água inferiores aos restantes *clusters* (2_S e 3_S), já no caso do período observado o *Cluster* 2_T apresenta uma maior variabilidade.

Ao longo desta análise os *Clusters* 3_T, 3_S e 3_H destacaram-se por apresentar valores de azoto, condutividade e nitratos elevados, o que os classifica pela negativa, já que uma qualidade de água *Excelente* está associada a valores reduzidos das variáveis em questão. Por outro lado, os *Clusters* 1_T, 1_S e 1_H destacaram-se por apresentar valores reduzidos de condutividade, pH e temperatura. À exceção da temperatura, os valores das variáveis associam-se a uma qualidade de água *Excelente* quando os seus valores são reduzidos. Por sua vez, a temperatura sendo uma variável cuja qualidade de água *Boa* e *Excelente* se associa ao intervalo de valores [6,5;25,5] e cujo *Cluster* 1_S possui unicamente valores dentro deste intervalo, considera-se que os *Clusters* 1_T, 1_S e 1_H estão associados a uma melhor qualidade de água comparativamente com os restantes, pelo menos quando considerados os dados dos meses secos. Finalmente, os *Clusters* 2_T, 2_S e 2_H que apresentam valores mais elevados de CQO e são semelhante aos restantes *clusters* relativamente às outras variáveis aparentam ter uma qualidade de água classificada entre as qualidades dos *Clusters* 1_T, 1_S e 1_H e 3_T, 3_S e 3_H.

Em suma, as conclusões retiradas vão de encontro ao conhecimento atual em que as estações de monitorização que compõem os *Clusters* 1_T, 1_S e 1_H caracterizam-se por uma qualidade de água *Excelente*, as estações que constituem os *Clusters* 2_T, 2_S e 2_H caracterizam-se por uma qualidade de água *Boa* e as estações que formam os *Clusters* 3_T, 3_S e 3_H caracterizam-se por uma qualidade de água *Razoável*.

Azoto

Considerando apenas a variável de qualidade de água AZOTO, observam-se na Figura 6.8 os dendrogramas obtidos considerando a totalidade dos dados, os meses secos e os meses húmidos. Em todos os casos são obtidos dois *clusters* que diferem de acordo com os dados considerados, o que significa que a concentração de azoto amoniacal varia consoante a estação de monitorização e o período do ano (húmido ou seco). Uma análise conjunta dos *boxplots* com os valores do azoto nos diferentes *clusters*, apresentados na Figura 6.8, e das medidas descritivas de cada *cluster*, apresentadas na Tabela 6.4, pode evidenciar as diferenças que os distinguem.

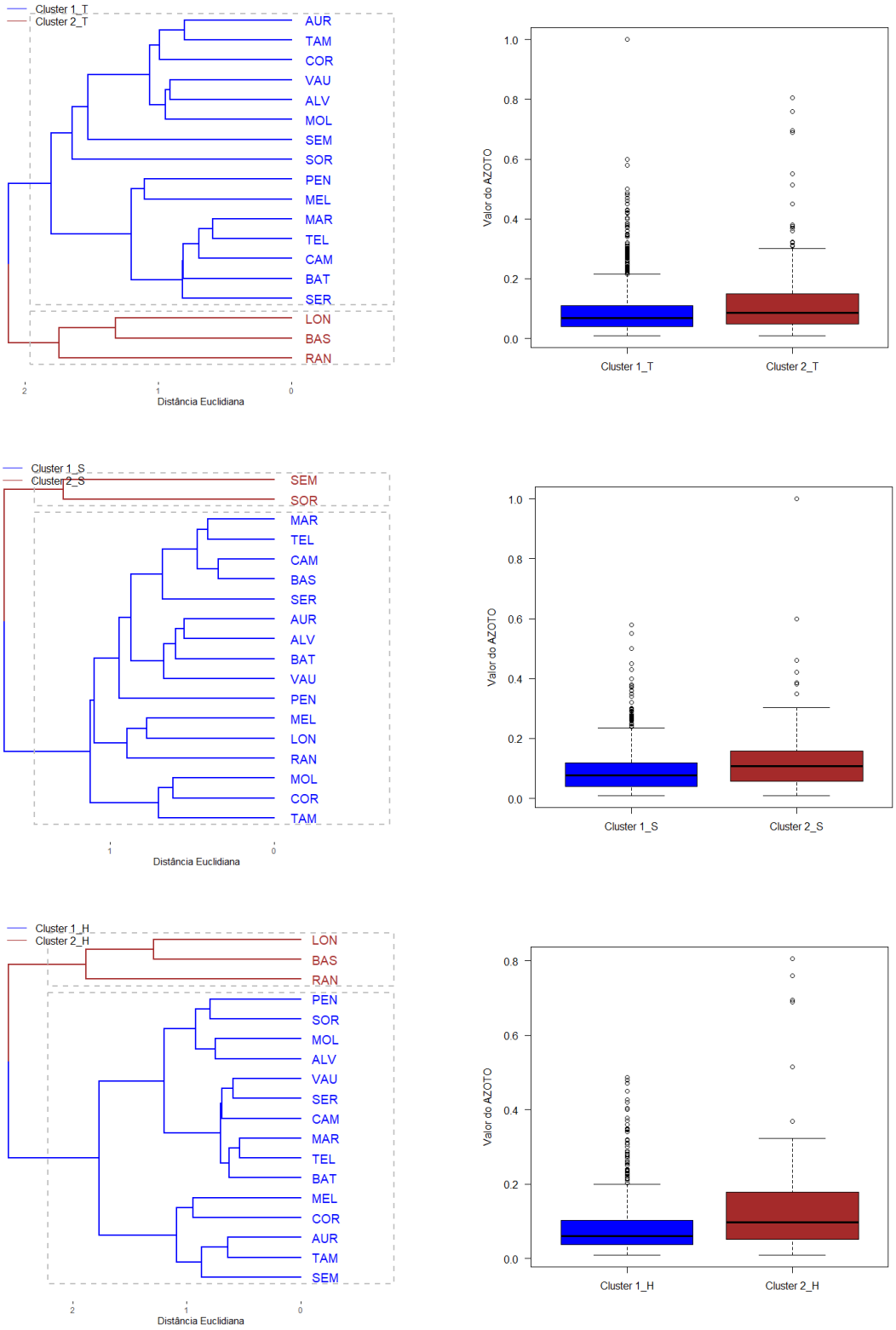


Figura 6.8: Dendrogramas da variável AZOTO (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

Tabela 6.4: Medidas descritivas da variável AZOTO, para os *clusters* obtidos.

		Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Total	*Cluster* 1_T	0,010 - 1,000	0,040	0,069	0,089	0,110	0,074	84,042
	Cluster 2_T	0,010 - 0,805	0,048	0,085	0,114	0,150	0,101	88,794
Secos	*Cluster* 1_S	0,010 - 0,580	0,040	0,076	0,093	0,119	0,07	76,143
	Cluster 2_S	0,010 - 1,000	0,057	0,108	0,135	0,157	0,129	95,793
Húmidos	*Cluster* 1_H	0,010 - 0,487	0,038	0,060	0,082	0,103	0,069	84,317
	Cluster 2_H	0,010 - 0,805	0,052	0,097	0,126	0,178	0,112	88,914

Os *boxplots* apresentados indicam que a distribuição do azoto é sempre assimétrica, independentemente do *cluster* ou conjunto de dados considerado. Contudo, existem diferenças notórias entre os dois *clusters* formados seja qual for o conjunto de dados considerado. De facto, os *Clusters* 2_T, 2_S e 2_H possuem uma amplitude interquartil e uma mediana superior comparativamente aos *Clusters* 1_T, 1_S e 1_H, respetivamente.

Cruzando as informações fornecidas pelos dendrogramas e pelos *boxplots*, concluiu-se ainda que a observação máxima foi registada num mês seco e na estação SEM ou SOR.

Comparando ainda os valores das médias e medianas dos *clusters* dos meses secos (*Clusters* 1_S e 2_S) com os respetivos valores dos *clusters* dos meses húmidos (*Clusters* 1_H e 2_H) repara-se que os valores de azoto são ligeiramente inferiores no período húmido.

Carência Bioquímica de Oxigénio a 5 dias

Por sua vez, na Figura 6.9 encontram-se, à semelhança da Figura 6.8, os dendrogramas obtidos utilizando a totalidade dos dados e os dados separados em períodos (seco e húmido), mas desta vez tendo por base a variável CBO_5. Para esta variável, em todos os casos, foram obtidos dois *clusters*, sendo um deles de dimensão reduzida, apenas com duas estações de monitorização, que consequentemente se destacam em relação às restantes. Em particular, nos meses secos tem-se as estações SEM e AUR e nos meses húmidos as estações MOL e RAN.

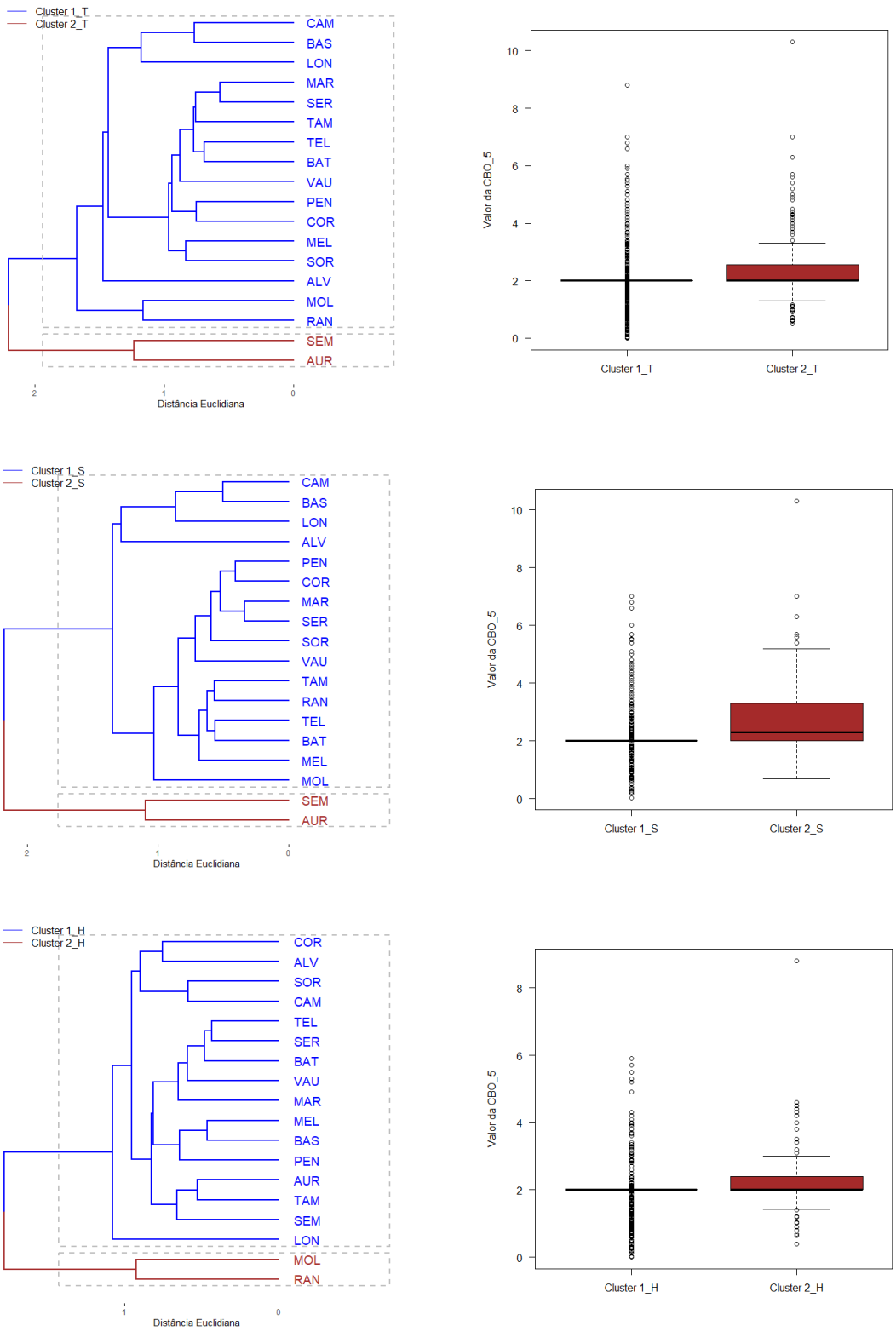


Figura 6.9: Dendrogramas da variável CBO_5 (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

Tabela 6.5: Medidas descritivas da variável CBO_5, para os *clusters* obtidos.

		Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Total	*Cluster* 1_T	0,010 - 8,800	2,000	2,000	2,071	2,000	0,677	32,681
	Cluster 2_T	0,500 - 10,300	2,000	2,000	2,389	2,525	1,073	44,919
Secos	*Cluster* 1_S	0,040 - 7,000	2,000	2,000	2,148	2,000	0,761	35,425
	Cluster 2_S	0,700 - 10,300	2,000	2,300	2,844	3,300	1,375	48,327
Húmidos	*Cluster* 1_H	0,010 - 5,900	2,000	2,000	2,000	2,000	0,565	28,259
	Cluster 2_H	0,390 - 8,800	2,000	2,000	2,187	2,400	0,842	38,516

A Figura 6.9 evidencia diferenças notórias entre o *Cluster* 1_S e o *Cluster* 2_S. De facto, as estações que constituem este segundo *cluster* (SEM e AUR) possuem valores de CBO_5 que se destacam mesmo quando a totalidade dos dados é considerada (notar que os *Clusters* 2_T e 2_S coincidem). O *Cluster* 2_H destaca-se igualmente em relação ao *Cluster* 1_H. Estas observações decorrem do facto de os *Clusters* 1_t, 1_S e 1_H possuírem amplitude interquartil nula (ver Tabela 6.5).

Carência Química de Oxigénio

Na Figura 6.10 encontram-se os dendrogramas gerados considerando os três períodos (total, seco e húmido) e a variável CQO. Em todos os períodos consideram-se dois *clusters*. Em particular, quando considerada a totalidade dos dados a estação de monitorização ALV destaca-se. De acordo com o conhecimento atual, esta estação possui uma qualidade de água *Excelente*, logo, à partida as informações fornecidas pela Figura 6.10 que exibe os *boxplots* dos *clusters* e pela Tabela 6.6 que apresenta as suas medidas descritivas devem confirmar esta conclusão.

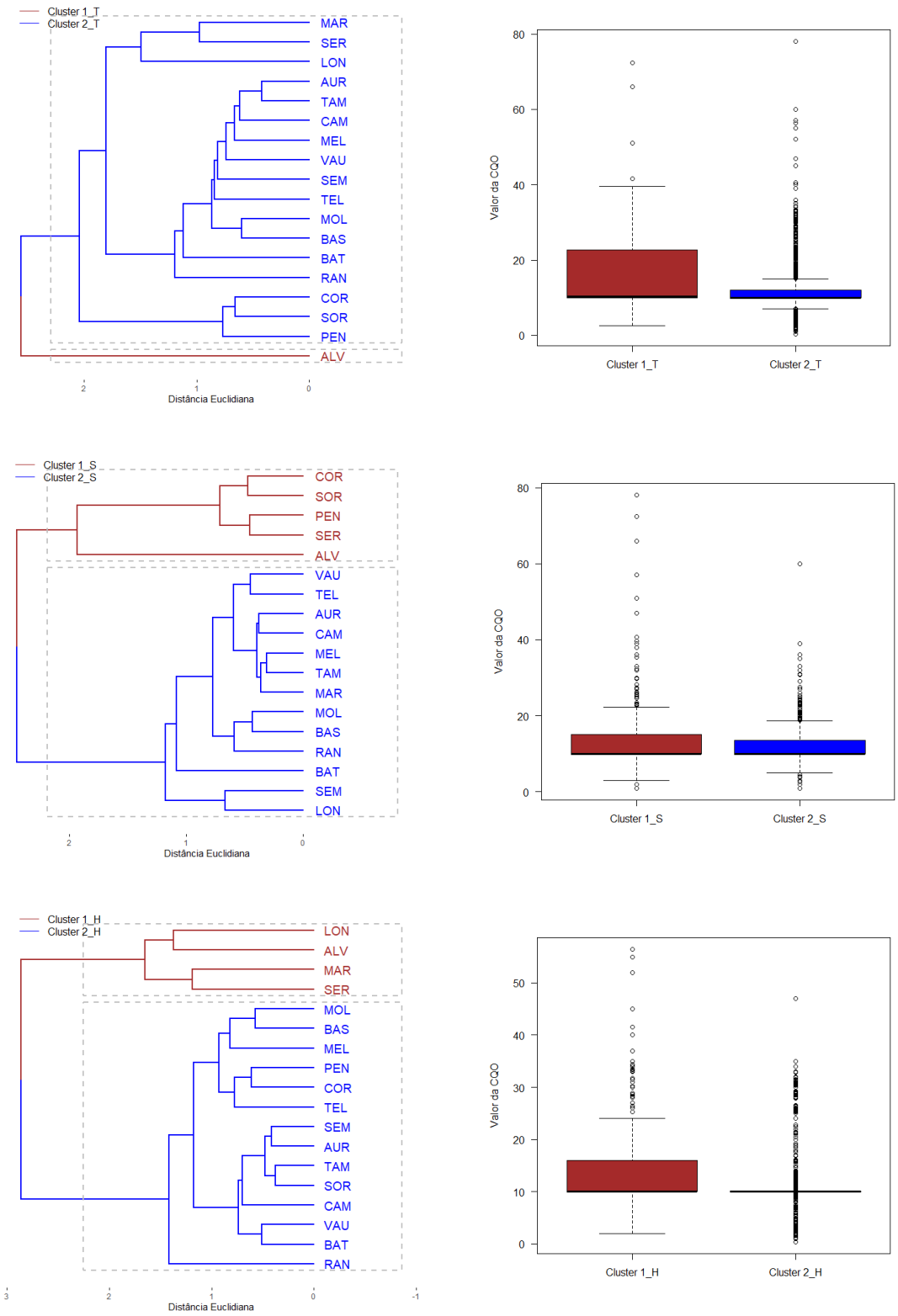


Figura 6.10: Dendrogramas da variável CQO (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

Tabela 6.6: Medidas descritivas da variável CQO, para os *clusters* obtidos.

		Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Total	*Cluster* 1_T	2,640 - 72,400	10,000	10,260	16,860	22,430	11,548	68,492
	Cluster 2_T	0,412 - 78,000	10,000	10,000	12,045	12,000	7,026	58,333
Secos	*Cluster* 1_S	1,000 - 78,000	10,000	10,000	13,810	15,000	10,158	73,579
	Cluster 2_S	1,000 - 60,000	10,000	10,000	11,960	13,450	5,949	49,729
Húmidos	*Cluster* 1_H	2,000 - 56,400	10,000	10,000	14,420	16,000	9,034	62,647
	Cluster 2_H	0,412 - 47,000	10,000	10,000	11,562	10,068	6,801	58,823

Observando os *boxplots* da Figura 6.10 verifica-se que os *Clusters* 1_T, 1_S e 1_H possuem medianas e amplitudes interquartis superiores comparativamente aos *Clusters* 2_T, 2_S e 2_H, respetivamente. Em particular, o *Cluster* 1_T composto pela a estação do Alvão (ALV) apresenta valores de CQO superiores, contrariamente ao que seria de esperar.

O mesmo se verifica para os restantes casos. Nos meses secos os maiores valores são registados nas estações ALV, COR, PEN, SER e SOR e nos meses húmidos nas estações ALV, LON, MAR e SER.

Condutividade

Na Figura 6.11 encontram-se os dendrogramas obtidos considerando os mesmos três períodos (total, seco e húmido) e a variável COND. Em todos eles são considerados dois *clusters*. Em particular, os *clusters* formados considerando a totalidade dos dados ou apenas os meses secos são iguais. Para além disso, esta figura apresenta ainda a representação gráfica, *boxplots*, dos valores de COND em cada *cluster*. Já a Tabela 6.7 apresenta as medidas descritivas de cada *cluster*.

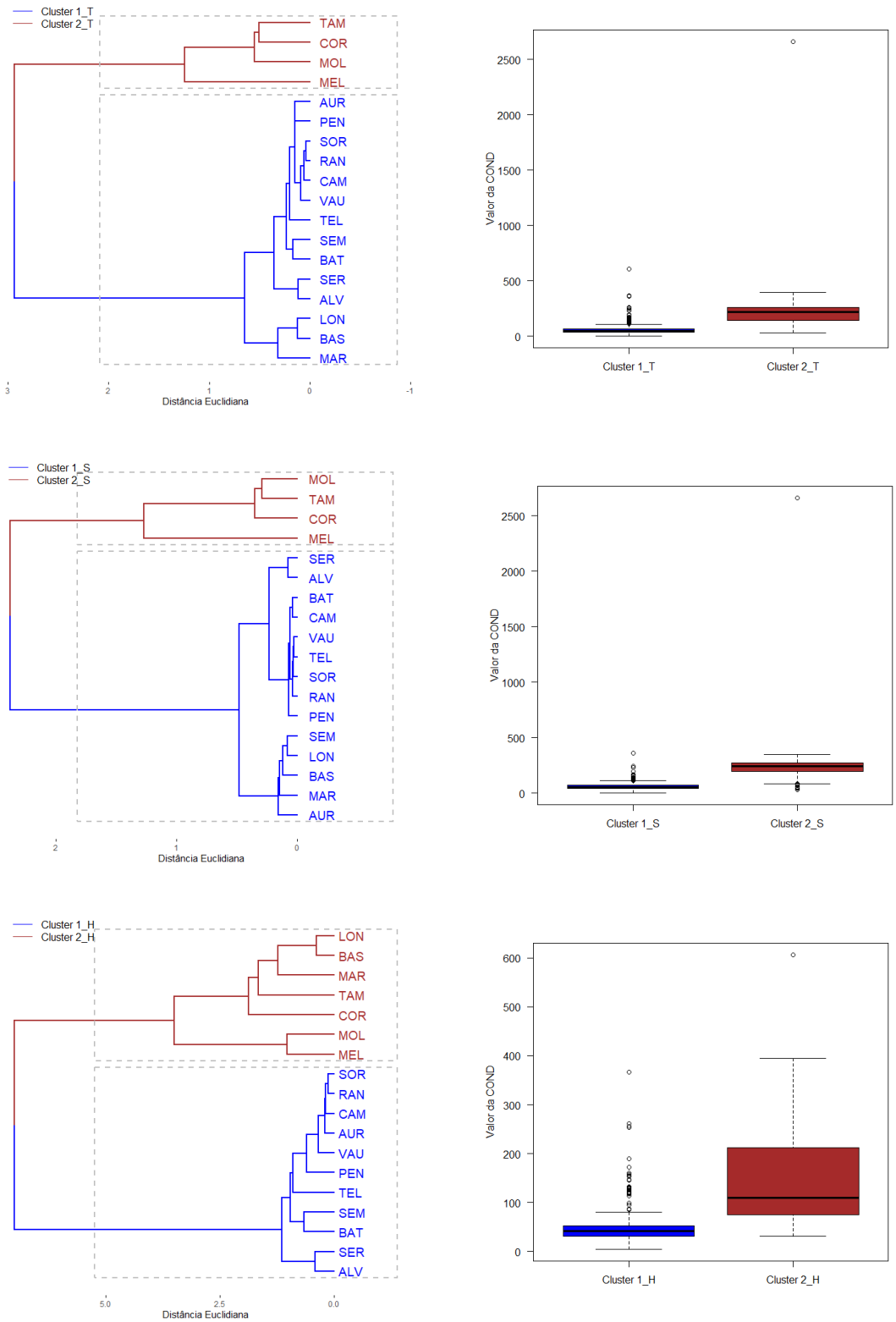


Figura 6.11: Dendrogramas da variável COND (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

Tabela 6.7: Medidas descritivas da variável COND, para os *clusters* obtidos.

		Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Total	*Cluster* 1_T	5,000 - 606,000	37,000	50,000	55,800	66,750	34,927	62,591
	Cluster 2_T	29,300 - 2660,000	144,300	220,000	203,800	259,600	130,106	63,852
Secos	*Cluster* 1_S	5,000 - 361,310	41,870	53,000	58,790	71,000	32,217	54,798
	Cluster 2_S	29,300 - 2660,000	199,000	245,300	239,600	274,200	171,317	71,496
Húmidos	*Cluster* 1_H	5,000 - 367,000	31,000	42,000	43,650	52,000	28,252	64,722
	Cluster 2_H	31,000 - 606,000	75,000	110,200	140,500	212,600	79,436	56,521

Observando a Figura 6.11 verifica-se que os *Clusters* 2_T, 2_S e 2_H possuem valores e amplitudes interquartis superiores, em relação aos *Clusters* 1_T, 1_S e 1_H, respetivamente. Para além disso, deteta-se a existência de *outliers* em todos os *clusters*, em particular, os *Clusters* 2_T e 2_S possuem *outliers* severos.

Os valores da medidas descritivas, presentes na Tabela 6.7, revelam que os valores de COND são inferiores no período húmido.

Nitratos

Analisando os dendrogramas da Figura 6.12 referentes à variável NI, concluiu-se que a estação de Moledo (MOL) se destaca das demais quando considerada a totalidade dos dados ou os dados dos meses húmidos. Por sua vez, a estacão de Foz Corgo (COR) destaca-se nos meses secos.

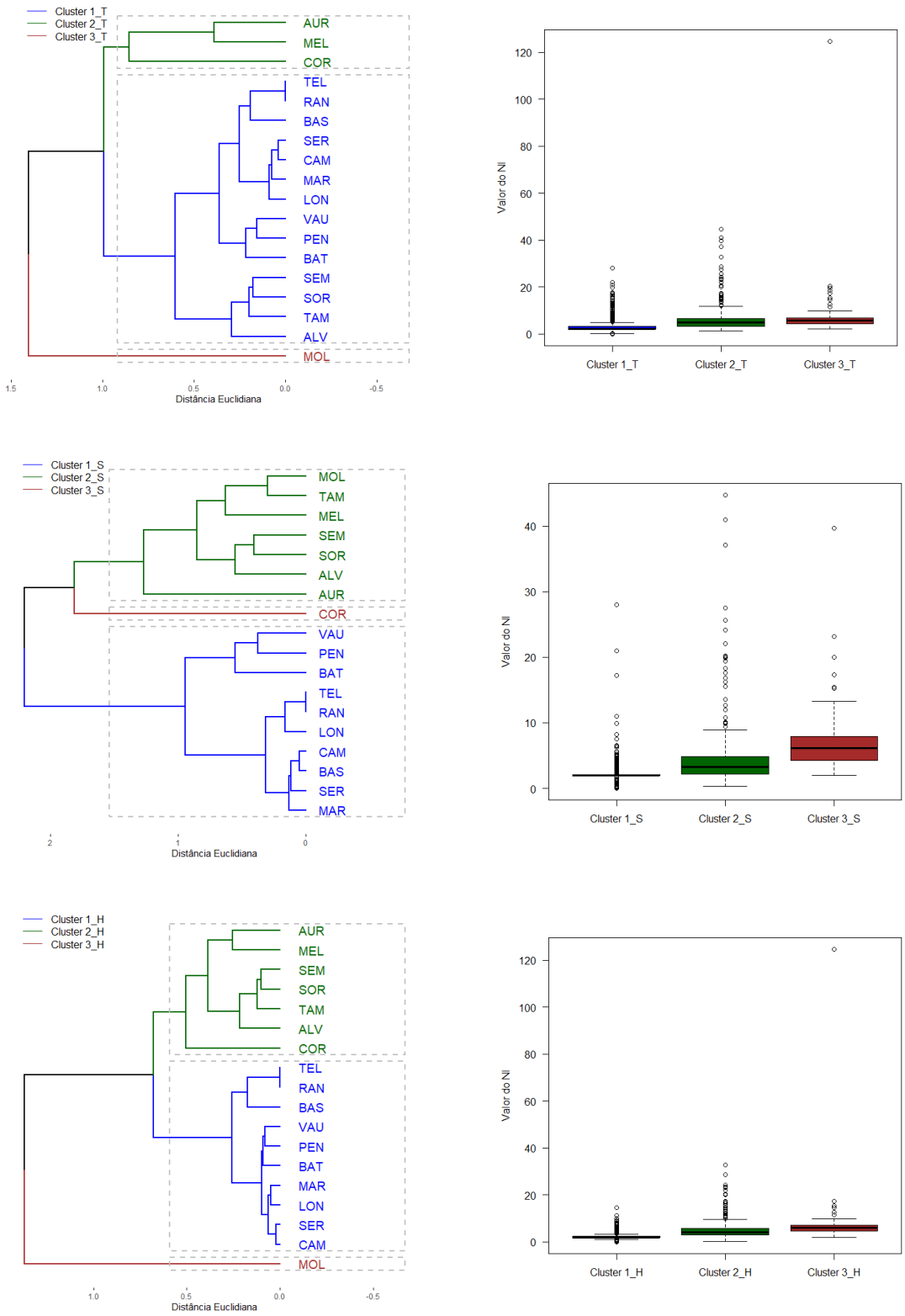


Figura 6.12: Dendrogramas da variável NI (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

Tabela 6.8: Medidas descritivas da variável NI, para os *clusters* obtidos.

		Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Total	*Cluster* 1_T	0,050 - 28,000	2,000	2,000	2,852	3,200	1,969	69,049
	Cluster 2_T	1,400 - 44,700	3,200	4,755	6,084	6,625	5,489	90,213
	Cluster 3_T	2,000 - 124,500	4,300	5,800	6,879	6,789	10,335	150,243
Secos	*Cluster* 1_S	0,050 - 28,000	2,000	2,000	2,374	2,065	1,780	74,982
	Cluster 2_S	0,340 - 44,700	2,151	3,300	4,503	4,900	4,662	103,535
	Cluster 3_S	2,000 - 39,700	4,292	6,100	7,728	7,910	6,050	78,285
Húmidos	*Cluster* 1_H	0,100 - 14,485	2,000	2,000	2,480	2,571	1,261	50,840
	Cluster 2_H	0,200 - 32,869	3,000	4,100	4,863	5,700	3,417	70,269
	Cluster 3_H	2,000 - 124,500	4,575	6,150	7,693	7,186	13,170	171,183

A análise dos *boxplots* da Figura 6.12 determina que, nos meses húmidos a estação MOL (*Cluster* 3_H) regista valores de nitratos superiores aos registados nas restantes estações (*Clusters* 1_H e 2_H), observe-se os valores das médias e medianas da Tabela 6.8. Esta estação destaca-se, mesmo quando a totalidade dos dados é considerada. Para além disso, regista-se nesta estação durante um mês húmido o maior valor de nitratos (entre todos os registos recolhidos), que é um valor atípico. Relativamente aos meses secos, a estação detentora dos maiores registos é a estação de COR (comparar os valores das médias e medianas dos *Clusters* 1_S, 2_S e 3_S).

Oxigénio Dissolvido

O oxigénio dissolvido (OD) é uma variável cuja concentração é afetada pelas concentrações de alguns dos outros compostos presentes na água, assim sendo a sua variação é natural. Tal facto pode ser observado através dos dendrogramas da Figura 6.13, uma vez que os *clusters* formados não são iguais de acordo com os dados considerados (período observado, meses secos e meses húmidos).

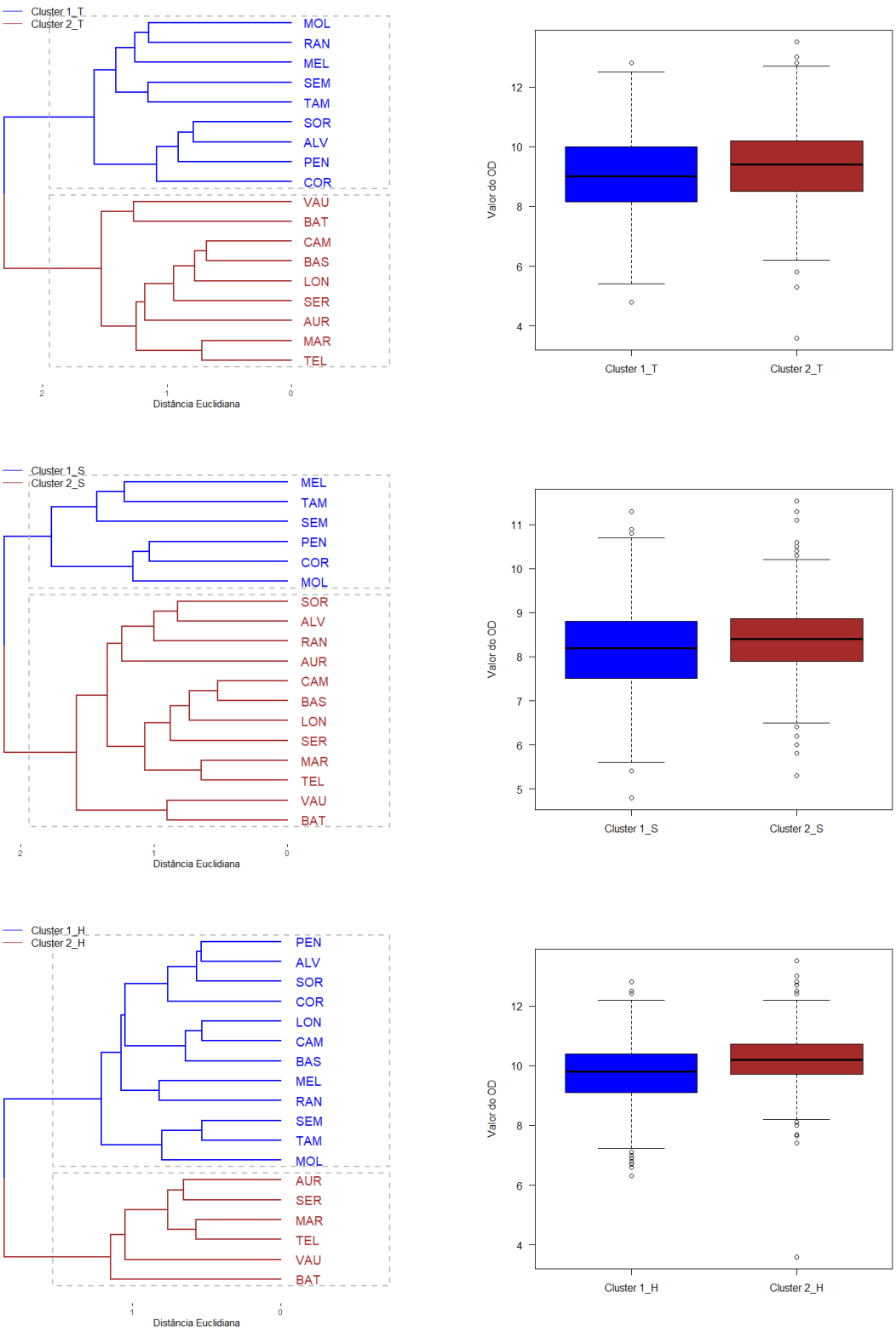


Figura 6.13: Dendrogramas da variável OD (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

Tabela 6.9: Medidas descritivas da variável OD, para os *clusters* obtidos.

		Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Total	*Cluster* 1_T	4,800 - 12,800	8,150	9,000	9,042	10,000	1,233	13,638
	Cluster 2_T	3,600 - 13,500	8,500	9,400	9,375	10,200	1,193	12,722
Secos	*Cluster* 1_S	4,800 - 11,300	7,500	8,200	8,155	8,800	1,028	12,604
	Cluster 2_S	5,300 - 11,530	7,900	8,400	8,388	8,854	0,806	9,605
Húmidos	*Cluster* 1_H	6,300 - 12,800	9,100	9,800	9,687	10,389	0,992	10,242
	Cluster 2_H	3,600 - 13,500	9,700	10,200	10,180	10,710	0,954	9,371

Analisando a Figura 6.13 e a Tabela 6.9 deteta-se que os *Clusters* 1_T, 1_S e 1_H possuem uma amplitude interquartil ligeiramente superior comparativamente aos *Clusters* 2_T, 2_S e 2_H, respetivamente. No entanto, possuem valores ligeiramente inferiores (ver médias e medianas). Comparando agora os *clusters* dos meses secos com os *clusters* dos meses húmidos verifica-se que nos meses secos os valores de OD são inferiores, de facto, as médias e medianas dos *Clusters* 1_S e 2_S são inferiores às médias e medianas dos *Clusters* 1_H e 2_H.

pH

Na Figura 6.14 observam-se os dendrogramas obtidos utilizando a totalidade dos dados e os dados separados em períodos seco e húmido, tendo por base a variável pH. Em todos os casos são considerados três *clusters* que poderão indicar se estas estações possuem um carácter ácido, básico ou neutro, após uma análise mais aprofundada dos dados.

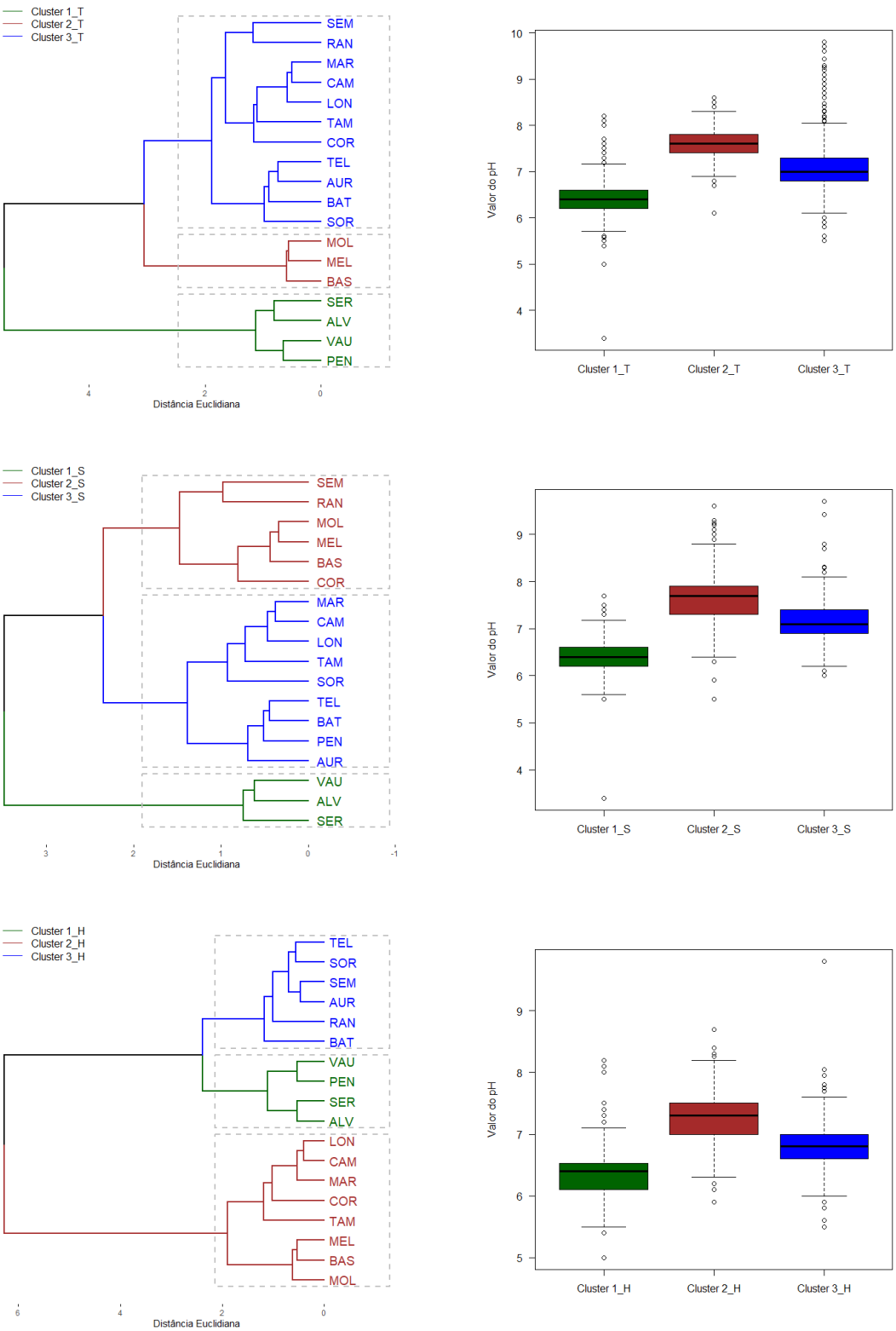


Figura 6.14: Dendrogramas da variável pH (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

Tabela 6.10: Medidas descritivas da variável pH, para os *clusters* obtidos.

		Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Total	*Cluster* 1_T	3,400 - 8,200	6,200	6,400	6,419	6,600	0,423	6,586
	Cluster 2_T	6,100 - 8,600	7,400	7,603	7,637	7,800	0,306	4,005
	Cluster 3_T	5,500 - 9,800	6,800	7,000	7,080	7,300	0,492	6,951
Secos	*Cluster* 1_S	3,400 - 7,700	6,200	6,400	6,437	6,604	0,430	6,678
	Cluster 2_S	5,500 - 9,600	7,300	7,700	7,634	7,900	0,538	7,044
	Cluster 3_S	6,000 - 9,700	6,900	7,100	7,153	7,400	0,446	6,231
Húmidos	*Cluster* 1_H	5,000 - 8,200	6,100	6,400	6,353	6,518	0,407	6,412
	Cluster 2_H	5,900 - 8,700	7,000	7,300	7,282	7,500	0,394	5,415
	Cluster 3_H	5,500 - 9,800	6,600	6,800	6,775	7,000	0,369	5,445

A Figura 6.14 sugere que os *Clusters* 1_T, 1_S e 1_H possuem, comparativamente aos restantes *clusters* observados em cada período um carácter mais ácido. Assim, de um modo geral, as estações de ALV, PEN, SER e VAU possuem os menores valores de pH registados (mais ácidos).

Ao mesmo tempo os *Clusters* 2_T, 2_S e 2_H possuem, respetivamente em cada período um carácter mais básico e por fim os *Clusters* 3_T, 3_S e 3_H possuem, em cada período um carácter aparentemente básico (ver valores médios e medianas, Tabela 6.10).

Temperatura

Por fim, na Figura 6.15 é possível observar os dendrogramas sugeridos pela variável TEMP, em todos os períodos (total, seco e húmido) e os correspondentes *boxplots*.

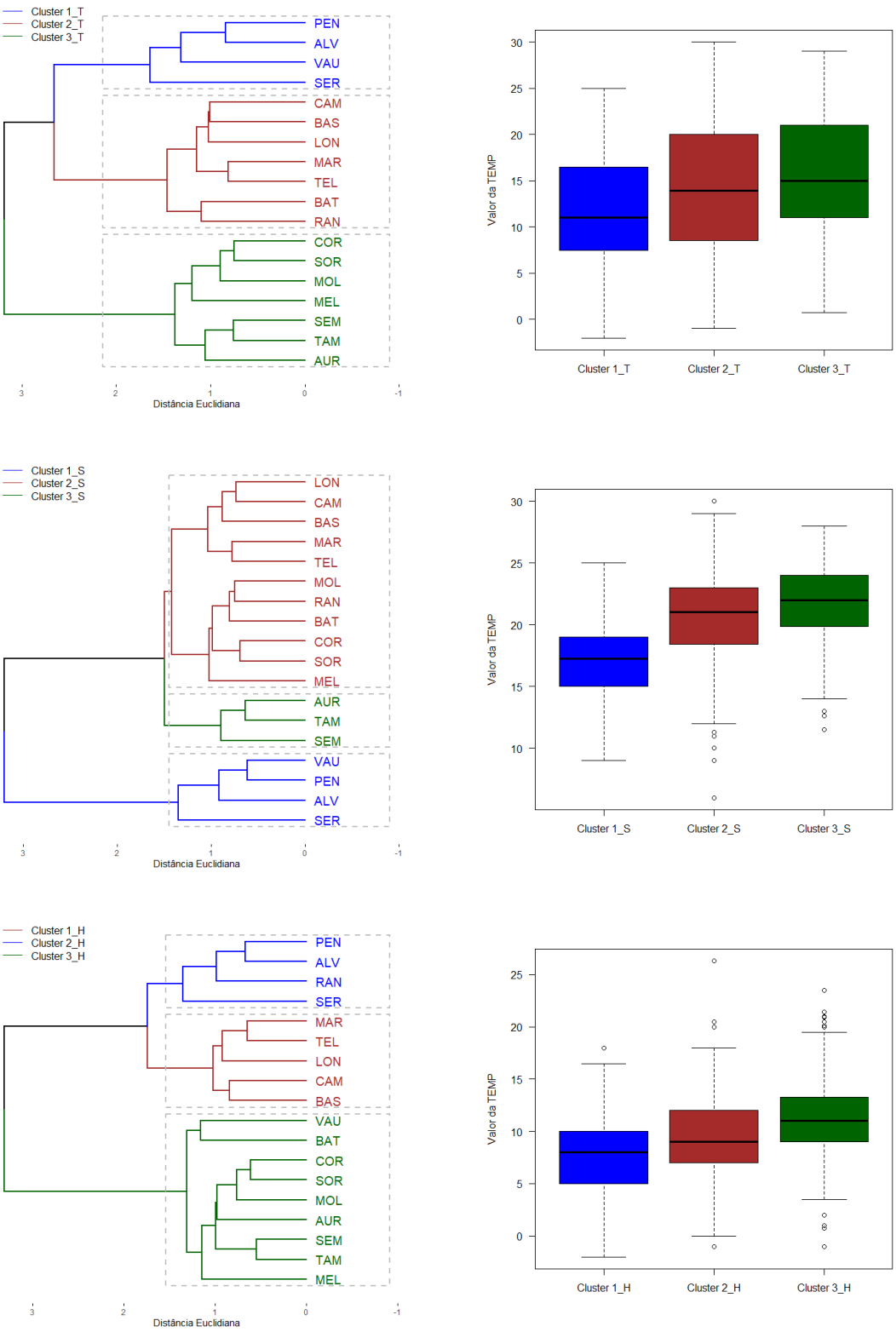


Figura 6.15: Dendrogramas da variável TEMP (à esquerda) e respetivos *boxplots* (à direita), para o período observado (em cima), para os meses secos (ao meio) e para os meses húmidos (em baixo).

Tabela 6.11: Medidas descritivas da variável TEMP, para os *clusters* obtidos.

		Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Total	*Cluster* 1_T	-2,000 - 25,000	7,500	11,000	11,720	16,430	5,643	48,151
	Cluster 2_T	-1,000 - 30,000	8,500	13,900	14,180	20,000	6,629	46,749
	Cluster 3_T	0,700 - 29,000	11,000	15,000	15,560	21,000	5,728	36,820
Secos	*Cluster* 1_S	9,000 - 25,000	15,000	17,250	17,000	19,000	3,258	19,169
	Cluster 2_S	6,000 - 30,000	18,450	21,000	20,590	23,000	3,389	16,455
	Cluster 3_S	11,500 - 28,000	19,880	22,000	21,480	24,000	3,500	16,295
Húmidos	*Cluster* 1_H	-2,000 - 18,000	5,000	8,000	7,747	10,000	3,576	46,162
	Cluster 2_H	-1,000 - 26,300	7,000	9,000	9,349	12,000	4,078	43,620
	Cluster 3_H	-1,000 - 23,500	9,000	11,000	11,450	13,220	3,535	30,875

Analisando a Figura 6.15 deteta-se que os *Clusters* 1_T, 1_S e 1_H possuem, respetivamente em cada período, os menores valores de temperatura. Ora, uma vez que estes *clusters* são constituídos pelas mesmas estações conclui-se que as estações do ALV, PEN, SER e VAU apresentam as menores temperaturas. Simultaneamente, deteta-se que os *Clusters* 2_T, 2_S e 2_H possuem, respetivamente, em cada período temperaturas compreendidas entre os restantes *clusters*e que os *Clusters* 3_T, 3_S e 3_H possuem em cada período as maiores temperaturas.

Para além disso, as medidas descritivas presentes na Tabela 6.11 indicam que a temperatura é superior nos meses secos comparativamente aos meses húmidos.

6.4 Classificação dos *Clusters* através de Medidas de Risco

O trabalho desenvolvido até ao momento permitiu comparar pontualmente os diversos *clusters*, no entanto, não é evidente que as pequenas diferenças observadas em alguns casos se traduzam em níveis de poluição da água distintos. Neste sentido, pretende-se quantificar o risco de poluição por forma a diferenciar estes casos. Para o efeito, foram selecionadas as variáveis CBO_5, COND, OD e TEMP e calculados os valores das medidas de risco simples apresentadas no Capítulo 2.

Segundo Bahnemann (2015), uma variável baseada numa população finita ou numa amostra finita de uma população maior, tem sempre uma distribuição discreta. Deste modo e por forma a se determinar a distribuição empírica ou distribuição de frequências de cada variável, o conjunto de dados de cada uma delas foi organizado tendo por base uma partição da sua amplitude.

Para cada um dos casos considerados (totalidade dos dados, meses secos ou meses húmidos) seja $S = \{S_1, \ldots, S_i\}$ o conjunto dos *clusters* de uma única variável, onde $S_i, i = 1, \ldots, I$ representam os *clusters* obtidos e I pode ser 1, 2 ou 3. O

mínimo e máximo absolutos de S (respetivamente, r_{min} e r_{max}) foram calculados e o intervalo por eles formado dividido em N subintervalos de igual amplitude, onde N é determinado através da equação (2.18) (Secção 2.3) e T é a dimensão máxima de $S_i, i = 1, \ldots I$. Posteriormente, foi calculada a probabilidade (P) e média (M) das observações de cada subintervalo, obtendo-se para cada *cluster* uma distribuição de frequências. Na Figura 6.16 é representado o histograma da variável OD do *Cluster* 1_T que ilustra a distribuição de frequências apresentada na Tabela 6.12.

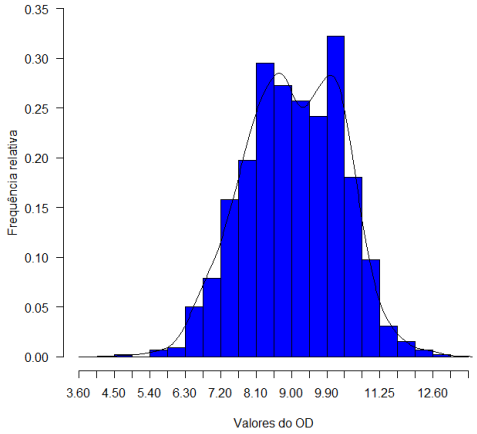


Figura 6.16: Histograma e curva de densidade da variável OD no *Cluster* 1_T.

Tabela 6.12: Distribuição de frequências do OD no *Cluster* 1_T.

M	4,800	5,550	6,182	6,571	6,927	7,411	7,868	8,313	8,763
P	0,001	0,003	0,008	0,018	0,035	0,071	0,089	0,133	0,123
M	9,205	9,652	10,102	10,520	10,941	11,429	11,867	12,450	12,800
P	0,116	0,109	0,145	0,081	0,044	0,014	0,007	0,003	0,001

Esta distribuição permitiu o cálculo das diversas medidas de risco referentes a esta variável e a este *cluster*. Analogamente, as restantes distribuições cumpriram o mesmo propósito.

Daqui em diante, repare-se que, sempre que possível, a probabilidade de perda e a função de excesso médio foram calculadas para o limiar que separa as classes *Excelente* e *Bom*, de acordo com a Tabela 6.2 (Secção 6.1). Nos casos em que não existe este limite, estas medidas foram calculadas para o limiar que distingue os níveis de qualidade *Bom* e *Razoável*.

Note-se ainda que, no cálculo da probabilidade de perda foram utilizadas duas abordagens, uma que utiliza a distribuição de frequências e outra que utiliza os

dados originais. A segunda abordagem gera uma estimativa baseada no número de observações que excede o limiar e surge como uma alternativa à primeira, uma vez que, na grande maioria dos casos e dependendo do valor para o qual a probabilidade de perda foi calculada, a primeira abordagem gera estimativas imprecisas. Tal facto deve-se ao número reduzido de classes envolvido no cálculo.

Por fim, especifica-se ainda que, o nível de confiança considerado para o cálculo do valor em risco foi 95%.

Carência Bioquímica de Oxigénio a 5 dias

Segundo a legislação em vigor, os valores da variável CBO_5 relacionam-se inversamente com o nível da qualidade da água, ou seja, quanto maiores os valores da CBO_5, pior a qualidade da água. Em concreto, o valor 3 mg/l distingue o nível da qualidade da água *Excelente* e *Bom*, dos rios e albufeiras.

Na Tabela 6.13 são apresentados os valores de diversas medidas de risco calculadas para cada *cluster* e nas Tabelas 6.14 e 6.15 as ordens dos seus valores (*rankings*), com o intuito de complementar e simplificar as possíveis conclusões a serem retiradas em termos de risco. Esta ordenação (crescente ou decrescente) dos valores de cada medida é dependente da definição da medida de risco em causa e da interpretação, em termos de qualidade de água, da variável em estudo. Por exemplo, no caso dos valores médios da CBO_5 valores superiores estão associados a um maior risco, uma vez que valores superiores a 3 mg/l se associam a níveis de qualidade reduzida (não *Excelente*).

Na Tabela 6.14 as ordens de cada medida de risco são determinadas de acordo com o conjunto de dados considerado (totalidade dos dados, meses secos e meses húmidos), já na Tabela 6.15 as ordens são determinadas independentemente do conjunto considerado, ou seja, todos os *clusters* obtidos são comparados entre si. Tal distinção permite uma visão completa de toda a informação que cada medida de risco transmite.

Para além disso, nestas duas tabelas é apresentada a soma das ordens (SO) por *cluster* que será utilizada como um índice de risco e permitirá a generalização da classificação dos *clusters* tendo em conta as diversas medidas de risco e consequente comparação. No seu cálculo e no que toca à probabilidade de perda, apenas as ordens da abordagem pela distribuição de frequências foram consideradas.

Na Tabela 6.15 é ainda exposta a soma das ordens de cada medida considerando cada conjunto de dados, isto é, para o período observado, para os meses secos e para os meses húmidos tem-se, respetivamente, 9 valores que representam cada uma das

medidas de risco e por fim a sua soma, SO. Por exemplo, para a média tem-se três valores 7, 10 e 4 que correspondem à soma dos *rankings* da totalidade dos dados, dos meses secos e húmidos, nesta ordem. O mesmo sucede para as restantes medidas de risco e no final tem-se as SO: 68,5, 84 e 36,5, para a totalidade dos dados, para os meses secos e para os meses húmidos, respetivamente. Este acréscimo permite a comparação dos períodos considerados, a título de exemplo, para a variável CBO_5, conclui-se que a classificação obtida para os meses secos é pior em termos de risco do que a classificação obtida para os meses húmidos.

Tabela 6.13: Medidas de risco por *cluster* para a variável CBO_5.

		M	V	DP	DA	SS	PP*	E	FEM	VaR	TVaR
Total	*Cluster* 1_T	2,071	0,453	0,673	0,336	0,311	0,062 / 0,062	1,271	0,061	3,237	0,176
	Cluster 2_T	2,389	1,142	1,069	0,703	0,886	0,163 / 0,167	1,691	0,215	4,378	0,281
Secos	*Cluster* 1_S	2,148	0,568	0,754	0,436	0,417	0,079 / 0,086	1,344	0,093	3,905	0,150
	Cluster 2_S	2,844	1,862	1,364	0,986	1,355	0,283 / 0,308	2,023	0,442	5,360	0,244
Húmidos	*Cluster* 1_H	2,000	0,315	0,561	0,238	0,179	0,042 / 0,043	1,049	0,033	2,839	0,160
	Cluster 2_H	2,135	0,438	0,662	0,440	0,295	0,077 / 0,083	1,622	0,061	3,650	0,155

* pela distribuição de frequências/pelos dados originais

Tabela 6.14: Ordens dos valores de cada medida de risco para a variável CBO_5 e para cada *cluster* no período observado, nos meses secos e nos meses húmidos.

		M	V e DP	DA	SS	PP*	E	FEM	VaR	TVaR	**SO**
Total	*Cluster* 1_T	1	1	1	1	1 / 1	1	1	1	1	**9**
	Cluster 2_T	2	2	2	2	2 / 2	2	2	2	2	**18**
Secos	*Cluster* 1_S	1	1	1	1	1 / 1	1	1	1	1	**9**
	Cluster 2_S	2	2	2	2	2 / 2	2	2	2	2	**18**
Húmidos	*Cluster* 1_H	1	1	1	1	1 / 1	1	1	1	2	**10**
	Cluster 2_H	2	2	2	2	2 / 2	2	2	2	1	**17**

* pela distribuição de frequências/pelos dados originais

Tabela 6.15: Ordens dos valores de cada medida de risco para os *clusters* da variável CBO_5.

		M	V e DP	DA	SS	PP*	E	FEM	VaR	TVaR	**SO**	**OF**
Total	*Cluster* 1_T	2	3	2	3	2 / 2	2	2,5**	2	4	**22,5**	2
	Cluster 2_T	5	5	5	5	5 / 5	5	5	5	6	**46**	5
	SO	7	8	7	8	7 / 7	7	7,5	7	10	**68,5**	**2**
Secos	*Cluster* 1_S	4	4	3	4	4 / 4	3	4	4	1	**31**	4
	Cluster 2_S	6	6	6	6	6 / 6	6	6	6	5	**53**	6
	SO	10	10	9	10	10 / 10	9	10	10	6	**84**	**3**
Húmidos	*Cluster* 1_H	1	1	1	1	1 / 1	1	1	1	3	**11**	1
	Cluster 2_H	3	2	4	2	3 / 3	4	2,5**	3	2	**25,5**	3
	SO	4	3	5	3	4 / 4	5	3,5	4	5	**36,5**	**1**

* pela distribuição de frequências/pelos dados originais
** média aritmética

Os valores médios dos *clusters* dos meses secos, presentes na Tabela 6.13 indicam uma pior qualidade de água quando comparados com os valores médios dos *clusters*

dos meses húmidos. De facto, para além destes serem superiores também as restantes medidas calculadas o são. Assim, é possível concluir que o risco nos meses secos é superior ao risco nos meses húmidos. Tal como as somas das ordens dos dois períodos confirmam (ver Tabela 6.15).

Analisando os valores da probabilidade de perda dos *Clusters* 1_T e 2_T conclui-se que existem observações superiores a três, ou seja, nem sempre a qualidade da água é *Excelente* tendo em conta esta variável. No entanto, para o *Cluster* 1_T o seu valor é bastante reduzido, o que significa que a probabilidade de perda avalia o risco das estações que constituem este *cluster* como praticamente nulo.

A Tabela 6.14 evidencia que o risco é inferior nos *Clusters* 1_T, 1_S e 1_H comparativamente aos *Clusters* 2_T, 2_S e 2_H, respetivamente (ver SO).

Como se pode observar na Tabela 6.15, existem casos de empate dos valores de uma determinada medida de risco, isto é, uma medida de risco possui o mesmo valor para diferentes *clusters*. Nestes casos, o *ranking* considerado para estes *clusters* corresponde à média aritmética das ordens dos *clusters* envolvidos. Por exemplo, a função de excesso médio toma o mesmo valor para o *Cluster* 1_T e para o *Cluster* 2_H. Como estes *clusters* correspondem às ordens 2 e 3 e não podendo as ordens ser discriminadas, considera-se o valor médio das mesmas, 2,5.

Condutividade

Relativamente à variável COND, sabe-se que o aumento dos seus valores é inversamente proporcional ao nível da qualidade da água.

A Tabela 6.16 exibe os valores de diversas medidas de risco aplicadas a cada *cluster* para a variável COND e as Tabelas 6.17 e 6.18 apresenta as ordenações destes valores e respetivas somas por *cluster*.

Tabela 6.16: Medidas de risco por *cluster* para a variável COND.

		M	V	DP	DA	SS	PP*	E	FEM	VaR	TVaR
Total	*Cluster* 1_T	55,803	737,517	27,157	12,397	695,786	0,001 / 0,003	0,298	0,290	123,128	1,405
	Cluster 2_T	203,763	16263,232	127,527	69,911	12437,786	0,554 / 0,299	1,157	9,349	253,046	19,084
Secos	*Cluster* 1_S	58,792	253,828	15,932	3,050	251,477	0,001 / 0,001	0,063	0,133	57,250	2,155
	Cluster 2_S	228,536	3842,242	61,986	37,913	1208,969	0,212 / 0,454	0,901	11,846	305,745	0,000
Húmidos	*Cluster* 1_H	43,651	730,236	27,023	14,345	578,855	0,005 / 0,005	0,969	0,157	70,969	5,019
	Cluster 2_H	140,542	6235,111	78,963	67,456	4013,868	0,131 / 0,109	2,117	3,920	285,496	5,279

* pela distribuição de frequências/pelos dados originais

Tabela 6.17: Ordens dos valores de cada medida de risco para a variável COND e para cada *cluster* no período observado, nos meses secos e nos meses húmidos.

		M	V e DP	DA	SS	PP*	E	FEM	VaR	TVaR	**SO**
Total	*Cluster* 1_T	1	1	1	1	1 / 1	1	1	1	1	**9**
	Cluster 2_T	2	2	2	2	2 / 2	2	2	2	2	**18**
Secos	*Cluster* 1_S	1	1	1	1	1 / 1	1	1	1	2	**10**
	Cluster 2_S	2	2	2	2	2 / 2	2	2	2	1	**17**
Húmidos	*Cluster* 1_H	1	1	1	1	1 / 1	1	1	1	1	**9**
	Cluster 2_H	2	2	2	2	2 / 2	2	2	2	2	**18**

* pela distribuição de frequências/pelos dados originais

Tabela 6.18: Ordens dos valores de cada medida de risco para os *clusters* da variável COND.

		M	V e DP	DA	SS	PP*	E	FEM	VaR	TVaR	**SO**	**OF**
Total	*Cluster* 1_T	2	3	2	3	1,5** / 2	2	3	3	2	**21,5**	2
	Cluster 2_T	5	6	6	6	6 / 5	6	5	4	6	**50**	6
	SO	7	9	8	9	7,5 / 7	8	8	7	8	**71,5**	**3**
Secos	*Cluster* 1_S	3	1	1	1	1,5** / 1	1	1	1	3	**13,5**	1
	Cluster 2_S	6	4	4	4	5 / 6	3	6	6	1	**39**	4
	SO	9	5	5	5	6,5 / 7	4	7	7	4	**52,5**	**1**
Húmidos	*Cluster* 1_H	1	2	3	2	3 / 3	4	2	2	4	**23**	3
	Cluster 2_H	4	5	5	5	4 / 4	5	4	5	5	**42**	5
	SO	5	7	8	7	7 / 7	9	6	7	9	**65**	**2**

* pela distribuição de frequências/pelos dados originais
** média aritmética

De acordo com a Tabela 6.2 (Secção 6.1), a variável COND não possui um limite que distinga os níveis de qualidade *Excelente* e *Bom*, assim as medidas: probabilidade de perda e função de excesso médio, só permitem detetar a presença de observações que se enquadram no nível de qualidade *Razoável* (>250) que, de acordo com a Tabela 6.16 existem efetivamente em todos os *clusters*.

Os valores presentes na Tabela 6.17 indicam que os *Clusters* 2_T, 2_S e 2_H possuem riscos superiores aos riscos dos *Clusters* 1_T, 1_S e 1_H, respetivamente. De facto, todas as medidas de risco evidenciam esta questão. Por exemplo, os valores em risco determinam que o risco do *Cluster* 2_T é muito superior ao risco do *Cluster* 1_T, já que para além de os valores serem bastante distantes, o valor do *Cluster* 2_T ultrapassa 250 (ver Tabela 6.16). O mesmo se verifica para os *Clusters* 2_S e 2_H.

Apesar dos *clusters* formados nos meses secos e húmidos serem distintos, é possível concluir, através da soma das ordens da média destes períodos, que os valores de condutividade observados nos meses secos são superiores aos valores observados nos meses húmidos. No entanto, analisando a SO conclui-se que, no geral, as medidas de risco consideradas avaliam de pior forma os meses húmidos (S0=65) do que os meses secos (S0=52,5). Tal facto, pode significar que existem meses classificados como húmidos que atingem níveis de risco superiores.

Oxigénio Dissolvido

A Tabela 6.19 exibe os valores das diversas medidas de risco aplicadas aos diferentes *clusters* para a variável OD. Esta variável, ao contrário das restantes, relaciona-se diretamente com os níveis de qualidade de água, isto é, quanto maior a sua concentração, melhor a qualidade da água. Assim sendo, contrariamente aos casos vistos até ao momento, existem medidas de risco cujos *rankings*, presentes nas Tabelas 6.20 e 6.21, foram atribuídos considerando a ordem decrescente, um exemplo é o valor médio.

Tabela 6.19: Medidas de risco por *cluster* para a variável OD.

		M	V	DP	DA	SS	PP*	E	FEM	VaR	VTaR
Total	*Cluster* 1_T	9,042	1,503	1,226	1,017	0,780	0,225 / 0,198	2,411	0,142	10,941	0,289
	Cluster 2_T	9,375	1,405	1,185	0,980	0,708	0,139 / 0,112	2,377	0,065	10,966	0,572
Secos	*Cluster* 1_S	8,155	1,043	1,021	0,809	0,528	0,489 / 0,411	2,402	0,328	9,818	0,379
	Cluster 2_S	8,388	0,637	0,798	0,608	0,308	0,375 / 0,276	2,162	0,150	9,834	0,290
Húmidos	*Cluster* 1_H	9,687	0,967	0,983	0,771	0,547	0,044 / 0,051	2,147	0,032	10,843	0,553
	Cluster 2_H	10,178	0,890	0,943	0,706	0,478	0,008 / 0,008	2,072	0,011	11,368	0,605

* pela distribuição de frequências/pelos dados originais

Tabela 6.20: Ordens dos valores de cada medida de risco para a variável OD e para cada *cluster* no período observado, nos meses secos e nos meses húmidos.

		M	V e DP	DA	SS	PP*	E	FEM	VaR	TVaR	**SO**
Total	*Cluster* 1_T	2	2	2	2	2 / 2	2	2	2	2	**18**
	Cluster 2_T	1	1	1	1	1 / 1	1	1	1	1	**9**
Secos	*Cluster* 1_S	2	2	2	2	2 / 2	2	2	2	1	**17**
	Cluster 2_S	1	1	1	1	1 / 1	1	1	1	2	**10**
Húmidos	*Cluster* 1_H	2	2	2	2	2 / 2	2	2	2	2	**18**
	Cluster 2_H	1	1	1	1	1 / 1	1	1	1	1	**9**

* pela distribuição de frequências/pelos dados originais

Tabela 6.21: Ordens dos valores de cada medida de risco para os *clusters* da variável OD.

		M	V e DP	DA	SS	PP*	E	FEM	VaR	TVaR	**SO**	**OF**
Total	*Cluster* 1_T	4	6	6	6	4 / 4	6	4	3	6	**45**	6
	Cluster 2_T	3	5	5	5	3 / 3	4	3	2	2	**32**	4
	SO	7	11	11	11	7 / 7	10	7	5	8	**77**	**3**
Secos	*Cluster* 1_S	6	4	4	3	6 / 6	5	6	6	4	**44**	5
	Cluster 2_S	5	1	1	1	5 / 5	3	5	5	5	**31**	3
	SO	11	5	5	4	11 / 11	8	11	11	9	**75**	**2**
Húmidos	*Cluster* 1_H	2	3	3	4	2 / 2	2	2	4	3	**25**	2
	Cluster 2_H	1	2	2	2	1 / 1	1	1	1	1	**12**	1
	SO	3	5	5	6	3 / 3	3	3	5	4	**37**	**1**

* pela distribuição de frequências/pelos dados originais

De acordo com a Tabela 6.2 (Secção 6.1), o valor 8 mg/l distingue os níveis de qualidade *Excelente* e *Bom*, logo a probabilidade de perda e a função de excesso

médio foram calculadas para este valor. Dada a relação da variável OD com a qualidade da água, a probabilidade de perda corresponde à probabilidade das observações que são inferiores a 8. Por sua vez, a função de excesso médio identifica a média das observações inferiores a 8.

A soma das ordens dos valores médios dos meses secos e meses húmidos presentes na Tabela 6.21 revelam que a concentração de OD é superior nos meses húmidos (menor risco). Já a soma das ordens da variância e/ou do desvio padrão e/ou do desvio absoluto demonstram que a variabilidade de valores não é influenciada pela altura do ano (chuvosa ou não). No geral, as medidas de risco calculadas classificam os meses secos com um maior risco comparativamente aos meses húmidos, veja-se as SO correspondentes (75 e 37).

Temperatura

Por fim, a Tabela 6.19 apresenta os valores das diversas medidas de risco calculadas para a TEMP e analogamente aos restantes casos as Tabelas 6.20 e 6.23 apresentam os *rankings* de cada medida e as respetivas somas por *cluster* e por conjunto de dados considerado.

Tabela 6.22: Medidas de risco por *cluster* para a variável TEMP.

		M	V	DP	DA	SS	SI	PPS*	PPI*	E	FEMS	FEMI	VaR	TVaR
	Cluster 1_T	11,719	31,606	5,622	4,766	16,292	15,313	0,000 / 0,000	0,238 / 0,196	2,653	0,000	0,482	20,027	1,070
Total	*Cluster* 2_T	14,149	43,476	6,594	5,725	22,558	20,918	0,019 / 0,019	0,081 / 0,116	2,768	0,015	0,252	24,417	0,496
	Cluster 3_T	15,556	32,615	5,711	4,982	17,457	15,158	0,021 / 0,019	0,007 / 0,019	2,609	0,022	0,024	24,409	0,553
	Cluster 1_S	16,999	10,434	3,230	2,630	4,698	5,736	0,000 / 0,000	0,000 / 0,000	2,232	0,000	0,000	22,309	0,408
Secos	*Cluster* 2_S	20,595	11,338	3,367	2,727	4,631	6,708	0,035 / 0,036	0,002 / 0,002	2,286	0,038	0,001	25,044	0,927
	Cluster 3_S	21,480	12,059	3,473	2,800	4,604	7,456	0,072 / 0,072	0,000 / 0,000	2,247	0,090	0,000	26,370	0,467
	Cluster 1_H	7,747	12,585	3,548	2,785	6,487	6,098	0,000 / 0,000	0,384 / 0,366	2,247	0,000	0,897	13,552	0,464
Húmidos	*Cluster* 2_H	9,349	16,423	4,053	3,239	8,597	7,826	0,002 / 0,002	0,236 /0,219	2,347	0,002	0,541	16,338	0,378
	Cluster 3_H	11,448	12,301	3,507	2,754	7,067	5,235	0,000 / 0,000	0,053 / 0,044	2,209	0,000	0,077	18,170	0,921

* pela distribuição de frequências/pelos dados originais

Tabela 6.23: Ordens dos valores de cada medida de risco para a variável TEMP e para cada *cluster* no período observado, nos meses secos e nos meses húmidos.

		M	V e DP	DA	SS	PPS*	E	FEMS	VaR	TVaR	**SO**
	Cluster 1_T	1	1	1	1	1 / 1	2	1	1	3	**12**
Total	*Cluster* 2_T	2	3	3	3	2 / 2,5**	3	2	3	1	**22**
	Cluster 3_T	3	2	2	2	3 / 2,5**	1	3	2	2	**20**
	Cluster 1_S	1	1	1	3	1 / 1,5	1	1	1	1	**11**
Secos	*Cluster* 2_S	2	2	2	2	2 / 3	3	2	2	3	**20**
	Cluster 3 $_S$	3	3	3	1	3 / 1,5	2	3	3	2	**23**
	Cluster 1_H	1	2	2	1	1,5** / 3	2	1,5**	1	2	**14**
Húmidos	*Cluster* 2_H	2	3	3	3	3 / 2	3	3	2	1	**23**
	Cluster 3_H	3	1	1	2	1,5** / 1	1	1,5**	3	3	**17**

* pela distribuição de frequências/pelos dados originais
** média aritmética

Tabela 6.24: Ordens dos valores de cada medida de risco para os *clusters* da variável TEMP.

		M	V e DP	DA	SS	PPS*	E	FEMS	VaR	TVaR	**SO**	**OF**
Total	*Cluster* 1_T	4	7	7	8	2,5** / 2,5**	8	2,5**	4	9	**52**	5
	Cluster 2_T	5	9	9	9	6 / 6,5**	9	6	7	5	**65**	9
	Cluster 3_T	6	8	8	7	7 / 6,5**	7	7	6	6	**62**	8
	SO	15	24	24	24	15,5 / 15,5	24	15,5	17	20	**179**	**3**
Secos	*Cluster* 1_S	7	1	1	2	2,5** / 2,5**	2	2,5**	5	2	**25**	1
	Cluster 2_S	8	2	2	4	8 / 8	5	8	8	8	**53**	6
	Cluster 3_S	9	3	5	5	9 / 9	3,5	9	9	4	**56,5**	7
	SO	24	6	8	11	19,5 / 19,5	10,5	19,5	22	14	**134,5**	**2**
Húmidos	*Cluster* 1_H	1	5	4	3	2,5 ** / 2,5**	3,5	2,5**	1	3	**25,5**	2
	Cluster 2_H	2	6	6	6	5 / 5	6	5	2	1	**39**	4
	Cluster 3_H	3	4	3	1	2,5** / 2,5**	1	2,5**	3	7	**27**	3
	SO	6	15	13	10	10 / 10	10,5	10	6	11	**91,5**	**1**

* pela distribuição de frequências/pelos dados originais
** média aritmética

Segundo a legislação em vigor, para que a qualidade da água seja *Excelente* ou *Boa*, a temperatura da mesma deve variar entre 6,5 e 25,5^{o}C e não ser inferior ou superior a estes valores, respetivamente. Esta condição implica um cuidado redobrado na atribuição das ordens, recorde-se os casos dos valores médios das variáveis CBO_5 e OD, em que os *rankings* foram atribuídos considerando a ordem crescente e decrescente, respetivamente.

Uma vez que não é possível atribuir ordens que classifiquem simultaneamente o risco tendo por base o limite 6,5 e 25,5 (a interpretação de algumas medidas de risco varia consoante o limite de referência), as Tabelas 6.23 e 6.24 focam-se no limite 25,5. As ordens que classificam o risco tendo por base o limite 6,5 seriam obtidas de forma análoga, mas considerado as medidas SI, PPI e FEMI em vez das medidas SS, PPS e FEMS.

Analisando os valores da probabilidade de perda inferior e superior apresentados na Tabela 6.22 conclui-se que existem valores fora do intervalo de excelência tanto nos meses secos como nos meses húmidos. Apenas o *Cluster* 1_S possui todos os valores dentro deste intervalo, como já tinha sido verificado.

De notar ainda que, as medidas de risco calculadas classificam os meses secos com um maior risco comparativamente aos meses húmidos, veja-se as SO correspondentes (134,5 e 91,5, respetivamente).

Com o objetivo de complementar a análise efetuada nesta subsecção pretende-se avaliar o risco mensal, por forma a determinar se a divisão em meses secos e meses húmidos estabelecida é apropriada para avaliar o risco.

6.5 Classificação dos Meses através de Medidas de Risco

Nesta subsecção pretende-se aplicar uma metodologia que permita classificar os meses através de medidas de risco, em particular pretende-se avaliar se as variações naturais dos valores de cada variável alteram o nível de risco de poluição ao longo dos meses. Para o efeito, à semelhança da abordagem anterior, foram consideradas as variáveis CBO_5, COND, OD e TEMP e aplicadas as diversas medidas de risco simples aos dados de poluição mensais.

Novamente, pretende-se determinar a distribuição empírica de cada variável, mas neste caso para cada mês.

Seja $S = \{S_1, \ldots, S_{12}\}$ o conjunto de dados referentes a uma única variável, a título de exemplo CBO_5, onde $S_i, i = 1, \ldots 12$ representam os dados dessa mesma variável em cada um dos 12 meses. O mínimo e máximo absolutos de S foram calculados e o intervalo por eles formado dividido em subintervalos de igual amplitude, com o auxílio da regra de Rice. Posteriormente, foi calculada a probabilidade e média das observações de cada subintervalo, obtendo-se assim uma distribuição de frequências para cada mês.

Recorde-se que no cálculo do valor em risco foi considerado um nível de 95% de confiança e que a probabilidade de perda e a função de excesso médio foram calculadas para o limiar que separa os níveis de qualidade de água *Excelente* e *Bom* sempre, de acordo com a Tabela 6.2 (Secção 6.1), sempre que tal existe. Para além disso, apenas as ordens da probabilidade de perda calculadas através da distribuição de frequências foram consideradas para efeitos de soma das ordens e em todos os casos em que se verificaram empates dos valores de uma determinada medida de risco, isto é, sempre que uma medida de risco possui o mesmo valor em diferentes meses, foi considerada a média aritmética das ordens.

Carência Bioquímica de Oxigénio a 5 dias

Os valores da variável CBO_5 relacionam-se inversamente com o nível da qualidade da água, ou seja, quanto maiores os valores da CBO_5, pior a qualidade da água. Em concreto, o valor 3 mg/l distingue o nível da qualidade da água *Excelente* e *Bom*, dos rios e albufeiras (Tabela 6.2, Secção 6.1).

Na Tabela 6.25 são apresentados os valores de diversas medidas de risco aplicadas a esta variável nos diferentes meses, já na Tabela 6.26 são apresentadas as suas ordens e ainda a soma das mesmas por mês.

Tabela 6.25: Medidas de risco por mês para a variável CBO_5.

	M	V	DP	DA	SS	PP*	E	FEM	VaR	TVaR
Janeiro	1,923	0,297	0,545	0,279	0,136	0,009 / 0,014	0,694	0,019	2,016	0,152
Fevereiro	2,065	0,375	0,612	0,301	0,230	0,028 / 0,074	0,898	0,035	2,986	0,119
Março	2,036	0,550	0,741	0,309	0,389	0,028 / 0,051	0,908	0,056	2,897	0,139
Abril	2,066	0,197	0,443	0,207	0,140	0,014 / 0,046	0,674	0,019	2,929	0,060
Maio	2,058	0,509	0,714	0,325	0,322	0,042 / 0,060	0,922	0,061	2,956	0,186
Junho	2,336	1,157	1,076	0,657	0,941	0,208 / 0,139	1,185	0,190	4,700	0,166
Julho	2,201	0,604	0,777	0,470	0,452	0,074 / 0,106	0,963	0,103	3,881	0,127
Agosto	2,393	0,859	0,927	0,659	0,667	0,130 / 0,167	1,150	0,184	4,725	0,137
Setembro	2,138	0,517	0,719	0,377	0,389	0,060 / 0,083	0,939	0,090	3,820	0,182
Outubro	2,149	0,435	0,659	0,389	0,331	0,069 / 0,093	0,955	0,078	3,846	0,108
Novembro	1,946	0,309	0,556	0,279	0,154	0,009 / 0,023	0,723	0,022	2,032	0,161
Dezembro	1,960	0,237	0,487	0,249	0,099	0,019 / 0,032	0,705	0,013	2,911	0,069

* pela distribuição de frequências/pelos dados originais

Tabela 6.26: Ordens dos valores de cada medida de risco para a variável CBO_5.

	M	V e DP	DA	SS	PP*	E	FEM	VaR	TVaR	**SO**	OF
Janeiro	1	3	4	2	1,5** / 1	2	3	1	8	**25,5**	3
Fevereiro	6	5	5	5	5,5 ** / 7	5	5	7	4	**47,5**	5
Março	4	9	6	8	5,5** / 5	6	6	3	7	**54,5**	6
Abril	7	1	1	3	3 / 4	1	2	5	1	**24,0**	2
Maio	5	7	7	6	7 / 6	7	7	6	12	**64,0**	7
Junho	11	12	11	12	12 / 11	12	12	11	10	**103,0**	12
Julho	10	10	10	10	10 / 10	10	10	10	5	**85,0**	10
Agosto	12	11	12	11	11 / 12	11	11	12	6	**97,0**	11
Setembro	8	8	8	9	8 / 8	8	9	8	11	**77,0**	9
Outubro	9	6	9	7	9 / 9	9	8	9	3	**69,0**	8
Novembro	2	4	3	4	1,5** / 2	4	4	2	9	**33,5**	4
Dezembro	3	2	2	1	4 / 3	3	1	4	2	**22,0**	1

* pela distribuição de frequências/pelos dados originais
** média aritmética

Os valores médios dos diferentes meses indiciam uma qualidade de água *Excelente*, já que todos os valores são inferiores a 3. No entanto, a análise exclusiva desta medida não nos permite retirar conclusões efetivas. As restantes medidas calculadas auxiliam-nos neste sentido. Por exemplo, as probabilidades de perda evidenciam que, existem em todos os meses observações superiores a 3. Por sua vez, os valores da função de excesso médio indicam que, na grande maioria dos meses essas observações são muito próximas de 3. Deste modo, conclui-se que, nem sempre a qualidade da água é *Excelente*.

Uma análise geral da Tabela 6.26 destaca os meses de verão (junho, julho e agosto) pela negativa. O valor da cauda em risco é a única medida que não concentra os seus *rankings* mais elevados nos meses de maior calor. Apesar disto, é de esperar que a soma das ordens obtenha os maiores valores nos meses de verão, situação que,

de facto, se confirma. O mês de outubro, que tipicamente é considerado um mês húmido, possui um valor de soma de *rankings* muito mais próximo dos valores dos meses secos do que dos valores dos meses húmidos.

Condutividade

Relativamente à variável COND, sabe-se que o aumento dos seus valores é inversamente proporcional ao nível da qualidade da água. Ao contrário da variável CBO_5, esta variável não possui um limite que distinga os níveis *Excelente* e *Bom*.

A abordagem adotada na análise da variável CBO_5, foi também aplicada aqui, assim sendo, a Tabela 6.27 exibe os valores de diversas medidas de risco aplicadas a esta variável nos diferentes meses e a Tabela 6.28 a ordenação destes mesmos valores e ainda a sua soma geral, por meses.

Tabela 6.27: Medidas de risco por mês para a variável COND.

	M	V	DP	DA	SS	PP*	E	FEM	VaR	TVaR
Janeiro	73,464	1710,177	41,354	17,379	1631,002	0,046 / 0,019	0,187	0,517	64,353	12,091
Fevereiro	75,062	2114,430	45,983	21,872	1987,173	0,060 / 0,019	0,227	0,407	256,769	0,000
Março	75,862	2512,134	50,121	22,962	2372,571	0,056 / 0,046	0,215	1,806	282,517	0,000
Abril	78,141	3636,882	60,307	34,961	3300,133	0,093 / 0,056	0,308	1,568	266,930	0,000
Maio	80,320	3882,773	62,312	36,921	3505,281	0,097 / 0,065	0,319	1,964	270,200	0,000
Junho	100,438	34604,730	186,023	62,660	33482,938	0,125 / 0,083	0,397	12,635	262,273	12,315
Julho	42,334	9269,502	96,278	68,405	7896,243	0,148 / 0,097	0,419	3,437	273,202	0,000
Agosto	106,205	5950,654	77,140	58,128	4931,329	0,171 / 0,134	0,458	4,432	275,876	0,000
Setembro	109,645	5779,563	76,023	58,485	4736,031	0,181 / 0,130	0,472	3,901	271,603	0,000
Outubro	105,485	6160,871	78,491	56,516	5197,308	0,171 / 0,106	0,479	3,503	261,129	2,806
Novembro	84,558	2737,653	52,323	29,640	2496,841	0,088 / 0,042	0,298	0,267	253,037	0,000
Dezembro	76,743	1921,232	43,832	20,080	1814,497	0,056 / 0,032	0,215	0,415	257,466	0,000

* pela distribuição de frequências/pelos dados originais

Tabela 6.28: Ordens dos valores de cada medida de risco para a variável COND.

	M	V e DP	DA	SS	PP*	E	FEM	VaR	TVaR	**SO**	OF
Janeiro	2	1	1	1	1 / 1,5**	1	4	1	11	**23**	1
Fevereiro	3	3	3	3	4 / 1,5**	4	2	3	5**	**30**	3
Março	4	4	4	4	2,5** / 5	2,5**	6	12	5**	**44**	5
Abril	6	6	6	6	6 / 6	6	5	7	5**	**53**	6
Maio	7	7	7	7	7 / 7	7	7	8	5**	**62**	7
Junho	9	12	11	12	8 / 8	8	12	6	12	**90**	12
Julho	1	11	12	11	9 / 9	9	8	10	5**	**76**	8
Agosto	11	9	9	9	11 / 12	10	11	11	5**	**86**	11
Setembro	12	8	10	8	12 / 11	11	10	9	5**	**85**	10
Outubro	10	10	8	10	10 / 10	12	9	5	10	**84**	9
Novembro	8	5	5	5	5 / 4	5	1	2	5**	**41**	4
Dezembro	5	2	2	2	2,5** / 3	2,5**	3	4	5**	**28**	2

* pela distribuição de frequências/pelos dados originais
** média aritmética

Não existindo, de acordo com a legislação em vigor, um limiar que distinga os valores da condutividade a partir dos quais a qualidade da água é *Boa* ao invés de *Excelente*, utilizar-se-á por base para esta análise o valor 250 que distingue os níveis de qualidade *Bom* e *Razoável*. Notoriamente, todos os valores médios são muito inferiores a 250, mas as variâncias dos diversos meses são muito elevadas, o que para além de indicar que os valores de condutividade variam bastante ao longo de cada mês, sugere que a qualidade da água pode ser considerada *Razoável* em alguns momentos. Os valores da probabilidade de perda confirmam esta mesma questão, comprovando que existem, ainda que poucas, observações superiores a 250 em todos os meses.

Novamente, os meses de verão parecem ser os detentores de elevados *rankings*, no entanto, os meses de setembro e outubro também se destacam neste sentido. De facto, e novamente, o mês de outubro possui uma soma das ordens idêntica aos meses de verão, o que sugere uma partição alternativa entre os meses secos e húmidos.
De notar ainda que, contrariamente ao espectável, o mês que detém o valor em risco mais elevado, é o mês de março.

Oxigénio Dissolvido

Contrariamente às restantes variáveis em estudo, os valores da variável OD são diretamente proporcionais ao nível da qualidade da água. Em particular, o valor 8 mg/l limita os intervalos de valores entre os quais a água dos rios e albufeiras é classificada como *Excelente* e *Boa* (Tabela 6.2, Secção 6.1).

Na Tabela 6.29 são apresentados os valores de diversas medidas de risco aplicadas, já na Tabela 6.30 é apresentada a sua ordenação e a respetiva soma global por meses.

Tabela 6.29: Medidas de risco por mês para a variável OD.

	M	V	DP	DA	SS	PP*	E	FEM	VaR	TVaR
Janeiro	10,455	0,369	0,607	0,463	0,168	0,000 / 0,000	1,180	0,000	11,322	0,339
Fevereiro	10,597	0,508	0,713	0,471	0,206	0,000 / 0,000	1,291	0,000	12,164	0,241
Março	10,095	0,437	0,661	0,508	0,193	0,000 / 0,000	1,245	0,000	11,332	0,227
Abril	9,676	0,391	0,625	0,436	0,202	0,005 / 0,009	1,230	0,002	10,487	0,265
Maio	9,021	0,526	0,725	0,529	0,245	0,042 / 0,060	1,342	0,030	10,511	0,106
Junho	8,379	0,636	0,798	0,642	0,329	0,213 / 0,306	1,419	0,169	9,700	0,102
Julho	8,018	0,608	0,780	0,596	0,301	0,361 / 0,440	1,428	0,292	8,814	0,418
Agosto	8,002	0,571	0,755	0,556	0,293	0,343 / 0,458	1,392	0,277	8,852	0,454
Setembro	8,131	0,668	0,817	0,569	0,395	0,273 / 0,343	1,454	0,249	9,662	0,051
Outubro	8,551	0,939	0,969	0,736	0,489	0,190 / 0,213	1,605	0,163	9,649	0,406
Novembro	9,446	0,547	0,739	0,629	0,273	0,005 / 0,028	1,354	0,002	10,576	0,259
Dezembro	10,134	0,589	0,768	0,610	0,273	0,005 / 0,005	1,380	0,003	11,403	0,282

* pela distribuição de frequências/pelos dados originais

Tabela 6.30: Ordens dos valores de cada medida de risco para a variável OD.

	M	V e DP	DA	SS	PP*	E	FEM	VaR	TVaR	**SO**	OF
Janeiro	2	1	2	1	3 / 3	1	3	4	4	**21**	1
Fevereiro	1	4	3	4	2 / 2	4	2	1	8	**29**	2
Março	4	3	4	2	1 / 1	3	1	3	9	**30**	3
Abril	5	2	1	3	6 / 5	2	4	7	6	**36**	4
Maio	7	5	5	5	7 / 7	5	7	6	10	**57**	6,5*
Junho	9	10	11	10	9 / 9	9	9	8	11	**86**	11
Julho	11	9	8	9	12 / 11	10	12	12	2	**85**	9,5*
Agosto	12	7	6	8	11 / 12	8	11	11	1	**75**	8
Setembro	10	11	7	11	10 / 10	11	10	9	12	**91**	12
Outubro	8	12	12	12	8 / 8	12	8	10	3	**85**	9,5*
Novembro	6	6	10	7	5 / 6	6	5	5	7	**57**	6,5*
Dezembro	3	8	9	6	4 / 4	7	6	2	5	**50**	5

* pela distribuição de frequências/pelos dados originais
** média aritmética

Uma análise preliminar da Tabela 6.29 permite observar que os meses de julho e agosto possuem valores médios extremamente próximos de oito. Por sua vez, também os meses de junho, setembro e outubro apresentam concentrações médias relativamente próximas deste valor. Tal facto, indica que nestes meses o nível de qualidade da água poderá não ser sempre *Excelente*, o que se confirma através da função de excesso médio. Esta mesma função evidencia que nos meses de janeiro, fevereiro e março a qualidade da água, no que toca a esta variável, se mantém no nível *Excelente*.

A análise anterior é corroborada pelos *rankings*, uma vez que os três primeiros meses do ano possuem as OF mais baixas e os meses de junho a outubro as OF mais altas. Os meses de maio e novembro aparentam possuir uma qualidade "intermédia"e equiparada, já que os índices de risco OF tomam o valor 6,5 em ambos os casos.

Temperatura

A temperatura da água, $TEMP$, afeta a qualidade da água, mas ao contrário das restantes variáveis, não só valores altos se associam a perdas de qualidade, os valores baixos também o fazem.

De acordo com a Tabela 6.2 (Secção 6.1), esta variável não possui um valor que distinga os níveis de qualidade *Excelente* e *Bom*, no entanto sabe-se que valores inferiores a 6,5^{o}C e valores superiores a 25,5^{o}C estão associados a uma qualidade de água *Razoável*.

Deste modo, as Tabelas 6.31 e 6.32 apresentam, à semelhança das anteriores, os valores das diversas medidas de risco aplicadas a esta variável nos diferentes meses

e a sua ordenação e soma final por meses, respetivamente. No entanto, neste caso, a ordem dos *rankings* pode ser interpretada em dois sentidos, as ordens apresentadas apenas consideram o limite 25,5. O estudo para o limite inferior, 6,5, seria obtido de forma análoga como explicado anteriormente.

Tabela 6.31: Medidas de risco por mês para a variável TEMP.

	M	V	DP	DA	SS	SI	PPS*	PPI*	E	FEMS	FEMI	VaR	TVaR
Janeiro	7,095	6,504	2,550	1,774	2,708	3,796	0,000 / 0,000	0,236 / 0,347	1,380	0,000	0,747	9,757	0,404
Fevereiro	7,576	6,245	2,499	1,828	2,461	3,784	0,000 / 0,000	0,194 / 0,273	1,345	0,000	0,574	9,826	0,500
Março	10,115	6,665	2,582	1,667	2,250	4,415	0,000 / 0,000	0,042 / 0,060	1,340	0,000	0,204	14,482	0,000
Abril	12,459	8,005	2,829	1,910	3,564	4,441	0,000 / 0,000	0,023 / 0,023	1,531	0,000	0,081	17,457	0,465
Maio	15,662	7,670	2,769	2,118	4,016	3,653	0,000 / 0,000	0,000 / 0,000	1,492	0,000	0,000	20,138	0,628
Junho	20,206	10,471	3,236	2,589	4,591	5,879	0,005 / 0,032	0,000 / 0,000	1,652	0,012	0,000	25,435	0,130
Julho	22,417	6,245	2,499	1,978	2,634	3,611	0,014 / 0,093	0,000 / 0,000	1,411	0,049	0,000	25,479	0,403
Agosto	22,042	6,145	2,479	2,035	2,384	3,761	0,009 /0,046	0,000 / 0,000	1,370	0,023	0,000	25,470	0,259
Setembro	19,390	7,611	2,759	2,181	3,129	4,481	0,000 / 0,000	0,005 / 0,005	1,498	0,000	0,002	22,775	0,347
Outubro	15,142	9,583	3,096	2,301	4,976	4,607	0,005 / 0,005	0,000 / 0,000	1,581	0,004	0,000	20,221	0,231
Novembro	10,265	11,882	3,447	2,568	4,980	6,902	0,000 / 0,000	0,093 / 0,130	1,661	0,000	0,329	14,769	0,241
Dezembro	7,647	7,013	2,648	2,122	3,203	3,810	0,000 / 0,000	0,245 / 0,315	1,459	0,000	0,642	12,043	0,067

* pela distribuição de frequências/pelos dados originais

Tabela 6.32: Ordens dos valores de cada medida de risco para a variável TEMP.

	M	Var. e DP	DA	SS	PPS*	E	FEMS	VaR	TVaR	**SO**	OF
Janeiro	1	4	2	5	4,5** / 4,5**	4	4,5**	1	9	**35,0**	3
Fevereiro	2	2	3	3	4.5** / 4,5**	2	4,5**	2	11	**34,0**	2
Março	4	5	1	1	4,5** / 4,5**	1	4,5**	4	1	**26,0**	1
Abril	6	9	4	8	4,5** / 4,5**	9	4,5**	6	10	**61,0**	5
Maio	8	8	7	9	4,5** / 4,5**	7	4,5**	7	12	**67,0**	8
Junho	10	11	12	10	9,5** / 10	11	10	10	3	**86,5**	12
Julho	12	3	5	4	12 / 12	5	12	12	8	**73,0**	10
Agosto	11	1	6	2	11 / 11	3	11	11	6	**62,0**	6
Setembro	9	7	9	6	4,5** / 4,5**	8	4,5**	9	7	**64,0**	7
Outubro	7	10	10	11	9,5** / 9	10	9	8	4	**78,5**	11
Novembro	5	12	11	12	4,5** / 4,5**	12	4,5**	5	5	**71,0**	9
Dezembro	3	6	8	7	4,5** / 4,5**	6	4,5**	3	2	**44,0**	4

* pela distribuição de frequências/pelos dados originais
** média aritmética

Analisando as temperaturas médias dos diferentes meses vê-se que, existem médias relativamente próximas de ambos os limiares.

Já observando a probabilidade de perda e/ou a função de excesso médio superiores verifica-se que, apenas os meses de junho, julho, agosto e outubro registam temperaturas superiores a 25,5^{o}C. Por sua vez, observando a probabilidade de perda e/ou a função de excesso médio inferiores deteta-se que nenhum destes meses possui observações inferiores a 6,5. Comparando os diferentes valores em risco destes quatro meses demostra-se que, entre eles, o mês de junho é o que representa a situação de maior risco. Inversamente, deteta-se que os restantes meses, à exceção do mês de maio, não possuem observações superiores a 25,5, mas possuem observações inferiores a 6,5.

Neste caso e uma vez que as ordens apresentadas consideram apenas o limite 25,5 pretende-se analisar segundo esta referência qual é o mês que apresenta o maior risco. Segundo a soma das ordens, o nível de risco varia ao longo de todos os meses, não havendo uma distinção entre meses secos e meses húmidos. De facto, o período de abril a novembro possui as maiores SO e não o período de maio a setembro. Já o período de dezembro a abril é o que possui as menores SO e, portanto, segundo esta análise estes são os meses mais seguros.

De um modo geral, o trabalho desenvolvido nesta secção é resumido pelos *rankings* finais de cada variável e permite, para cada uma delas, identificar os meses mais ou menos críticos. As variáveis CBO_5 e COND classificam os meses de maio a outubro como os que possuem o maior risco. Por sua vez, a variável OD exclui desta classificação o mês de maio, isto é, considera que o período de junho a outubro é o pior. Por fim, a variável TEMP destaca os meses de abril a novembro pela negativa. Assim, apenas a variável TEMP sugere uma partição, diga-se bastante, diferente. Tal facto, poderá estar relacionado com a dualidade desta variável, por outras palavras, com a associação de valores baixos (inferiores a 6,5) e altos (superiores a 25,5) ao nível de qualidade de água *Razoável*.

Recorde-se que a divisão adotada anteriormente considerava como meses secos os meses de maio a setembro e como meses húmidos os meses de outubro a abril. Deste modo, pode dizer-se que o conjunto das variáveis selecionadas para classificar os meses através de medidas de risco sugere uma partição alternativa dos dados. Futuramente, esta partição poderá ser explorada e os resultados obtidos comparados com os resultados obtidos através da divisão considerada no presente estudo.

6.6 Análise das Séries Temporais dos *Clusters*

Nesta secção apresentam-se processos de modelação de séries temporais aplicados aos *clusters* obtidos. Nestes processos não se consideram as séries das variáveis CBO_5 e TEMP, uma vez que a variável CBO_5 possui muitas observações de valor dois e a variável TEMP possui dois valores de referência (consequentemente, duas classificações distintas). Assim, as séries temporais analisadas são da variável COND e OD, considerando a totalidade dos dados. Tais processos não foram aplicados aos restantes *clusters* dado o período de tempo em que foram observados. Repare-se que, considerando apenas os dados dos meses secos ou húmidos não é possível obter séries temporais (por exemplo, para os meses secos tem-se cinco observações consecutivas e sete observações ausentes, por ano).

O principal objetivo desta aplicação é diferenciar os *clusters* em termos de poluição e identificar o que representa um maior risco. Posteriormente, pretende-se comparar os resultados obtidos com a respetiva classificação obtida através de medidas de risco.

Para identificar o *cluster* que possui o maior nível de poluição ir-se-á comparar as previsões pontuais obtidas por cada modelo de previsão aplicado. No entanto, para que tal seja possível pretende-se primeiramente identificar o modelo que produz previsões mais precisas. Neste estudo, serão aplicados os modelos SARIMA (modelos paramétricos) e os métodos de Holt-Winters (modelos não paramétricos) a cada um dos *clusters*.

Recorde-se que os *clusters* em estudo estão representados, respetivamente, nas Figuras 6.11 e 6.13 (Secção 6.3). O *Cluster* 1_T de COND é composto por quatro estações e o *Cluster* 2_T por catorze, já os *clusters* de OD são constituídos por nove estações cada. As respetivas séries temporais foram obtidas através da aplicação da média em cada momento, isto é, para cada mês, $(t = 1, \ldots, 144)$ foi considerada a média das observações das estações que constituem cada *cluster*. Cada uma delas foi posteriormente dividida em série de treino e teste, a série de treino com 132 observações para identificar e estimar os modelos e a série de teste com as últimas 12 observações para avaliar a qualidade de previsão dos modelos.

Nas Figuras 6.17, 6.18, 6.19 e 6.20 representam-se, respetivamente, as séries de COND dos *Clusters* 1_T e 2_T e as séries de OD dos *Clusters* 1_T e 2_T.

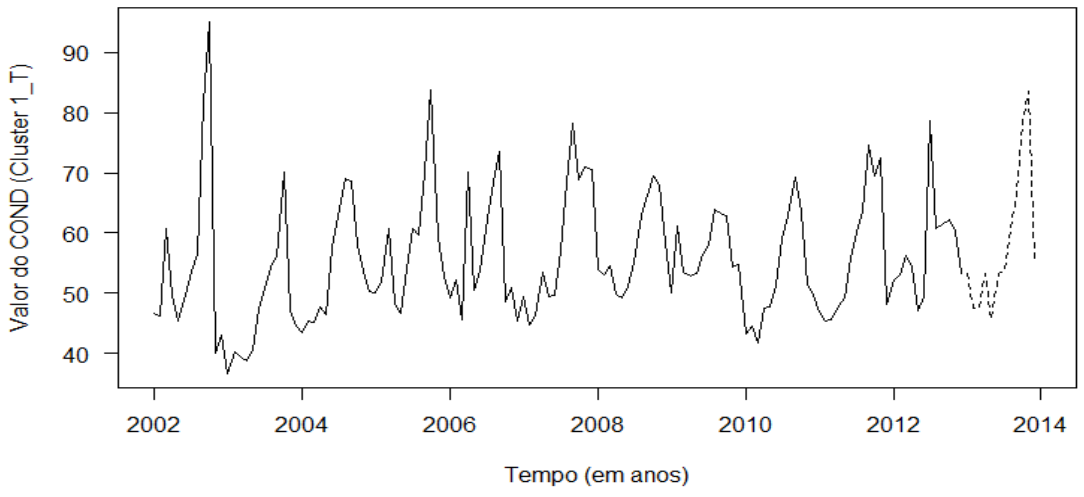


Figura 6.17: Série temporal de COND do *Cluster* 1_T.

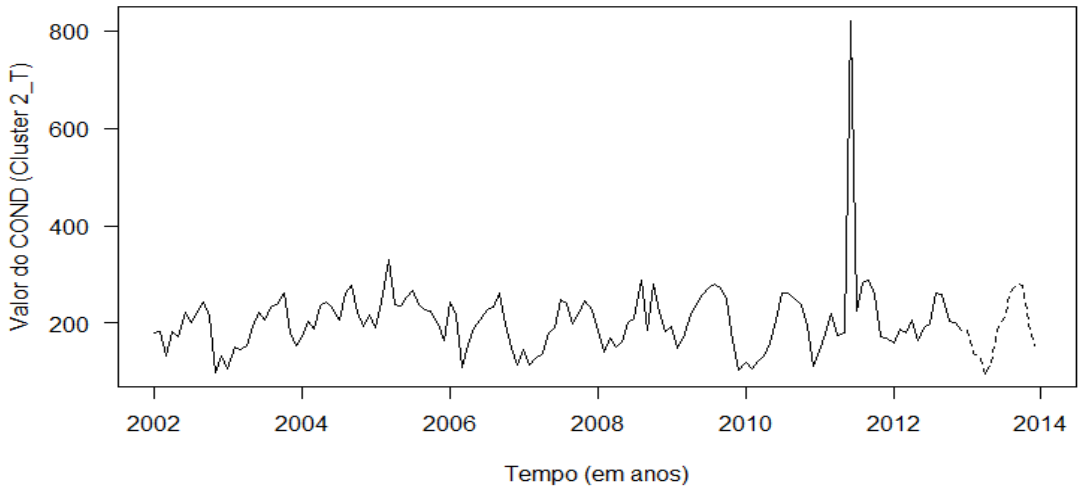


Figura 6.18: Série temporal de COND do *Cluster* 2_T.

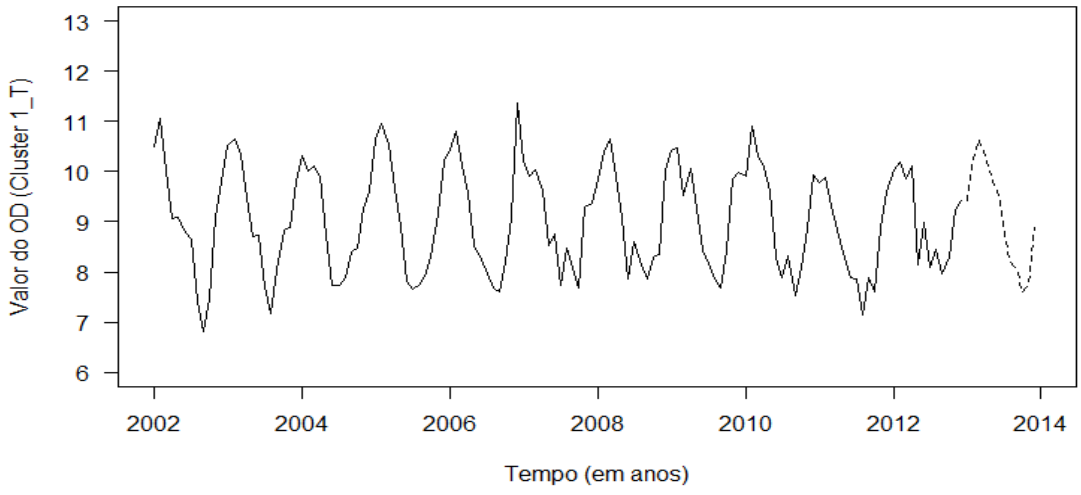


Figura 6.19: Série temporal de OD do *Cluster* 1_T.

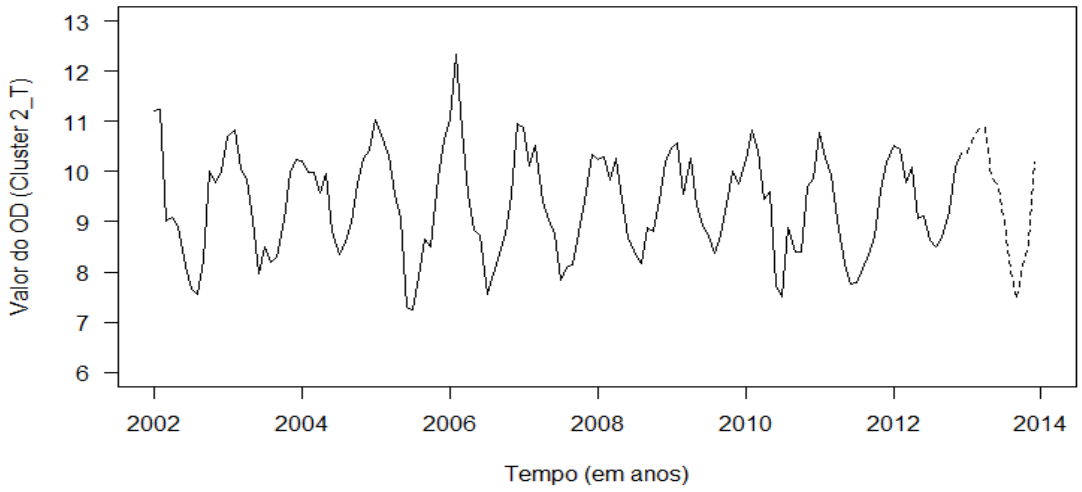


Figura 6.20: Série temporal de OD do *Cluster* 2_T.

6.6.1 Análise Exploratória das Séries

Nesta subsecção, pretende-se realizar uma análise inicial das séries em estudo, para o efeito são calculadas as suas medidas descritivas e apresentados os respetivos *boxplots*.

Condutividade

Na Tabela 6.33 apresentam-se as medidas descritivas das séries de COND dos *Clusters* 1_T e 2_T.

Tabela 6.33: Medidas descritivas das séries de COND, dos *Clusters* 1_T e 2_T.

	Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Cluster 1_T	36,680 - 95,070	48,110	53,540	55,800	61,550	10,445	18,717
Cluster 2_T	93,250 - 820,500	163,750	199,440	203,760	237,940	71,406	35,044

De acordo com a Tabela 6.33 os valores da média, mediana, amplitude e desvio padrão são superiores no *Cluster* 2_T, comparativamente ao *Cluster* 1_T. Assim, à partida a qualidade da água das estações que constituem este *Cluster* (2_T) é inferior à qualidade da água das estações que constituem o *Cluster* 1_T.

Considerando os *boxplots* que representam graficamente a distribuição dos valores de COND do *Cluster* 1_T e 2_T (apresentados na Figura 6.21) verifica-se a existência

de um *outlier* severo no *Cluster* 2_T. Com o intuito de melhorar a precisão do modelo selecionado na seguinte análise, este valor foi substituído pela média das observações do mês em que foi observado, no caso, este foi observado em julho de 2011, logo foi substituído pela média das observações registadas neste mês ao longo do período de observação. A representação gráfica desta série após esta substituição e a representação gráfica da série de COND do *Cluster* 1_T noutra escala (referentes às representações das Figuras 6.18 e 6.17) estão presentes nas Figuras B.2 e B.1 (Apêndice B).

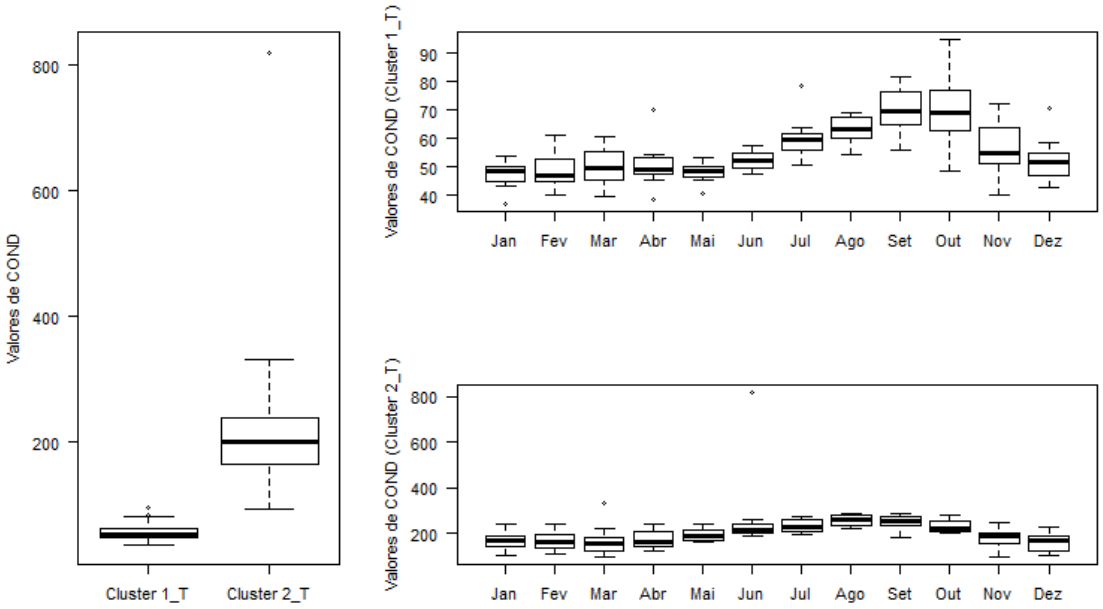


Figura 6.21: *Boxplots* das série de COND dos *Clusters* 1_T e 2_T (à esquerda) e *boxplots* dos valores de COND dos *Clusters* 1_T e 2_T divididos por meses (à direita).

Ainda relativamente a esta variável apresentam-se na Figura 6.21 os *boxplots* por meses dos *Clusters* 1_T e 2_T. A sua análise evidencia que no *Cluster* 1_T os valores de COND aumentam a partir do mês de maio, atingindo o pico no mês de outubro e começando a decrescer a partir desse mês. No *Cluster* 2_T verifica-se o mesmo comportamento, no entanto iniciado no mês de março e atingindo o pico no mês de setembro.

Oxigénio Dissolvido

Na Tabela 6.34 apresentam-se as medidas descritivas das séries de OD dos *Clusters* 1_T e 2_T.

Tabela 6.34: Medidas descritivas das séries de OD, dos *Clusters* 1_T e 2_T.

	Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Cluster 1_T	6,806 - 11,367	8,125	8,954	9,042	9,953	1,048	11,592
Cluster 2_T	7,233 - 12,333	8,585	9,428	9,375	10,222	1,037	11,058

A sua análise indica que os valores da média, mediana e amplitude são inferiores para o *Cluster* 1_T, comparativamente ao *Cluster* 2_T, no entanto o desvio padrão é ligeiramente superior.

Os *boxplots* que complementam esta análise e representam graficamente a distribuição dos valores de COND em ambos os *clusters*, apresentados na Figura 6.22, indicam que ambas as distribuições são ligeiramente assimétricas.

Na Figura (6.22) apresentam-se também os *boxplots* por meses dos *Clusters* 1_T e 2_T. A sua análise evidencia que a variação dos valores de OD ao longo dos meses é semelhante em ambos os *clusters*. De facto, estes diminuem até ao mês de agosto e julho, respetivamente, prosseguindo-se um aumento.

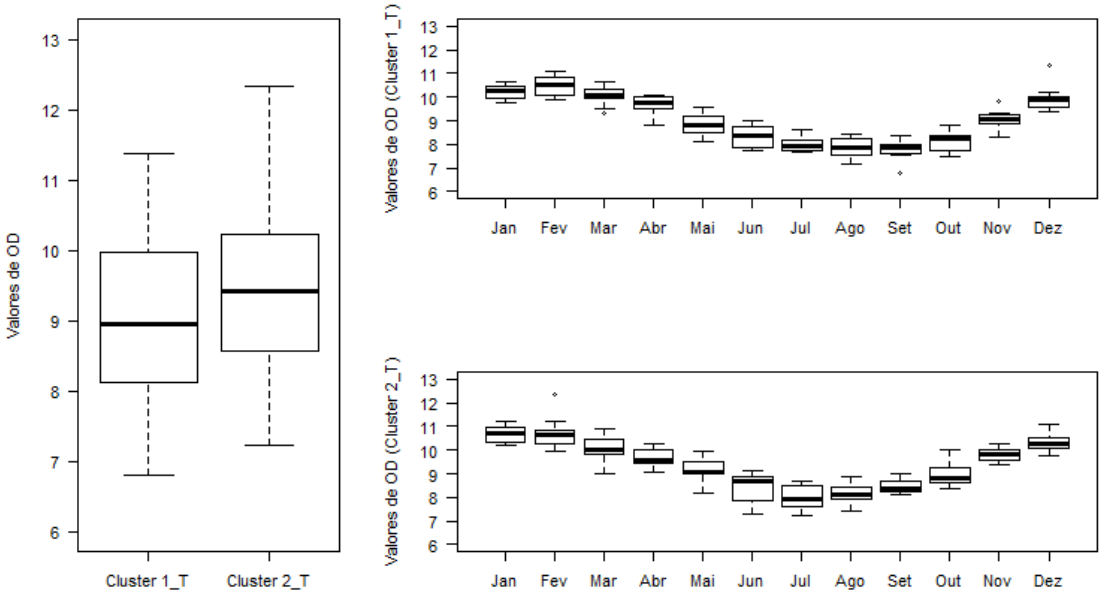


Figura 6.22: *Boxplots* das série de OD dos *Clusters* 1_T e 2_T (à esquerda) e *boxplots* dos valores de OD dos *Clusters* 1_T e 2_T divididos por meses (à direita).

Terminada esta análise, pode-se agora proceder à modelação das quatro séries temporais.

6.6.2 Condutividade

Cluster 1_T **- Modelo SARIMA**

O primeiro passo para identificar um modelo SARIMA que descreva a série em estudo passa por tornar a mesma estacionária, caso não o seja. A representação gráfica da série de COND do *Cluster* 1_T (presente na Figura 6.17) indica que a mesma não é estacionária em média, nem em variância. Assim sendo, aplicou-se uma transformação logarítmica para tornar a série estacionária em variância e de seguida uma diferenciação regular, para tornar a série estacionária em média. Esta diferenciação demonstrou ser suficiente para tornar a série estacionária.

Uma vez que a série observada já foi transformada numa série estacionária, procede-se ao ajuste da parte sazonal do modelo. Para o efeito, determina-se o valor da sazonalidade, que, de acordo com a análise da Figura B.3 (Apêndice B) é 12, $s = 12$. Os restantes valores de P, D e Q são determinados através da exploração e comparação de várias possibilidades, em que P, D e Q tomam o valor 0, 1 ou 2.

Na prática, o melhor modelo é selecionado através da comparação dos valores de AIC, BIC e EQM em que, recorde-se diferenças de AIC inferiores a dois valores são insignificativas, ou seja, nestes casos, os modelos são igualmente bons. O modelo selecionado (melhor modelo) corresponde ao modelo em que P, D e Q são, respetivamente, 1, 0 e 1.

Terminado o ajuste da parte sazonal do modelo resta agora identificar as ordens da parte regular, ou seja, de p, d e q. Tipicamente, estas ordens tomam os valores 0, 1 ou 2.

Novamente, as ordens são determinadas através da comparação de vária possibilidades. Comparando os valores de AIC, BIC e EQM, determina-se que o melhor modelo corresponde ao modelo em que p, d e q tomam todos o valor 1. Deste modo, encontrou-se o modelo final, que corresponde a um SARIMA$(1,1,1)(1,0,1)_{12}$. As estimativas dos parâmetros e respetivos desvios padrões estão presentes na Tabela 6.35.

Tabela 6.35: Resultados do ajustamento do modelo SARIMA à série dos logarítmos da COND do *Cluster* 1_T.

Modelo final: SARIMA$(1,1,1)(1,0,1)_{12}$				
Parâmetro	ϕ_1	θ_1	ν_1	η_1
Estimativa	0,425	-0,990	0,959	-0,760
Desvio Padrão	0,086	0,018	0,044	0,140

Terminada a estimação do modelo SARIMA e de forma a validar o mesmo, deve-

se proceder à análise dos resíduos, que devem cumprir os pressupostos de normalidade, média nula, variância constante e não correlação. A Figura 6.23 exibe o histograma, a curva de densidade dos resíduos, e as respetivas FAC e FACP.

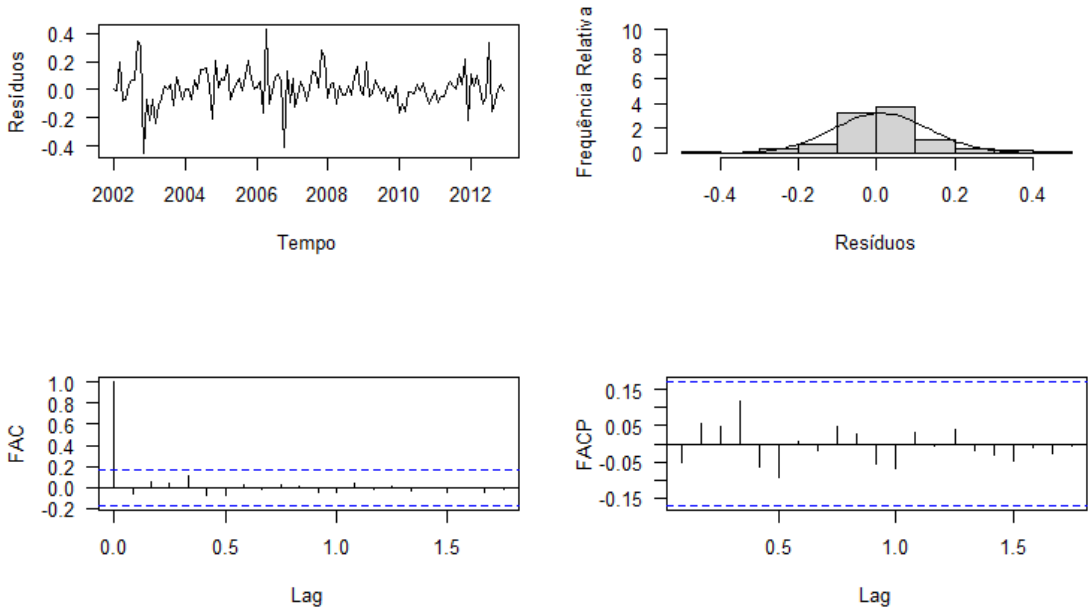


Figura 6.23: Série dos resíduos, respetivo histograma, curva de densidade, FAC e FACP após a aplicação do modelo SARIMA à série de COND do *Cluster* 1_T.

A análise dos mesmos evidencia uma distribuição aleatória em torno de zero, uma curva de densidade simétrica e semelhante à de uma distibuição Normal e a ausência de autocorrelação dos resíduos. De facto, a hipótese de normalidade é confirmada através da aplicação do teste de Kolmogorov-Smirnov, com valor de prova 0,230, a hipótese de média nula através da aplicação do teste t-Student com valor de prova 0,299 e a hipótese de independência através da aplicação do teste de Teste de Ljung-Box, para o qual se obteve o valor de prova 0,989. Apenas a hipótese de variância constante não é verificada (observar a representação gráfica), o que sucede devido à grande variabilidade que esta variável apresenta.

Mesmo assim, foram calculadas as previsões pontuais e intervalares para a série de teste. Na Figura 6.24 encontram-se representadas as previsões (pontuais e intervalares) no período de teste (de janeiro de 2013 a dezembro de 2013), as estimativas pontuais no período de treino (de janeiro de 2002 a dezembro de 2012) e a série observada. A mesma representação noutra escala é apresentada na Figura B.4 (Apêndice B).

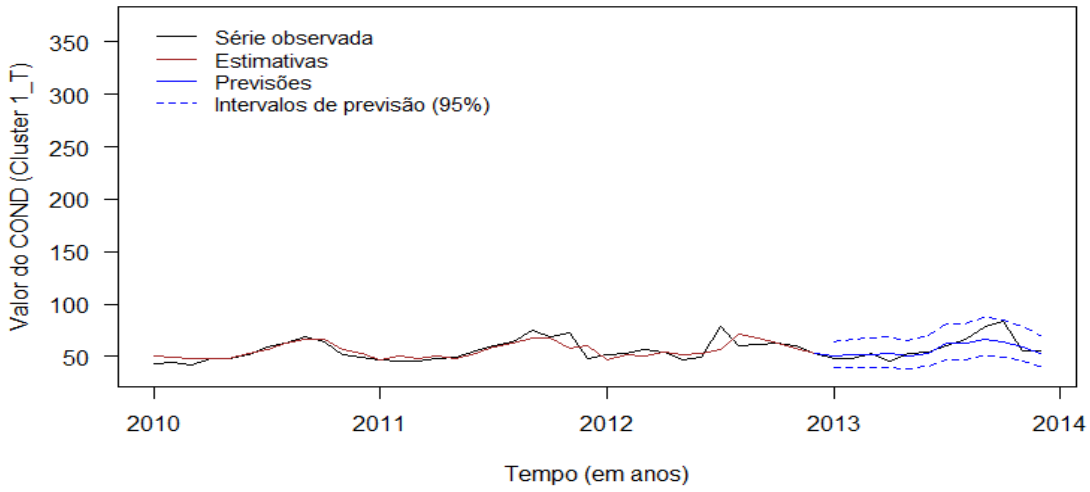


Figura 6.24: Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do modelo SARIMA para a COND do *Cluster* 1_T.

A respetiva taxa de cobertura, ou seja, a percentagem de valores observados que pertencem ao intervalo de previsão, é de 100%.

Cluster 1_T - Método de Holt-Winters

O primeiro passo para modelar uma série temporal através de um método de alisamento exponencial passa por identificar se a série em estudo possui tendência e/ou sazonalidade. No caso, a série em estudo possui uma forte componente sazonal, logo o método mais adequado para a estudar é o método de Holt-Winters.

O método de Holt-Winters distingue o tipo de sazonalidade (aditivo ou multiplicativo), por este mesmo motivo, ambos os modelos foram considerados e estimados. A seleção do mais eficaz é feita através da comparação do erro quadrático médio, EQM, escolhendo-se aquele que possui o valor mais baixo. No caso, o modelo aditivo demonstrou ser o mais adequado. As suas estimativas iniciais para o nível, declive e fatores sazonais são apresentadas na Tabela 6.36, para além das estimativas das constantes de alisamento.

Tabela 6.36: Resultados do ajustamento do método aditivo de Holt-Winters à série de COND do *Cluster* 1_T.

Modelo final: Aditivo				
$\hat{\alpha} = 0,271$	$\hat{\beta} = 0,021$	$\hat{\gamma} = 0,586$	$\hat{l}_1 = 51,415$	$\hat{b}_1 = -0,087$
$\hat{s}_1 = -2,536$	$\hat{s}_2 = -1,732$	$\hat{s}_3 = -1,211$	$\hat{s}_4 = -1,012$	$\hat{s}_5 = -4,048$
$\hat{s}_6 = 0,271$	$\hat{s}_7 = 16,012$	$\hat{s}_8 = 8,492$	$\hat{s}_9 = 13,751$	$\hat{s}_{10} = 12,741$
$\hat{s}_{11} = 1,394$	$\hat{s}_{12} = 0,162$			

Estimado o modelo procedeu-se ao cálculo das previsões pontuais calculadas a 1-passo, para a série de teste. A validade dos intervalos de previsão do método de Holt-Winters depende do cumprimento do pressuposto de normalidade dos erros das previsões pontuais. Deste modo, aplicou-se o teste de Kolmogorov-Smirnov e comprovou-se que estes possuem uma distribuição Normal (valor de prova = 0,941).

Posto isto, pode-se agora proceder ao cálculo das previsões intervalares, calculadas a 1-passo, para a série de teste. Na Figura 6.25 representam-se as previsões pontuais e intervalares, no período de teste, e as estimativas pontuais, no período de treino, sobrepostas à série observada. A mesma representação numa escala mais adequada é apresentada na Figura B.5 (Apêndice B).

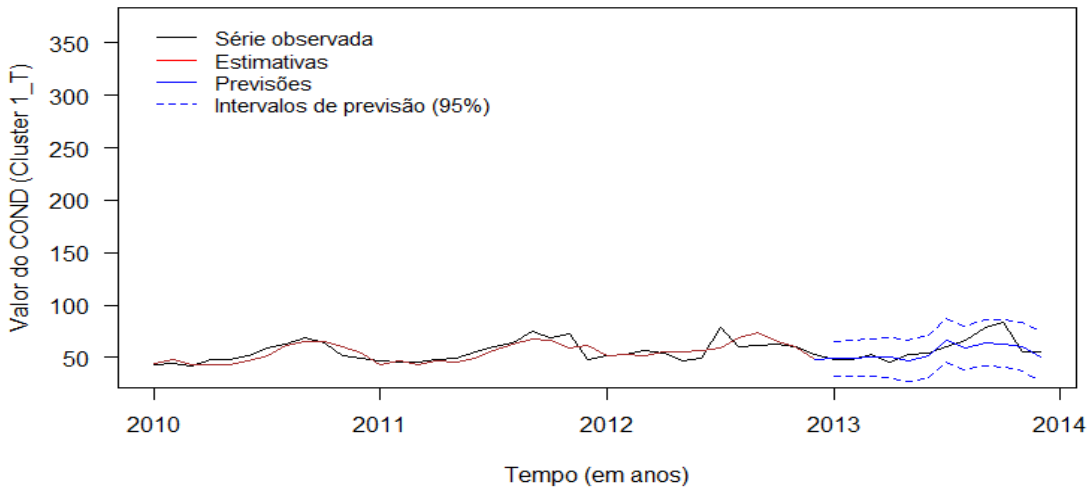


Figura 6.25: Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do método de Holt-Winters para a COND do *Cluster* 1_T.

As previsões obtidas apresentam uma taxa de cobertura de 100%. Os valores observados e os valores das previsões pontuais e intervalares (no período de teste)

deste método e do modelo SARIMA são apresentadas na Tabela C.1 (Apêndice C).

Cluster 2_T - Modelo SARIMA

A representação gráfica da série da COND do *Cluster* 2_T indica que a mesma é estacionária em média, mas não é estacionária em variância. Logo, à semelhança da série da COND do *Cluster* 1_T, aplicou-se uma transformação logarítmica, para tornar a série estacionária em variância. Note-se que, uma vez que a série é estacionária em média, não é necessário estabilizar a tendência, ou seja no modelo final deve ter-se que $d = 0$.

Posto isto, procedeu-se à determinação das ordens da parte sazonal do modelo. Para o efeito, determina-se o valor da sazonalidade, $s = 12$ (ver Figura B.6 - Apêndice B), os restantes valores de P, D e Q são determinados através da exploração e comparação de várias possibilidades. Os valores de AIC, BIC e EQM determinam que o melhor modelo corresponde ao modelo em que P=0, D=1 e Q=1.

De forma análoga encontram-se os melhores valores para p, d e q, que são, respetivamente, 0, 0 e 2, onde d já era conhecido. Deste modo, determinou-se o modelo final, que corresponde a um SARIMA$(0,0,2)(0,1,1)_{12}$. As estimativas dos parâmetros e respetivos desvios padrões estão presentes na Tabela 6.37.

Tabela 6.37: Resultados do ajustamento do modelo SARIMA à série dos logarítmos de COND do *Cluster* 2_T.

Modelo final: SARIMA$(0,0,2)(0,1,1)_{12}$			
Parâmetro	θ_1	θ_2	η_1
Estimativa	0,477	0,234	-1,000
Desvio Padrão	0,089	0,081	0,261

Determinado o modelo final procedeu-se à análise dos resíduos com o intuito de validar o mesmo. A Figura 6.26 apresenta o histograma, a curva de densidade dos resíduos e as respetivas FAC e FACP. A sua análise evidencia uma distribuição aleatória em torno de zero, uma curva de densidade ligeiramente assimétrica à esquerda e a ausência de autocorrelação dos resíduos. Apesar da assimetria observada a hipótese de normalidade não é rejeitada pelo teste de Kolmogorov-Smirnov (valor de prova = 0,284). Assim, e comprovando os restantes pressupostos através do teste t-Student e do teste de Ljung-Box, com valores de prova 0,498 e 0,476, conclui-se que os resíduos seguem uma distribuição Normal, de média nula, variância constante e que são independentes entre si.

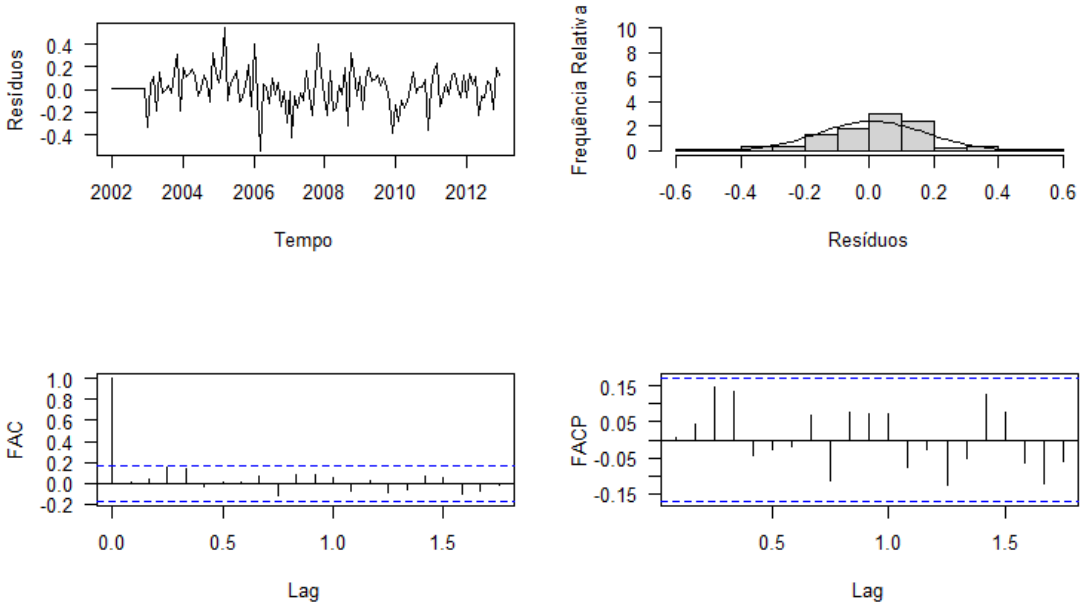


Figura 6.26: Série dos resíduos, respetivo histograma, curva de densidade, FAC e FACP após a aplicação do modelo SARIMA à série de COND do *Cluster* 2_T.

Posto isto, calculam-se as previsões pontuais e intervalares para a série de teste. As previsões (pontuais e intervalares) no período de teste, as estimativas pontuais no período de treino e a série observada encontram-se representadas na Figura 6.27.

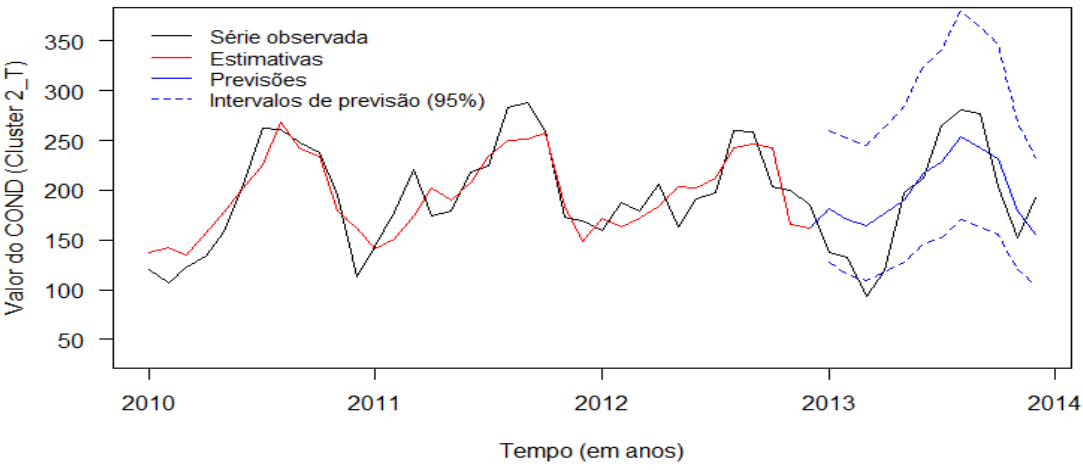


Figura 6.27: Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do modelo SARIMA para a COND do *Cluster* 2_T.

As previsões obtidas apresentam uma taxa de cobertura de 92%.

Cluster 2_T - Método de Holt-Winters

O primeiro passo para modelar uma série temporal através de um método de alisamento exponencial passa por identificar se a série em estudo possui tendência e/ou sazonalidade. Uma vez mais, ambos os modelos, aditivo e multiplicativo foram estimados e a sua seleção é feita através da comparação do EQM, escolhendo-se o modelo aditivo por possuir o valor mais baixo. As estimativas iniciais para o nível, declive e fatores sazonais e estimativas das constantes de alisamento do mesmo são apresentadas na Tabela 6.38.

Tabela 6.38: Resultados do ajustamento do método aditivo de Holt-Winters à série de COND do *Cluster* 2_T.

Modelo final: Aditivo				
$\hat{\alpha} = 0,414$	$\hat{\beta} = 0,000$	$\hat{\gamma} = 0,381$	$\hat{l}_1 = 217,555$	$\hat{b}_1 = 0,601$
$\hat{s}_1 = -45,498$	$\hat{s}_2 = -39,052$	$\hat{s}_3 = -29,286$	$\hat{s}_4 = -25,864$	$\hat{s}_5 = -20,024$
$\hat{s}_6 = 17,548$	$\hat{s}_7 = 35,864$	$\hat{s}_8 = 69,641$	$\hat{s}_9 = 52,607$	$\hat{s}_{10} = 30,742$
$\hat{s}_{11} = -14,794$	$\hat{s}_{12} = -46,335$			

Como referido, o pressuposto de normalidade dos erros das previsões pontuais é necessário para o cálculo de previsões intervalares. Assim, depois de calculadas as previsões pontuais a 1-passo, aplicou-se o teste de Kolmogorov-Smirnov e verificou-se que os seus erros seguem uma distribuição Normal, já que o valor de prova é de 0,947.

Assim, pode-se agora efetuar o cálculo das previsões intervalares, calculadas a 1-passo, para a série de teste. Na Figura 6.28 representam-se as previsões (pontuais e intervalares) no período de teste, as estimativas pontuais no período de treino e a série observada.

Neste caso, as previsões obtidas apresentam uma taxa de cobertura de 92%.

Os valores observados, as previsões pontuais e os limites das previsões intervalares (no período de teste) deste método e do modelo SARIMA (aplicados à série de COND do *Cluster* 2_T) são apresentadas na Tabela C.2 (Apêndice C).

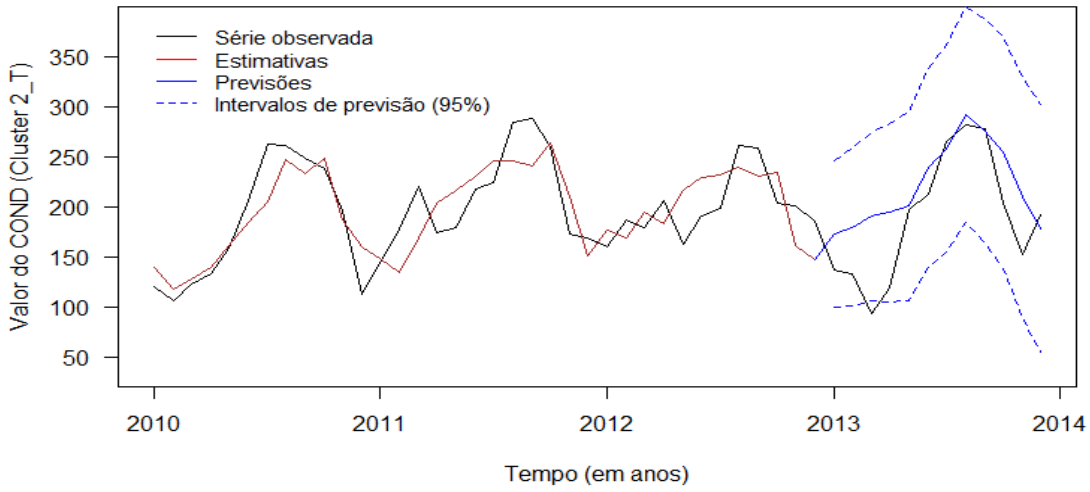


Figura 6.28: Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através o método de Holt-Winters para a COND do *Cluster* 2_T.

Comparação dos Métodos (Condutividade)

A Tabela 6.39 apresenta os valores de várias medidas de avaliação calculadas para a série de treino e para a série de teste de cada *cluster* de COND, depois de aplicado cada tipo de modelo. O intuito do cálculo destas medidas é comparar a eficácia de ambos os modelos e identificar o que produz previsões mais precisas.

Tabela 6.39: Medidas de avaliação de cada um dos modelos aplicados a cada um dos *clusters* de COND.

		Série de Treino				Série de Teste			
	Modelo	EQM	REQM	EPAM	U-Theil	EQM	REQM	EPAM	U-Theil
Cluster 1_T	SARIMA	53,669	7,326	8,602	0,812	**53,583**	**7,320**	**8,143**	**0,745**
	Holt-Winters	73,776	8,589	10,744	1,064	69,133	8,315	10,113	0,881
Cluster 2_T	SARIMA	975,180	31,228	12,475	0,767	**1491,591**	**38,621**	**23,121**	**1,006**
	Holt-Winters	1353,895	36,795	14,387	0,886	2107,395	45,906	26,501	1,283

Comparando os valores das diversas medidas verifica-se que, tanto para a análise do *Cluter* 1_T como para a análise do *Cluster* 2_T, os modelos SARIMA são mais exatos comparativamente aos métodos de Holt-Winters (as várias medidas assumem valores inferiores). Assim, as previsões comparadas serão as calculadas através dos modelos SARIMA para ambos os *clusters*.

6.6.3 Oxigénio Dissolvido

Cluster 1_T - Modelo SARIMA

O primeiro passo para identificar um modelo SARIMA que descreva a série em estudo passa por tornar a mesma estacionária, caso não o seja. A representação gráfica da série de OD do *Cluster* 1_T (presente na Figura 6.19) indica que a mesma é estacionária em média, mas não é estacionária em variância. Logo, aplicou-se uma transformação logarítmica para ultrapassar esta questão. Repare-se que, uma vez que a série é estacionária em média, futuramente, considerar-se-á $d = 0$.

Feita a transformação necessária deve-se agora proceder ao ajuste do modelo, começando-se pela parte sazonal e posteriormente parte regular e comparando as diversas combinações de possibilidades para P, D e Q e p, d e q. Recorde-se que os parâmetros da parte sazonal e regular tomam tipicamente valores de 0, 1 ou 2 (valores superiores a estes são extremamente raros). Analisando a Figura B.7 (Apêndice B) determina-se que $s = 12$.

Em cada uma das etapas, os valores de AIC, BIC e EQM permitiram avaliar o modelo mais apropriado, que se determinou ser um modelo SARIMA$(1,0,0)(1,0,1)_{12}$. As estimativas dos parâmetros e respetivos desvios padrões estão presentes na Tabela 6.40.

Tabela 6.40: Resultados do ajustamento do modelo SARIMA à série dos logarítmos do OD do *Cluster* 1_T.

Modelo final: SARIMA$(1,0,0)(1,0,1)_{12}$				
Parâmetro	ϕ_1	ν_1	η_1	μ
Estimativa	0,275	0,910	-0,966	2,195
Desvio Padrão	0,092	0,001	0,084	0,033

Terminada a estimação do modelo, procedeu-se à análise dos resíduos, que devem cumprir os pressupostos de normalidade, média nula, variância constante e não correlação. Na Figura 6.29 estão representados todos os componentes gráficos que sustentam esta avaliação. Como se pode observar a distribuição dos mesmos é aleatória em torno de zero, a respetiva curva de densidade é ligeiramente assimétrica à direita e os resíduos não são correlacionados. De facto, a hipótese de normalidade é confirmada através da aplicação do teste de Kolmogorov-Smirnov, cujo valor de prova é 0,800, a hipótese de média nula é verificada através do teste t-Student com valor de prova 0,890 e a hipótese de independência é verificada através da aplicação do teste de Teste de Ljung-Box, com o valor de prova 0,093. Assim, concluiu-se que os resíduos verificam todos os pressupostos.

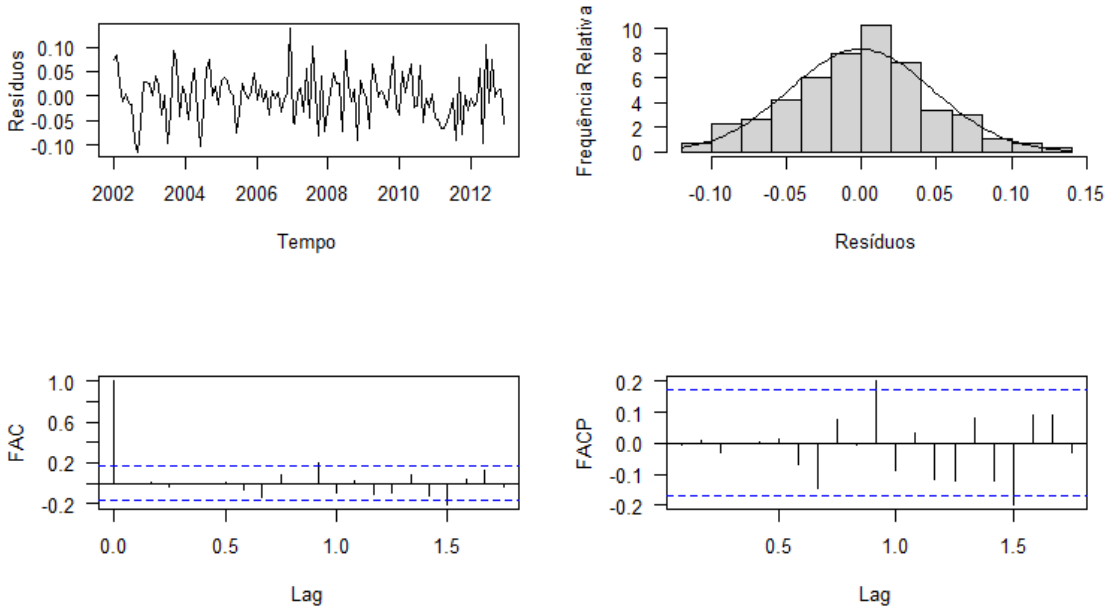


Figura 6.29: Série dos resíduos, respetivo histograma, curva de densidade, FAC e FACP após a aplicação do modelo SARIMA à série de OD do *Cluster* 1_T.

Posto isto, reúnem-se agora todas as condições para que as previsões pontuais e intervalares para a série de teste sejam calculadas. Na Figura 6.30 apresentam-se as previsões (pontuais e intervalares) no período de teste e as estimativas pontuais no período de treino sobrepostas à série observada. As previsões obtidas apresentam uma taxa de cobertura de 100%.

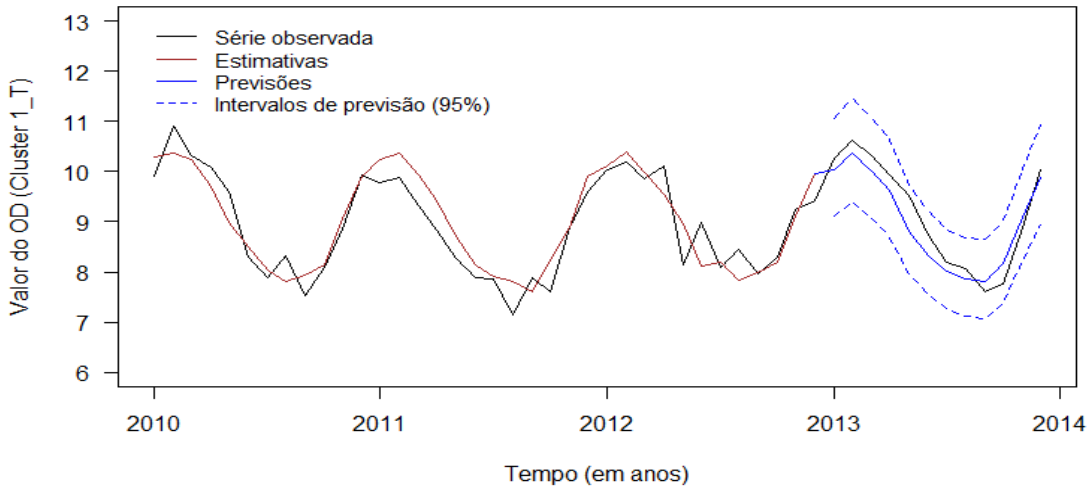


Figura 6.30: Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do modelo SARIMA para o OD do *Cluster* 1_T.

Cluster 1_T - Método de Holt-Winters

Novamente, ambos os métodos de Holt-Winters (com sazonalidade aditiva e multiplicativa) foram estimados e o mais eficaz selecionado através da comparação do erro quadrático médio, EQM, escolhendo-se aquele que possui o valor mais baixo, que no caso demonstrou ser o modelo aditivo. As suas estimativas iniciais para o nível, declive e fatores sazonais são apresentadas na Tabela 6.41, para além das estimativas das constantes de alisamento.

Tabela 6.41: Resultados do ajustamento do método aditivo de Holt-Winters à série de OD do *Cluster* 1_T.

Modelo final: Aditivo				
$\hat{\alpha} = 0,099$	$\hat{\beta} = 0$	$\hat{\gamma} = 0,339$	$\hat{l}_1 = 9,195$	$\hat{b}_1 = 0,007$
$\hat{s}_1 = 0,933$	$\hat{s}_2 = 1,206$	$\hat{s}_3 = 0,786$	$\hat{s}_4 = 0,639$	$\hat{s}_5 = -0,452$
$\hat{s}_6 = -0,669$	$\hat{s}_7 = -1,088$	$\hat{s}_8 = -1,135$	$\hat{s}_9 = -1,313$	$\hat{s}_{10} = -1,070$
$\hat{s}_{11} = -0,062$	$\hat{s}_{12} = 0,580$			

As previsões pontuais para a série de teste foram calculadas a 1-passo. Posteriormente, com o intuito de calcular previsões intervalares, verificou-se se os seus erros seguem uma distribuição Normal através da aplicação do teste de Kolmogorov-

Smirnov. Neste caso, obteve-se um valor de prova de 0,827, concluindo-se que, os erros das previsões pontuais seguem uma distribuição Normal.

Assim, estão agora reunidas todas as condições para efetuar o cálculo das previsões intervalares a 1-passo para a série de teste. As previsões (pontuais e intervalares) no período de teste, as estimativas pontuais no período de treino e a série observada encontram-se representadas na Figura 6.31.

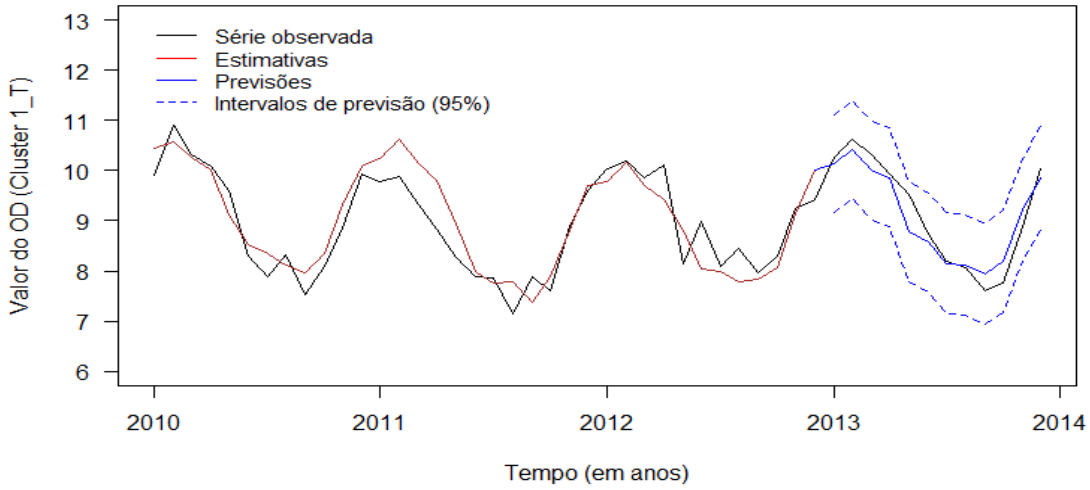


Figura 6.31: Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do método de Holt-Winters para o OD do *Cluster* 1_T.

As previsões obtidas apresentam uma taxa de cobertura de 100%.

Os valores observados, as previsões pontuais e os limites das previsões intervalares (no período de teste) deste método e do modelo SARIMA (aplicados à série de OD do *Cluster* 1_T) são apresentadas na Tabela C.3 (Apêndice C).

Cluster 2_T - Modelo SARIMA

A representação gráfica da série de OD do *Cluster* 2_T (presente na Figura 6.20) indica que a mesma é estacionária em média, mas não é estacionária em variância. Logo, aplicou-se uma transformação logarítmica e novamente considerar-se-á que $d = 0$, uma vez que não é necessário estabilizar a tendência.

Feita a transformação necessária, procede-se à determinação das ordens da parte sazonal do modelo. Para o efeito, determina-se o valor da sazonalidade, $s = 12$ (ver Figura B.8 - Apêndice B) e os valores de P, D e Q são determinados através da exploração e comparação de várias possibilidades (modelos) e respetivos valores

de AIC, BIC e EQM. No caso, determina-se que o melhor modelo corresponde ao modelo em que P, D e Q são, respetivamente, 2, 0 e 0.

De forma análoga encontra-se os melhores valores para p, d e q, que são, respetivamente, 1, 0 e 0, onde d já era conhecido. Deste modo, determinou-se o modelo final, que corresponde a um SARIMA$(1,0,0)(2,0,0)_{12}$. As estimativas dos parâmetros e respetivos desvios padrões estão presentes na Tabela 6.42.

Tabela 6.42: Resultados do ajustamento do modelo SARIMA à série dos logarítmos do OD do *Cluster* 2_T.

Modelo final: SARIMA$(1,0,0)(2,0,0)_{12}$				
Parâmetro	ϕ_1	ν_1	ν_2	μ
Estimativa	0,490	0,497	0,241	2,232
Desvio Padrão	0,087	0,089	0,092	0,029

Determinado o modelo final procedeu-se à análise dos resíduos com o intuito de validar o mesmo. A Figura 6.32 exibe a representação gráfica, histograma e curva de densidade dos resíduos, para além das respetivas FAC e FACP.

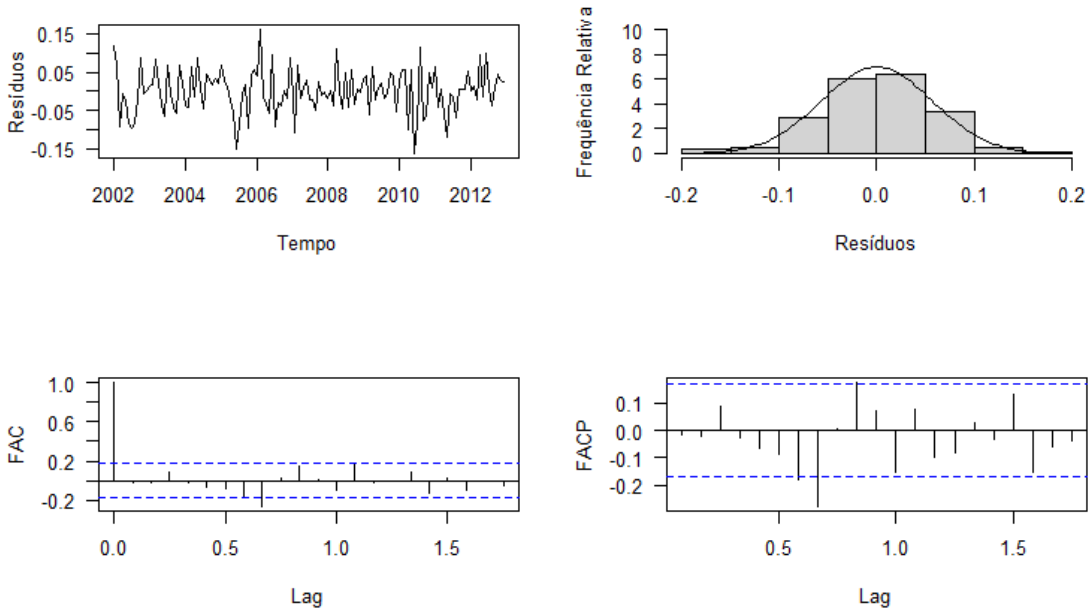


Figura 6.32: Série dos resíduos, respetivo histograma, curva de densidade, FAC e FACP após a aplicação do modelo SARIMA à série de OD do *Cluster* 2_T.

A sua análise revela uma distribuição aleatória em torno de zero, uma curva de densidade simétrica e resíduos correlacionados. A hipótese de normalidade é comprovada pela aplicação do teste de Kolmogorov-Smirnov cujo valor de prova é 0,986. Os pressupostos relativos à média e independência são, respetivamente, avaliados através do teste t-Student e do teste de Ljung-Box, obtendo-se os valores de prova 0,988 e 0,013. Assim, conclui-se que os resíduos seguem uma distribuição Normal, de média nula, variância constante, mas dependentes entre si.

No entanto, foram calculadas as previsões (pontuais e intervalares) para a série de teste. As previsões (pontuais e intervalares) no período de teste, as estimativas pontuais no período de treino e a série observada encontram-se representadas na Figura 6.33. As previsões obtidas apresentam uma taxa de cobertura de 92%.

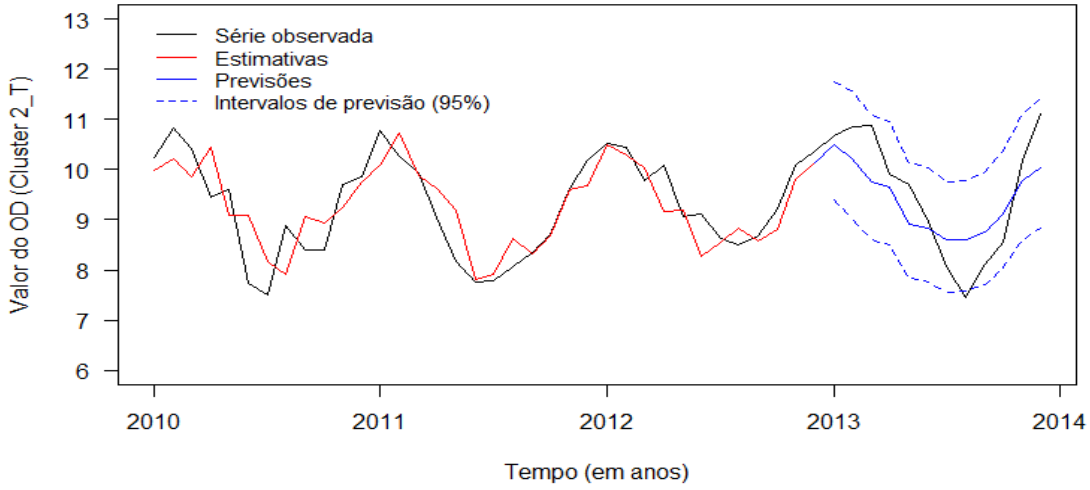


Figura 6.33: Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através do modelo SARIMA para o OD do *cluster* 2_T.

Cluster 2_T - Método de Holt-Winters

Uma vez mais, ambos os modelos, aditivo e multiplicativo foram estimados e a sua seleção feita através da comparação do EQM, escolhendo-se o modelo aditivo. As suas estimativas iniciais para o nível, declive e fatores sazonais e estimativas das constantes de alisamento são apresentadas na Tabela 6.43.

Tabela 6.43: Resultados do ajustamento do método aditivo de Holt-Winters à série de OD do *Cluster* 2_T.

Modelo final: Aditivo				
$\hat{\alpha} = 0,233$	$\hat{\beta} = 0,004$	$\hat{\gamma} = 0,291$	$\hat{l}_1 = 9,784$	$\hat{b}_1 = 0,015$
$\hat{s}_1 = 1,150$	$\hat{s}_2 = 1,144$	$\hat{s}_3 = 0,620$	$\hat{s}_4 = 0,322$	$\hat{s}_5 = -0,369$
$\hat{s}_6 = -1,012$	$\hat{s}_7 = -1,379$	$\hat{s}_8 = -1,225$	$\hat{s}_9 = -1,044$	$\hat{s}_{10} = -0,582$
$\hat{s}_{11} = 0,270$	$\hat{s}_{12} = 0,669$			

Novamente, procedeu-se ao cálculo das previsões pontuais a 1-passo para a série de teste e comprovou-se o pressuposto de normalidade dos seus erros através da aplicação do teste de Kolmogorov-Smirnov, obtendo-se um valor de prova de 0,975.

De seguida, procedeu-se ao cálculo das previsões intervalares a 1-passo para a série de teste. As previsões (pontuais e intervalares) no período de teste, as estimativas pontuais no período de treino e a série observada encontram-se representadas na Figura 6.34.

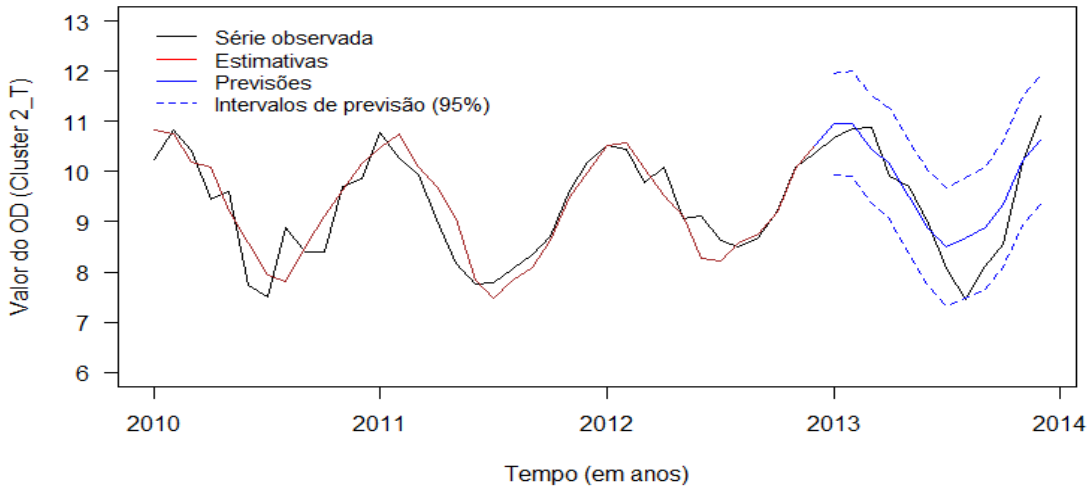


Figura 6.34: Estimativas pontuais (no período de treino) e previsões pontuais e intervalares (no período de teste) obtidas através o método de Holt-Winters para o OD do *Cluster* 2_T.

À semelhança do *Cluster* 1_T, os previsões obtidas apresentam uma taxa de cobertura de 92%. Os valores observados, as previsões pontuais e os limites das previsões intervalares (no período de teste) deste método e do modelo SARIMA (aplicados à série de OD do *Cluster* 2_T) são apresentadas na Tabela C.4 (Apêndice C).

Comparação dos Métodos (Oxigénio Dissolvido)

A Tabela 6.44 exibe os valores das medidas de avaliação, mas calculadas para a série de treino e teste de cada *cluster* de OD após a aplicação do modelo SARIMA e método de Holt-Winters.

Tabela 6.44: Medidas de avaliação de cada um dos modelos aplicados a cada um dos *clusters* de OD.

		Série de Treino				Série de Teste			
	Modelo	EQM	REQM	EPAM	U-Theil	EQM	REQM	EPAM	U-Theil
Cluster 1_T	SARIMA	0,179	0,423	3,695	0,577	0,108	0,329	3,220	0,503
	Holt-Winters	0,246	0,496	4,197	0,708	**0,098**	**0,315**	**2,777**	**0,501**
Cluster 2_T	SARIMA	0,280	0,529	4,457	0,739	0,509	0,714	6,732	0,908
	Holt-Winters	0,269	0,519	4,350	0,776	**0,297**	**0,545**	**4,884**	**0,768**

Observando a mesma verifica-se, para ambos os *clusters*, que os valores das várias medidas são inferiores para o método de Holt-Winters. Assim, os métodos de Holt-Winters são mais exatos que os modelos SARIMA para ambos os *clusters* e, portanto, as previsões comparadas serão as calculadas através deste método.

6.7 *Clusters*: Medidas de Risco *vs* Séries Temporais

Nesta secção pretende-se em primeiro lugar identificar através das previsões obtidas, para cada variável, qual o *cluster* que possui o maior risco de poluição. Em segundo lugar, pretende-se comparar os resultados obtidos com a classificação anteriormente obtida através de medidas de risco. Esta comparação poderá evidenciar se a classificação através de medidas de risco obtida pode ser interpretada como um índice de risco futuro.

Condutividade

A Figura 6.35 apresenta as representações gráficas das previsões pontuais calculadas para o *Cluster* 1_T e 2_T de COND, para além da linha horizontal que identifica o valor 250. Este valor distingue (segundo a Tabela 6.2, utilizada pela APA) o nível de qualidade *Razoável* dos níveis de qualidade *Excelente* e *Bom*. Em particular, valores superiores a 250 associam-se ao nível de qualidade *Razoável* e valores inferiores a 250 associam-se aos níveis de qualidade *Excelente* ou *Bom*.

Observando a Figura 6.35 deteta-se que as previsões pontuais do *Cluster* 2_T são sempre superiores às previsões pontuais do *Cluster* 1_T. Para além disso, o *Cluster* 2_T possui uma previsão pontual superior ao valor de referência. Assim, diz-se que o *Cluster* 2_T (estações: COR, TAM, MEL e MOL) possui um maior nível de poluição.

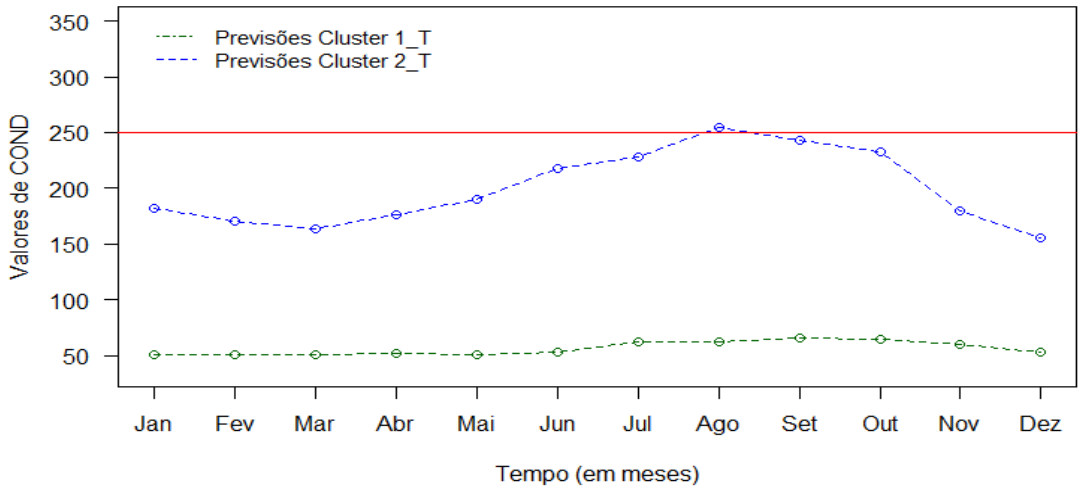


Figura 6.35: Previsões pontuais calculadas através dos modelos SARIMA para os *Clusters* 1_T e 2_T de COND.

Para complementar esta análise apresentam-se na Tabela 6.45 para cada *cluster* os valores médios da totalidade das previsões pontuais e os valores médios das previsões correspondentes aos meses secos e húmidos, juntamente com os respetivos *rankings* (ordens-O) de classificação do risco, que devem ser interpretados como quanto maiores maior o risco de poluição. Para além disto e apesar da limitação desta comparação, uma vez que os *clusters* referentes aos meses secos ou húmidos são distintos, apresentam-se as ordens das médias (O-M) e ordens finais (OF) anteriormente obtidas para a classificação de *clusters* através de medidas de risco. Repare-se que a comparação exata dos mesmos *clusters*, isto é, *Cluster* 1_T, 2_T, 1_S, 2_S, 1_H e 2_H através da análise de séries temporais e análise de risco é impossível, já que os conjuntos de dados referentes aos meses secos/húmidos (*Clusters* 1_S, 2_S, 1_H e 2_H) não podem ser modelados através de séries temporais.

Para além disto, note-se que as ordens referentes à abordagem de medidas de risco (O-M e OF) foram obtidas após o cálculo das médias dos valores de cada *cluster* de COND para o período observado (janeiro de 2002 a dezembro de 2013), ao passo que as ordens referentes à abordagem de séries temporais (O) foram obtidas após o cálculo das médias das previsões de cada *cluster* de COND para o período de teste (janeiro de 2013 a dezembro de 2013).

Tabela 6.45: Ordens dos *clusters* de COND calculadas através de séries temporais e medidas de risco.

		M	**O**		**O-M**	OF
Total	*Cluster* 1_T	56,367	2	*Cluster* 1_T	2	2
	Cluster 2_T	199,433	5	*Cluster* 2_T	5	6
	OF		**2**		**2**	**3**
Secos	*Cluster* 1_T	58,824	3	*Cluster* 1_S	3	1
	Cluster 2_T	226,457	6	*Cluster* 2_S	6	4
	OF		**3**		**3**	**1**
Húmidos	*Cluster* 1_T	54,612	1	*Cluster* 1_H	1	3
	Cluster 2_T	180,130	4	*Cluster* 2_H	4	5
	OF		**1**		**1**	**2**

Comparando as ordens das médias, O-M, dos *Clusters* 1_T e 2_T, obtidos através de medidas de risco verifica-se que o risco do *Cluster* 2_T é superior ao risco do *Cluster* 1_T, logo conclui-se que ambas as abordagens (medidas de risco e séries temporais) retiram conclusões iguais. Aliás, analisando as ordens obtidas através de cada uma das abordagens verifica-se que as mesmas coincidem em todos os casos.

Observando as ordens das médias e as ordens finais (ambas obtidas através de medidas de risco) verifica-se que, estas não coincidem na maioria dos casos. Tal facto destaca a influência das restantes medidas de risco (semivariância, entropia, etc) na ordenação final dos riscos e sugere uma nova proposta de trabalho futuro que será explicitada na Secção 7.1.

Oxigénio Dissolvido

Relativamente à variável OD, a Figura 6.36 exibe as representações gráficas das previsões pontuais calculadas para os *Clusters* 1_T e 2_T de OD, para além da linha horizontal que identifica o valor 8 (valor de referência que distingue níveis de qualidade da água).

Recorde-se que os valores de OD se relacionam indiretamente com o nível de qualidade da água, isto é, valores inferiores associam-se a uma menor qualidade de água. Em particular, valores inferiores a 8 associam-se a níveis de qualidade *Bons* ou *Razoáveis*.

Observando a Figura 6.36 repara-se que as previsões pontuais do *Cluster* 1_T são sempre inferiores às previsões pontuais do *Cluster* 2_T. Para além disso, o *Cluster* 1_T possui uma previsão pontual inferior a oito. Logo, conclui-se que o *Cluster* 1_T (estações: MOL, RAN, MEL, SEM, TAM, SOR, ALV, PEN e COR) possui um maior nível de poluição.

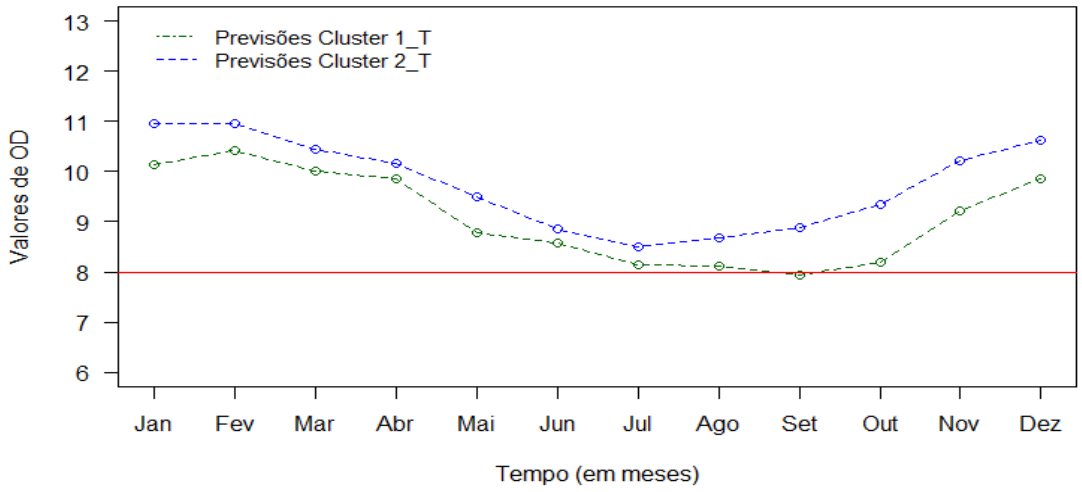


Figura 6.36: Previsões pontuais calculadas através do método de Holt-Winters para os *Clusters* 1_T e 2_T de OD.

À semelhança da análise feita para a variável COND, são apresentadas na Tabela 6.46 as médias da totalidade das previsões pontuais e as médias das previsões correspondentes aos meses secos e húmidos, juntamente com as respetivas ordens de classificação do risco (O), para além das ordens das médias(O-M) e ordens finais (OF) anteriormente obtidas na classificação dos *clusters* através de medidas de risco. As limitações referidas para a variável COND verificam-se também no caso desta variável.

Tabela 6.46: Ordens dos *clusters* de OD calculadas através de séries temporais e medidas de risco.

		M	O		**O-M**	OF
Total	*Cluster* 1_T	9,100	4	*Cluster* 1_T	4	6
	Cluster 2_T	9,760	2	*Cluster* 2_T	3	4
	OF		**2**		**2**	**3**
Secos	*Cluster* 1_T	8,309	6	*Cluster* 1_S	6	5
	Cluster 2_T	8,881	5	*Cluster* 2_S	5	3
	OF		**3**		**3**	**2**
Húmidos	*Cluster* 1_T	9,665	3	*Cluster* 1_H	2	2
	Cluster 2_T	10,388	1	*Cluster* 2_H	1	1
	OF		**1**		**1**	**1**

Analisando as ordens das médias, dos *Clusters* 1_T e 2_T, obtidos através das duas abordagens (medidas de risco e séries temporais) verifica-se que o risco do *Cluster* 1_T é superior ao risco do *Cluster* 2_T. Por sua vez, analisando somente as ordens das

médias da abordagem de séries temporais verifica-se que o *Cluster* 1_T possui sempre uma ordem superior à do *Cluster* 2_T, quer se considere a totalidade das previsões pontuais ou apenas as previsões referentes ao período seco ou húmido. A abordagem de medidas de risco concorda com esta conclusão. Além disso, ambas as abordagens estão em concordância no que toca à distinção entre período seco ou húmido, no caso, ambas determinam que o risco é superior nos meses secos (ver SO).

Comparando agora as ordens das médias e as ordens finais (ambas obtidas através de medidas de risco) verifica-se que, as mesmas não coincidem na maioria dos casos, destando-se novamente a influência das restantes medidas de risco. No entanto, destaca-se igualmente o período seco pela negativa.

Assim, conclui-se que, a avaliação da qualidade de água, através do cálculo de medidas de risco ou através da aplicação de modelos de séries temporais é, no geral, muito semelhante e, portanto, a classificação através de medidas de risco anteriormente obtida pode ser interpretada como um índice de risco futuro.

Capítulo 7

Conclusão

O presente trabalho utilizou conceitos de teoria de risco e conceitos de séries temporais, com o objetivo comum e principal de analisar e classificar a qualidade da água de superfície da bacia hidrográfica do rio Douro.

Uma análise preliminar dos dados revelou uma evidente sazonalidade. Por este motivo e dado que estudos anteriores comprovaram que a variação da precipitação afeta a qualidade da água o presente estudo considerou e aplicou as metodologias a três casos distintos: à totalidade dos dados, ao período seco e ao período húmido.

A análise de *clusters* aplicada inicialmente, quer ao conjunto das variáveis de qualidade quer a cada uma delas em particular, revelou que os três casos considerados geram *clusters* distintos entre si resultados que comprovam a ideia de que os níveis de poluição variam ao longo do tempo. Esta característica dificultou a comparação dos *clusters*, em termos de poluição.

Para combater estas questões elaborou-se uma classificação de *clusters* baseada em várias medidas de risco, considerando os *clusters* obtidos através das variáveis: carência bioquímica de oxigénio a 5 dias, condutividade, oxigénio dissolvido e temperatura da água. Os resultados permitiram não só comparar os níveis de risco de poluição dos *clusters* de cada período (total, seco e húmido), como inferir, de um modo geral, que o risco de poluição é superior no período seco comparativamente ao período húmido.

No entanto, as classificações obtidas para a variável condutividade via medidas de risco indicaram que o risco de poluição é superior no período húmido. Assim, elaborou-se uma classificação de risco, semelhante à anterior, não aplicada aos *cluster* mas sim aos meses com o intuito de avaliar se as conclusões obtidas através desta variável são corretas ou resultaram de uma partição em períodos (seco e húmido) inadequada. O que se verificou foi que, de facto, esta classificação sugere uma partição

alternativa dos dados, destacando-se os meses de maio a outubro como os detentores de maior risco de poluição e não os meses de maio a setembro correspondentes ao período seco considerado anteriormente.

Para além disso, esta classificação permitiu identificar os meses do ano mais críticos (que de acordo com a metodologia aplicada são aqueles que possuem os maiores índices de risco). Em particular, as variáveis carência bioquímica de oxigénio e condutividade classificaram os meses de maio a outubro como os que possuem o maior risco, a variável oxigénio dissolvido excluiu desta classificação o mês de maio, ou seja, considerou que o período de junho a outubro é o mais crítico e, por fim, a variável temperatura destacou os meses de abril a novembro.

Adicionalmente, foi elaborado um estudo na área de séries temporais dedicado aos *clusters* das variáveis condutividade e oxigénio dissolvido, com o objetivo de produzir previsões futuras para ambos e comparar se o *cluster* identificado com o maior risco em cada caso, era efetivamente o *cluster* identificado com o maior risco pela classificação de medidas de risco anteriormente elaborada. Para o efeito e dada a natureza das séries em questão, foram aplicadas duas abordagens: os modelos SARIMA e métodos de alisamento exponencial. No caso da variável condutividade, os modelos SARIMA demonstraram ser os mais precisos, já relativamente à variável oxigénio dissolvido os métodos de Holt-Winters foram os melhores. Por fim, as previsões obtidas para ambos os *clusters* foram comparadas e de acordo com os limites de qualidade relativos a cada variável, identificaram-se os *clusters* (estações de monitorização) que podem possuir um maior risco futuro, comprovando-se, em ambos os casos, que este era o mesmo previamente identificado através da classificação de medidas de risco no caso da totalidade dos dados. Assim, a variável condutividade assinalou as estações de Corgo, Melres, Moledo e Tâmega pela negativa, ou seja, considerou que, entre as estações em estudo, estas são as que possuem um maior risco de poluição. Já a variável oxigénio dissolvido identificou as estações Alvão, Corgo, Melres, Moledo, Penude, Ranhados, Semealho, Tâmega e Sordo.

Uma vez que as ordenações, em particular da média e finais, obtidas por ambas as abordagens (medidas de risco e séries temporais) para os *clusters* são aproximadamente iguais, conclui-se que o índice de risco baseado em medidas de risco é apropriado para avaliar a qualidade da água.

7.1 Trabalho Futuro

Como trabalho futuro sugere-se a elaboração de um estudo de classificação baseado em medidas de risco equivalente ao considerado no atual trabalho e aplicado aos mesmos *clusters*, mas que considere a partição maio - outubro (correspondente ao período seco) e novembro - abril (correspondente ao período húmido). A ideia é comparar os resultados obtidos em ambos os casos (futuro e atual), dando especial atenção aos resultados obtidos através da variável condutividade.

Relativamente à classificação através de medidas de risco baseada na variável temperatura (caso dos *clusters* e meses), recomenda-se a construção das Tabelas de ordens análogas às apresentadas, mas focadas no limite 6,5. O objetivo é retirar conclusões que englobem a informação fornecida por ambas. Alternativamente, poder-se-á produzir duas novas tabelas, uma que considera somente para o período de maio a setembro e avalia o risco tendo por base o limiar 25,5 e outra que considera somente o período de outubro a abril e se baseia no limiar 6.5. Notoriamente para além da partição maio - setembro / outubro - abril poder-se-á considerar a partição maio - outubro / novembro - abril no caso desta produzir melhores resultados depois de aplicada a primeira abordagem sugerida nas propostas de trabalho futuro.

Para além do descrito, sugere-se à APA e salvaguardando que tal é possível, a elaboração de duas tabelas que determinem as fronteiras de qualidade das variáveis físico-químicas aplicáveis em rios e albufeiras de acordo com o período do ano (seco ou húmido), isto é, uma tabela que defina diferentes valores de referência para a mesma variável nos diferentes períodos, já que os valores registados nos períodos seco e húmido são distintos. Estas tabelas iriam substituir a atualmente utilizada e permitir uma análise mais específica da qualidade. Em particular, como continuação deste trabalho estas seriam utilizadas para produzir novas classificações de risco, baseadas nas mesmas variáveis de qualidade de água, mas naturalmente distintas e mais específicas.

No que diz respeito à análise de séries temporais sugere-se a aplicação de outros métodos, como métodos de regressão com erros correlacionados. O objetivo é encontrar um método mais eficaz do que os utilizados para cada uma das variáveis e simultaneamente capaz de gerar boas previsões para qualquer uma das variáveis. O principal desafio desta sugestão é lidar com a variabilidade característica de alguma das variáveis.

Já relativamente à comparação de ambas as abordagens, medidas de risco e séries temporais, sugere-se o estudo de séries de medidas de risco, isto é, de séries de variância, entropia, etc dos *clusters* de condutividade e oxigénio dissolvido. O

objetivo é obter ordens para cada uma das medidas (através das previsões pontuais) e posteriormente comparar a sua soma, ordem global, por *cluster*, com as ordens finais obtidas no presente trabalho. Alternativamente, sugere-se a aplicação de medidas de risco às previsões pontuais resultantes de processos de modelação SARIMA ou Holt-Winters. A aplicação destas duas propostas requer a existência de mais dados em estudo, nomeadamente, na primeira proposta pretende-se que cada *cluster* seja constituído por mais estações de monitorização e na segunda proposta pretende-se que o período de teste seja mais extenso.

Por fim, sugere-se a aplicação de todo o trabalho realizado a uma zona, rio ou bacia hidrográfica, onde os níveis de poluição atuais sejam muito elevados. A ideia é, não somente reforçar que todo o trabalho realizado é uma nova e boa proposta, como identificar claramente as estações mais poluídas comparativamente com as restantes, já que, no presente trabalho, alguns dos *clusters* possuem observações e medidas descritivas semelhantes.

Bibliografia

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716–723.

[2] Alpuim, T. (1997). *Séries Temporais* (2ª ed.). Lisboa: Associação de Estudantes da Faculdade de Ciências de Lisboa.

[3] Arya, F. K. & Zhang, L (2015). Time series analysis of water quality parameters at Stillaguamish River using order series method. *Stochastic Environmental Research and Risk Assessment*, 29, 227–239.

[4] Associação Sistema Terrestre Sustentável - ZERO (2023). *Água e Oceanos*. Consultado em 24 de janeiro de 2023. Disponível em `https://zero.ong/areas-tematicas/agua-e-oceanos/`

[5] Bahnemann D. (2015). *Distributions for Actuaries*. Casualty Actuarial Society.

[6] Box, G. e Jenkins, G. (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day.

[7] Brachinger, H. W. & Weber, M. (1997). Risk as a primitive: A survey of measures of perceived risk. *Operations-Research-Spektrum*, 19, 235-250.

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

[9] Brito, I. (2023). A portfolio stock selection model based on expected utility, entropy and variance. *Expert Systems with Applications*, 213, 118896.

[10] Brown, R. (1959). *Statistical forecasting for inventory control*. NewYork: McGraw - Hill.

[11] Chapra, S. C. & Canale, R. P. (2015). *Numerical methods for engineers* (7ª ed.). New York: McGraw-Hill Education.

[12] Chatfield, C. (2004). The Analysis of Time Series: An Introduction (6ª ed.). Chapman and Hall/CRC.

[13] Cleveland, R. B., Cleveland, W. S., McRae, J. E. & Terpenning I (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, *6*(1), 3-73.

[14] Cordeiro C. M. H. (2011). *Métodos de Reamostragem em Modelos de Previsão*. Ph.D. thesis, Universidade Técnica de Lisboa.

[15] Costa, M. & Goncalves, A. (2011). Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stochastic Environmental Research and Risk Assessment*. *25*, 151–163.

[16] Costa, M. & Monteiro, M. (2015). Discrimination of water quality monitoring sites in River Vouga using a mixed-effect state space model. *Stochastic Environmental Research and Risk Assessment*, 30, 607-619.

[17] Costa, M. & Monteiro, M. (2018). A Time Series Model Comparison for Monitoring and Forecasting Water Quality Variables. *Hidrology*, 5, 37.

[18] Cox L. A. (2008). What's Wrong with Risk Matrices?. *Risk Analysis*, *28*(2), 497–512.

[19] Dickey, D. A. & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, *74*(366), 427-431.

[20] *Fondriest Environmental Inc.* (2023). Consultado em 3 de março de 2023. Disponível em `https://www.fondriest.com/`

[21] Galton, F. (1877). Typical Laws of Heredity. *Nature*, 492-514.

[22] Gonçalves, A. M. & Alpuim, T. (2011). Water quality monitoring using cluster analysis and linear models. Environmetrics, 22, 933–945.

[23] Gooijer, J. G. & Hyndman R. J. (2006). 25 Years of Time Series Forecasting. *International journal of forecasting*, 22(3), 443–473.

[24] Holt, C. (1957). Forecasting trends and seasonal by exponentially weighted averages. *Office of Naval Research Memorandum*, 20.

[25] Hyndman, R. J. & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2ª ed.). Consultado a 2 de fevereiro de 2023. Disponível em `https://otexts.com/fpp2/`

[26] Kaas, R., Goovaerts, M., Dhaene, J. & Denuit, M. (2008). *Modern Actuarial Risk Theory* (2ª ed.). Springer.

[27] Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2019). *Loss Models: From Data to Decisions* (5ª ed.). Wiley.

[28] Knight, F. H. (1921). *Risk, Uncertainty, and Profit*. Chicago: University of Chicago Press.

[29] Kwiatkowski, D., Phillips, P. C., Schmidt, P. & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. *Journal of econometrics*, 54, 159–178.

[30] Kwiatkowski, S. (2018). *Entropy is a measure of uncertainty*. Consultado em 30/12/2022. Disponível em `https://towardsdatascience.com/entropy-is-a-measure-of-uncertainty-e2c000301c2c$#:\sim$:text=Entropy`

[31] Ljung, G. M. & Box G. E. P. (1978). On a Measure of a Lack of Fit in Time Series Models. *Biometrika*, 65(2), 297–303.

[32] MacKenzie, C. A. (2014). Summarizing Risk Using Risk Measures and Risk Indices. *Risk Analysis*, *34* (12), 2143-2162.

[33] Mano, A.P. (1989). Contribuição para o Estudo da Aplicabilidade de Índices de Qualidade da Água, Dissertação de Mestrado, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Lisboa, Portugal.

[34] Makridakis, S., Wheelwright, S., & Hyndman, R.(1998). *Forecasting: Methods and Applications* (3). NewYork: John Wiley and Sons.

[35] Markowitz H. M. (1959). Portfolio Selection: Efficient Diversification of Investment. *Yale University Press*.

[36] Milligan G. W. & Cooper, M. C. (1988). A Study of Standardization of Variables in Cluster Analysis. *Journal of Classification*, 5, 181-204.

[37] Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M. & Stork, J. (2015). Comparison of diferent Methods for Univariate Time Series imputation in R. *Research Paper*, 1-20.

[38] Pearsons, W. (1919). Indices of business conditions. *Review of Economic Statistics*, 1, 5-107.

[39] Phillips, P. C. & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75, 335–346.

[40] Pollatsek, A. & Tversky, A. (1970). A Theory of Risk. *Journal of Mathematical Psychology*, 7, 540-553.

[41] Reis, E. (1997). *Estatística Multivariada Aplicada*. Lisboa: Edições Sílabo.

[42] Said, S. E. & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71, 599–607.

[43] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6, 461–464.

[44] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

[45] Shumway R. H. & Stoffer D. S. (2015). *Time Series Analysis and Its Applications With R Examples*. (EZ ed.).

[46] SNIRH (2023) - Sistema Nacional de Informação de Recursos Hídricos. `https://snirh.apambiente.pt/`

[47] Sokal, R. & Rohlf, F. (1962). The comparison of dendograms by objective methods. *Taxon*, 11, 33-40.

[48] Tsay R. S. (2010). *Analysis of Financial Time Series* (3ª ed.). John Wiley and Sons.

[49] Winters, P (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6, 324-342.

[50] *World Wide Fund For Nature* - WWF (2023). *Dia Mundial da Água*. Consultado em 24 de janeiro de 2023. Disponível em `https://www.wwf.org.br`

[51] Worrall, F. & Burt, T. P. (1999). A univariate model of river water nitrate time series *Journal of Hydrology*, 214, 74-90.

[52] Yule, G. U. (1927). On the method of investigating periodicities in disturbed series, with special reference to Wölfer's sunspot numbers. *Philosophical Transactions of the Royal Society*, 226, 267–298.

Legislação:

Directiva n.º 2000/60/CE do Parlamento Europeu e do Conselho, de 23 de outubro de 2000.

Lei n.º 58/2005 de 29 de dezembro. Diário da República n.º 249/2005 - I Série - A

Decreto-Lei n.º 347/2007 de 19 de outubro. Diário da República n.º 202/2007 - I Série

Apêndice A

Medidas Descritivas da Análise de *Clusters* utilizando o Conjunto das Variáveis de Qualidade

Tabela A.1: Medidas descritivas das variáveis de qualidade, para os *clusters* obtidos considerando o período observado.

			Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Total	AZOTO	*Cluster* 1_T	0,010 - 0,580	0,040	0,060	0,080	0,100	0,065	81,045
		Cluster 2_T	0,010 - 0,760	0,037	0,060	0,083	0,100	0,076	91,124
		Cluster 3_T	0,010 - 1,000	0,050	0,085	0,107	0,136	0,089	83,005
	CBO_5	*Cluster* 1_T	0.010 - 5.700	2.000	2.000	2.035	2.000	0.653	32.073
		Cluster 2_T	0,110 - 7,000	2,000	2,000	2,054	2,000	0,675	32,856
		Cluster 3_T	0,170 - 10,300	2,000	2,000	2,183	2,125	0,815	37,344
	CQO	*Cluster* 1_T	1,000 - 72,400	10,000	10,000	12,630	11,150	8,696	68,875
		Cluster 2_T	0,412 - 55,000	10,000	10,000	12,617	13,400	7,308	57,920
		Cluster 3_T	1,000 - 78,000	10,000	10,000	11,930	12,000	6,587	55,233
	COND	*Cluster* 1_T	5,000 - 261,000	17,700	32,000	34,460	44,000	26,090	75,709
		Cluster 2_T	27,000 - 606,000	53,250	68,000	76,780	99,000	37,666	49,061
		Cluster 3_T	29,000 - 2660,000	49,800	73,000	130,000	221,000	119,013	91,538
	NI	*Cluster* 1_T	0,050 - 28,000	2,000	2,000	2,632	2,610	2,150	81,670
		Cluster 2_T	0,060 - 14,485	2,000	2,000	2,288	2,000	1,234	53,929
		Cluster 3_T	0,200 - 124,500	2,800	4,100	5,057	5,800	5,325	105,288
	OD	*Cluster* 1_T	3,600 - 13,500	8,500	9,500	9,375	10,248	1,238	13,203
		Cluster 2_T	5,300 - 13,000	8,395	9,200	9,252	10,177	1,187	12,827
		Cluster 3_T	4,800 - 12,800	8,200	9,100	9,077	10,000	1,225	13,497
	pH	*Cluster* 1_T	3,400 - 9,800	6,200	6,500	6,498	6,770	0,456	7,025
		Cluster 2_T	5,500 - 8,800	7,000	7,200	7,220	7,482	0,387	5,355
		Cluster 3_T	5,500 - 9,700	6,800	7,200	7,235	7,600	0,567	7,831
	TEMP	*Cluster* 1_T	-2,000 - 25,000	8,000	12,000	12,290	17,000	5,752	46,789
		Cluster 2_T	-1,000 - 30,000	8,000	13,450	14,140	20,500	6,877	48,636
		Cluster 3_T	0,700 - 29,000	10,470	14,600	15,360	21,000	5,824	37,922

Tabela A.2: Medidas descritivas das variáveis de qualidade, para os *clusters* obtidos considerando os meses secos.

			Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Secos	AZOTO	*Cluster* 1_S	0,010 - 0,580	0,040	0,072	0,089	0,109	0,071	79,734
		Cluster 2_S	0,010 - 0,550	0,040	0,070	0,084	0,103	0,066	78,786
		Cluster 3_S	0,010 - 1,000	0,060	0,092	0,114	0,140	0,093	81,576
	CBO_5	*Cluster* 1_S	0,040 - 5,700	2,000	2,000	2,131	2,000	0,794	37,267
		Cluster 2_S	0,300 - 7,000	2,000	2,000	2,141	2,000	0,752	35,103
		Cluster 3_S	0,170 - 10,300	2,000	2,000	2,364	2,500	1,010	42,701
	CQO	*Cluster* 1_S	1,000 - 72,400	10,000	10,000	13,100	14,000	9,354	71,432
		Cluster 2_S	1,000 - 39,000	10,000	10,000	12,240	14,000	5,960	48,705
		Cluster 3_S	1,000 - 78,000	10,000	10,000	12,230	13,120	6,923	56,586
	COND	*Cluster* 1_S	5,000 - 160,000	18,000	39,000	37,980	52,850	22,844	60,143
		Cluster 2_S	31,600 - 361,310	49,000	64,150	73,780	96,550	32,377	43,883
		Cluster 3_S	29,300 - 2660,000	59,750	149,500	164,140	254,000	157,006	95,655
	NI	*Cluster* 1_S	0,050 - 28,000	2,000	2,000	2,900	2,900	2,870	98,986
		Cluster 2_S	0,060 - 8,200	2,000	2,000	2,033	2,000	0,862	42,390
		Cluster 3_S	1,100 - 44,700	2,700	4,000	5,185	5,600	5,034	97,069
	OD	*Cluster* 1_S	5,400 - 11,530	7,900	8,400	8,436	9,000	0,902	10,698
		Cluster 2_S	5,300 - 10,400	7,800	8,300	8,263	8,700	0,765	9,259
		Cluster 3_S	4,800 - 11,300	7,700	8,300	8,261	8,800	0,976	11,810
	pH	*Cluster* 1_S	3,400 - 8,300	6,300	6,600	6,602	6,900	0,446	6,754
		Cluster 2_S	6,100 - 9,250	7,100	7,300	7,330	7,600	0,437	5,964
		Cluster 3_S	5,500 - 9,700	7,015	7,500	7,500	7,900	0,575	7,668
	TEMP	*Cluster* 1_S	9,000 - 25,000	15,000	18,000	17,590	20,000	3,434	19,527
		Cluster 2_S	6,000 - 30,000	19,000	21,000	20,770	23,120	3,474	16,728
		Cluster 3_S	10,000 - 29,000	19,000	21,600	20,920	23,500	3,426	16,377

Tabela A.3: Medidas descritivas das variáveis de qualidade, para os *clusters* obtidos considerando os meses húmidos.

			Amplitude	1º Quartil	Mediana	Média	3º Quartil	Desvio Padrão	Coeficiente de Variação
Húmidos	AZOTO	*Cluster* 1_H	0,010 - 0,805	0,040	0,062	0,088	0,113	0,078	88,802
		Cluster 2_H	0,010 - 0,760	0,036	0,058	0,087	0,110	0,085	97,358
		Cluster 3_H	0,010 - 0,470	0,046	0,078	0,100	0,130	0,077	76,911
	CBO_5	*Cluster* 1_H	0,010 - 5,700	2,000	2,000	2,012	2,000	0,584	29,037
		Cluster 2_H	0,110 - 5,900	2,000	2,000	2,005	2,000	0,591	29,491
		Cluster 3_H	0,440 - 8,800	2,000	2,000	2,075	2,000	0,688	33,167
	CQO	*Cluster* 1_H	1,000 - 56,400	10,000	10,000	11,860	10,530	7,343	61,932
		Cluster 2_H	0,412 - 55,000	10,000	10,000	12,906	13,000	8,036	62,268
		Cluster 3_H	2,000 - 33,000	10,000	10,000	12,150	10,400	6,683	55,013
	COND	*Cluster* 1_H	5,000 - 317,000	30,000	42,000	50,280	52,000	44,324	88,146
		Cluster 2_H	27,000 - 606,000	53,690	66,000	75,320	96,250	40,649	53,970
		Cluster 3_H	31,000 - 395,000	147,000	214,800	194,800	246,800	74,732	38,356
	NI	*Cluster* 1_H	0,100 - 32,869	2,000	2,650	3,315	3,900	2,470	74,515
		Cluster 2_H	0,100 - 14,485	2,000	2,000	2,491	2,218	1,419	56,955
		Cluster 3_H	2,000 - 124,500	4,700	6,100	6,977	7,203	8,123	116,412
	OD	*Cluster* 1_H	3,600 - 13,500	9,398	9,960	9,862	10,500	1,004	10,179
		Cluster 2_H	7,800 - 13,000	9,400	10,000	9,943	10,600	0,915	9,203
		Cluster 3_H	6,800 - 12,800	8,900	9,900	9,658	10,400	1,129	11,695
	pH	*Cluster* 1_H	5,000 - 9,800	6,400	6,600	6,627	6,900	0,464	7,002
		Cluster 2_H	5,500 - 8,700	6,900	7,200	7,151	7,400	0,366	5,119
		Cluster 3_H	6,100 - 8,300	7,200	7,500	7,431	7,700	0,391	5,265
	TEMP	*Cluster* 1_H	-2,000 - 20,514	7,500	10,000	9,879	12,000	3,927	39,750
		Cluster 2_H	-1,000 - 26,300	7,000	9,000	9,349	12,000	4,078	43,620
		Cluster 3_H	1,000 - 23,500	9,000	11,000	11,740	14,000	3,622	30,839

Apêndice B

Representações Gráficas das Séries Temporais dos *Clusters*

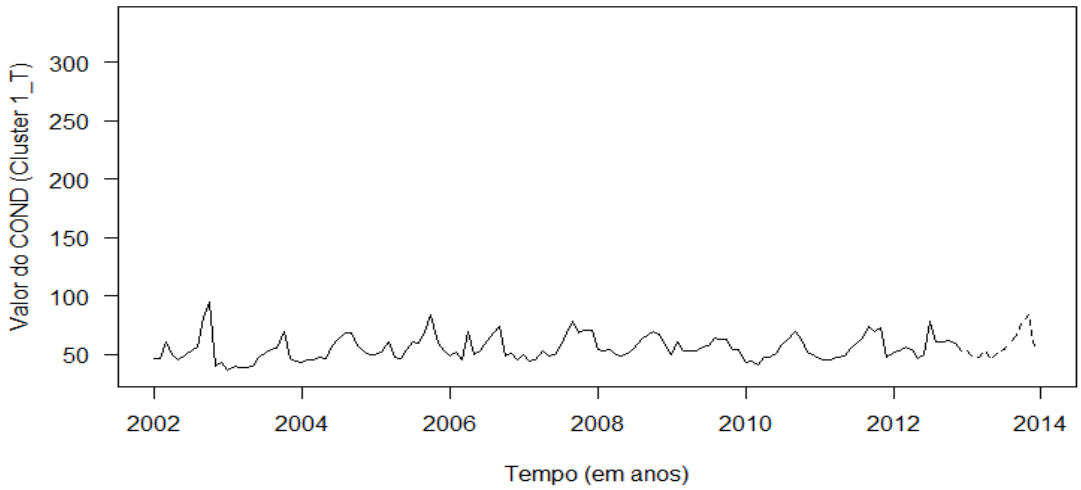


Figura B.1: Série temporal de COND do *Cluster* 1_T.

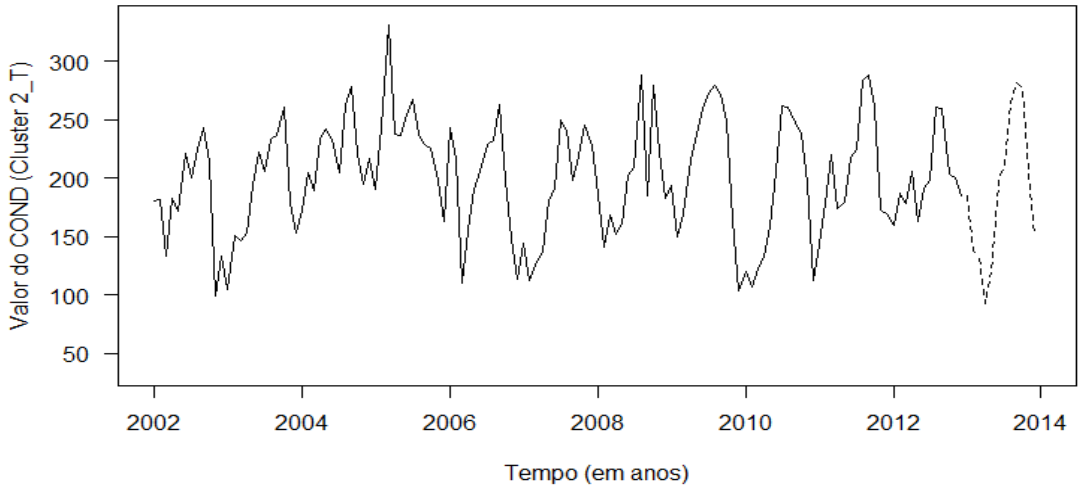


Figura B.2: Série temporal de COND do *Cluster* 2_T.

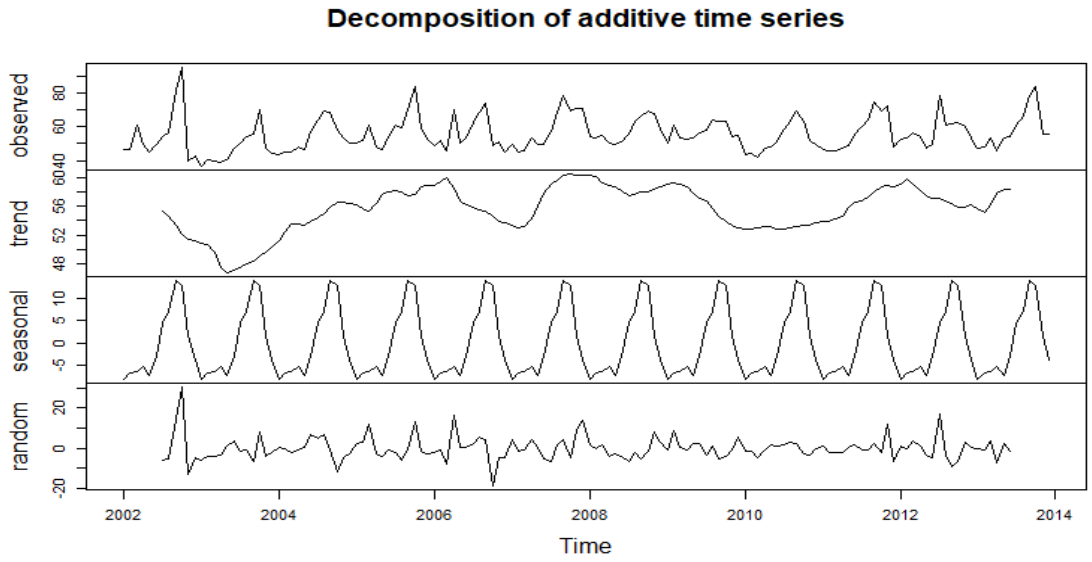


Figura B.3: Decomposição aditiva da série de COND do *Cluster* 1_T.

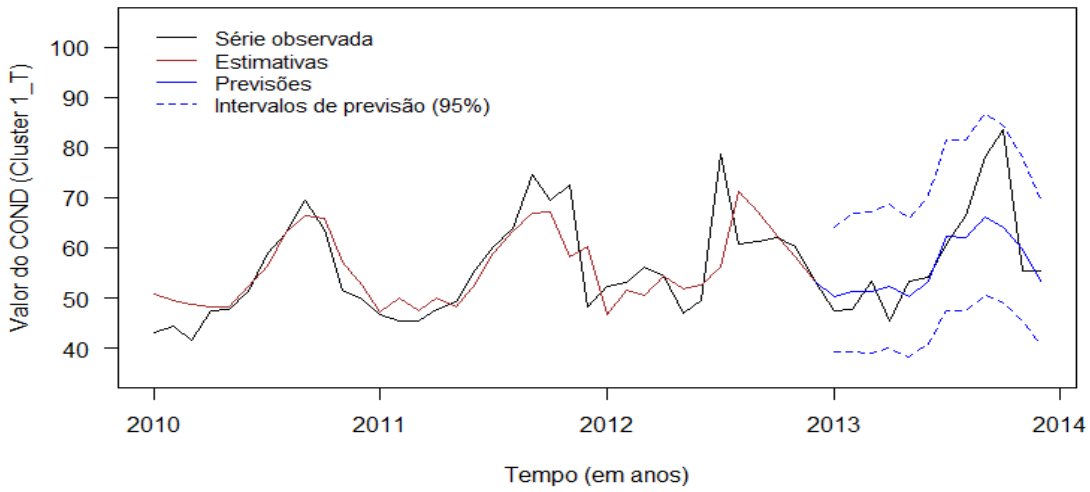


Figura B.4: Estimativas e previsões pontuais e intervalares obtidas para a COND do *Cluster* 1_T através do modelo SARIMA.

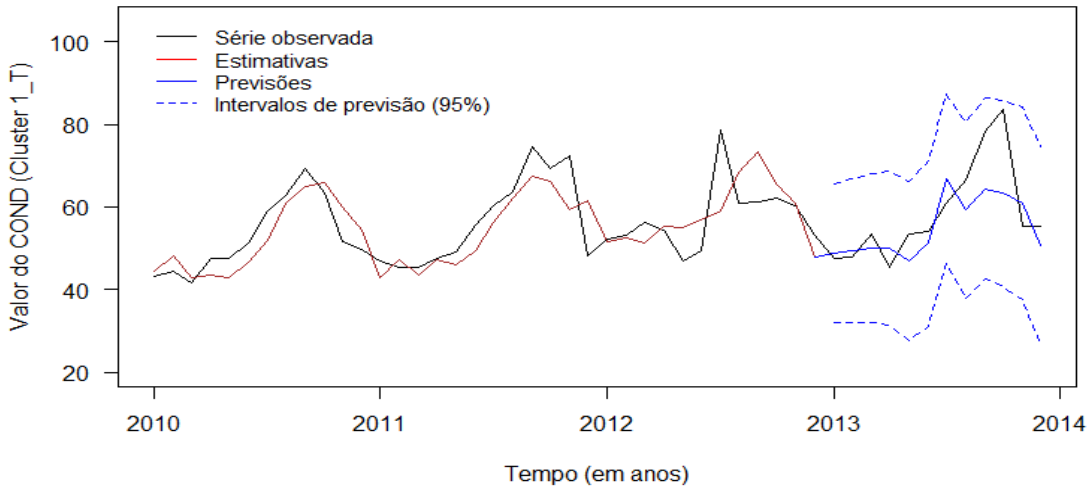


Figura B.5: Estimativas e previsões pontuais e intervalares obtidas para a COND do *Cluster* 1_T através do método de Holt-Winters.

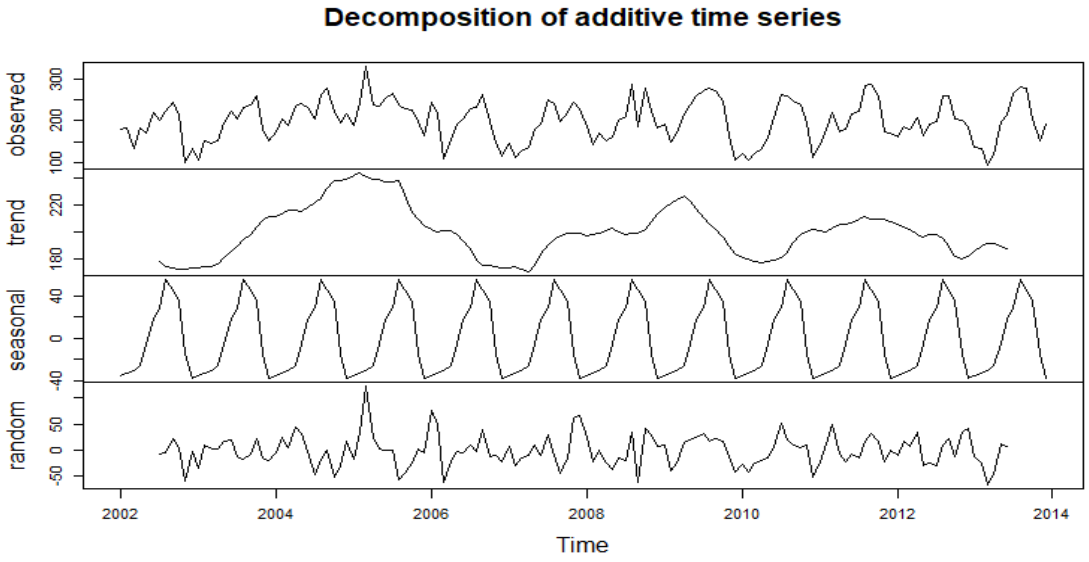


Figura B.6: Decomposição aditiva da série de COND do *Cluster* 2_T.

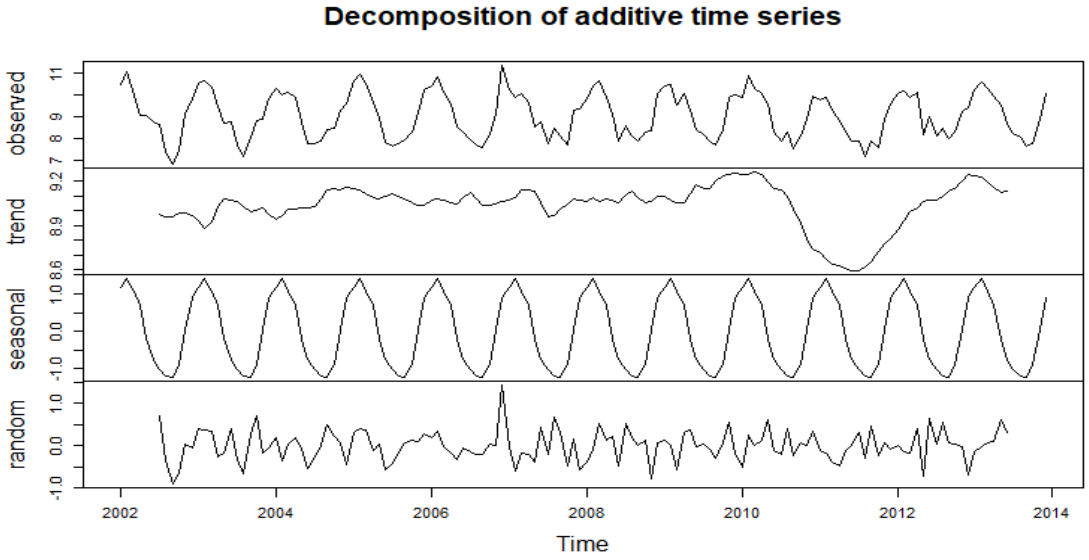


Figura B.7: Decomposição aditiva da série de OD do *Cluster* 1_T.

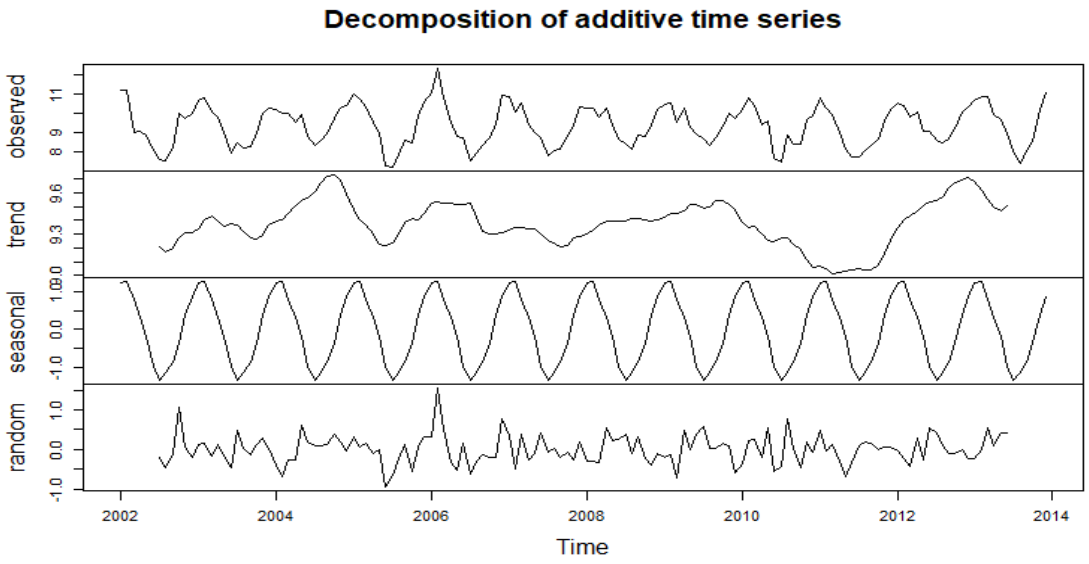


Figura B.8: Decomposição aditiva da série de OD do *Cluster* 2_T.

Apêndice C

Previsões Pontuais e Intervalares do Conjunto de Métodos Aplicados

Tabela C.1: Valores observados e previsões pontuais e intervalares resultantes dos processos de modelação SARIMA e Holt-Winters para o *Cluster* 1_T de COND.

	Modelo SARIMA			Método Holt-Winters		
Valor observado	Valor previsto	Limite inferior	Limite superior	Valor previsto	Limite inferior	Limite superior
47,429	50,228	39,373	64,075	48,793	32,002	65,583
47,786	51,203	39,261	66,777	49,510	32,087	66,933
53,357	51,204	39,096	67,062	49,944	31,886	68,002
45,528	52,419	39,985	68,719	50,057	31,361	68,753
53,357	50,143	38,239	65,755	46,933	27,597	66,270
54,000	53,458	40,761	70,108	51,166	31,185	71,146
60,778	62,172	47,403	81,542	66,820	46,193	87,447
66,500	62,159	47,391	81,528	59,213	37,936	80,490
78,286	66,189	50,462	86,818	64,386	42,455	86,317
83,500	64,210	48,951	84,225	63,289	40,701	85,877
55,429	59,653	45,476	78,251	60,855	37,606	84,103
55,286	53,368	40,683	70,010	50,536	26,623	74,449

Tabela C.2: Valores observados e previsões pontuais e intervalares resultantes dos processos de modelação SARIMA e Holt-Winters para o *Cluster* 2_T de COND.

	Modelo SARIMA			Método Holt-Winters		
Valor observado	Valor previsto	Limite inferior	Limite superior	Valor previsto	Limite inferior	Limite superior
137,500	182,139	127,641	259,905	172,659	100,247	245,071
132,750	170,485	115,014	252,711	179,706	101,334	258,078
93,250	164,122	109,772	245,381	190,073	106,163	273,982
119,250	176,444	118,014	263,804	194,095	104,991	283,199
197,500	189,989	127,073	284,054	200,537	106,525	294,549
212,500	217,435	145,431	325,090	238,709	140,033	337,385
264,500	227,611	152,237	340,304	257,627	154,498	360,756
281,000	254,046	169,917	379,827	292,005	184,607	399,402
277,500	243,204	162,666	363,618	275,572	164,069	387,075
203,750	232,224	155,322	347,201	254,308	138,845	369,770
152,250	179,456	120,036	268,291	209,373	90,082	328,664
192,000	156,038	104,397	233,223	178.433	55,433	301,432

Tabela C.3: Valores observados e previsões pontuais e intervalares resultantes dos processos de modelação SARIMA e Holt-Winters para o *Cluster* 1_T de OD.

	Modelo SARIMA			Método Holt-Winters		
Valor observado	Valor previsto	Limite inferior	Limite superior	Valor previsto	Limite inferior	Limite superior
10,265	10,045	9,126	11,058	10,135	9,162	11,107
10,611	10,377	9,394	11,463	10,414	9,437	11,391
10,311	10,019	9,067	11,070	10,001	9,019	10,983
9,944	9,630	8,715	10,641	9,860	8,874	10,846
9,522	8,800	7,964	9,723	8,775	7,784	9,766
8,722	8,332	7,541	9,207	8,565	7,569	9,561
8,200	8,021	7,259	8,863	8,153	7,153	9,153
8,056	7,865	7,118	8,690	8,113	7,108	9,117
7,611	7,821	7,078	8,642	7,941	6,931	8,950
7,767	8,161	7,385	9,017	8,191	7,177	9,205
8,878	9,080	8,217	10,033	9,206	8,187	10,224
10,044	9,884	8,945	10,921	9,851	8,828	10,874

Tabela C.4: Valores observados e previsões pontuais e intervalares resultantes dos processos de modelação SARIMA e Holt-Winters para o *Cluster* 2_T de OD.

	Modelo SARIMA			Método Holt-Winters		
Valor observado	Valor previsto	Limite inferior	Limite superior	Valor previsto	Limite inferior	Limite superior
10,678	10,503	9,395	11,740	10,949	9,934	11,964
10,856	10,213	9,022	11,562	10,958	9,915	12,000
10,889	9,757	8,595	11,077	10,448	9,378	11,517
9,917	9,652	8,496	10,965	10,165	9,069	11,262
9,711	8,915	7,846	10,129	9,489	8,366	10,611
8,956	8,822	7,764	10,024	8,860	7,712	10,008
8,067	8,590	7,559	9,760	8,508	7,334	9,681
7,444	8,598	7,567	9,770	8,677	7,478	9,875
8,111	8,756	7,706	9,950	8,872	7,648	10,096
8,544	9,123	8,029	10,366	9,349	8,101	10,598
10,200	9,773	8,601	11,105	10,216	8,943	11,488
11,100	10,038	8,834	11,406	10,629	9,333	11,926