



Análise de Séries Temporais de Dados Meteorológicos da Cidade do Porto

Ana Catarina Pinheiro Monteiro

Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos
Departamento de Ciência de Computadores
2021

Orientador

Sérgio Crisóstomo, Professor Auxiliar, Faculdade de Ciências da Universidade do Porto

Coorientador

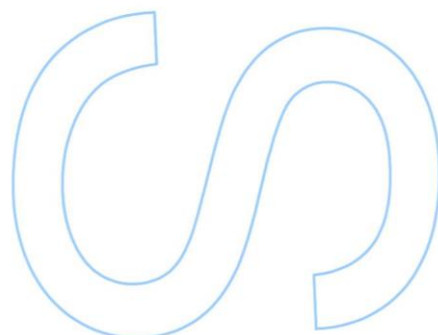
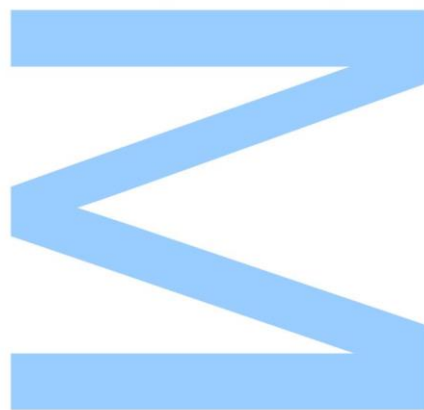
Rita Ribeiro, Professor Auxiliar, Faculdade de Ciências da Universidade do Porto



Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ____/____/____



Resumo

O Observatório Meteorológico da Serra do Pilar, atual Instituto Geofísico da Universidade do Porto (IGUP), tem como um dos seus objetivos, o registo de dados meteorológicos, informação fundamental para a investigação climatológica. Estas observações iniciaram-se em 1887, tendo sido realizadas de forma regular até 2007. O estudo destes dados afigura-se como fundamental para a caracterização e estudo da evolução do clima da cidade do Porto, nomeadamente, através da análise exploratória de registos horários e do estudo da homogeneidade dessas séries temporais.

Esta dissertação tem como objectivo, a análise exploratória das séries temporais de um sub-conjunto das variáveis meteorológicas registadas pelo observatório meteorológico do IGUP. As séries temporais referem-se à segunda parte do século XX e primeira década do século XXI.

Considerando os dados fornecidos pelo IGUP, foi realizada, numa primeira fase, uma análise exploratória, essencial para a compreensão dos variáveis meteorológicas a estudar e aferir a qualidade do registo das mesmas. Foram identificados, valores em falta, erros de consistência e *outliers* globais e contextuais. De seguida, procedeu-se ao estudo das respetivas séries temporais, tendo em vista a análise do seu comportamento no período em causa, com vista à caracterização do clima do Porto. Depois de analisar as variáveis da temperatura, precipitação, pressão atmosférica e velocidade do vento, no período de tempo dos registos, verificamos que a temperatura apresentou valores mais baixos nos anos iniciais, e evidenciou uma tendência crescente nos últimos anos da série. Podemos também destacar o valor elevado da precipitação em 2001, cujo valor se evidenciou claramente face aos restantes anos da série. Finalmente, as séries temporais foram submetidas a testes de homogeneidade. A temperatura foi a única variável meteorológica que registou a presença de não-homogeneidades, classificando a sua série temporal de classe 3 ou *suspect*.

Abstract

The *Observatório Meteorológico da Serra do Pilar*, currently *Instituto Geofísico da Universidade do Porto (IGUP)*, has as one of its objectives, the registration of meteorological data, which is fundamental for climatological research. These observations began in 1887, having been carried out regularly until 2007. The study of these data is essential for the characterization and study of the evolution of the climate of the city of Porto, namely through the exploratory analysis of hourly records and the study of the homogeneity of these time series. However, these data still lack essential analyses for a better understanding of the evolution of the climate, namely the exploratory analysis of hourly records and the study of the homogeneity of these time series.

This dissertation aims at the exploratory analysis of the time series of a subset of meteorological variables recorded by the IGUP meteorological observatory. The time series refer to the second part of the 20th century and the first decade of the 21st century.

Considering the data provided by IGUP, an exploratory analysis was carried out in a first stage, essential for the understanding of the meteorological variables to be studied and to assess the quality of their record. Missing values, consistency errors and global and contextual outliers were identified. Next, the study of the respective time series was carried out, in order to the analysis of their behavior in the period in question, in order to characterize the climate of Oporto. The aim was to establish possible correlations between them and how they might influence future research. After analyzing the variables of temperature, precipitation, atmospheric pressure and wind speed, in the time period of the records, we verified that the temperature presented lower values in the initial years, and showed an increasing trend in the last years of the series. We can also highlight the high value of precipitation in 2001, whose value clearly stood out compared to the other years of the series. Finally, the time series were submitted to homogeneity tests. The temperature was the only meteorological variable that registered the presence of non-homogeneities, classifying its time series as class 3 or *suspect*.

Aos meus orientadores por todo o conhecimento transmitido e tempo dispendido,
aos meus pais, a quem tudo devo, pelo apoio e suporte incondicional, à minha
irmã por ser a minha maior inspiração e motivação e aos meus avós que sempre
acreditaram em mim. Os meus sinceros agradecimentos.

Conteúdo

Resumo	i
Abstract	iii
Conteúdo	viii
Lista de Tabelas	x
Lista de Figuras	xv
1 Introdução	1
1.1 Motivação	1
1.2 Objetivos	2
1.3 Organização	2
2 Contextualização	3
2.1 Variáveis meteorológicas	3
2.2 Dados climáticos	7
2.3 Análise exploratória de dados	7
2.3.1 Gráficos para análise univariada	10
2.3.2 Gráficos para análise de variáveis climáticas	13
2.4 Séries temporais	17
2.4.1 Medidas de dependência	19
2.4.2 Testes de correlação de variáveis	20

2.4.3	Estacionaridade	21
2.4.4	Modelos	23
2.4.5	<i>Moving averages</i>	24
2.5	Homogeneização	25
2.5.1	Métodos para detetar e corrigir <i>breaks</i> não climáticos	27
2.5.2	Métodos de homogeneização	27
2.5.3	Processo de homogeneização	33
2.5.4	Classificação e avaliação qualitativa dos testes	35
2.6	Ferramentas	36
3	Análise Exploratória de Dados Meteorológicos	37
3.1	Caracterização do conjunto de dados	37
3.2	Pré-processamento de dados	38
3.2.1	Valores desconhecidos	39
3.2.2	Erros de consistência	40
3.2.3	<i>Outliers</i> globais e contextuais	40
3.3	Análise de dados univariada	42
3.3.1	Temperatura	43
3.3.2	Precipitação	52
3.3.3	Pressão atmosférica	58
3.3.4	Velocidade do vento	62
3.4	Análise de dados multivariada	67
3.5	Discussão	70
4	Testes de Homogeneidade	73
4.1	Aplicação de testes de homogeneidade	73
4.1.1	Temperatura	74
4.1.2	Precipitação	76
4.1.3	Pressão atmosférica	78

4.1.4	Velocidade do vento	79
4.2	Discussão	80
5	Conclusões	83
5.1	Contribuições	83
5.2	Limitações do estudo e trabalho futuro	84
A	Resultados da análise da qualidade dos dados	85
A.1	Erros de consistência	86
A.1.1	Temperatura	86
A.1.2	Precipitação	87
A.2	Outliers	88
A.2.1	<i>Outliers</i> globais	88
A.2.2	Temperatura	88
A.2.3	Pressão atmosférica	89
A.2.4	Velocidade do vento	91
A.3	Ausência de registos	91
A.3.1	Temperatura	91
A.3.2	Pressão atmosférica	91
A.3.3	Precipitação	92
A.3.4	Velocidade do vento	92
B	Resultados da análise exploratória de dados	93
B.1	Análise univariada	93
B.1.1	Temperatura	93
B.1.2	Precipitação	97
B.1.3	Pressão atmosférica	99
B.1.4	Velocidade do Vento	103
B.2	Análise multivariada	109

C	Resultados da aplicação de testes de homogeneização	111
C.1	Tabelas de valores críticos para testes de homogeneização	111
C.2	Resultados dos testes de homogeneização	112
C.2.1	Precipitação	112
C.2.2	Pressão atmosférica	113
C.2.3	Velocidade do vento	114
D	Ferramentas	115
	Bibliografia	121

Lista de Tabelas

2.1	Índices de temperatura utilizados em [9] e [48].	8
2.2	Índices de precipitação utilizados em [56] e [28].	8
2.3	Técnicas EDA sugeridas face ao tipo de dados [31].	9
2.4	Técnicas EDA sugeridas face ao objetivo pretendido [31].	10
2.5	Sumário das principais características de alguns métodos de homogeneização, classificados por tipo de abordagem (paramétricos, não-paramétricos, baseado em probabilidade ou de regressão) [49].	34
3.1	Variáveis e respetivos períodos de registo de observações.	38
3.2	Tamanho da série temporal antes e depois de corrigidos os <i>missing values</i>	39
3.3	Erros de consistência na velocidade do vento.	40
3.4	Limites dos <i>outliers</i> globais para cada variável.	41
3.5	<i>Outliers</i> globais encontrados para cada variável.	41
3.6	Valores máximos e mínimos da temperatura média anual, diurna e noturna.	46
3.7	Índices de temperatura diária mínima, por década. (TN_n - Noite mais fria; TN_x - Noite mais quente; FD - Dias de frio intenso; TN - Noites tropicais; CWDI - Duração da onda de frio; CSDI - duração de períodos frios)	49
3.8	Índices de temperatura diária máxima, por década. (TX_n - Dia mais frio; TX_x - Dia mais quente; SU - Dias de verão; TX35 - Dias muito quentes; HWDI - Duração da onda de calor; WSDI - Duração de períodos quentes)	50
3.9	Índices de precipitação e respetivos valores. (DD - Dias secos; R10 - Dias de forte precipitação; R20 - Dias de precipitação muito forte; R25 - Dias de precipitação extremamente forte)	57

4.1	Comparação dos resultados obtidos para as séries anuais de temperatura, mínimas (T_{min}), máximas (T_{max}), intervalo (DTR) e médias (T_{mean}).	76
4.2	Comparação dos resultados obtidos para as séries anuais de pressão atmosférica mínimas (P_{min}), máximas (P_{max}), intervalo ($P_{intervalo}$) e médias (P_{mean}).	79
4.3	Comparação dos resultados obtidos para as séries anuais de velocidade do vento mínimas (V_{min}), máximas (V_{max}), intervalo ($V_{intervalo}$) e médias (V_{mean}).	80
A.1	Erros de consistência encontrados ao importar os dados da Temperatura.	86
A.2	Erros de consistência encontrados ao importar os dados da Precipitação.	87
A.3	<i>Outliers</i> globais	88
A.4	<i>Outliers</i> encontrados por mês do ano, e respetivo valor sugerido, para valores horários de temperatura (1960-2007).	88
A.5	<i>Outliers</i> encontrados por mês do ano, e respetivo valor sugerido, para valores horários de pressão atmosférica (1999-2007).	89
A.6	<i>Outliers</i> encontrados por mês do ano, e respetivo valor sugerido, para valores horários de velocidade do vento (1994-2007).	91
A.7	Ausência de registos de temperatura identificados através do heatmap.	91
A.8	Ausência de registos de pressão atmosférica identificados através do heatmap.	91
A.10	Avárias registadas nos instrumentos de medição da precipitação.	92
A.11	Ausência de registos de velocidade do vento identificados através do heatmap.	92
B.1	Precipitação acumulada (mm) por estação do ano e total (1998-2005).	97
C.1	Valores críticos de 5% e 1% para os testes de homogeneização Buishand Range Test, Pettit Test e Standard Normal Homogeneity Test (SNHT), para conjuntos de dados de diferentes tamanhos [21].	111
C.2	Valores críticos de 1% e 5% para o teste de homogeneização Von Neumann [21].	111
D.1	<i>Packages</i> e funções da linguagem R utilizadas e respetivas descrições.	117
D.2	<i>Packages</i> e funções da linguagem R utilizadas em séries temporais, e respetivas descrições.	118
D.3	<i>Packages</i> e funções da linguagem R utilizadas na homogeneização de séries temporais e respetiva descrição.	119

Lista de Figuras

2.1	Perfil vertical da pressão atmosférica em função da altitude [50].	5
2.2	Relação entre a pressão de vapor saturante e a temperatura [50].	6
2.3	Exemplo de um gráfico de linhas [27]	10
2.4	Exemplo de um gráfico de barras [6]	11
2.5	Exemplo de um histograma [6]	12
2.6	Exemplo de um <i>boxplot</i> [20].	12
2.7	<i>Warming Stripes</i> referentes às temperaturas médias anuais em Portugal de 1850 a 2020 [18].	13
2.8	Exemplo de um climograma [7].	14
2.9	Exemplo de um gráfico de anomalias com desvios percentuais positivos e negativos em relação à média mensal de precipitação [51].	15
2.10	Exemplo de um gráfico de anomalias contextuais de precipitação por mês durante o ano de 2018 em Santiago de Compostela usando um período de 1971-2010 como referência [51].	16
2.11	Exemplo de um <i>Quadrant Type Scatter Plot</i> para a identificação de anomalias temperatura e precipitação durante o inverno no Sul de Tenerife no período de 1981 a 2020 [52].	17
2.12	Exemplo de um <i>heatmap</i> ilustrando as temperaturas máximas diárias por mês [3].	18
2.13	Exemplo de uma matriz de correlação de <i>Pearson</i> [4].	21
2.14	Exemplo de gráfico de <i>cross correlation</i> com correlação significativa [1].	22
2.15	Exemplo de uma série temporal estacionária [59].	22
2.16	Exemplo de uma série temporal não estacionária, devido à tendência [59].	22
2.17	Diferentes <i>moving averages</i> aplicadas à mesma série temporal [27].	25

2.18	Exemplo de decomposição de uma série temporal com modelo aditivo [27].	25
2.19	Processo de homogeneização. No topo da figura encontram-se os passos iniciais e em baixo o processo de correção de não-homogeneidades não climáticas [39].	35
3.1	Representação da temperatura onde é possível identificar outliers.	42
3.2	Boxplot da temperatura por mês para deteção de <i>outliers</i> contextuais	42
3.3	Valores da Temperatura (1960-2007).	44
3.4	Evolução dos valores médios anuais de temperatura mínima, média e máxima (1960-2007).	45
3.5	<i>Warming stripes</i> das temperaturas médias máximas e mínimas anuais (°C) (1960-2007).	46
3.6	Evolução das temperaturas médias mensais por mês, e linha de tendência (1960-2007).	47
3.7	Distribuição da temperatura máxima e mínima diária, por mês.	48
3.8	Evolução das temperaturas médias anuais diurnas e noturnas, e linha de tendência. (1960-2007).	49
3.9	Evolução das temperaturas médias anuais por estação do ano, e linha de tendência (1960-2007).	49
3.10	Temperatura hora a hora (°C) com sazonalidade diária (frequência = 24).	52
3.11	Gráfico da precipitação, registos de hora a hora (1997-2006).	53
3.12	Total de precipitação por ano (1998-2005).	53
3.13	Heatmap da precipitação mensal (<i>mm</i>).	54
3.14	Total de precipitação por mês (1998-2005).	55
3.15	Boxplot das precipitações mensais (1998-2005).	55
3.16	Evolução da precipitação anual acumulada nas estações do ano e seu <i>boxplot</i> (1998-2005).	56
3.17	Anomalias de precipitação dos anos 2002 e 2005.	57
3.18	Decomposição da série temporal da precipitação: médias diárias (<i>mm</i>) com sazonalidade anual (frequência = 365).	58
3.19	Série temporal da pressão atmosférica (<i>mmHg</i>) e linha da tendência (1999-2007).	59
3.20	<i>Boxplot</i> da pressão atmosférica onde é possível identificar outliers extremos (1999-2007).	59

3.21	Pressão atmosférica diária diurna e noturna com <i>moving average</i> de ordem 365.	60
3.22	Gráfico das médias mensais da pressão atmosférica ao longo dos anos (1999-2007).	61
3.23	Distribuição da pressão atmosférica média diária, por mês.	62
3.24	Decomposição da série temporal da pressão atmosférica: médias diárias (<i>mmHg</i>) com sazonalidade anual (frequência = 365).	62
3.25	Distribuição de <i>missing values</i> na velocidade do vento (1994-2007).	63
3.26	<i>Boxplot</i> da velocidade do vento onde é possível identificar <i>outliers</i>	64
3.27	médias anuais da velocidade do vento (Km/h)	65
3.28	Médias mensais da velocidade do vento por mês.	65
3.29	Distribuição da velocidade do vento média diária, por mês.	66
3.30	Diferentes <i>moving averages</i> aplicadas à série temporal da velocidade do vento diária (2000).	66
3.31	Decomposição da série temporal da velocidade do vento (1994-2007): médias diárias (<i>km/h</i>) com sazonalidade anual (frequência = 365).	67
3.32	Matriz de correlação entre as variáveis meteorológicas (2000-2005).	68
3.33	Gráfico de dispersão entre a precipitação e a velocidade do vento, linha de tendência e respectivo valor R e <i>p-value</i> (2000-2005).	69
3.34	Gráfico de dispersão entre a pressão atmosférica e a velocidade do vento, linha de tendência e respectivo valor R e <i>p-value</i> (2000-2005).	69
3.35	Gráfico de dispersão entre a pressão atmosférica e a precipitação , e linha de tendência e respectivo valor R e <i>p-value</i> (2000-2005).	70
3.36	Climograma (2000-2006).	70
3.37	<i>Quadrant Type Scatter Plot</i> das anomalias de temperatura e precipitação, no inverno (2000-2006).	71
3.38	<i>Normalized cross-correlation</i> entre a precipitação e a velocidade do vento.	71
4.1	Gráficos dos testes de homogeneidade às séries temporais de temperatura anual, valores médios (Tmean), máximos (Tmax), mínimos (Tmin), intervalo (DTR) e níveis de confiança de 95 e 99%.	75
4.1	Gráficos dos testes de homogeneidade às séries temporais de temperatura anual, valores médios (Tmean), máximos (Tmax), mínimos (Tmin), intervalo (DTR) e níveis de confiança de 95 e 99%.	76

4.2	Gráficos de ACF e PACF da série temporal da precipitação anual (1998-2005).	77
4.3	<i>Standard Normal Homogeneity Test</i> aplicados às series anuais de pressão atmosférica mínimas (P_{min}), máximas (P_{max}), intervalo (P_{int}) e médias (P_{mean}) e níveis de confiança de 95 e 99%.	78
4.4	<i>Buishand range test</i> aplicado às series anuais de velocidade do vento mínimas (V_{min}), máximas (V_{max}), intervalo (V_{int}), médias (V_{mean}) e níveis de confiança de 95 e 99%.	80
A.1	<i>Printscreen</i> de folha de Excel de junho de 2007 - velocidade do vento.	85
B.1	<i>Warming stripes</i> das temperaturas médias anuais ($^{\circ}\text{C}$) (1960-2007).	93
B.2	<i>Warming stripes</i> das temperaturas médias anuais por estação do ano ($^{\circ}\text{C}$) (1960-2007).	94
B.3	Heatmap das temperaturas de hora a hora de 1992 e 2001($^{\circ}\text{C}$).	95
B.4	Decomposição da série temporal da temperatura (1960-2007): hora a hora com sazonalidade anual (frequência = 8766).	96
B.5	Decomposição da série temporal da temperatura (1960-2007): Médias diárias com sazonalidade anual (frequência = 365).	96
B.6	Decomposição da série temporal da precipitação (1998-2005): hora a hora (mm) com sazonalidade anual (frequência = 8766).	97
B.7	Decomposição da série temporal da precipitação (1998-2005): médias mensais de precipitação (mm) (frequência = 12).	98
B.8	Distribuição de <i>missing values</i> na pressão atmosférica e submetidos a correção (1999-2007).	99
B.9	Distribuição da pressão atmosférica máxima e mínima diária, por mês.	100
B.10	Heatmap da Pressão Atmosférica horária dos anos 2005 e 2007	101
B.11	Decomposição das série temporal da pressão atmosférica: hora a hora (mmHg) com sazonalidade diária (frequência = 24).	102
B.12	Decomposição das série temporal da pressão atmosférica: hora a hora (mmHg) com sazonalidade anual (frequência = 8766).	102
B.13	Distribuição de <i>missing values</i> na velocidade do vento antes de qualquer correção (1994-2007).	103
B.14	Médias mensais da velocidade do vento ao longo dos anos e linha de tendência.	103

B.15	Série temporal da velocidade do vento horária e linha da tendência (1994-2007).	104
B.16	Distribuição da velocidade do vento máxima e mínima diária, por mês.	105
B.17	Heatmap das velocidade do vento de hora a hora.	106
B.17	Heatmap das velocidade do vento de hora a hora.	107
B.18	Decomposição da série temporal da velocidade do vento (1994-2007): hora a hora (km/h) com sazonalidade diária (frequência = 24).	107
B.19	Decomposição da série temporal da velocidade do vento (1994-2007): hora a hora (km/h) com sazonalidade anual (frequência = 8766).	108
B.20	<i>Quadrant Type Scatter Plot</i> das anomalias de temperatura e precipitação, no outono (2000-2006).	109
B.21	<i>Quadrant Type Scatter Plot</i> das anomalias de temperatura e precipitação, na primavera (2000-2006).	109
B.22	<i>Quadrant Type Scatter Plot</i> das anomalias de temperatura e precipitação, no verão (2000-2006).	110
C.1	Gráficos de ACF e PAFC da série temporal da precipitação mensal (1998-2005).	112
C.2	<i>Pettitt test</i> aplicados às series anuais de pressão atmosférica mínimas (P_{min}), máximas (P_{max}), intervalo (P_{int}), médias (P_{mean}) e níveis de confiança de 95 e 99%.113	113
C.3	<i>Buishand range test</i> aplicados às series anuais de pressão atmosférica mínimas (P_{min}), máximas (P_{max}), intervalo (P_{int}), médias (P_{mean}) e níveis de confiança de 95 e 99%.	113
C.4	<i>Pettitt test</i> aplicado às series anuais de velocidade do vento mínimas (V_{min}), máximas (V_{max}), intervalo (V_{int}), médias (V_{mean}) e níveis de confiança de 95 e 99%.114	114
C.5	<i>Standard Normal Homogeneity Test</i> aplicado às series anuais de velocidade do vento mínimas (V_{min}), máximas (V_{max}), intervalo (V_{int}), médias (V_{mean}) e níveis de confiança de 95 e 99%.	114

Capítulo 1

Introdução

Ao longo dos últimos anos temos assistido a uma preocupação crescente com o tema das alterações climáticas. Muitos dos estudos realizados neste âmbito partem da análise de uma vasta coleção de dados meteorológicos. O estudo de parâmetros como a temperatura do ar, a precipitação, ou a pressão atmosférica, têm-se revelado importantes para o estudo da evolução do clima e impacto em setores primários como a agricultura, a pesca ou a florestação.

Dada a atualidade do tema, esta dissertação tem como objetivo a análise de séries temporais de dados meteorológicos da cidade do Porto, fornecidos pelo Observatório Meteorológico da Serra do Pilar, atual Instituto Geofísico da Universidade do Porto (IGUP) e que dizem respeito ao período de tempo compreendido entre 1960 e 2007 (47 anos).

1.1 Motivação

O IGUP, também conhecido como Observatório Meteorológico da Serra do Pilar, foi fundado em 1883, sendo, por isso, um dos primeiros postos meteorológicos criados em Portugal. Faz parte da Rede Meteorológica Nacional e tem como principal função a obtenção e registo de dados meteorológicos tendo em vista a previsão do estado do tempo, tão necessária a variados setores da economia, mas também servindo de base para a realização de estudos climáticos.

Inicialmente a recolha destes dados começou por ser manual, e enviada para o Instituto Meteorológico de Lisboa várias vezes ao dia. Atualmente, encontra-se instalada no IGUP uma estação meteorológica automática que regista estas observações de forma cada vez mais completa.

O estudo aqui apresentado pretende, a partir dos dados fornecidos pelo IGUP, fazer a análise de séries temporais, perceber a evolução das variáveis meteorológicas ao longo do respetivo período de tempo, assim como estudar a sua homogeneidade e classificar a série com base nos resultados obtidos. Atendendo à quantidade de dados recolhidos pelo IGUP, e dada a importância do seu estudo, é de extrema relevância a sua validação e correção no que concerne a possíveis erros no processo de medida ou imprecisões no seu registo. A análise das séries temporais permite

estudar possíveis alterações do clima no período de tempo considerado. Nesse seguimento, é também crucial entender se as séries são homogêneas ou não (e de que modo diferentes testes podem verificar essas características), para determinar possíveis erros sistemáticos, derivados, por exemplo, da troca de instrumentos de medida nos períodos de tempo em análise.

1.2 Objetivos

O principal objetivo desta dissertação consiste no tratamento e análise exploratória dos dados meteorológicos fornecidos pelo IGUP, no período de tempo registrado por cada variável. Serão aplicados diferentes testes de homogeneidade, capazes de identificar se as séries em questão são ou não homogêneas e, caso o resultado seja negativo, onde se situam no tempo as não-homogeneidades. Face ao objetivo estabelecido, este estudo foca-se em quatro pontos principais:

- pré-processamento do conjunto de dados meteorológicos registrados no IGUP, identificando valores em falta, ou com possíveis erros;
- análise exploratória dos dados que permita a sua visualização, de forma fácil e intuitiva, e análise de possíveis correlações entre as variáveis;
- análise e estudo das séries temporais, procurando encontrar resultados relevantes ao entendimento da evolução das variáveis meteorológicas ao longo do tempo;
- aplicação de testes de homogeneidade, capazes de identificar não-homogeneidades na série temporal, procurando perceber se a mesma é ou não homogênea.

1.3 Organização

Posto o capítulo introdutório, a dissertação encontra-se organizada pelos seguintes capítulos. No Capítulo 2 é feita uma breve introdução às variáveis meteorológicas analisadas neste trabalho, assim como aos índices climáticos derivados das séries de temperatura e precipitação. De seguida, é dada uma introdução de conceitos fundamentais na análise exploratória de dados e no estudo de séries temporais. São introduzidos os testes de homogeneidade utilizados, e contextualizado o processo de correção de séries não homogêneas. Finalmente, são ainda identificadas as ferramentas utilizadas neste estudo. No Capítulo 3 é caracterizado o conjunto de dados, e descrito o pré-processamento dos mesmos. De seguida, é realizada a análise exploratória das séries temporais das diferentes variáveis meteorológicas. É também realizada uma análise multivariada de correlação entre elas. No Capítulo 4 são aplicados diferentes testes de homogeneidade nas séries temporais em estudo e discutidos os resultados obtidos pelos mesmos. Por último, no Capítulo 5 são apresentadas as principais contribuições desta dissertação e limitações encontradas ao longo da elaboração da mesma, assim como sugestões para estudos futuros com base no estudo realizado.

Capítulo 2

Contextualização

Neste capítulo apresentamos as variáveis meteorológicas analisadas neste trabalho. Abordamos algumas noções referentes à análise exploratória de dados assim como diferentes técnicas utilizadas na mesma. Serão ainda mencionados alguns conceitos característicos do estudo e análise de séries temporais. Expor-se-ão alguns testes de homogeneidade e possível método de correção de séries temporais. Finalmente, são identificados e descritos os pacotes e funções mais relevantes, utilizados neste projeto.

2.1 Variáveis meteorológicas

O clima refere-se a um conjunto de condições médias meteorológicas que variam tanto no espaço como no tempo, ou, mais especificamente, a uma descrição estatística, de modo quantitativo, das características do estado do tempo, durante uma certa duração. Nesse sentido, recorreremos a observações meteorológicas, que registam estes atributos em cada local, com o objectivo de caracterizar o clima [16].

São então os meteorologistas quem se ocupa deste estudo, denominado por meteorologia, atendendo às propriedades físicas da atmosfera, focando a sua atenção nas variáveis mais impactantes nas atividades humanas e com maior relevância no estudo da evolução do clima, prevendo assim estados futuros [33].

As medições de variáveis meteorológicas são realizadas utilizando instrumentos específicos. Algumas dessas variáveis são a temperatura do ar, a precipitação, a pressão atmosférica, a humidade relativa, a direção e velocidade do vento, os quais serão abordados de seguida, salientando as suas características principais.

Temperatura do ar

A temperatura é conhecida por ser uma das variáveis mais comuns para caracterizar o estado do tempo. Esta pode ser definida como uma quantidade física caracterizada pelo movimento aleatório médio das moléculas de um corpo físico, ou seja, pelo comportamento onde dois corpos em contacto térmico tendam a uma temperatura igual, pode afirmar-se que a temperatura representa o estado termodinâmico de um corpo cujo valor é determinado pela direção do fluxo líquido de calor entre dois corpos [74], Ainda assim, interessa, no âmbito climático, estudar a temperatura do ar.

A temperatura pode variar entre o dia e a noite, entre uma estação e outra, assim como entre a localização geográfica [50]. Por este motivo, é importante realçar que é preciso ter em conta a superfície onde vai ser colocado o instrumento de medição, instrumento esse conhecido como termómetro, [50, 74] e que pode diferir em vários modelos, nomeadamente termómetros de máxima e mínima, capazes de medir, de modo específico, as temperaturas máximas e mínimas, respetivamente, registadas num certo período de tempo [50]. A medição da temperatura requer um acondicionamento adequado do termómetro, para evitar a exposição solar direta e garantir uma boa proteção e ventilação para assegurar a fiabilidade dos dados recolhidos [23].

Existem três unidades de medida distintas da temperatura do ar, consoante a função de escala eleita: *Celsius* ($^{\circ}\text{C}$), *Fahrenheit* ($^{\circ}\text{F}$) ou *Kelvin* (K) [33, 50].

Precipitação

O termo precipitação é definido como sendo o produto líquido ou sólido da condensação do vapor de água caído das nuvens, ou depositada do ar, no solo e pode incluir chuva, granizo, neve, orvalho, geada e nevoeiro [74]. Como tal, a sua origem dá-se nas nuvens, em altas altitudes, com a formação de cristais de gelo, devido às temperaturas baixas de quase -40°C [50], que crescem ao juntar-se a outras gotículas de água (criando o granizo), ou aglomerando-se a outros cristais, formando os flocos de neve. Ao sair da nuvem, devido ao peso e à gravidade, e durante a sua queda até ao solo, estes flocos de neve, ou granizo, poderão encontrar uma camada de ar quente, derretendo e formando deste modo a precipitação em forma líquida [50].

Podemos medir a quantidade de precipitação, num certo intervalo de tempo, assim como a intensidade da mesma, expressa em termos de profundidade vertical da água, ou equivalente em forma sólida [74]. Ambas podem ser medidas diretamente através de instrumentos, ainda que, na devida ordem, a segunda derive da primeira. O dito instrumento trata-se do pluviómetro e é capaz de medir a quantidade de precipitação caída num certo lugar e intervalo de tempo [50].

Para caracterizar a mesma, devemos ter em conta a sua distribuição ao longo do tempo, seja uma média anual, mensal ou semanal, ou da quantidade de precipitação acumulada, assim como da frequência climática de precipitações intensas e dos horários destas medições, (que devem ser feitas em horários fixos comuns a toda a rede ou redes de interesse) [16, 74]. A unidade de

precipitação geralmente utilizada é milímetros (volume por unidade de área) ou em quilogramas (massa por unidade de área), se for em forma líquida [74].

Pressão atmosférica

Ainda que não seja visível, o ar que nos rodeia exerce força em todos os corpos com os quais contacta. Essa força é chamada pressão atmosférica, e trata-se da força por unidade de área exercida pelo peso da atmosfera acima [50, 53]. Formalmente, podemos afirmar que a pressão é igual ao peso da coluna de ar vertical acima de uma dada superfície horizontal, e que se estende até ao limite externo da atmosfera [74].

A pressão atmosférica varia temporalmente e espacialmente, e é influenciada por outras variáveis meteorológicas, tais como a temperatura, a humidade. Varia também de forma exponencial com a altitude, tal como podemos verificar na Figura 2.1. À medida que a altitude aumenta, a pressão atmosférica diminui [50, 53].

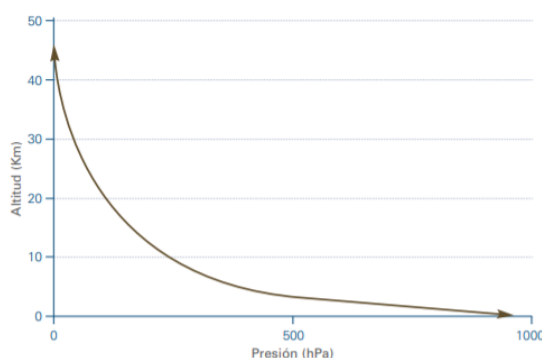


Figura 2.1: Perfil vertical da pressão atmosférica em função da altitude [50].

O instrumento de medição desta variável meteorológica é o barómetro. No entanto existem várias unidades de medida para a pressão atmosférica [74]. As mais comuns são: milímetros de mercúrio (*mmHg*), milibares (*mbar*) e pascais (*Pa*) [50]. Ainda que a unidade SI de pressão seja o pascal, é habitual que os meteorologistas expressem mais frequentemente esta variável em hectopascais (*hPa*) (e, por sua vez, que 1 *hPa* é igual a 1 *mbar* [23, 74]).

Humidade relativa

Sendo a água um dos principais componentes da atmosfera, podemos encontrá-la nos estados líquido, gasoso, ou sólido, tendo em conta as condições físicas envolventes. Ora, a humidade consiste na quantidade de vapor de água que o ar contém, quantidade essa não constante e dependente, mais uma vez, de diversos fatores circundantes, tais como a proximidade do mar ou de plantas [50]. Podemos referir-mo-nos à humidade na atmosfera como sendo humidade absoluta (g/m^3), humidade específica (g/kg), ou, a mais usualmente utilizada: humidade relativa [50]. Esta última indica-nos o quão perto está o ar de estar saturado. Isto é, trata-se da relação entre a

quantidade de vapor de água presente no ar e a quantidade máxima de vapor de água necessária para que ocorra saturação naquela temperatura em particular [5]. Podemos assim concluir que que a mesma se altera quando a temperatura varia, cf. Figura 2.2.

Para medir a humidade relativa do ar é utilizado um psicrómetro, e expressamos o seu valor em percentagem (%), recorrendo à seguinte fórmula:

$$h = \frac{e}{E}100 \quad (2.1)$$

onde e representa o conteúdo de vapor da massa de ar e E a sua capacidade máxima de armazenamento, mais conhecida como pressão de vapor saturante, que mostra a quantidade máxima de vapor de água que uma massa de ar pode conter antes de passar ao seu estado líquido, por outras palavras, a chamada saturação [50]. Diz-se então que, se o ar tiver uma humidade relativa de 100%, está saturado [5].

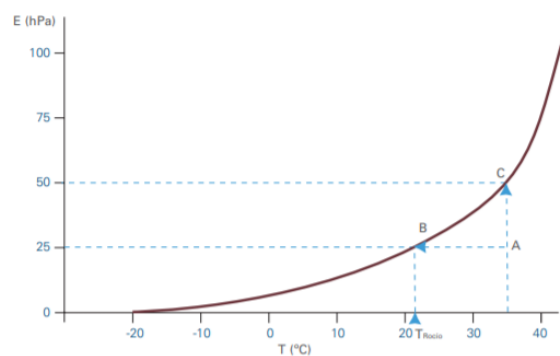


Figura 2.2: Relação entre a pressão de vapor saturante e a temperatura [50].

Através da Figura 2.2, pode verificar-se que à medida que a temperatura aumenta, a pressão de vapor de saturação também, levando a um aumento na capacidade de vapor de água no ar, e causando, deste modo, mudanças na humidade relativa [5, 50].

Direção e velocidade do vento

O vento pode ser definido como o movimento de ar de uma zona para outra, originado pela diferença de pressão entre pontos com localizações distintas. Neste caso, é possível afirmar que tende a mover-se de zonas de alta pressão para zonas de baixa pressão [50]. Ainda que não seja visível, pode caracterizar-se esta variável de acordo com a sua direção e velocidade. A direção do vento é determinada atendendo ao movimento dos objetos à medida que o ar passa por eles, [5] e pode medir-se utilizando cata-ventos, que indicam a sua direção. No que concerne à velocidade, o instrumento utilizado é o anemómetro, e as suas unidades de medida são em km/h ou m/s . Para identificar alterações repentinas, de velocidade e direção do vento, existem ainda sensores de vento de resposta rápida que dão essas informações [23].

2.2 Dados climáticos

De um modo geral podemos entender o termo clima como sendo uma síntese das condições meteorológicas durante um determinado período de tempo, num certo local, e que depende fortemente de condições externas (naturais ou antropogénicas) [16].

Para se poder caracterizar o clima recorre-se a observações meteorológicas, utilizadas para registar as condições climáticas de cada lugar assim como a sua evolução, tendo em vista o interesse subjacente em analisar a sua duração, frequência, valores máximos e mínimos e outras estatísticas relevantes à sua caracterização [16].

Dada a atualidade deste assunto, o impacto da evolução destas observações, ao longo do tempo tem sido alvo de estudo e muita atenção, por toda a comunidade científica. O estudo destas alterações tem-se vindo a mostrar cada vez mais completo face à disponibilidade e quantidade de dados recolhidos por várias estações, localizadas em diversas regiões em todo o mundo. Todavia, por vezes, estes estudos acabam por ser comprometidos pela falta de homogeneidade, resultante da falta de rotina na recolha das observações, ou pela falta de qualidade de muitos conjuntos de dados. Consequentemente, a confiabilidade dos dados precisa de ser testada para evitar resultados ambíguos ou chegar a conclusões que possam comprometer análises científicas anteriores [28].

Ainda que todas as variáveis meteorológicas sejam pertinentes a este assunto, a temperatura do ar e a precipitação têm-se mostrado alvo de um maior interesse dado o impacto direto que, a alteração de ambas, provoca nos ecossistemas naturais e a nível socioeconómico, afetando as mais diversas áreas (nomeadamente os setores primários) inclusive a saúde pública [14].

Em Portugal, foram realizados alguns estudos relacionados com a temperatura do ar em [9] e [48] assim como referentes à precipitação em [56] e [15], todos eles com o objetivo principal de avaliar alterações na intensidade frequência e duração de temperaturas e precipitação extremas. Outros estudos foram realizados, nomeadamente em [65] e [28] tendo em conta dados diários de precipitação acumulada, temperatura média, mínima e máxima procurando explicar a variabilidade climática [28]. Para justificar e entender estas variações, das duas variáveis em causa, foram utilizados índices (maioritariamente derivados da Organização Meteorológica do Mundo), tais como podemos observar nas Tabelas 2.1 e 2.2. Os índices descritos nas tabelas foram aplicados numa escala de tempo anual e sazonal e ajudam a entender as características do clima assim como visam facilitar a sua compreensão.

2.3 Análise exploratória de dados

Quando em contacto com um conjunto de dados, sejam eles categóricos ou numéricos, e em que formato seja, a sua análise estatística começa, tipicamente, com uma análise exploratória de dados ou *Exploratory Data Analysis* (EDA). Esta consitui uma primeira análise crítica dos dados, e que, habitualmente, ocorre antes da sua análise formal ou confirmatória [58].

Tabela 2.1: Índices de temperatura utilizados em [9] e [48].

Index	Descrição	Definição	Unidade
TX	Temperatura máxima	Temperatura máxima diária	$^{\circ}C$
Mean TX	Temperatura máxima	Média da temperatura máxima	$^{\circ}C$
SU/TX25	Dias de verão	Número de dias cuja temperatura máxima $25^{\circ}C$	dias
TX35	Dias muito quentes	Número de dias cuja temperatura máxima $35^{\circ}C$	dias
ID	Dias gelados	Número de dias cuja temperatura máxima $0^{\circ}C$	dias
TX_x	Dia mais quente	Valor máximo da temperatura máxima diária	$^{\circ}C$
TX_n	Dia mais frio	Valor mínimo da temperatura máxima diária	$^{\circ}C$
TX90		Número de dias cuja temperatura máxima percentil 90	dias
HWDI	Duração da onda de calor	Número de dias, em intervalos de pelo menos 6 dias consecutivos, cuja $TX > \text{média} + 5^{\circ}C$	dias
WSDI	Duração de períodos quentes (<i>warm spell</i>)	Número de dias, por ano onde, em intervalos de pelo menos 6 dias consecutivos onde $TX > \text{percentil } 90$	dias
TN	Temperatura mínima	Temperatura mínima diária	
Mean TN	Temperatura mínima	Média da temperatura mínima	$^{\circ}C$
FD	Dias de frio intenso	Número de dias cuja temperatura mínima $0^{\circ}C$	dias
TN/TN20	Noites tropicais	Número de dias cuja temperatura mínima $20^{\circ}C$	dias
TN_n	Noite mais fria	Valor mínimo da temperatura mínima diária	$^{\circ}C$
TN_x	Noite mais quente	Valor máximo da temperatura mínima diária	$^{\circ}C$
TN90		Número de dias cuja temperatura mínima percentil 90	dias
CWDI	Duração da onda de frio	Número de dias, em intervalos de pelo menos 6 dias consecutivos, cuja $TN < \text{média} - 5^{\circ}C$	dias
CSDI	duração de períodos frios (<i>cold spell</i>)	Número de dias, por ano onde, em intervalos de pelo menos 6 dias consecutivos onde $TN < \text{percentil } 90$	dias
DTR	Intervalo de temperatura diurna	Diferença média entre temperatura máxima e mínima	$^{\circ}C$
ETR	Intervalo de temperaturas extremas	Diferença entre TX_x e TN_n	$^{\circ}C$

Tabela 2.2: Índices de precipitação utilizados em [56] e [28].

Index	Descrição	Definição	Unidade
RR	Precipitação Total	Soma dos totais diários de precipitação	milímetros
DD	Dias secos	Número de dias com precipitação $< 1 \text{ mm}$	dias
LRD	Dias de pouca chuva	Número de dias com precipitação $< 3 \text{ mm}$	dias
DP	Períodos secos	Duração média dos períodos de seca (seguimento de DD)	dias
DP_x	Períodos de seca mais longo	Duração máxima dos períodos de seca	dias
CWD	Dias consecutivos de chuva	Duração máxima do período de chuva ($RR \geq 1 \text{ mm}$)	dias
R10	Dias de forte precipitação	Número de dias por ano com $RR \geq 10 \text{ mm}$	dias
R20	Dias de precipitação muito forte	Número de dias por ano com $RR \geq 20 \text{ mm}$	dias
R25	Dias de precipitação extremamente forte	Número de dias por ano com $RR \geq 25 \text{ mm}$	dias
RX1D	Maior quantidade de precipitação num 1 dia	Precipitação máxima anual em intervalos de 1 dia	milímetros
RX5D	Maior quantidade de precipitação em 5 dias	Precipitação máxima anual em intervalos de 5 dias	milímetros
SDII	Índice de intensidade diária simples	Quantidade média de precipitação num dia chuvoso ($RR > 1 \text{ mm}$)	milímetros
PrecTot	Precipitação total num ano em dias húmidos	Precipitação total anual em dias $\geq 1 \text{ mm}$	milímetros

Se, por um lado, a análise confirmatória se baseia na aplicação de modelos estatísticos formais e inferência, isto é, representações matemáticas ideais às características observadas [58], a análise exploratória de dados resume-se à aplicação de métodos cujo principal objetivo visa maximizar a compreensão do conjunto de dados, detetar possíveis erros (valores significativamente diferentes das demais observações), verificar hipóteses, determinar relações entre as variáveis, extraindo e criando outras, eventualmente, relevantes, determinar preliminarmente quais os modelos mais adequados e avaliar de que modo e direção se relacionam as variáveis e os resultados [31].

Tipicamente, um conjunto de dados apresenta um formato de tabela onde cada coluna representa uma variável ou atributo e cada linha representa uma observação ou exemplo. Cada variável poderá ser numérica ou categórica. As variáveis numéricas poderão ser contínuas ou discretas - no caso do domínio ser um conjunto numérico e discreto. As variáveis categóricas poderão ser nominais ou ordinais - no caso de existir uma relação de ordem entre as categorias.

A aplicação de técnicas de análise exploratória é primordial para que se possam evidenciar aspetos, eventualmente através da ocultação de parte dos dados, que de outro modo não seriam visíveis [58]. Os métodos de EDA podem ser vista sob duas perspetivas distintas: gráficos ou não-gráficos, e univariados ou multivariados. No que diz respeito à visualização, os métodos gráficos resumem os dados de forma pictórica ou recorrendo a diagramas e mostram os dados com mais ou menos detalhe, conforme o que já se aprendeu sobre os mesmos [37]. Os métodos não gráficos envolvem o cálculo de estatísticas resumidas fornecendo uma visão sobre a distribuição da variável de interesse [31, 58]. Os métodos univariados examinam uma variável de cada vez, e os multivariados examinam duas ou mais variáveis para explorar de que modo se relacionam. Destes últimos métodos, a análise bivariada, que examina duas variáveis em simultâneo, é das análises mais comuns [58].

As Tabelas 2.3 e 2.4 apresentam algumas sugestões de técnicas EDA, descritas posteriormente, a aplicar face ao tipo de dados em questão ou do objetivo pretendido. Contudo, e apesar de existirem diretrizes que nos indiquem de que modo devemos explorar os dados, apenas através de tentativas e experiência conseguimos perceber quais as melhores técnicas EDA a aplicar dadas as circunstâncias.

Tabela 2.3: Técnicas EDA sugeridas face ao tipo de dados [31].

Tipo de dados	Técnica sugerida
Categórico	Estatísticas descritivas
Univariado Contínuo	Gráfico de linhas, histograma
Bivariada Contínuo	2D <i>scatter plot</i>
Matriz de 2 dimensões	<i>Heatmap</i>
Multivariada	2D/3D <i>scatter plot</i> com uma 3ª variável com cor, forma ou tamanho diferente
Vários grupos	<i>side-by-side boxplot</i>

O objetivo de recorrer a representações gráficas dos dados é responder a perguntas sobre as variáveis, de um modo simples e de fácil percepção, que de outro modo não seria tão evidente. Os

Tabela 2.4: Técnicas EDA sugeridas face ao objetivo pretendido [31].

Objetivos	Técnica sugerida
Encontrar <i>outliers</i>	Histograma <i>scatter plots</i> , <i>boxplots</i>
Ideia da distribuição de uma variável	Histograma
Visualizar dados de grande dimensão	2D <i>scatter plot</i>

gráficos das variáveis permitem estudar as relações, tendências e padrões que os dados tenham assim como estabelecer uma relação de familiaridade com os mesmos numa primeira instância [2]. Neste estudo iremos focar-nos nos gráficos mais utilizados no estudo de variáveis climáticas. Em particular iremos fazer a descrição de *Warming Stripes*, o climograma, gráficos de anomalias, *quadrant type scatter plot*, ou o *heatmap*.

2.3.1 Gráficos para análise univariada

Gráfico de linhas

Um dos gráficos mais utilizados no contexto de séries temporais é o gráfico de linhas, uma vez que permite visualizar os dados ao longo do tempo [17]. Este tipo de gráfico possibilita observar os tipos de relações entre dados contínuos de uma ou mais variáveis, ligando os dados de modo a identificar determinadas características e padrões [36]. Na Figura 2.3 é possível observar o exemplo de um gráfico de linhas cujo eixo das abscissas representa uma escala quantitativa com igual desfasamento entre as observações.



Figura 2.3: Exemplo de um gráfico de linhas [27]

É possível encontrar, também no contexto de séries temporais, gráficos de linhas com uma linha de regressão que mostra a tendência nos dados. Estes gráficos são úteis para ajudar a identificar um determinado padrão nos dados, que nem sempre é obvio [36]. Nomeadamente, ao exprimir os valores observados de uma variável meteorológica ao longo do tempo, é possível identificar períodos de tempo com valores em falta ou se os mesmos descrevem uma tendência monotónica crescente ou decrescente.

Gráfico de barras

Os gráficos de barras são utilizados para visualizar quantidades associadas a grandes volumes de dados uma vez que são de fácil compreensão e intuitivos. Utilizados para dados categóricos, nestes gráficos, a altura das barras retangulares representa o valor do conjunto de dados representados e é proporcional a esses valores. Na Figura 2.4 é possível observar o exemplo de um gráfico de barras que representa o número de observações de um determinado conjunto de dados [29].

No que diz respeito ao estudo de dados climáticos, este tipo de gráfico é útil, por exemplo, para descrever a quantidade de precipitação acumulada ao longo de um determinado ano (onde cada barra representa um determinado mês), ou a média anual de velocidade do vento durante um determinado período de tempo, onde cada barra representa um certo ano.

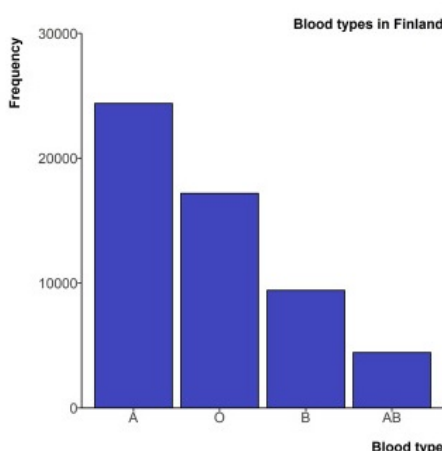


Figura 2.4: Exemplo de um gráfico de barras [6]

Histograma

Semelhantemente ao gráfico de barras, também o histograma é um tipo de gráfico utilizado para representar a distribuição da frequência de um conjunto de dados, neste caso, quantitativos. Para representar um histograma são utilizadas barras retangulares, tal como nos gráficos de barras, mas sem os intervalos de espaço entre as mesmas. No eixo horizontal, do x, são representadas as classes, previamente definidas, do conjunto de dados em intervalos, que diferem consoante os casos em estudo [6]. Na Figura 2.5 é possível visualizar o exemplo de um histograma que ilustra o peso em *kg* dos pacotes de um carteiro. Analisando a figura atrás mencionada, pode observar-se, por exemplo, que existem dois pacotes com peso entre um e dois *kg*, e dezoito deles com peso compreendido entre um e dois *kg*. De um modo geral, valores que coincidem com os limites dos intervalos considerados, são incluídos na barra anterior à considerada, mas nem sempre esta regra é seguida [58].

Gráficos como o histograma são de fácil compreensão [6], pelo que no estudo de variáveis climáticas podem revelar-se muito úteis. É exemplo o uso destes gráficos quando se procura

representar a frequência diária da temperatura do ar, ou no caso da frequência mensal da humidade relativa de modo percentual.

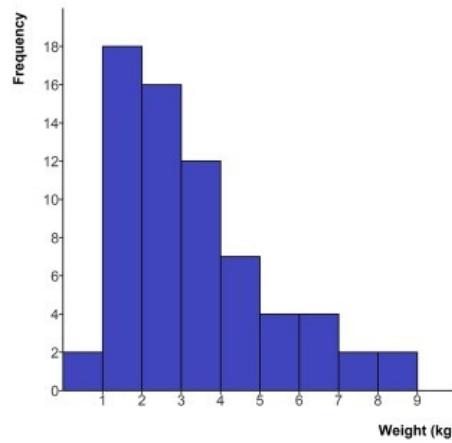


Figura 2.5: Exemplo de um histograma [6]

Boxplot

Uma das representações gráficas mais populares para visualizar a distribuição de uma variável contínua é o boxplot. O boxplot, ou também conhecido como gráfico de bigodes, tem como principal objetivo mostrar informações sobre a localização e a dispersão dos dados. O boxplot assume uma distribuição normal dos dados e utiliza um resumo de cinco números para a caracterizar: a mediana (Q_2 ou percentil 50), que consiste na mediana do conjunto de dados, o primeiro (Q_1 ou percentil 25) e terceiro (Q_3 ou percentil 75) quartil e o valor mínimo e máximo. Tal como podemos observar na Figura 2.6, depois de traçada a mediana, é desenhada a caixa que une o Q_1 ao Q_3 , representando esta o *interquartile range* (IQR), que consiste em 50% dos dados e na diferença entre os quartis mencionados atrás; de seguida são desenhados os chamados “bigodes”, representados na imagem a azul, que unem os valores extremos definidos como os limites mínimo e máximo [20, 25]. Todos os valores que excedam o mínimo e máximo representados, (retratados na imagem por pontos verdes) são considerados *outliers*.

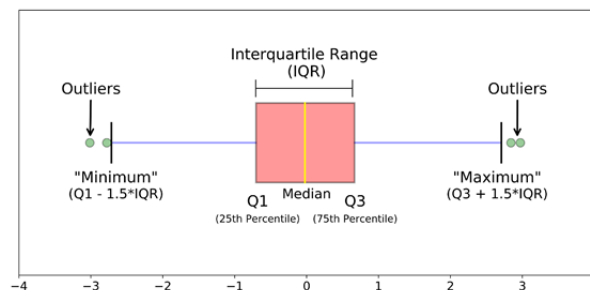


Figura 2.6: Exemplo de um *boxplot* [20].

Os *outliers* podem ser definidos como sendo um subconjunto de observações que aparentam ser inconsistentes com as demais que seguem uma distribuição hipotética. Ora, atendendo à

construção do *boxplot*, *outliers* vão ser todas as observações que excedam os limites inferior e superior do *boxplot*. O *interquartile range* é definido como $(Q_3 - Q_1)$ e os limites, mínimo e máximo, são calculados na forma $Q_1 - k * IQR$ e $Q_3 + k * IQR$, respetivamente, sendo o valor de k habitualmente igual a 1.5 ou 3 [61]. Dados que se encontrem fora dos limites $(Q_1 - 1.5 * IQR)$ e $(Q_3 + 1.5 * IQR)$ mas que não ultrapassem $(Q_1 - 3 * IQR)$ ou $(Q_3 + 3 * IQR)$ são considerados *mild outliers*; já no caso de ultrapassarem estes últimos limites, são considerados *extreme outliers* [13].

Ainda que esta técnica seja frequentemente utilizada para a deteção de *outliers*, por vezes os mesmos são declarados erroneamente, é exemplo disso quando os dados se apresentam enviesados. Para estas situações são propostas algumas alternativas tais como o *adjusted boxplot* [25].

Gráficos como o *boxplot* podem ser muito úteis quando, por exemplo, se procuram identificar valores extremos de temperatura, precipitação ou de outras variáveis climáticas em relação a um determinado período de tempo.

2.3.2 Gráficos para análise de variáveis climáticas

Warming Stripes

Os gráficos de *Warming Stripes*, criados por Ed Hawkins, cientista climático da Universidade de Reading, no Reino Unido, em 2018 [64, 76], são um tipo de representação gráfica capaz de descrever as mudanças, na temperatura média anual, de um local específico ao longo de um período de tempo. Tal como se pode visualizar no exemplo apresentado na Figura 2.7, cada faixa desse gráfico representa, a temperatura média ao longo de um ano e é colorida de acordo com um gradiente de azul-escuro a vermelho-escuro, facilitando, desse modo, o reconhecimento de anos mais frios e mais quentes, no período de tempo considerado. Estas faixas devem ser lidas da esquerda, do primeiro ano do intervalo temporal analisado para a direita, até ao último ano considerado [30].

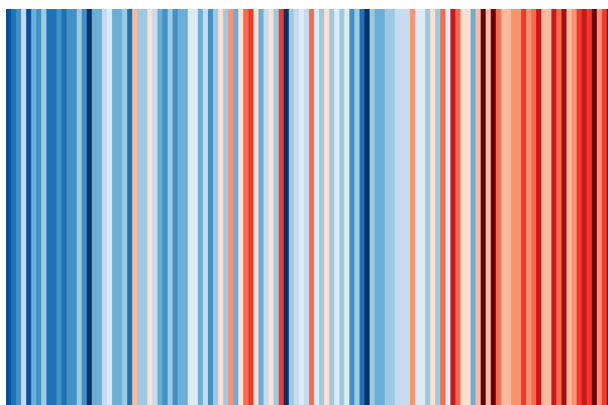


Figura 2.7: *Warming Stripes* referentes às temperaturas médias anuais em Portugal de 1850 a 2020 [18].

Este tipo de representação gráfica foi proposto com o propósito de ser simples mas capaz de retratar, de um modo ilustrativo e eficaz, as alterações climáticas, nomeadamente as tendências

de temperatura de uma determinada região ao longo de um conjunto de anos. Face à quantidade de informação com detalhes específicos acerca das mudanças de temperatura, estes gráficos acabam por preencher uma lacuna no estudo da temperatura, permitindo a sua compreensão [18].

Climograma

O climograma refere-se à representação gráfica da relação entre duas variáveis climáticas, como por exemplo a temperatura e a precipitação [24]. Este gráfico permite compreender de um modo simples como se caracteriza o perfil climático de uma região e a sua variação climática. Nele encontramos, representadas graficamente, as variações médias mensais pluviométricas assim como da temperatura, ao longo de um certo período de tempo, e o modo como elas se relacionam.

Para construir este tipo de gráfico, conforme o exemplo da Figura 2.8, consideramos no eixo das abcissas (eixo horizontal) a escala temporal, normalmente os meses, e no eixo das ordenadas (eixo vertical), de um lado, as médias mensais de precipitação, habitualmente em (*mm*) e, do outro, as médias mensais de temperatura do ar, em graus *Celsius* ou *Fahrenheit*. A representação destas duas variáveis no mesmo gráfico é feita utilizando a sobreposição de um gráfico de barras com os valores de precipitação mensais, a um gráfico de linear que une os doze valores da temperatura mensal [7].

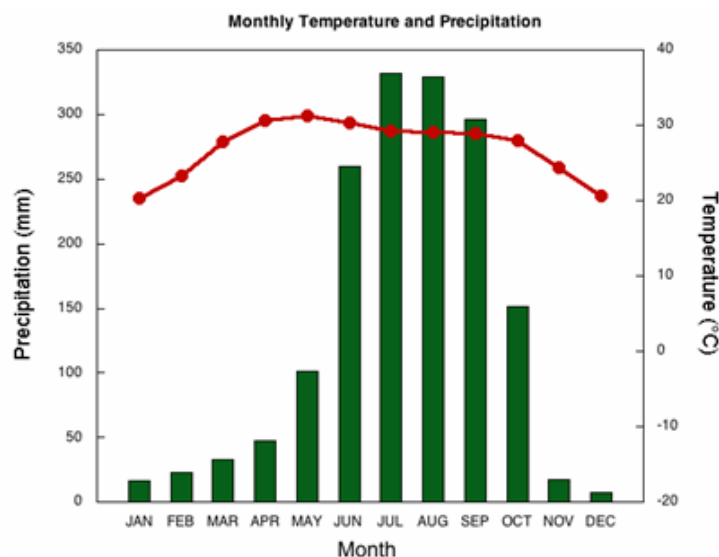


Figura 2.8: Exemplo de um climograma [7].

Através do climograma é possível identificar épocas mais húmidas ou secas, onde a precipitação é maior ou menor, respetivamente, e analisar, ao longo da escala temporal, de que modo ambas se relacionam.

Gráfico de Anomalias

Um tipo de gráfico habitualmente utilizado aquando do estudo de uma determinada variável climática, ao longo de um ano, é o gráfico de anomalias. A anomalia de uma variável climática corresponde a um desvio face a um valor de referência ou à média observada no período de tempo considerado. As mesmas podem ser consideradas positivas se o seu valor for superior em relação à média, e negativas se forem inferiores. Para analisar de que modo elas ocorrem, recorre-se a um simples gráfico de barras, com valores positivos e negativos, em relação à média, e com as cores vermelho e azul, respetivamente, tal como podemos verificar no exemplo da Figura 2.9 [51].

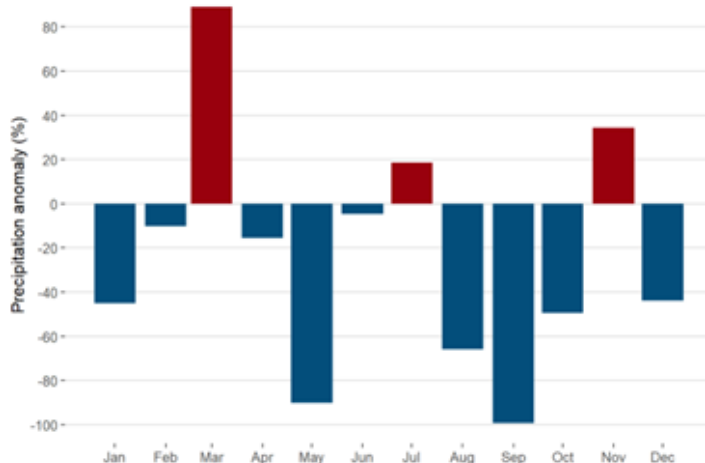


Figura 2.9: Exemplo de um gráfico de anomalias com desvios percentuais positivos e negativos em relação à média mensal de precipitação [51].

No entanto, este gráfico não consegue explicar o contexto geral dos valores anómalos, nomeadamente qual o valor extremo mais alto e mais baixo em cada mês. Para tal, por vezes, pode ser útil criar um gráfico que possibilite visualizar a distribuição da variável, intervalo interquartil e respetivos valores extremos nesse ano. Para criar este gráfico calculam-se as anomalias de cada mês, tendo em conta o valor observado em relação à média desse mês, (no conjunto de anos considerados). Depois, subtrai-se 100 ao valor da anomalia obtida, para definir a média em 0 e, deste modo, representar as mesmas em forma de percentagem. Posto isto, calculam-se as estatísticas restantes: o máximo, o mínimo, os quartis 25%, 75% e o intervalo interquartil por mês de toda a série temporal. Desta forma, em cada mês, do ano considerado, é possível observar a anomalia (positiva a vermelho e negativa a azul), o valor extremo mais alto e mais baixo, representado por uma barra cinza claro, e o *boxplot*, visível em cada barra, na cor cinza mais escuro, que mostra a distribuição da série temporal. O resultado final será semelhante ao demonstrado na Figura 2.10. Neste gráfico é possível perceber, por exemplo, que no mês de março os valores de precipitação foram mais elevados face à média do mês de março no período considerado de toda a série temporal e que no mês de setembro a precipitação foi cerca de 100% mais baixa o esperado, face à média desse mesmo mês coincidindo com a anomalia mais baixa da série temporal em questão.

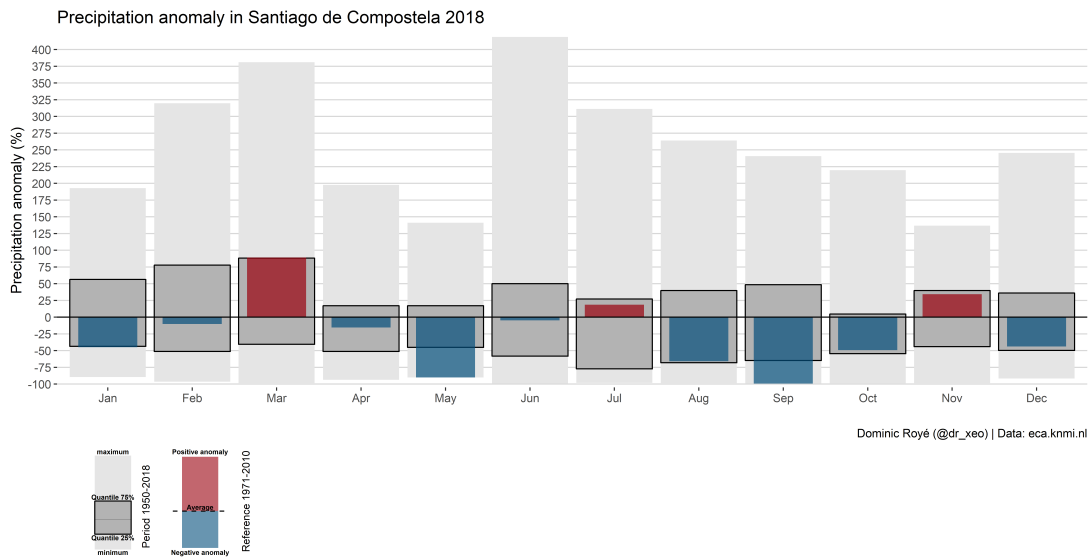


Figura 2.10: Exemplo de um gráfico de anomalias contextuais de precipitação por mês durante o ano de 2018 em Santiago de Compostela usando um período de 1971-2010 como referência [51].

Quadrant Type Scatter Plot

Por vezes pode ser útil visualizar as anomalias, da precipitação e da temperatura, num único gráfico que seja capaz de identificar numa dada região, se uma determinada estação ou mês foi seco-quente ou húmido-frio, e comparar com os anos do intervalo de tempo considerado [52]. A este tipo de gráfico, ilustrado na Figura 2.11 chamamos *Quadrant Type Scatter Plot*.

Para proceder à construção deste gráfico é necessário ter em conta os dados relativos às médias diárias de temperatura e precipitação, e respetiva estação do ano. No que concerne ao cálculo das anomalias, no caso da temperatura, estas são efetuadas do modo, tipicamente, mais conhecido: subtraindo ao valor da temperatura em causa o valor da média desta variável; já no caso da precipitação, tendo em vista facilitar a sua interpretação gráfica, as anomalias são expressas em percentagem com a média fixada em 0% em vez de 100%. Isto é, tomando, como exemplo, um caso onde a média de pluviosidade seja de 60mm e no ano em causa o valor de precipitação seja de 0mm, nesse ano, a anomalia será de -100%, se tiver chovido 60mm será de 100% e no caso de 120mm será de 200% e assim sucessivamente. Posto isto, e para criar este gráfico, são definidos quatro quadrantes: no canto superior direito, o 1º quadrante, com os pontos relativos a anos seco-quentes, e no canto inferior esquerdo, o 3º quadrante, com os pontos relativos a anos frio-húmidos. Em cada um destes quadrantes são anotados os pontos referentes a anos anómalos no que respeita a conjugação da precipitação com a temperatura.

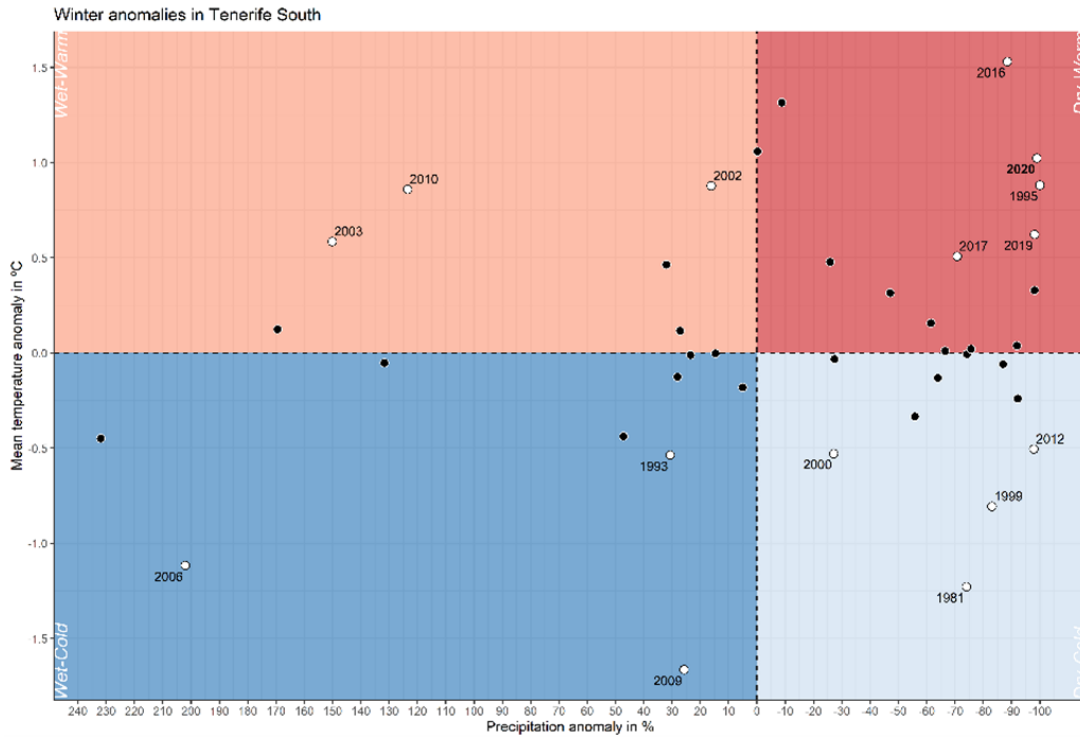


Figura 2.11: Exemplo de um *Quadrant Type Scatter Plot* para a identificação de anomalias temperatura e precipitação durante o inverno no Sul de Tenerife no período de 1981 a 2020 [52].

Heatmap

Atualmente, e com quantidades de dados cada vez maiores, é importante encontrar soluções de visualização gráfica capazes de extrair informação relevante. O *heatmap* é um tipo de representação gráfica que procura facilitar o estudo e análise de dados, frequentemente em grandes volumes. Este tipo de gráfico trata-se de uma matriz com várias células, cuja representação dos dados é feita tendo em conta um sistema de codificação de cores onde, habitualmente, cada linha representa um objeto e cada coluna uma condição, ponto no tempo, instância ou propriedade, e a cor dessa célula varia consoante o valor dos dados associados [32, 73]. Representando os dados através de um *heatmap* é possível compactar uma densa quantidade de informações e identificar padrões relativos mais rapidamente ao invés de recorrer a extensos relatórios analíticos, e observar de que modo os dados se comportam face ao intervalo de tempo utilizado, que será diferente caso seja de anos, meses, dias, horas ou outro. Na Figura 2.12 encontra-se um exemplo de um *heatmap* onde se encontram as temperaturas máximas diárias por mês ao longo de um ano (a ausência de dados é representada por células em branco).

2.4 Séries temporais

Nas diversas áreas que envolvem o dia-a-dia da população, existem variáveis medidas sequencialmente ao longo do tempo, sendo estas relativas a ciências físicas, ambientais, engenharia,

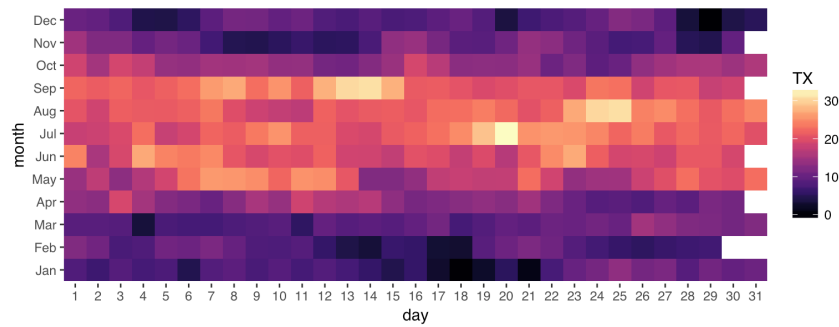


Figura 2.12: Exemplo de um *heatmap* ilustrando as temperaturas máximas diárias por mês [3].

economia, etc.. [59]. A análise desses dados, observados em diferentes momentos, remete para problemas de inferência e estatística atuais. No entanto, esta análise pode ser restringida pela dependência entre os dados, causada pela sua adjacência no tempo, pondo assim em causa a aplicabilidade de métodos estatísticos, que assumem uma sequência de variáveis aleatórias independentes e distribuídas de forma idêntica [11, 59]. Tendo como base esta condição, podemos referir-nos a esta sequência de variáveis aleatórias como um processo estocástico de variáveis discretas [11], ou, mais habitualmente, como uma série temporal, à qual se podem aplicar análises capazes de responder a questões matemáticas e estatísticas colocadas pela correlação dos seus dados, no tempo.

No estudo de uma série temporal procura-se explicar a correlação entre as observações para conseguir compreender o passado, obter as melhores previsões futuras ou executar procedimentos inferenciais corretos nos dados. Deste modo, o primeiro passo, deve passar por uma análise cuidadosa sobre os dados registados e observados ao longo do tempo, com a finalidade de encontrar o melhor método de análise e estatísticas aplicáveis, que possam ser utilizadas para resumir a informação dos dados [59].

Antes de aprofundar alguns métodos estatísticos específicos, é importante mencionar duas abordagens diferentes na análise de séries temporais frequentemente identificadas como abordagens no domínio do tempo e da frequência. A primeira baseia-se na importância da linha temporal para investigação de relacionamentos desfasados, enquanto que a última se sustenta na importância e comportamento dos ciclos [59].

Notação

Uma série temporal é representada como uma coleção de variáveis aleatórias, indexadas de acordo com a ordem histórico-temporal, isto é, uma sequência de variáveis aleatórias $y_1, y_2, y_3, \dots, y_t$, indexadas por t , sendo este discreto e variável ao longo de um conjunto, ou subconjunto de números inteiros, ou algum índice semelhante. Sempre que não seja necessário especificar o comprimento da série temporal, a mesma pode ser denominada por y_t [11, 59]. O modelo de uma série temporal é uma sequência de variáveis aleatórias e a série observada é considerada uma realização do modelo [11]. Após definir o modelo, é possível fazer previsões no tempo, isto

é, dada a série y_t , $y_{t+k|t}$ representa a previsão para um valor futuro, previsto no tempo $t + k$, sendo k o intervalo de tempo [11].

No que respeita a representação gráfica de uma série temporal, é comum traçar os valores das variáveis aleatórias no eixo vertical das ordenadas, e a escala do tempo no eixo das abcissas [59].

Componentes

Ao lidar com dados de séries temporais é importante medir a dependência entre as observações, entendendo a dinâmica dos dados ou, de outro modo, a sua evolução temporal. Se entendermos a dinâmica, podemos utilizar modelos matemáticos para a representar, mesmo que aproximadamente, e reproduzi-la e projetá-la no futuro. De um modo geral, e de acordo com [27], podemos afirmar que as três principais componentes de uma série temporal são:

- **Tendência** – mudança de longo prazo na média dos dados, que não precisa necessariamente de ser linear, podendo, por vezes, ser referida como uma “mudança de direção”, quando pode ir de uma tendência crescente para uma tendência decrescente, por exemplo;
- **Sazonalidade** – ciclos dentro de um ano geralmente relacionados com as estações do ano; é uma componente fixa e com frequência conhecida;
- **Ciclos** – contrariamente à sazonalidade, um ciclo ocorre quando existem subidas e descidas nos dados, mas com frequência desconhecida;

Em alguns casos, podemos ainda afirmar a existência de uma quarta componente, sendo esta intitulada como componente aleatória ou restante. A mesma é constituída pelo que resta, depois da tendência e da sazonalidade serem removidas, sendo esta, uma sequência de variáveis aleatórias correlacionadas cuja média é igual a zero, e algumas propriedades desejáveis, as quais serão abordadas mais tarde [27].

2.4.1 Medidas de dependência

Existem diferentes medidas que descrevem o comportamento de um dado processo à medida que ele avança no tempo, das quais podemos dar destaque:

Função média - é uma função de t , e denota o valor esperado de uma variável. É exemplo, a média de temperaturas altas mensais, numa certa cidade, ao longo de um dado intervalo de tempo [59]. No contexto de séries temporais podemos referir à mesma como:

$$\mu_t = E[y_t] \quad (2.2)$$

Função auto-covariância - esta função mede a dependência linear de duas variáveis, numa série temporal, mas observadas em momentos diferentes s e t , da seguinte forma:

$$\gamma(s, t) = cov(y_s, y_t) = E[(y_s - \mu_s)(y_t - \mu_t)] \quad (2.3)$$

É uma função comutativa, e portanto,

$$\gamma(s, t) = \gamma(t, s) \quad (2.4)$$

para todo o tempo s e t [59].

Função de Auto-correlação (ACF) - esta função mede a relação linear entre valores defasados numa série temporal, isto é, revela de que modo se altera a correlação entre dois valores, à medida que o seu intervalo temporal se altera [41] da seguinte forma:

$$\rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}} \quad (2.5)$$

Função de Auto-correlação Parcial (PACF) - através desta função consegue-se medir a correlação entre y_t e y_{t-k} , depois de remover os efeitos nos desfasamentos $1, 2, 3, \dots, k-1$ [11]. Isto porque se y_0 está correlacionado com y_1 , e y_1 está correlacionado com y_2 , então y_0 encontra-se necessariamente correlacionado com y_2 também. Assim sendo, no caso de não existir nenhum desfasamento para remover, a ACF e a PACF terão o mesmo resultado [27]. Cada PACF poderá ser estimada como o último coeficiente de um modelo autoregressivo [11].

2.4.2 Testes de correlação de variáveis

Perceber a correlação entre duas variáveis desempenha um papel importante ao fazer a análise exploratória dos dados. Uma correlação mede a relação entre duas variáveis, isto é, de que modo elas se relacionam entre si, e, no caso de séries temporais, permite aferir de que modo elas evoluem no tempo: se na mesma direção, se em direções opostas ou se são independentes [31]. Para este efeito, existem alguns métodos e representações gráficas que permitem visualizar e aferir a correlação entre variáveis.

2.4.2.1 Correlação de *Pearson*

Para medir a correlação entre duas variáveis é comum aplicar testes de correlação entre elas. Para observar a força e dependência linear entre duas ou mais variáveis contínuas, é frequente utilizar o teste de correlação de *Pearson*. O valor R , do coeficiente de correlação, pode variar entre -1 e 1 e, quanto mais próximo de 1 ou de -1 for o coeficiente mais forte será a relação entre as variáveis. Atendendo a esse valor, se a correlação for igual a zero indica dependência entre as duas variáveis, isto é, se uma variável aumentar ao longo do tempo não existe tendência na outra variável de diminuir ou aumentar. Relativamente à direção, uma correlação negativa implica que as duas variáveis em consideração variam em direções opostas, ou seja, se uma variável aumentar, a outra diminui e vice-versa. Por outro lado, se a correlação observada for positiva, as variáveis variam na mesma direção, isto é, quando uma aumenta, a outra também aumenta e se diminuir a outra também diminui. O *p-value* deste teste, é o nível de significância atribuído ao resultado obtido, e, idealmente, para ser considerado estatisticamente significativo, o seu valor deverá ser inferior a 0.05 [4].

Para representar os resultados obtidos por este teste, é comum utilizar uma matriz de correlação onde se podem ver os coeficientes de correlação com uma cor associada ao seu valor. Na Figura 2.13 pode observar-se um exemplo de uma matriz de correlação com os respetivos resultados do teste de *Pearson*.

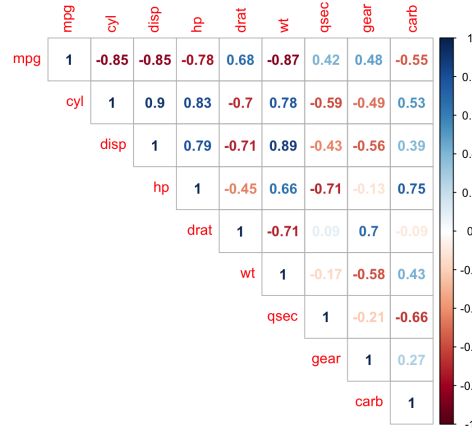


Figura 2.13: Exemplo de uma matriz de correlação de *Pearson* [4].

2.4.2.2 Normalized cross-correlation

Para detetar a correlação entre duas séries temporais também é comum recorrer aos cálculos de *cross-correlation*. No entanto, estes resultados são de difícil interpretação, e se existirem zeros nas séries ou se tiverem diferentes comprimentos, a correlação não é detetada. Para mitigar essas questões, recorre-se a *normalized cross-correlation*. Esta função mede a correlação entre as observações de duas séries temporais separadas por k unidades de tempo, i.e. a correlação entre x_t e y_{t+k} . *Lag* é o número de períodos de tempo que separa duas séries temporais.

Através de funções para calcular os valores de *cross correlation* [47], é possível visualizar a correlação entre as variáveis através de vários *lags* e perceber de que modo se correlacionam. Para facilmente identificar se uma correlação entre duas variáveis é significativa o seu valor absoluto deve ser superior a $\frac{2}{\sqrt{n-|k|}}$ onde n representa o número de observações e k o *lag* do valor absoluto observado. Na Figura 2.14 podemos observar o exemplo de um gráfico de *cross correlation*, para 15 observações, onde o valor de correlação entre as variáveis é significativo, uma vez que, para este exemplo, a correlação no *lag* -2 é aproximadamente 0.92, e maior que $\frac{2}{\sqrt{15-|2|}}$ [1].

2.4.3 Estacionaridade

Uma série temporal diz-se estacionária se não depender do tempo em que é feita a observação. Os dados de uma série temporal estacionária devem exibir comportamentos semelhantes ao longo do tempo enquanto que os dados de uma série temporal não-estacionária não asseguram essa regularidade [27]. Há que notar ainda que a presença da componente cíclica, tendo em conta a inexistência das componentes referidas atrás, não implica que a série não seja estacionária, uma

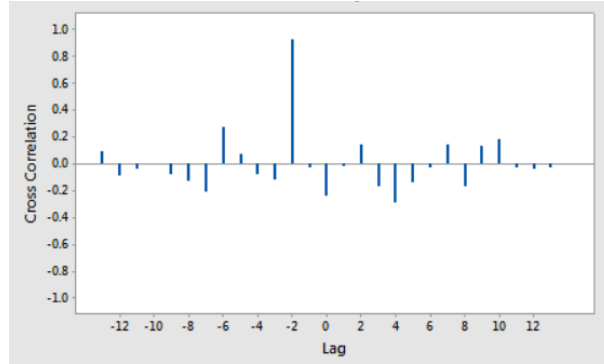


Figura 2.14: Exemplo de gráfico de *cross correlation* com correlação significativa [1].

vez que os ciclos apresentam uma frequência desconhecida, logo não conseguimos saber onde estarão os picos ou as depressões dos ciclos [27]. Na Figura 2.15 pode observar-se uma série estacionária, contrariamente à Figura 2.16 que ilustra a evolução da variável ao longo dos anos com uma tendência crescente e média não constante.

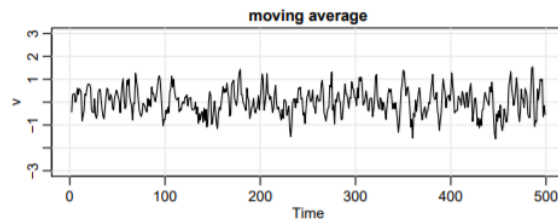


Figura 2.15: Exemplo de uma série temporal estacionária [59].

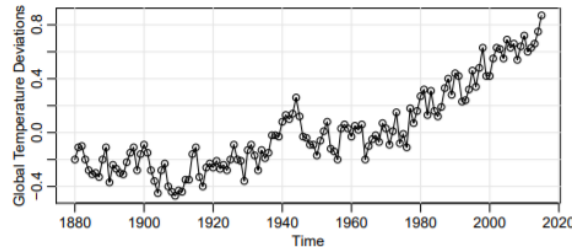


Figura 2.16: Exemplo de uma série temporal não estacionária, devido à tendência [59].

No que diz respeito às séries estacionárias, estas podem ser *strictly stationarity* (estacionariamente fortes) ou *weakly stationarity* (estacionariamente fracas). Uma série temporal é considerada forte se mostrar um comportamento probabilístico idêntico para todo o conjunto de valores da série, $y_{t_1}, y_{t_2}, \dots, y_{t_k}$ e valores *shifted*, isto é, $y_{t_1+h}, y_{t_2+h}, \dots, y_{t_k+h}$, sendo todo o $k = 1, 2, \dots$, todos os valores de tempo $t_1, t_2, \dots, t_k$, e todos os intervalos de tempo $h = 0, \pm 1, \pm 2, \dots$ [59]. Os dados deste tipo de séries temporais estacionárias devem mostrar um comportamento semelhante em diferentes intervalos de tempo [59]. Já no que toca a séries estacionárias fracas, podemos afirmar que as mesmas se caracterizam por ter uma média constante e independente do tempo t , $E(y_t) = \mu$, assim como a variância, $var(y_t) = E[(y_t - \mu)^2] = \sigma_y^2$, e a função de autocorrelação $\gamma(s, t)$ passa a depender apenas da distância entre s e t , $|s - t|$, e não mais do tempo [59].

Por norma, quando nos referimos a uma série temporal estacionária falamos deste último caso, diferenciando apenas quando a mesma se revela estritamente estacionária ou se for *Gaussian*, e, nesse caso usamos o termo *strictly stationarity* [59].

Existem alguns métodos, paramétricos e não paramétricos, propostos por diferentes investigadores, capazes de verificar se uma determinada série é ou não estacionária, nomeadamente, o teste "*Augmented Dickey-Fuller*", "*KPSS*" ou "*Zivot and Andrews*" [42].

2.4.4 Modelos

Tal como podemos observar no exemplo apresentado na Figura 2.16, existem muitos modelos dominados por uma tendência e/ou por sazonalidade. Normalmente, tendência e ciclo estão incluídos no mesmo componente: T_t .

Podemos encontrar modelos onde a tendência parece aproximadamente linear, e os ciclos sazonais têm constantes ao longo do tempo. Estes são **modelos lineares/aditivos** dados por:

$$Y_t = T_t + S_t + R_t \quad (2.6)$$

onde, Y_t é a série observada, T_t , S_t , e R_t , representam, as componentes tendência, efeito sazonal e restante, respetivamente [11].

No caso da tendência não ser linear e da amplitude dos ciclos sazonais crescerem com o aumento da tendência, o mais apropriado será um **modelo multiplicativo** que represente a relação entre os componentes:

$$Y_t = T_t S_t R_t \quad (2.7)$$

Neste último caso, podemos ainda notar que, se a variabilidade aumentar ao longo do tempo ou com a tendência, podemos utilizar variações *Box-Cox*, para estabilizar a variabilidade ao longo da série. Assim sendo, a Equação 2.7 torna-se:

$$\log(y_t) = \log(T_t) + \log(S_t) + \log(R_t) \quad (2.8)$$

Dependendo do propósito do estudo, o interesse na análise de uma série temporal pode residir em estudar a tendência, a sazonalidade, os ciclos ou o componente remanescente. Contudo, nem sempre é útil a presença destas componentes, prejudicando mesmo alguns métodos estatísticos, (é exemplo disso o cálculo de uma média na presença de tendência, que apenas faz sentido se os dados forem observações de variáveis subjacentes com uma média comum). Nessa sequência, é importante por vezes, aplicar algumas abordagens para estimar e/ou remover componentes de tendência e sazonalidade. Noutros casos é suficiente que as mesmas sejam filtradas, para que a componente restante se torne mais pronunciada.

No que concerne ao tratamento da tendência, podemos:

- estimar a mesma, através de uma função determinística linear, quadrática ou outra, e retirá-la;
- filtrar os dados com o *difference operator*, que consiste na diferença entre um ponto no tempo e o seu anterior, na série temporal;
- aplicar técnicas de suavização na série temporal, para descobrir certas características, como tendência de longo prazo, ou até para encontrar componentes sazonais. Dentro destas, podemos destacar *moving averages*, *kernel*, *engeneralized additive model*, e *lowess*, sendo esta última um método de suavização de regressão;

No caso de estimar/remover a sazonalidade as técnicas não diferem muito das anteriores:

- filtrar os dados com o operador de diferença sazonal
- tomar uma abordagem de regressão com variáveis *dummy* para os dias da semana ou meses ou série de Fourier

2.4.5 *Moving averages*

Tal como referido anteriormente, os dados de uma série temporal podem exibir uma variedade de padrões sendo, por isso, útil dividir a série em cada uma dessas componentes. Existem vários métodos para extrair essas componentes e decompor a série temporal. A transformação mais simples a aplicar é a chamada *moving averages* e tem como principal objetivo suavizar a série temporal relevando características e tendências de baixa frequência [59].

Uma *moving average* de ordem m no tempo t , m-MA [27], pode ser obtida da seguinte forma:

$$\hat{T}_t = \frac{1}{m} \sum_{j=-k}^k y_{t+j} \quad (2.9)$$

onde $m = 2k + 1$. Ou seja, a estimativa da tendência no tempo t é obtida através da média dos valores da série temporal em k períodos de t . As observações próximas no tempo têm um valor próximo, o que implica que a média elimine parte da aleatoriedade dos dados restando assim uma componente de tendência suave.

Na Figura 2.17 podem observar-se diferentes *moving averages* aplicadas à mesma série temporal. A linha da tendência, representada a vermelho, é mais suave que os dados originais e segue o movimento dos dados sem as principais flutuações. A transformação aplicada ao maior número de observações consecutivas (9-MA) corresponde à curva mais suave, pelo que se pode concluir que quanto maior o valor da média móvel, menos sensíveis são as alterações aos dados originais da série [27].

O algoritmo *Seasonal and Trend decomposition using Loess* (STL) é um algoritmo de decomposição de séries temporais versátil e robusto, que permite dividir a mesma nas componentes:



Figura 2.17: Diferentes *moving averages* aplicadas à mesma série temporal [27].

tendência, sazonalidade e restante. *Loess* é um método de suavização. Este tipo de decomposição foi desenvolvido para modelos aditivos cuja soma das três componentes resulta novamente na série original. Na Figura 2.18 encontra-se um exemplo de uma série temporal aditiva decomposta pelo método STL [8].

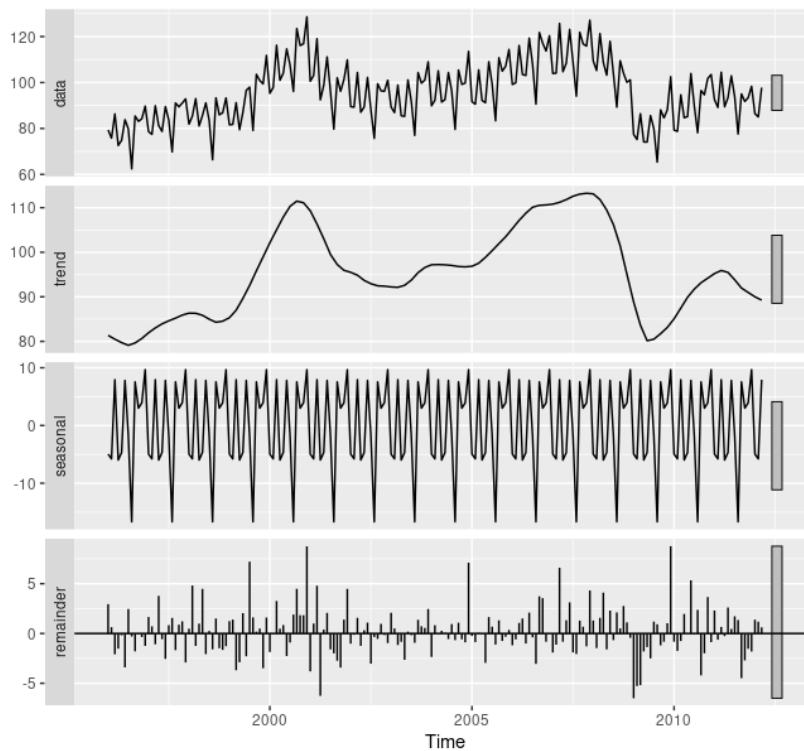


Figura 2.18: Exemplo de decomposição de uma série temporal com modelo aditivo [27].

2.5 Homogeneização

Uma série temporal climática é considerada homogênea se as variações registadas forem causadas apenas por variações no clima [55]. Estas irregularidades podem ser causadas por fatores climáticos ou relacionadas com o processo de recolha e registo dos dados. A fim de garantir a

validade de estudos realizados com base na análise de dados meteorológicos é vital considerar informação sobre a variabilidade do clima, tendências e ciclos [39]. As séries temporais deste tipo de dados apresentam, na maior parte das vezes, irregularidades dando origem a séries temporais não-homogêneas.

Como exemplo de influências climáticas, temos as erupções vulcânicas e consequente emissão de gases e partículas para a atmosfera, provocando alterações abruptas nas radiações solares e na circulação dos oceanos [39, 49]. Já no que diz respeito a causas não-climáticas podemos referir-nos a alterações nos instrumentos de medição, à localização das estações, e ao seu ambiente envolvente, assim como respetivas rotinas de observação e tratamento preliminar dos dados [39, 43]. É também de referir a influência das diferentes alturas do dia em que são registadas estas medições, e respetivas unidades de medida (no caso da temperatura, $^{\circ}C$ ou K), ou nas fórmulas aplicadas para obter médias das variáveis em estudo. Ainda no âmbito destas irregularidades, é importante considerar erros provenientes do processo de recolha, digitalização e processamento destas variáveis pela mão do homem [49].

Todos estes fatores podem levar à introdução das ditas discontinuidades/ não-homogeneidades nas séries temporais e influenciar erroneamente os resultados dos estudos realizados. Por este motivo, é imperativo detetá-las e corrigi-las, caso se verifiquem, para, só depois, poder usar as séries temporais em estudos climáticos [39, 49]. A este processo de deteção e recolha de irregularidades denominamos homogeneização [49].

Existem diferentes tipos de não-homogeneidades sendo as mais comuns [49]:

Erros pontuais - resultantes do processo de observação e recolha dos dados;

Break-points – alterações na média, responsáveis pelas mudanças de localização, instrumentos, alterações no meio envolvente ou práticas de observação dos dados;

Tendências – provocadas por uma possível descalibração de sensores ou devido ao crescimento urbano, nomeadamente a construção de imóveis ou a presença de árvores, perto da estação meteorológica, que influencie os registos das diferentes variáveis.

Destes três tipos de não-homogeneidades, a tendência é talvez a mais complicada de detetar, uma vez que se podem dever a tendências climáticas verdadeiras. Os *break-points* são talvez os mais frequentes tendo em conta que a maioria de mudanças ocorre de forma abrupta [49]. Ainda assim, o processo de identificação destas irregularidades não é simples nem objetivo. São muitas as técnicas (e *software* que as utilizam) que procuram obter os melhores resultados para detetar e corrigir as não-homogeneidades. Contudo, a maneira mais simples, numa fase inicial, será optar por uma análise visual que possibilite fornecer informações sobre períodos mais duvidosos [39].

No que diz respeito a estudos climáticos, a temperatura é a variável com maior incidência de estudo. Já no caso da precipitação esta tarefa mostra-se mais complicada devido à sua variabilidade. Aquando do estudo de não-homogeneidades são muitas as incertezas no caso

da precipitação, isto porque, muitas das flutuações climáticas na precipitação diária podem acabar por ser removidas como não-homogeneidades. Não são tidas em conta situações climáticas extremas, e os pluviómetros podem ser inclusive afetados pela força do vento. Por este motivo, no caso da precipitação é necessário adotar técnicas de homogeneização mais complexas e sofisticadas capazes de contornar estas adversidades [49].

2.5.1 Métodos para detetar e corrigir *breaks* não climáticos

A literatura mostra inúmeras abordagens no que diz respeito aos processos de homogeneização de dados climáticos. Todas elas diferem umas das outras nas suas características, dependendo do tipo de variável (temperatura, precipitação, pressão atmosférica, etc.), a frequência de medição dos dados (anuais, mensais, diários, etc.), da utilização de séries temporais de referência ou até a disponibilidade ou não de metadados (informações históricas da estação) [10]. Estas características dos métodos de homogeneização são discutidas em seguida.

- **Métodos de homogeneização diretos e indiretos** - os métodos diretos recorrem ao uso de metadados (informações históricas da estação), e a estudos dos efeitos de mudanças específicas nos instrumentos de medição; os indiretos utilizam métodos objetivos e subjetivos, utilizando diferentes técnicas estatísticas e gráficas, e considerando apenas o uso de dados da estação em causa (abordagem absoluta) [43, 49].
- **Métodos de homogeneização absolutos e relativos** - os métodos absolutos consideram apenas séries temporais de uma única estação, os relativos têm em conta dados de estações próximas, ou de referência, consideradas previamente homogêneas, para homogeneizar a estação em causa; a primeira mostra-se menos vantajosa na medida em que não permite identificar se a descontinuidade é natural ou não, sem um suporte ou registo histórico da estação [10, 49].
- **Técnicas de *multiple breakpoint*** – normalmente a série temporal em estudo apresenta mais do que um *break-point*, mas a maioria dos métodos apenas deteta um de cada vez; a maioria dos testes estatísticos aplica o mesmo método várias vezes, a subsets da série original, até não detetar mais nenhum *break* ou até o número de observações ser demasiado pequeno; existem ainda alguns métodos que identificam e corrigem vários *breakpoints* ao mesmo tempo ao invés de *step-by-step*, tais como o MASH ou ACMANT [49].

2.5.2 Métodos de homogeneização

Atualmente, são muitas as técnicas estatísticas propostas para homogeneizar dados climáticos e aceites pela comunidade científica. A aplicação de uma ou mais técnicas, à mesma série temporal, faz parte do processo de homogeneização. Nesta seção serão apresentadas as descrições de alguns métodos mais utilizados para deteção de não-homogeneidades e homogeneização de dados climáticos, e respetivas classificações tendo em conta as suas características. Serão ainda

descritos três métodos de regressão linear habitualmente referidos na literatura e um teste que verifica a presença de tendência na série temporal. (cf. Tabela 2.5) [10, 49].

A maioria dos testes pode ser integrada em três grupos distintos: métodos estatísticos paramétricos, não paramétricos baseados em probabilidade e métodos de regressão linear. Os mais frequentemente utilizados em estudos climáticos são: *Standard Normal Homogeneity Test* (SNHT), *Buishand Test*, *Pettitt Test*, entre outros. Estes testes informam-nos do nível de heterogeneidade da série em estudo e detetam o *break-point* mais provável. Outros testes, nomeadamente, *von Neumann ratio test*, não calculam a data do *break*, mas estimam o nível de não homogeneidade dos dados [39].

2.5.2.1 Métodos estatísticos

Para os seguintes testes, assume-se que Y_i (i o ano/mês de 1 a n) trata-se da série temporal a ser testada, $\bar{Y}$ a média e s o desvio padrão.

Buishand Range Test

O teste de *Buishand*, proposto por Buishand em 1982 [71], é considerado um teste paramétrico que pressupõe sob a hipótese nula que os valores testados são independentes e seguem uma distribuição normal idêntica. Sob a hipótese alternativa, este teste assume a presença de uma mudança gradual na média, é indicativa de um *break*. Este teste consegue identificar o mês ou ano onde esse *break* ocorre, mas é mais sensível aqueles que ocorram no meio da série.

Para este teste as somas parciais ajustadas podem ser definidas como:

$$S_0^* = 0 \text{ e } S_k^* = \sum_{i=1}^k (Y_i - \bar{Y}) \quad (2.10)$$

para $k = 1, \dots, n$

Se uma série for homogênea o valor de S_k^* será próximo de 0 tendo em conta a inexistência de desvios dos valores de Y_i em relação à sua média. No caso de existir um *break*, então S_k^* atinge um máximo ou mínimo perto do ano/mês $k = K$.

A relevância dos resultados do teste (R) é dada pela diferença entre o máximo e mínimo dos valores de S_k^* escalonados pelo desvio padrão da amostra.

$$R = \frac{(\max_{0 \leq k \leq n} S_k^* - \min_{0 \leq k \leq n} S_k^*)}{s} \quad (2.11)$$

Os valores críticos são dados por $R/\sqrt{n}$ e variam consoante o tamanho do conjunto de dados. Estes podem ser consultados na Tabela C.1, no apêndice C.

Pettitt Test

O *Pettitt test*, proposto por *Pettitt* em 1979 [71], trata-se de um teste não-paramétrico capaz de detetar o mês ou o ano onde um possível *break* tenha ocorrido. Tal como o anteriormente descrito, este teste assume a mesma hipótese nula e alternativa que o *Buishand Range Test* e apresenta-se, igualmente mais sensível, a *breaks* que ocorram no meio da série.

Os *ranks* $r_1, \dots, r_n$ de $Y_1, \dots, Y_n$ são utilizados para calcular as estatísticas:

$$X_k = 2 \sum_{i=1}^k r_i - k(n+1) \quad (2.12)$$

para $k = 1, \dots, n$.

Se ocorrer um *break* no ano/mês K , X_k assume um valor máximo ou mínimo perto do ano/mês $k = K$:

$$X_K = \max_{1 \leq k \leq n} |X_k| \quad (2.13)$$

Os valores críticos, consoante o tamanho do conjunto de dados, estão descritos na Tabela C.1 e podem ser consultados no Apêndice C.

Von Neumann Ratio Test

O *Von Neumann ratio test*, proposto por *Von Neumann* em 1941, trata-se de um teste não paramétrico cuja estatística se define como sendo a razão entre a diferença média quadrada sucessiva (ano-a-ano) e a variância [71]. A hipótese nula assume que os dados são valores aleatórios independentes, distribuídos de forma idêntica. O valor calculado é um indicador da presença de irregularidades na série. Ao contrário dos anteriores, este teste não dispõe de informações sobre a data da descontinuidade, e é, geralmente, usado em conjunto com outros testes de homogeneidade [49].

Assumindo que n é o comprimento do conjunto de dados, e Y_i o i -ésimo elemento desse conjunto, o *Von Neumann Ratio Test* pode ser definido como:

$$N = \sum_{i=1}^{n-1} (Y_i - Y_{i+1})^2 / \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (2.14)$$

Se $N = 2$ pode concluir-se que a série é homogênea, mas se existir um *break*, então o valor de N tende a ser inferior ao esperado. Já no caso de a amostra apresentar variações rápidas na média, então o valor de N pode ser superior a 2.

É ainda de notar que, em conjuntos de dados com *breaks* e variações de média semelhantes, o conjunto de dados com menor desvio padrão terá um N igualmente menor. Isto significa que

dados cuja resolução temporal sejam médias anuais devem ter valores de N menores que outros cujos valores sejam médias mensais. [39, 71].

Os valores críticos deste teste encontram-se na Tabela C.2, no Apêndice C.

2.5.2.2 Métodos probabilísticos

Standard Normal Homogeneity Test (SNHT)

SNHT é um dos testes mais populares quando nos referimos a testes de homogeneidade em estudos climáticos. É um teste de probabilidade e segue a mesma hipótese nula e alternativa que o *Buishand Range Test* com a exceção de que apresenta uma maior sensibilidade a *breaks* perto do início ou do fim da série.

Este teste é tipicamente aplicado a séries de referência compostas, isto é, séries com valores provenientes da estação e alguns valores de referência regionais, considerados previamente homogêneos [10]. Para um dado momento, são calculadas as médias antes e depois da série composta. Se a diferença entre ambas coincidir com um valor crítico, então a série é tida como não homogênea [49].

Mais tarde, *Alexandersson* e *Moberg* em 1997 melhorou o método deste teste para conseguir detetar tendências (*SNHT with trend*) e propôs uma estatística $T(k)$ Esta estatística compara as médias entre os primeiros k anos/meses registados com os últimos $n - k$ anos meses [39], do seguinte modo,

$$T(k) = k\bar{z}_1^2 + (n - k)\bar{z}_2^2 \quad (2.15)$$

para $k = 1, \dots, n$ e onde

$$\bar{z}_1 = \frac{1}{k} \frac{\sum_{i=1}^k (Y_i - \bar{Y})}{s} \quad (2.16)$$

e

$$\bar{z}_2 = \frac{1}{n - k} \frac{\sum_{i=k+1}^n (Y_i - \bar{Y})}{s} \quad (2.17)$$

onde

$$s = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (2.18)$$

Se ocorrer um *break* no ano/mês K , então $T(k)$ atinge um máximo perto do ano/mês $k = K$. Esse teste estatístico T_0 é definido como:

$$T_0 = \max_{1 \leq k \leq n} T(k) \quad (2.19)$$

A hipótese nula deste teste é rejeitada se T_0 se encontrar abaixo de um certo nível, tendo em conta o tamanho da amostra. Os valores críticos encontram-se na Tabela C.1, no apêndice C.

2.5.2.3 Métodos de regressão

O termo regressão, proposto por *sir Francis Galton* em 1885, consiste num método estatístico utilizado para modelar relações entre duas ou mais variáveis. Deste modo, é possível identificar quais apresentam um maior impacto relativamente a um tema de interesse, predizendo o valor da variável dependente (também conhecida como de resposta), partindo de um conjunto de variáveis dependentes (ou preditoras). Existem vários tipos de regressão, sendo a mais conhecida a regressão linear[34]. De seguida, são apresentados três métodos baseados em regressão, propostos para homogeneizar séries climáticas.

Two-Phase Regression

O modelo *Two-Phase Regression* (TPR) foi proposto por *Solow* em 1987, que criou uma técnica capaz de detetar um ponto de mudança na tendência de uma série temporal. É um método baseado em regressão, que surge frequentemente na literatura, como proposta de homogeneização de séries climáticas. Neste método, as linhas de regressão antes e depois do ano que está a ser testado são obrigadas a encontrar-se nesse ponto. Mais tarde, em 1995, *Easterling* e *Peterson* modificaram a técnica anterior para que as duas linhas de regressão não precisassem de se encontrar no ponto da descontinuidade: para um determinado ano, ou período de tempo, é ajustada uma linha de regressão à série de referência para o intervalo de tempo anterior desse ano, e a segunda linha é ajustada à segunda parte da série temporal. Este processo repete-se para todos os anos da série (com um mínimo de 5 anos em cada secção), e o ano com a soma residual mais baixa dos quadrados entre as duas funções de regressão é considerado o ano de uma possível descontinuidade. Caso a mesma seja significativa, a série é subdividida em duas, nesse ano, e cada uma é testada de modo semelhante até não serem encontradas descontinuidades significativas ou até que as séries sejam demasiado curtas para serem testadas [10, 43].

Multiple Linear Regression

Vincent em 1998 propôs o método, baseado em regressão, de homogeneização de *Multiple Linear Regression* (MLR), baseada em quatro modelos de regressão linear. O primeiro modelo determina se a série candidata é ou não homogénea durante o intervalo de tempo testado, e, se for (caso os resíduos da regressão forem variáveis independentes com média zero e variância constante), o teste termina e os restantes modelos não serão utilizados. Se porventura não forem encontradas homogeneidades, estima-se um segundo modelo que determine a existência de uma tendência geral na série em estudo e se a autocorrelação nos resíduos deste segundo modelo existir, então o modelo é descartado e um terceiro é calculado para aumentos sequenciais no momento em que uma etapa pode ocorrer. O último modelo é aplicado para definir a existência de tendências antes e depois do *break* identificado. Se a não homogeneidade encontrada no primeiro modelo não for uma tendência global, o terceiro modelo é aplicado para identificar a mudança de etapa única.

Na hipótese de os quatro modelos terem sido aplicados, existe a indicação de múltiplas não-homogeneidades na série testada. Nesse caso, a série temporal é subdividida na posição do *break* identificado e cada parte será testada separadamente a partir do primeiro modelo [10, 43]. O método de MLR é considerado como um dos mais robustos ainda que testes de eficiência tenham demonstrado, em comparação a outros métodos objetivos, uma capacidade de detecção inferior [49].

Cumulative Residuals

O método de *Cumulative Residuals* é, igualmente aos dois testes anteriormente referidos, baseado em regressão. Este relaciona conjuntos de dados de duas estações meteorológicas de referência: uma série temporal candidata (variável dependente) e dados de uma estação vizinha (variável independente) [10]. As séries candidatas podem ser consideradas homogêneas se os resíduos cumulativos do seu modelo simples de regressão linear não forem enviesados, isto é testado recorrendo a uma elipse que defina os limites de confiança: se os resíduos acumulados em relação ao tempo, e utilizando a escala temporal da variável em análise, estiverem dentro da elipse, então a hipótese de homogeneidade não é rejeitada para esse nível de significância. Para que isto se verifique vai depender do tamanho do conjunto de dados, do desvio padrão da amostra testada e da probabilidade usada para testar essa hipótese (normalmente de 80%) [10, 49].

Mann-Kendall Test

O teste de *Mann-Kendall*, proposto pela primeira vez por Mann em 1945 e estudado mais tarde por Kendall em 1975 [49], conhecido tipicamente por M-K, tem sido popularmente utilizado para detetar tendências e o seu grau de significância, em séries temporais de dados hidrológicos, nomeadamente de precipitação. Trata-se de um teste não paramétrico, recomendado pela Organização Meteorológica Mundial (WMO), uma vez que não obriga a que os dados sigam uma determinada distribuição.

O teste de *Mann-Kendall* analisa as diferenças entre os sinais de observações anteriores e posteriores. Ou seja, se uma tendência estiver presente, então os valores das observações aumentarão constantemente ou diminuirão de igual modo, comparando cada valor com aquele que o precede na série temporal.

A hipótese nula deste teste afirma a inexistência de uma tendência crescente ou decrescente ao longo do tempo. Já a hipótese alternativa indica a presença de uma tendência, podendo esta ser positiva, negativa ou não nula.

Para aplicar este teste é necessário: garantir que os dados não sejam dependentes, ou seja, que exista alguma correlação entre eles, caso haja, a capacidade para o teste avaliar corretamente a significância da tendência será afetada; e certificar que existe apenas uma observação por registo de tempo, caso contrário deve ser utilizado o valor da mediana.

Uma das limitações deste teste reside na hipótese de não detetar nenhuma tendência no caso de haver poucos dados, quando na verdade a encontraria caso o conjunto de dados disponíveis fosse maior. Assim sendo, quanto maior o número de dados, maior a probabilidade de encontrar uma tendência correta [75].

A hipótese nula deste teste diz que os dados são independentes e distribuídos de forma idêntica, enquanto que a hipótese alternativa assume que os dados seguem uma tendência monotónica. A estatística deste teste [44] é calculada da seguinte forma:

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^n \text{sgn}(Y_j - Y_k) \quad (2.20)$$

onde

$$\text{sign}(x) = \begin{cases} 1 & \text{se } x > 0 \\ 0 & \text{se } x = 0 \\ -1 & \text{se } x < 0 \end{cases} \quad (2.21)$$

e média de S igual a $E(S) = 0$.

Os processos descritos nesta secção pretendem representar a variedade de abordagens utilizadas para o mesmo propósito, tendo em conta que muitas são semelhantes ou variações umas das outras, diferindo em aspetos técnicos ou na resolução temporal dos dados. Na Tabela 2.5 pode observar-se uma descrição sumária de algumas características dos testes anteriormente abordados.

2.5.3 Processo de homogeneização

A homogeneização de uma série temporal pode ser definida como o processo de detetar e corrigir as não homogeneidades. No entanto, é consensual a importância de metadados, e especialistas na área dos dados originais, para se proceder à correção dessas irregularidades [49]. Uma das definições do processo de homogeneização, depois de ter sido considerada a presença de *outliers*, identificados de forma manual e visual, reside na aplicação dos seguintes passos [39]:

1. Deteção de possíveis *breaks* nos dados originais da série temporal, recorrendo a análise visual dos mesmos e à aplicação de testes como os referidos na Subsecção 2.5.2;
2. Comparação das datas, onde foram identificados os *breaks*, com metadados disponíveis e possíveis ocorrências de eventos climáticos, tais como, erupções vulcânicas, ou fenómenos como o *El Niño* (fenómeno natural que provoca o aquecimento das águas do Oceano Pacífico), procurando possíveis correspondências;
3. Seleção dos *breaks* de origem não climática, na série temporal, para serem sujeitos a correção;
4. Correção dos *breaks* não-climáticos.

Tabela 2.5: Sumário das principais características de alguns métodos de homogeneização, classificados por tipo de abordagem (paramétricos, não-paramétricos, baseado em probabilidade ou de regressão) [49].

Método	Tipo	Resolução temporal	Deteção de <i>breakpoints</i>	Observações
<i>Buishand Range Test</i> (1982)	Estatístico Paramétrico	-Mensal -Anual	- <i>Single break-point</i> -Identifica a data do <i>break</i>	-Mais sensível a <i>breaks</i> no meio da série temporal -Usado como método de deteção absoluto
<i>Pettitt Test</i> (1979)	Estatístico Não-paramétrico	-Mensal -Anual	- <i>Single break-point</i> -Identifica a data do <i>break</i>	-Mais sensível a <i>breaks</i> no meio da série temporal -Usado como método de deteção absoluto
<i>Standard Normal Homogeneity Test</i> (SNHT) (1986)	Probabilístico	-Mensal -Anual	- <i>Single break-point</i> -Identifica a data do <i>break</i>	-Um dos métodos mais populares como método de deteção -Comummente aplicado a séries compostas
<i>Standard Normal Homogeneity Test with Trend</i> (SNHT) (Alexandersson e Moberg 1997)	Probabilístico	-Mensal -Anual	- <i>Single break-point</i> -Identifica a data do <i>break</i>	-Deteção de não-homogeneidades do tipo tendência -Comparação baseada nas séries de referência
<i>Von Neumann</i> (1941)	Estatístico Não-paramétrico	-Mensal -Anual	- <i>Single break-point</i> -Não identifica a data do <i>break</i>	-Usado como método de deteção absoluto -Qualifica o nível de homogeneidade da série
<i>Two-phase regression</i> (1995)	Regressão	-Mensal -Anual	- <i>Single break-point</i> -Identifica a data do <i>break</i>	-Método de deteção relativo -Método hierárquico para identificar múltiplos <i>breakpoints</i> (algoritmo de corte) Alto desempenho na capacidade de deteção
<i>Multiple linear regression</i> (1998)	Regressão	-Mensal -Anual	- <i>Single break-point</i>	-Deteção de <i>breaks</i> e não-homogeneidades do tipo tendência -Método de deteção objetivo -Método de deteção relativo
<i>Method of cumulative residuals</i> (1998)	Regressão	-Mensal -Anual	- <i>Single break-point</i>	-Método de deteção relativo -Utilizado para análise exploratória -Qualifica o teste de homogeneidade
<i>Mann-Kendall test</i> (Mann, 1945; Kendall, 1975)	Estatístico Não-Paramétrico	-Mensal -Anual	-	-Deteta tendências significativas nos dados

No caso das séries temporais sob estudo, se basearem em observações mensais são sugeridos em [39] os seguintes passos para proceder à correção dos *breaks* não-climáticos:

1. Para cada *break* (t_{break}), começando no último, na escala temporal considerada, para o primeiro:
 - (a) Seleção de um intervalo de tempo Δt em torno do *break*, tendo em consideração a duração dos períodos homogêneos antes e depois desse *break*;
 - (b) Cálculo dos valores médios do parâmetro em causa ($\langle T \rangle$), para cada mês, separadamente, durante os dois intervalos de tempo: antes do *break* ($\langle T \rangle = t_{break} - \Delta t$) e depois do *break* ($\langle T \rangle = t_{break} + \Delta t$);
 - (c) Cálculo das correções (dT) para cada mês separadamente como sendo a diferença dessas médias, isto é, $dT = \langle T \rangle (t_{break} + \Delta t) - \langle T \rangle (t_{break} - \Delta t)$;
 - (d) Suavização dos 12 valores de correção mensais dT pela média adjacente de 3 meses, procurando alcançar uma variação razoável de dT ao longo do ano;
 - (e) São ignorados todos os dt cujo valor seja inferior a erros instrumentais;
 - (f) Correção dos dados nos períodos antes do *break* em causa, recorrendo ao dT para cada mês.

2. Depois de aplicados os passos referidos acima, procede-se de igual modo para o *break* anterior, no tempo;
3. Finalmente, é realizada uma análise visual e são aplicados novamente testes de homogeneização, aos dados já corrigidos.

Os passos anteriormente referidos, retratam o processo de homogeneização, encontram-se retratados na Figura 2.19.

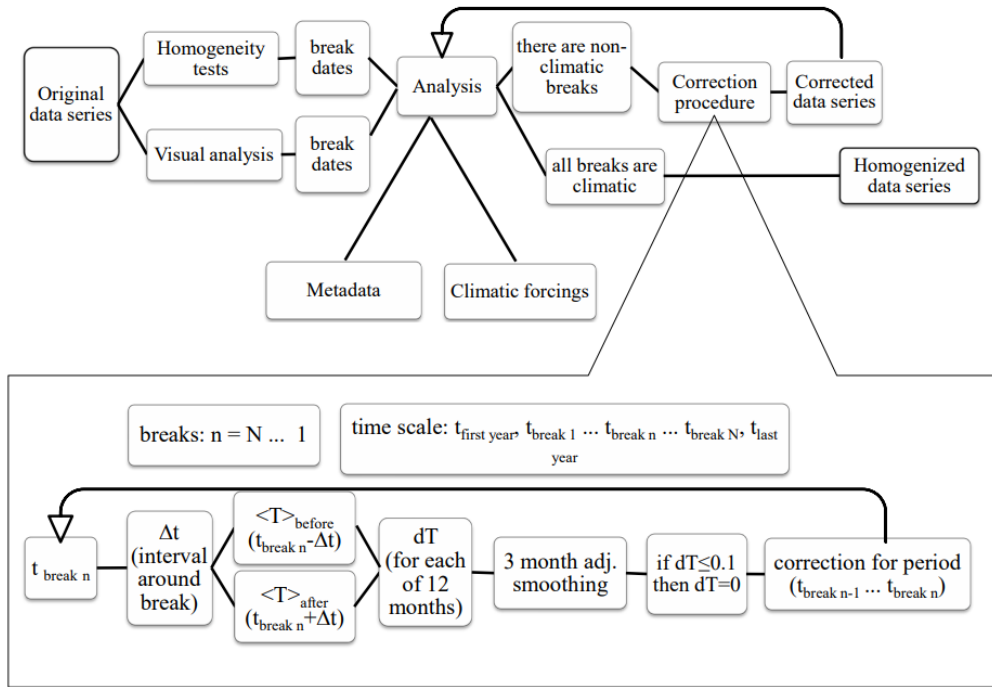


Figura 2.19: Processo de homogeneização. No topo da figura encontram-se os passos iniciais e em baixo o processo de correção de não-homogeneidades não climáticas [39].

2.5.4 Classificação e avaliação qualitativa dos testes

Depois de aplicados os diferentes testes a um determinado conjunto de dados segue-se uma avaliação qualitativa dos resultados obtidos pelos mesmos. De acordo com [71], a classificação pode ser feita dependendo do número de testes que rejeitem a hipótese nula. São distinguidas as seguintes categorias:

- Classe 1: *useful* – se um nenhum teste rejeitar a hipótese nula, no nível 1%;
- Classe 2: *doubtful* – se dois testes rejeitarem a hipótese nula, no nível 1%;
- Classe 3: *suspect* – se três ou mais testes rejeitarem a hipótese nula, no nível 1%.

Já no que diz respeito à interpretação qualitativa, [71] considerou:

- Classe 1: *useful* – Sem sinal aparente de não-homogeneidades da série, ou, se existirem, serão suficientemente pequenas para serem detetadas. A série mostra-se legível para aplicar uma análise de tendência;
- Classe 2: *doubtful* – implica a presença de não-homogeneidades na série. A análise da tendência deve ser feita com um ponto de vista crítico dada a anterior possibilidade;
- Classe 3: *suspect* – alta probabilidade de existirem não-homogeneidades na série. A análise da tendência deve ser considerada suspeita e apenas tendências muito grandes podem estar relacionadas com um sinal climático.

Séries que pertençam à classe 3 *suspect* não devem ser submetidas a testes que detetem tendências, como *Mann-Kendall*. Ainda assim, procurando melhorar a classificação obtida, existem métodos para corrigir as não-homogeneidades encontradas [71].

2.6 Ferramentas

Nesta dissertação será efetuada uma análise exploratória aos dados meteorológicos recolhidos pelo IGUP entre 1960 e 2007. Pretende-se efetuar um estudo das respetivas séries temporais, tendo em vista estabelecer correlações entre as mesmas e compreender melhor a evolução do clima da cidade do Porto. Para este efeito, será utilizada a linguagem R [46] em conjunto com algumas ferramentas que a mesma disponibiliza, facilitando todo o processo de estudo e análise atrás referido.

R é uma linguagem de programação *open-source*, intuitiva e focada na computação estatística e gráfica visando facilitar a manipulação, análise e visualização de dados. Todas as funções e conjuntos de dados do R encontram-se armazenados em *packages* com as mais diferentes funcionalidades. Atualmente a linguagem R tem mais de 150 000 *packages* disponíveis, desenvolvidos pela comunidade de utilizadores, e que potenciam e melhoram as funcionalidades base do R existentes ou adicionando novas, através das coleções de funções criadas. Existem alguns *packages* criados para a análise de séries temporais, sendo que alguns destes se focam no estudo de séries meteorológicas, disponibilizando funcionalidades úteis para o estudo do clima.

No Apêndice D são apresentados, de forma sucinta, todos os *packages* e elencadas as principais funções utilizados no contexto desta dissertação.

Capítulo 3

Análise Exploratória de Dados Meteorológicos da Cidade do Porto

Neste capítulo, é efectuada uma descrição do conjunto de dados alvo de estudo e dos principais passos do seu pré-processamento. Depois do pré-processamento os dados, é realizada uma análise exploratória apoiada na análise de gráficos que mostram o comportamento das variáveis, tendo em conta diferentes particularidades e estudando possíveis correlações. Por fim, realizamos um estudo do comportamento das principais componentes que caracterizam uma série temporal: sazonalidade, tendência e componente restante. Este tipo de análise tem como objetivo entender de que modo estas variáveis se comportam ao longo do tempo e caracterizar essa evolução.

3.1 Caracterização do conjunto de dados

O conjunto de dados a ser analisado foi registado pelo Instituto Geofísico da Universidade do Porto (IGUP). Este instituto acumula funções de investigação e ensino, integrando a Rede Meteorológica Nacional, para a qual fornece, diariamente dados sobre diversas variáveis meteorológicas [45].

Sob estudo estão dados relativos aos séculos XX e XXI, no período de 1960 a 2007 e foram disponibilizados dados das variáveis: temperatura do ar, precipitação, pressão atmosférica, e velocidade do vento, humidade relativa do ar, insolação, tensão do vapor, termómetro molhado. No entanto, neste projeto, a análise focou-se nas primeiras quatro variáveis referidas, procurando deste modo caracterizar, no período de tempo considerado, a sua evolução. Os diferentes conjuntos de dados encontram-se registados em formato Excel, correspondendo cada ficheiro a um ano de registos, de uma determinada variável. Nestes ficheiros, divididos por *sheets*, estão os meses de cada ano com os dados, de hora a hora, por cada variável. Na Figura A.1, no Apêndice A, encontra-se um printscreen de uma *sheet* de um ficheiro Excel, com informação da velocidade do vento relativa a julho de 2007.

Para cada variável, alguns dos ficheiros Excel continham, para além dos registos horários, médias diárias (das 0 às 24 horas), médias diurnas (das 9 às 20 horas) e noturnas (da 1 às 8, e das 21

às 24 horas, do mesmo dia), assim como observações máximas e mínimas diárias, e respetiva amplitude. Do mesmo modo, apenas em alguns ficheiros, referentes à mesma variável, foram registados alguns cálculos relativos ao mês em questão, especificamente extremos do mês, assim como médias mensais e diárias, e referentes a três décadas. No entanto, nem todas as variáveis de um determinado ano, ou, inclusive, anos completos, continham registos destas estatísticas. Por este motivo, estes valores foram posteriormente calculados, para as quatro variáveis em questão, com base nos dados de periodicidade horária já existentes.

O intervalo de tempo de dados disponíveis varia de variável para variável. A Tabela 3.1 indica o período de observações para cada variável.

Tabela 3.1: Variáveis e respetivos períodos de registo de observações.

Variável	Período de observações
Temperatura	1960-1976
	1980-1995
	2000-2007
Precipitação	1997-2003
	2005-2006
Pressão Atmosférica	1999-2007
Velocidade do vento	1994-2005
	2007

Ao longo de todo o conjunto de dados é possível identificar valores em falta em períodos de algumas horas, dias ou meses, por vezes justificada (por exemplo, por avarias de equipamento), outras vezes não sendo apresentada nenhuma justificação ou motivo para a ausência de dados.

3.2 Pré-processamento de dados

Numa primeira fase, procedeu-se ao pré-processamento dos dados fornecidos pelo IGUP, relativamente às variáveis anteriormente mencionadas, no respetivo período de tempo disponível. O pré-processamento de dados consiste numa das etapas mais importantes no que concerne ao estudo e análise de dados, uma vez que toma os dados "em bruto", isto é, tal como são recebidos, muitas vezes com erros, inconsistências e/ou incompletos, e altera e transforma-os de modo a serem estudados e prontos para se proceder à análise exploratória dos mesmos.

Começou-se por importar os dados criando uma série temporal para cada variável em estudo, reportando diferentes intervalos de tempo (cf. Tabela 3.1). A Tabela 3.2 mostra o número de observações de cada uma das séries temporais constituídas, antes serem feitas quaisquer alterações, a quantidade de *missing values* e respetiva percentagem, assim como tamanho e percentagem daqueles que foram corrigidos, e o tamanho das séries depois de feitas estas alterações. As observações registadas foram dispostas, em cada série temporal, tendo em conta o dia da observação assim como a sua hora respetiva, visto tratarem-se de registos de hora a hora. Todos

os valores obtidos foram convertidos em tipo numérico e os dados foram transformados de modo a que as datas e horas das observações se encontrem em intervalos de tempo regulares. No caso da pressão atmosférica, a todas as observações registadas (unidades milímetro de mercúrio (*mmHg*)), foi subtraído o valor 700, possivelmente por motivos de comodidade no registo em papel. Aquando da importação destes valores para o R, adicionou-se novamente o valor de 700 a cada observação.

Posto isto, procedeu-se então ao pré-processamento dos dados, onde foram detetados os *missing values*, erros de importação, e *outliers* globais, contextuais e extremos, nas séries temporais das variáveis em estudo. Nesse sentido, podem ser encontrados no Apêndice A algumas informações úteis e que complementem o referido no pré-processamento.

3.2.1 Valores desconhecidos

Habitualmente conhecidos como *not available values* (NAs), os valores desconhecidos ou *missing values* são aqueles cujo registo de uma certa variável se encontra em falta num determinado dia ou, neste caso, hora.

No que respeita à correção de NAs nas séries temporais consideradas, foram inicialmente removidos os casos que registassem mais de dois *missing values* consecutivos, uma vez que, na hipótese de ausência de registos durante dias, meses ou, inclusive anos completos, dificilmente seria possível corrigir estas situações corretamente. Posto isso, foi aplicada uma função capaz de corrigir os restantes valores em falta, que tomou em consideração a presença de sazonalidade na série temporal, através da decomposição STL (*Seasonal and Trend decomposition using Loess*), fazendo interpolação linear [12].

Na Tabela 3.2 pode observar-se o número de observações iniciais, considerando séries temporais regulares no tempo, e a respetiva quantidade e percentagem de *missing values* antes de qualquer transformação, assim como quantidade e percentagem de *missing values* que foram submetidos a correção e posterior total de observações que caracterizam a série temporal. A precipitação foi a única variável onde não foram realizadas alterações no que diz respeito aos *missing values* uma vez que não foi possível distinguir se seriam valores em falta ou valores de precipitação nula. Relativamente às restantes variáveis, a pressão atmosférica foi aquela onde mais *missing values* foram corrigidos (0.06 % face ao tamanho de dados iniciais).

Tabela 3.2: Tamanho da série temporal antes e depois de corrigidos os *missing values*.

Variável	Nº de observações inicial	Número (%) de <i>missing values</i>	Número (%) de <i>missing values</i> corrigidos	Nº de observações final
Temperatura	420768	65455 (15.6%)	93 (0.0262%)	355406
Precipitação	79632	-	-	79632
Pressão Atmosférica	78888	1726 (2.19%)	49 (0.0635%)	77211
Velocidade do Vento	122712	44950 (36.6%)	21 (0.027%)	77783

3.2.2 Erros de consistência

Aquando da importação dos dados em "bruto", dos ficheiros Excel para o R, foram encontrados alguns erros de consistência nos dados. Essas situações foram documentadas, para que depois pudessem ser corrigidas, salvaguardando a correta importação de todos os registos disponíveis, para cada variável.

Foram encontrados três tipos de erros de consistência: não correspondência do mês/ano escrito na *sheet* com o ano a que reportava o ficheiro; ausência ou excesso de dias num determinado mês; situações onde a ordem das *sheets*, com os nomes dos meses, não seguia a ordem cronológica correta.

Na Tabela 3.3 encontram-se o tipo de erros encontrados para a variável velocidade do vento, a *sheet*, assim como o ano do ficheiro, onde o mesmo se verifica

Tabela 3.3: Erros de consistência na velocidade do vento.

Ficheiro (ano)	Sheet	Tipo de erro
2001	3	o mês da sheet não coincide
2003	2	ausência de dias

A pressão atmosférica foi a única variável que não registou nenhum erro de consistência. Para as variáveis de temperatura e precipitação esses erros encontram-se reportados no Apêndice A.

3.2.3 Outliers globais e contextuais

Um *outlier* é considerado global se o seu valor se distanciar demasiado, da totalidade do conjunto de dados em que se insere. Para designar uma determinada observação como sendo um *outlier*, foi definido um intervalo aceitável de valores para cada variável (cf. Tabela 3.4). No caso de algum valor se encontrar acima ou abaixo desse intervalo esperado, esse registo é considerado um *outlier* global.

Apenas se registou a ausência de *outliers* globais nos registos da precipitação. Os restantes encontram-se sumariados na Tabela 3.5 e foram igualados a NAs para posteriormente serem tratados como os *missing values* e, se possível, corrigidos com o respetivo valor expectável. Pode observar-se, nessa tabela, que os *outliers* registados na pressão atmosférica no ano de 2005 e no dia 31 de julho de 2004 são, provavelmente, derivados de um erro de inserção nas folhas Excel. Tendo em conta que às observações desta variável foram subtraídas 700 unidades, possivelmente estes valores foram a exceção dessa operação, resultando numa inconsistência face aos demais.

Ainda atendendo a valores fora do expectável, uma determinada observação é considerada como *outlier* contextual se o seu valor se desviar significativamente das demais observações, no mesmo contexto. Esse mesmo valor, pode não ser considerado como outlier se ocorrer num contexto diferente. Isto é, para um contexto temporal, um determinado valor pode ser outlier em junho mas não em dezembro.

Tabela 3.4: Limites dos *outliers* globais para cada variável.

Variável	Limite Inferior	Limite Superior
Temperatura	-5°C	45°C
Velocidade do Vento	0	120
Precipitação	0 mm	100 mm
Pressão Atmosférica	712.6 mmHg	787.6 mmHg

Tabela 3.5: *Outliers* globais encontrados para cada variável.

Variável	Data	Hora	Valor Suspeito
Temperatura	2001/03/01	4h	74 °C
Velocidade do Vento	2000/10/04	16h	156 Km/h
	2004/02/19	3h	1207 mmHg
	2004/07/21	2h	1222.6 mmHg
	2004/07/31	12h	705.31 mmHg
	2004/10/10	4h	1149.8 mmHg
Pressão Atmosférica	2005/07/02	6h	705.7 mmHg
	2005/12/15	15h	787.68 mmHg
	2005/12/16	15h	789.33 mmHg
	2005/12/17	4h	788.6 mmHg
	2005/12/17	15h	787.6 mmHg

Também este tipo de *outliers* foi detetado nesta fase de pré-processamento. Representando as variáveis graficamente, torna-se perceptível identificar alguns desses casos. Observando o gráfico da Figura 3.1 é possível identificar o exemplo de um valor de temperatura de 44.9 °C, que, apesar de ser reportado no mês de agosto, trata-se de um valor muito acima dos habituais obtidos nessa altura do ano e, portanto, é considerado um *outlier* nesse contexto.

Considerando o mês como contexto, e com o objetivo de identificar *outliers* extremos, isto é, aqueles que excedem os limites $Q_1 - 3 \times IQR$ e $Q_3 + 3 \times IQR$, para cada variável, foram obtidos os *boxplots* da distribuição dos valores por mês. Na Figura 3.2 encontram-se os *boxplots* obtidos para a variável temperatura, agrupados por mês ao longo de todo o período de dados. Tal como se pode observar na figura, no caso da temperatura, é possível identificar *outliers* extremos nos meses de maio, junho, agosto, setembro e um, exceccionalmente, no mês de fevereiro.

Procurando ter em conta a sazonalidade presente nas séries temporais destas variáveis, foram identificados os outliers que, depois de aplicada uma decomposição STL, excedessem os limites $\pm 2(Q_{0.9} - Q_{0.1})$, onde Q_p é o p -quantile dos resíduos. Estes resultados tiveram em consideração a distribuição dos valores por mês, e podem ser encontrados no Apêndice A, para cada variável e a respetiva sugestão de correção que lhe corresponde [26]. Apesar de ser um método útil no que diz respeito à identificação de *outliers* e respetiva proposta de correção, não seria possível comprovar a sua veracidade, pelo que, neste projeto, optou-se apenas por identificá-los.

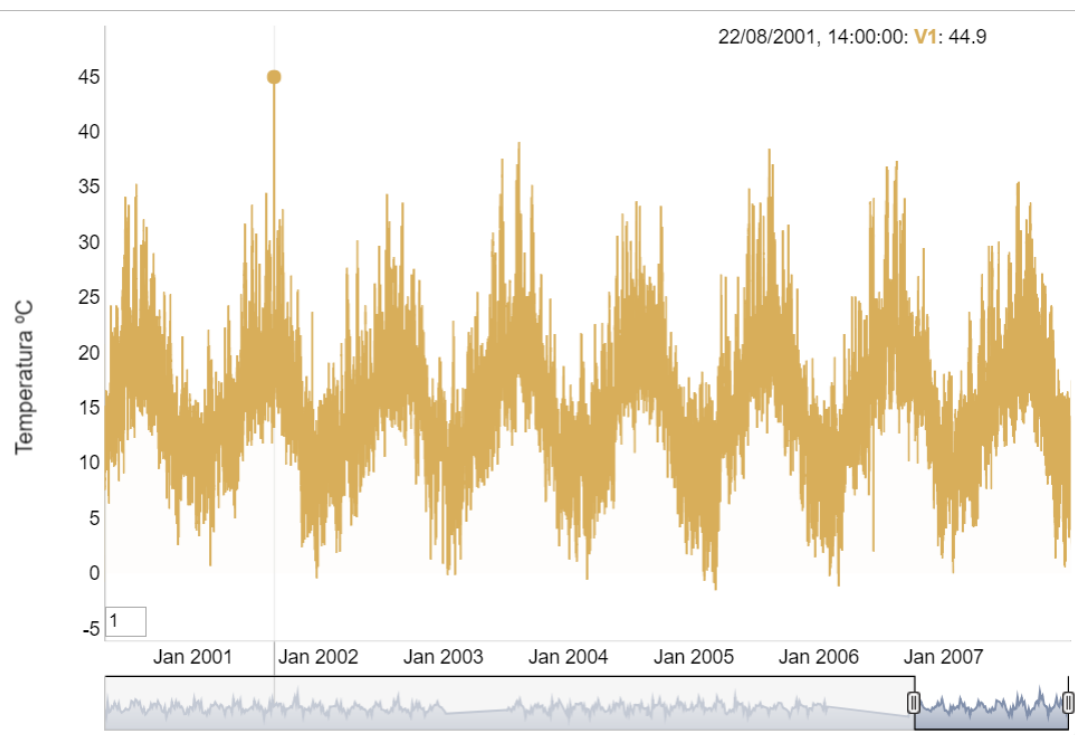


Figura 3.1: Representação da temperatura onde é possível identificar outliers.

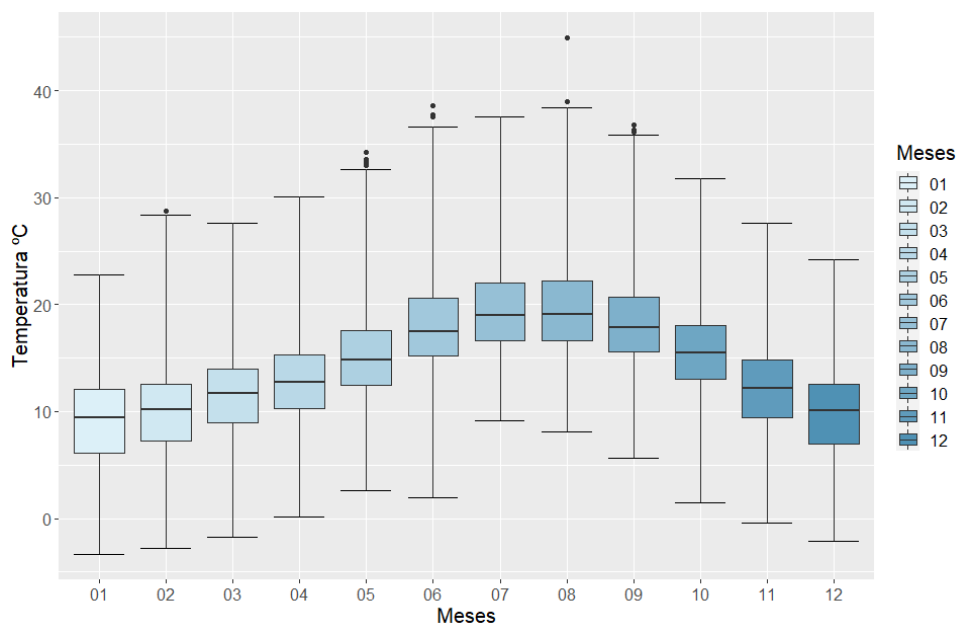


Figura 3.2: Boxplot da temperatura por mês para detecção de *outliers* contextuais

3.3 Análise de dados univariada

Procurando entender o comportamento de cada variável meteorológica ao longo do período de tempo registrado, foi realizada uma análise por cada uma das variáveis. Essa análise permite visualizar graficamente de que modo cada variável evolui ao longo do tempo, tendo em conta

diferentes particularidades que o caracterizam, nomeadamente, estações do ano, dias, meses, anos e respetivas médias.

Para cada variável foram calculados alguns atributos, com base nos dados já existentes, que permitem uma melhor caracterização das mesmas, nomeadamente a estação do ano, se o valor em causa é diurno ou noturno, e ainda o dia da semana a que corresponde. Para perceber se uma determinada observação corresponde ao dia ou à noite do respetivo dia, tiveram-se em conta as horas do nascer e do pôr do sol, consoante a localização dos ditos registos. Para tal, a latitude e longitude consideradas foram as da cidade do Porto.

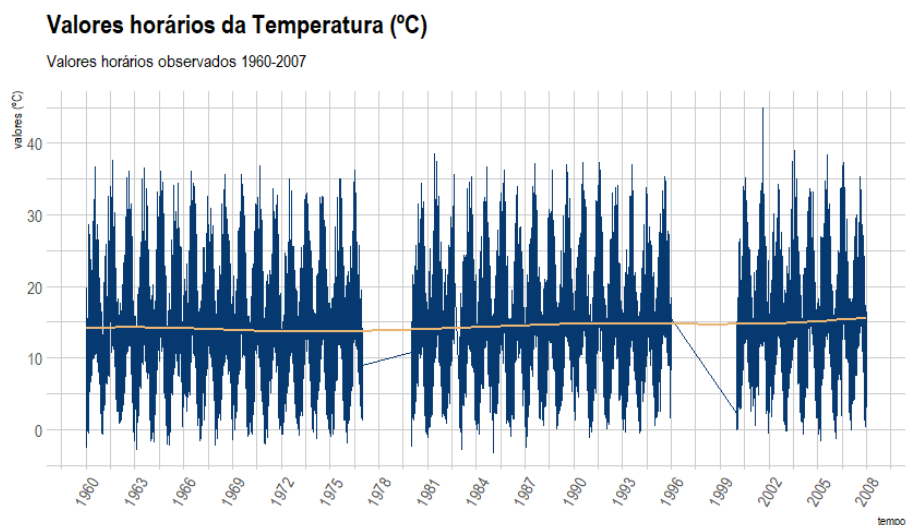
Em alguns gráficos foi desenhada uma linha que descreve a tendência. Para esse efeito, o método escolhido irá variar consoante o tamanho das observações: a suavização *loess* (*Locally Estimated Scatterplot Smoothing*) é utilizada apenas para conjuntos de dados com menos de 1000 observações, no entanto, este método tem complexidade $O(n^2)$. Assim, e para grandes conjuntos de dados, é utilizado o método GAM (*Generalized Additive Model*).

3.3.1 Temperatura

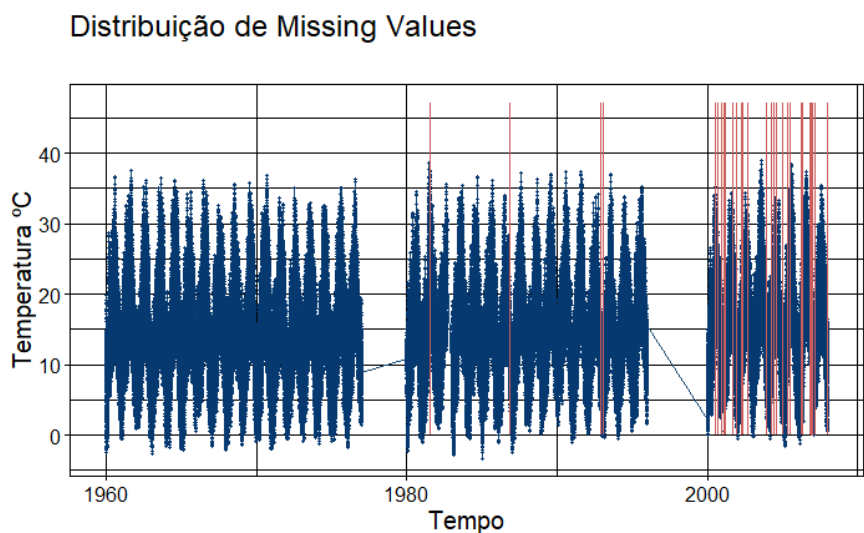
Inicialmente, foram representados os registos de temperatura de hora a hora (cf. Figura 3.3a), procurando perceber como os mesmos se distribuíam ao longo do tempo. Foi possível comprovar a ausência de registos entre 1976-1980 e 1995-2000. Ainda assim, e observando o comportamento da tendência, poucas são as conclusões perceptíveis pelos registos horários, pelo que se procedeu à visualização dos dados noutras escalas temporais.

Na Figura 3.3b pode observar-se, identificado a vermelho, a distribuição de *missing values* na série de temperatura, depois de desconsiderados períodos de maior ausência de registos, isto é, mais de duas observações consecutivas em falta. Estes *missing values* foram corrigidos, através de interpolação linear, tendo em conta a sazonalidade presente na série. Conforme o constatado na Tabela 3.2, foram corrigidos 93 *missing values*, aproximadamente 0.03% da série em questão.

Tendo em conta que os valores iniciais estão registados de hora a hora, criou-se ainda um *heatmap* capaz de representar os dados com essa frequência. Na Figura B.3, no Apêndice B estão representados os anos de 1992 e de 2001, com os valores horários de temperatura. Através do mesmo foi possível identificar a ausência de registos durante algumas horas, como é o caso do dia 19 e 20 de novembro de 1992, (Figura B.3a) que não possui qualquer registo desde as 20 horas do dia 19 às 8 horas do dia 20, e ainda confirmar a presença de alguns dos *outliers* identificados anteriormente na Tabela A.4, no Apêndice A. Estes *outliers* foram facilmente identificados (Figura B.3b) em consequência da mudança no gradiente das cores de um determinado dia, como é o caso do dia 22 de agosto de 2001, às 13 horas, com a cor amarela. Ainda que algumas das ausências de valores identificadas não sejam justificadas, outras comprovaram-se mais tarde, através da informação contida nos ficheiros Excel, estarem relacionadas com avarias nos instrumentos de medição. Estes casos encontram-se sumariados na Tabela A.7, no Apêndice A.



(a) Valores da Temperatura (1960-2007), de hora a hora e linha da tendência.



(b) Distribuição de *missing values* na série de temperatura, submetidos a correção (1960-2007).

Figura 3.3: Valores da Temperatura (1960-2007).

Na Figura 3.4 estão representados os valores médios anuais de temperatura, médias anuais das temperaturas mínimas diárias, médias anuais das temperaturas máximas diárias, e os respectivos valores máximos e mínimos de cada ano, desde 1960 a 2007. Nesta figura é possível identificar o ano que registou a temperatura mais elevada e a mais baixa, 44.9°C em 2001 e -3.3°C em 1985, respetivamente. A temperatura mais elevada corresponde a um possível *outlier*, identificado na Tabela A.4, no Apêndice A.

Ainda que, através da Figura 3.4 não sejam perceptíveis as diferenças na evolução das médias das temperaturas máximas e mínimas, ao observar-se a Figura 3.5 constata-se um aumento das mesmas ao longo intervalo de tempo em causa (1960-2007), com especial atenção aos últimos 5 anos cujo gradiente de cores apresenta um aumento dos valores da temperatura. O mesmo

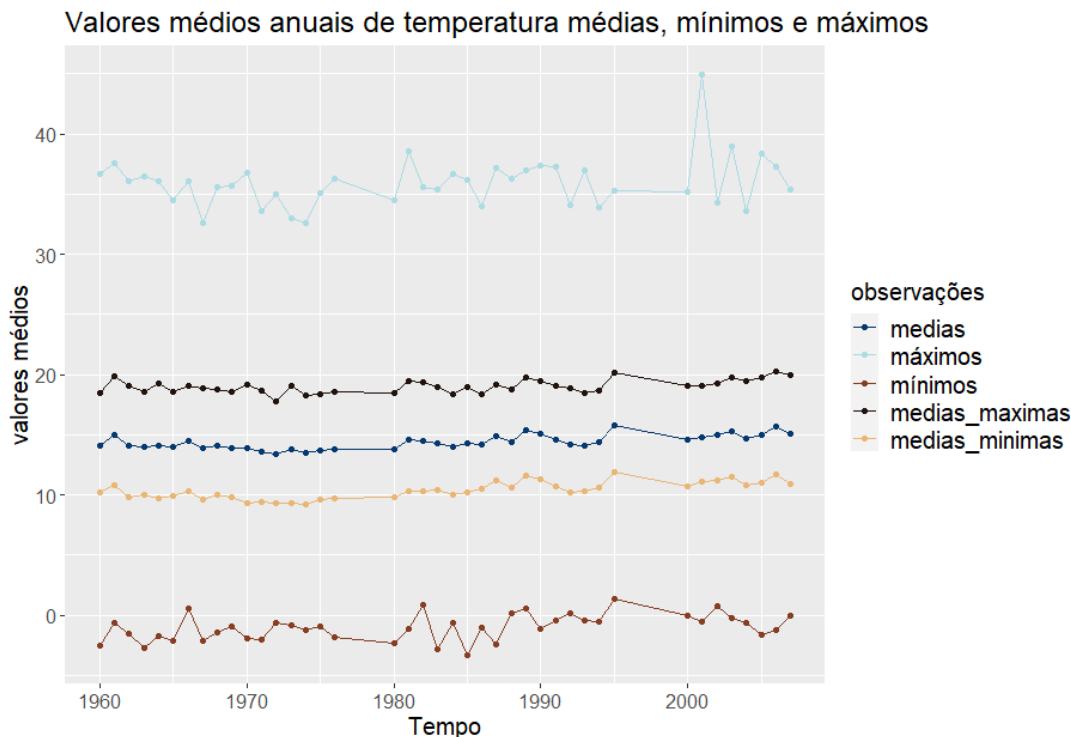


Figura 3.4: Evolução dos valores médios anuais de temperatura mínima, média e máxima (1960-2007).

se verifica quando se observa o gráfico de *warming stripes* respeitante às temperaturas médias anuais (cf. Figura B.1, no Apêndice B). Nas Figuras B.1 e 3.5, é possível perceber que o ano de 1961 regista temperaturas mais altas face às suas linhas vizinhas, podendo ser derivado à presença de *outliers* nesse ano, ou a alguma situação climática atípica. Independentemente disso, o aumento no gradiente de cores ao longo do período de tempo registado, sugere um aumento nos valores médios de temperatura ao longo dos anos.

Analisou-se de seguida, a série de um ponto de vista temporal mais reduzido, isto é, observando o comportamento dos valores de temperatura média, por mês no decorrer dos anos. Na Figura 3.6 encontram-se os gráficos que descrevem a distribuição dessas médias, e respetiva linha de tendência que as acompanha. Observando os mesmos, verificou-se um aumento de temperatura, de modo geral, em todos os meses ao longo dos anos, desde 1960 a 2007. No entanto, esse aumento é mais pronunciado nos meses mais quentes, nomeadamente junho, julho, agosto e setembro. Os meses de março, maio, setembro e dezembro, mostram um decréscimo na tendência, aproximadamente até meados de 1975, para depois aumentar até 2007.

Na Figura 3.7 está representada a distribuição das temperaturas mínimas e máximas diárias (Figuras 3.7a e 3.7b, respetivamente), para cada mês do período de tempo observado. No caso às temperaturas mínimas diárias, observa-se que os meses mais quentes apresentam uma distribuição de valores com menor variância do que os meses mais frios. Por outro lado, no caso das temperaturas máximas diárias, observa-se que os meses mais quentes apresentam uma distribuição de valores com maior variância do que os meses mais frios.

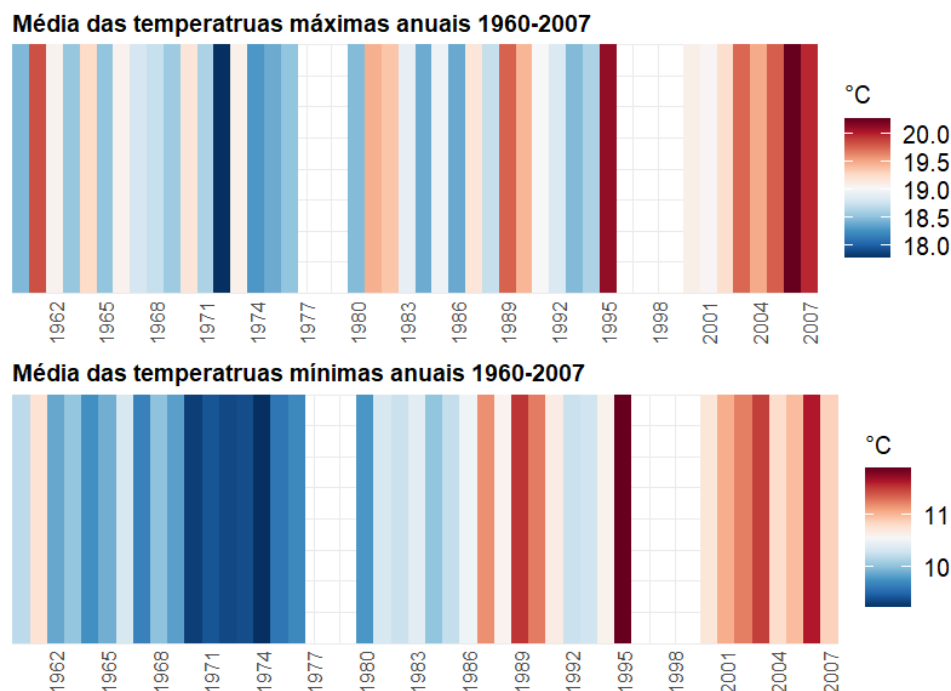


Figura 3.5: *Warming stripes* das temperaturas médias máximas e mínimas anuais (°C) (1960-2007).

Ao tentar perceber a influência dos parâmetros 'dia' e 'noite' nos valores observados de temperatura, as séries temporais com registos de hora a hora, diários e mensais, não foi possível obter resultados conclusivos. Posto isso, procurou perceber-se se, filtrando os valores por médias anuais, se as médias diurnas e noturnas se alteravam ao longo do período registado. Esses resultados encontram-se nas Figuras 3.8a e 3.8b, respetivamente, com os valores anuais médios de temperatura registados durante o dia e durante a noite. Atendendo aos mesmos, repara-se num decréscimo da tendência até 1975 (Figura 3.8a) e 1973 (Figura 3.8b) para, de seguida, aumentarem até 2007. Esta variação de tendência foi mais acentuada no caso das temperaturas noturnas. Os valores máximos e mínimos registados por estas séries podem ser observados na Tabela 3.6.

Tabela 3.6: Valores máximos e mínimos da temperatura média anual, diurna e noturna.

	Ano	Mínimo	Ano	Máximo
Temperatura Diurna	1972	14.9	1995	17.1
Temperatura Noturna	1974	11.5	1995	14.1

Observando a distribuição dos valores anuais de temperatura por estação do ano, foi possível aferir algumas alterações no comportamento da tendência destas séries. A Figura 3.9 mostra essas variações, ao longo do intervalo de tempo 1960 a 2007. Através da figura é possível notar uma descida da tendência até 1973 no inverno e outono, 1970 no caso do verão e 1975 na primavera, com uma descida mais significativa, para, de seguida, se reparar num aumento da mesma até final do período de tempo registado, à excepção do inverno que em 1995 volta a descer. Ainda assim, verificou-se uma variação mais acentuada da tendência no verão e no outono, que

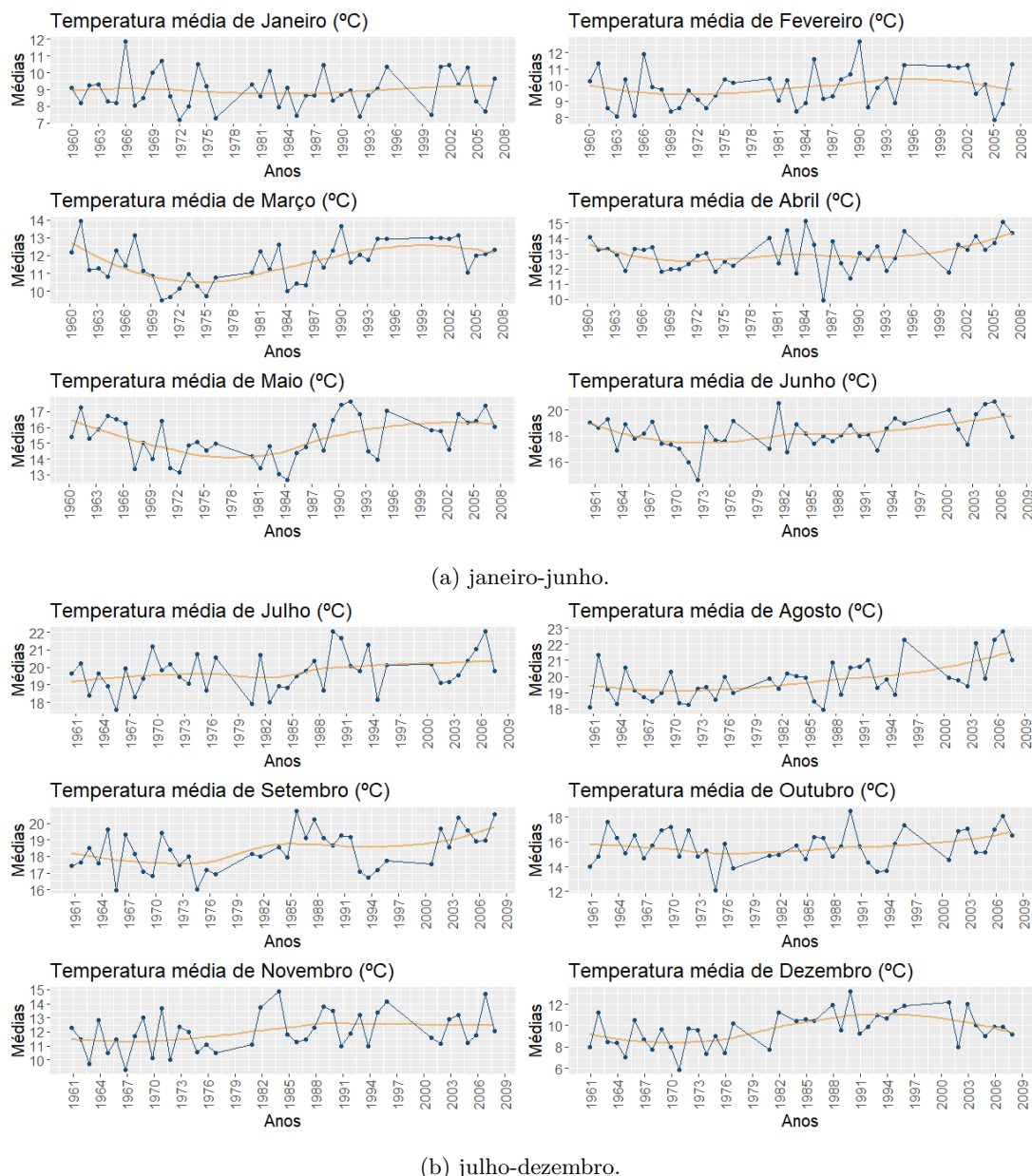
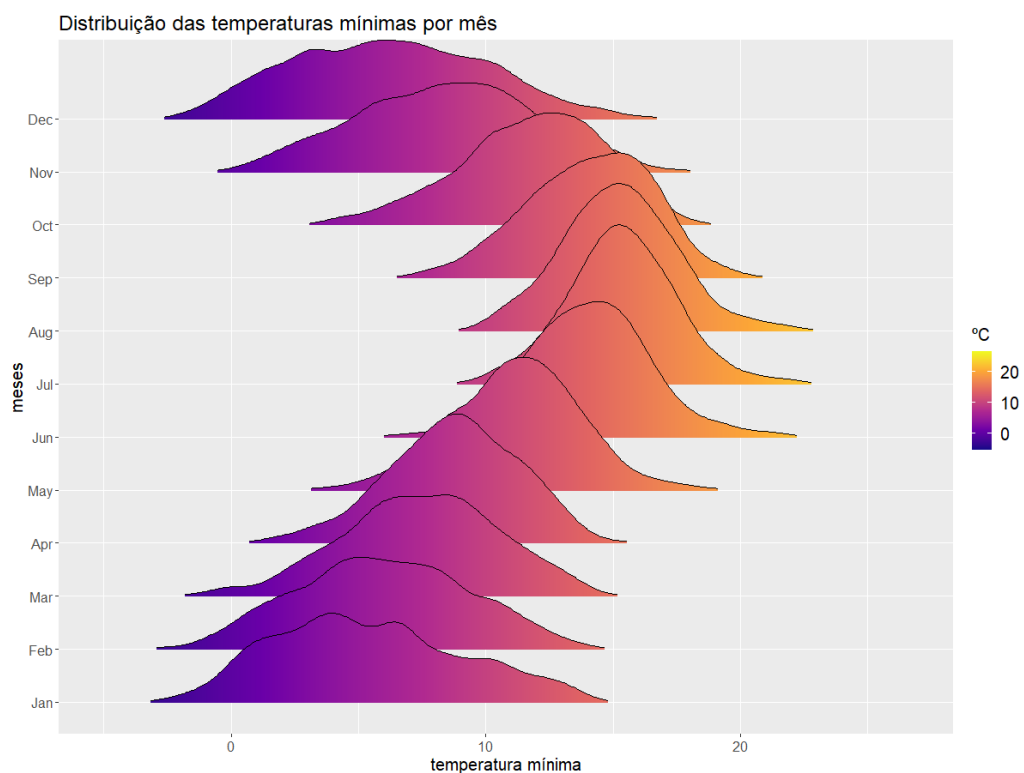


Figura 3.6: Evolução das temperaturas médias mensais por mês, e linha de tendência (1960-2007).

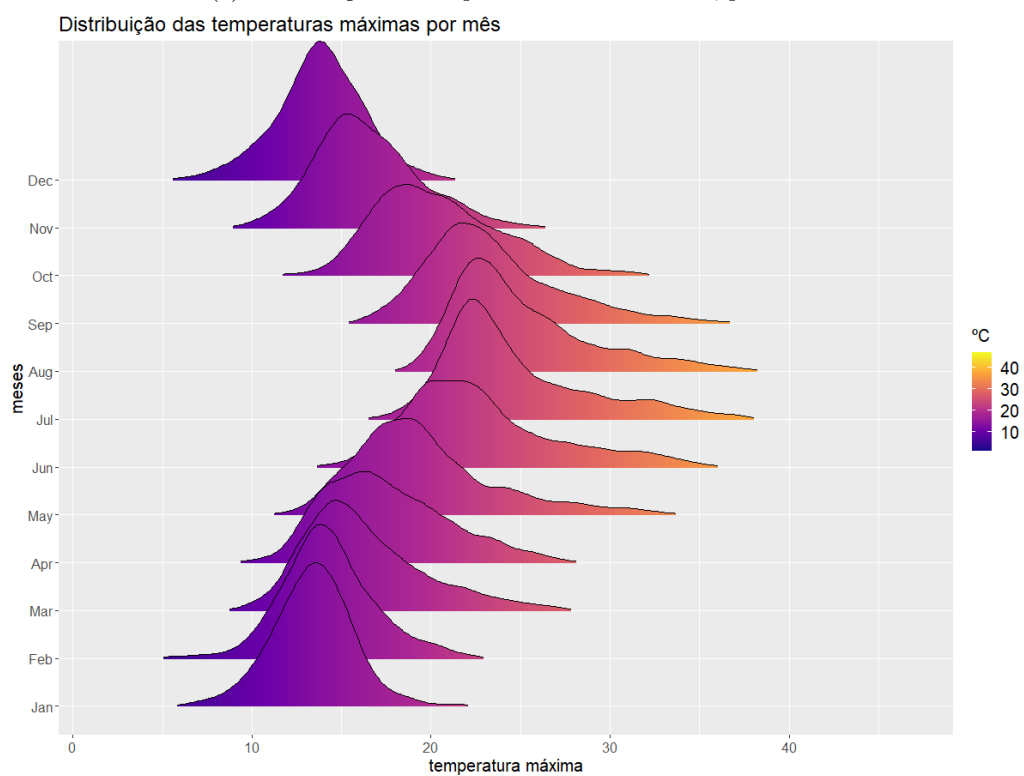
registraram o maior aumento de temperatura média, comparativamente às restantes.

As conclusões obtidas na Figura 3.9 podem ser comprovadas através da Figura B.2 (Apêndice B), que representa as médias anuais de temperatura por estação do ano através de um gráfico de *warming stripes* de 1960 a 2007. Mais uma vez, podemos verificar que existe um aumento de temperaturas ao longo do período de tempo considerado, com especial atenção aos anos de 2005 a 2007, tanto no verão como no outono, cujas cores correspondem a médias anuais de temperaturas mais altas face aos demais anos.

Para poder observar o comportamento da temperatura com base nalguns dos índices que caracterizam esta variável, agrupou-se esta variável por décadas e calcularam-se os respetivos

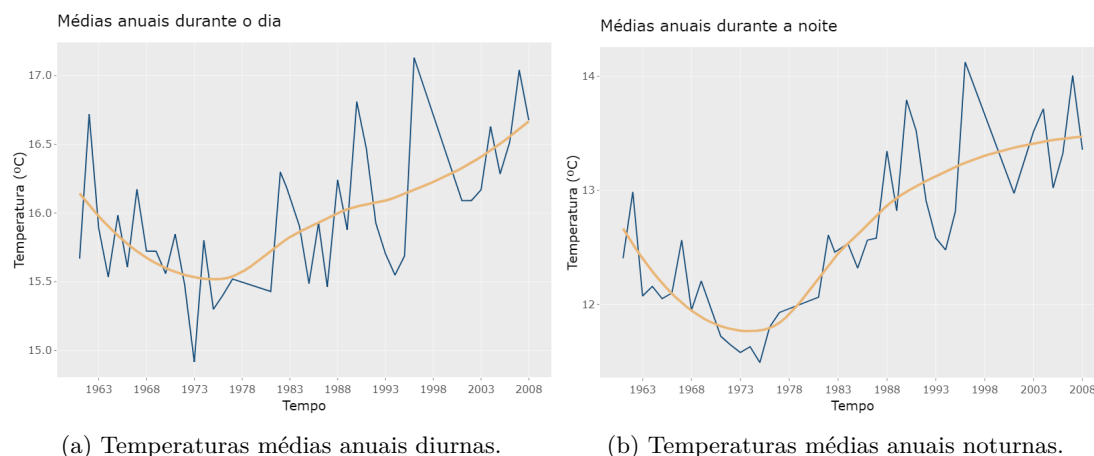


(a) Distribuição da temperatura mínima diária, por mês.



(b) Distribuição da temperatura máxima diária, por mês.

Figura 3.7: Distribuição da temperatura máxima e mínima diária, por mês.



(a) Temperaturas médias anuais diurnas.

(b) Temperaturas médias anuais noturnas.

Figura 3.8: Evolução das temperaturas médias anuais diurnas e noturnas, e linha de tendência. (1960-2007).

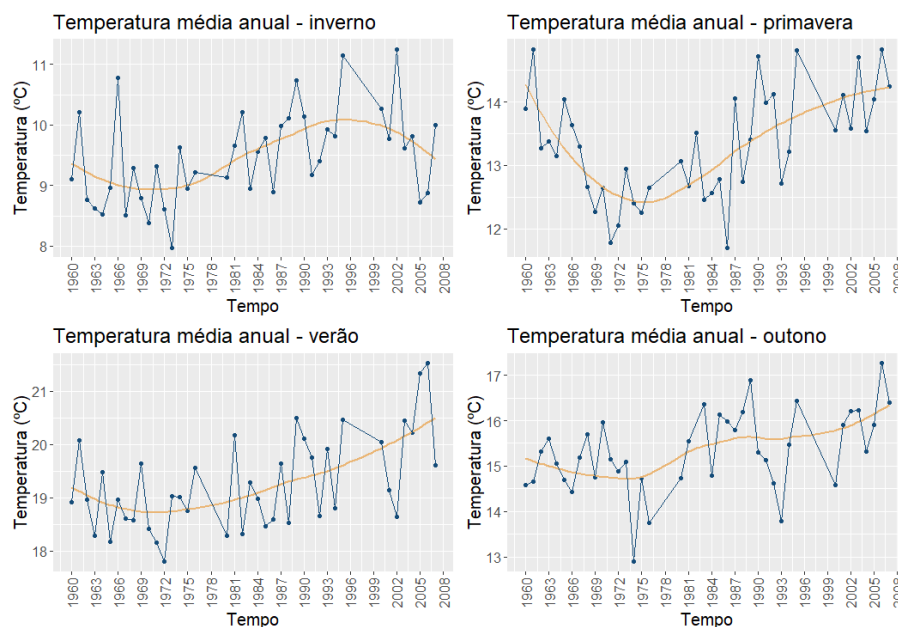


Figura 3.9: Evolução das temperaturas médias anuais por estação do ano, e linha de tendência (1960-2007).

índices para as temperaturas diárias mínimas e máximas. Esses resultados podem ser observados nas Tabelas 3.7 e 3.8, respectivamente.

Tabela 3.7: Índices de temperatura diária mínima, por década. (TN_n - Noite mais fria; TN_x - Noite mais quente; FD - Dias de frio intenso; TN - Noites tropicais; CWDI - Duração da onda de frio; CSDI - duração de períodos frios)

Período de tempo	TN_n	TN_x	FD	TN	CWDI	CSDI
1960-1970	-2.7 (1963)	21.3 (1964)	61	7	30	24
1970-1980	-2 (1971)	21.9 (1976)	35	3	23	50
1980-1990	-2.4 (1987)	22 (1987/89)	33	19	13	20
1990-2000	-1.1 (1990)	24.1 (1990)	7	13	10	13
2000-2007	-1.6 (2005)	24.9 (2006)	13	42	13	0
Total	-2.7 (1963)	24.9 (2006)	149	84	89	107

Tabela 3.8: Índices de temperatura diária máxima, por década. (TX_n - Dia mais frio; TX_x - Dia mais quente; SU - Dias de verão; TX35 - Dias muito quentes; HWDI - Duração da onda de calor; WSDI - Duração de períodos quentes)

Período de tempo	TX_n	TX_x	SU	TX35	HWDI	WSDI
1960-1970	3.5 (1962)	37.6 (1961)	472	9	63	30
1970-1980	4.3 (1971)	36.8 (1970)	290	5	57	21
1980-1990	3.6 (1987)	38.6 (1981)	367	13	47	70
1990-2000	4.8 (1994)	37.4 (1990)	296	7	27	16
2000-2007	6.2 (2001)	44.9 (2001)	430	15	47	26
Total	3.5 (1962)	44.9 (2001)	1855	49	241	163

Da Tabela 3.7 podemos observar que foi na década de 1960 que se registou a temperatura mínima mais baixa, e que esses valores têm vindo a aumentar ao longo do período referido. Também é possível verificar que o número de dias de frio intenso (dias cuja temperatura mínima diária é inferior a 0°C) têm vindo a diminuir, e que as décadas de 1990 e 2000 foram as que registaram menos dias de frio intenso. Contrariamente, o número de noites tropicais (dias em cuja temperatura mínima diária foi superior a 20°C) tem registado um aumento com maior incidência da década de 2000, que registou um total de 42 noites tropicais, tanto quanto a soma das restantes décadas que a antecederam. Deste modo concluímos que as tendências destes dois índices são o oposto. Relativamente à duração das ondas frias, estas foram calculadas através da média de temperatura mínima diária para cada ano, isto é, média de cada dia do ano em causa. Para considerar um determinado período de tempo como vaga fria foram tidos em conta os casos onde a temperatura mínima de um determinado dia fosse inferior à média diária respetiva menos 5°C . Se este evento se verificasse durante, pelo menos, 6 dias consecutivos, foi rotulado como vaga de frio. Observando a Tabela 3.7 é possível identificar a década de 1960 como aquela onde se registaram mais vagas de frio. Apesar das décadas registadas de temperatura do ar não serem regulares no tempo, estes índices corroboram o aumento de temperatura entre 1960 e 2007.

No caso da temperatura máxima diária, na Tabela 3.8 verifica-se que o valor mais alto ocorreu na década de 2000, e no ano de 2001 com 44.9°C , no entanto este valor coincide com um dos *outliers* identificados na Tabela A.4, Apêndice A. Ainda assim, observando os índices de dias de verão e de dias muito quentes (dias cuja temperatura máxima diária foi superior a 20°C e a 35°C , respetivamente), exceção da década de 1960, que registou 472 dias de verão, a década de 2000 mostrou-se mais quente que as restantes. Estas estatísticas comprovam o aumento de temperaturas máximas por dia, ao longo do período de tempo registado. Relativamente às ondas de calor, o método para calcular este índice foi idêntico ao das ondas frias, mas utilizando a temperatura máxima diária e somando 5°C à média diária correspondente. Foi na década de 1960 que se registaram mais vagas de calor, apesar das diferenças face às restantes décadas não ser muito grande e, possivelmente, registam valores mais baixos derivado à ausência de registos nas décadas de 1980 e 1990. Nas Tabelas 3.7 e 3.8 é possível observar outros índices característicos da temperatura. Nomeadamente, a duração de períodos frios (CSDI) e quentes (WSDI), para as temperaturas diárias mínimas e máximas, respetivamente. Estes índices equivalem ao número de dias, por ano, onde TN é inferior ao percentil 90 e TX é superior ao percentil 90 para médias

diárias de temperatura mínima e máxima.

Finalmente, observaram-se as três componentes da série temporal da temperatura. Começou por se confirmar se a série era ou não estacionária, aplicando o teste *Augmented Dickey-Fuller*, que se revelou positivo, com um *p-value* de 0.01, indicando que a série é estacionária. Procedeu-se então à decomposição da série nas suas três componentes.

Tendo em conta que os dados iniciais se encontravam registados de hora a hora, foram obtidas as componentes por dados de hora a hora, médias diárias, e médias anuais. Entende-se como frequência de uma série temporal o número de observações por ciclo (igual a 1 caso seja anual, 12 se se tratarem de dados por mês, 4 no caso de dados trimestrais ou 52 se forem registados semanalmente). No entanto, se a frequência for superior à frequência semanal, tal como é o caso dos dados em estudo, existe mais do que uma maneira de lidar com a frequência. Isto é, dados com observações diárias podem ter uma sazonalidade semanal (frequência = 7) ou uma sazonalidade anual (frequência = 365,25, geralmente igual a 365 uma vez que é mais comum o uso de frequências inteiras). Do mesmo modo, dados observados de hora a hora, podem ter uma sazonalidade diária (frequência = 24), semanal (frequência = 168) ou anual (frequência = 8766). A Figura 3.10 apresenta a decomposição da série temporal da temperatura considerando médias mensais, i.e. com frequência igual a 12. No Apêndice B, as Figuras 3.10, B.4 e B.5, apresentam as componentes, da série temporal da temperatura, nas diferentes frequências descritas anteriormente.

Em particular, e observando as Figuras 3.10 e B.4 no Apêndice B, relativas às observações com sazonalidade anual, podemos encontrar variações nas três componentes (tendência, sazonalidade e restante), com base na frequência aplicada a cada uma destas séries temporais. Tendo em conta o fator que as distingue é a sazonalidade, podemos concluir que a mesma é explícita em ambas as figuras ao longo do tempo considerado. O comportamento das componentes tendência e sazonalidade na Figura B.5 é idêntico ao da Figura B.4, o que se justifica, tendo em conta tratarem se dados com a mesma sazonalidade (anual). As barras cinzentas, que se encontram nas laterais direitas do gráfico de cada componente, mostram as escalas relativas de cada uma. Uma vez que as escalas de cada um desses gráficos varia, as barras também variam no seu tamanho. Uma barra cinzenta, na componente restante, com tamanho maior que a dos dados iniciais, mostra que a variação da componente restante é pequena em comparação com a dos dados iniciais.

Tendo em conta as Figuras, B.4 e B.5, presentes no Apêndice B, e observando as barras laterais que surgem ao lado de cada componente, ao considerar-se a componente tendência, a sua barra é maior que a das restantes, indicando que a variação atribuída à tendência é menor que a das outras componentes e consequentemente, apenas uma pequena parte da variação dos dados iniciais da série temporal. Posto isto, é necessário ter alguma precaução ao avaliar o comportamento da tendência. Ainda assim, na Figura 3.10, que mostra os dados de temperatura hora a hora, observa-se um decréscimo na tendência até 1972 para depois se revelar num aumento gradual nos anos seguintes, ainda que com algumas oscilações. O mesmo se pode observar nos gráficos dos valores de temperatura com sazonalidade anual.

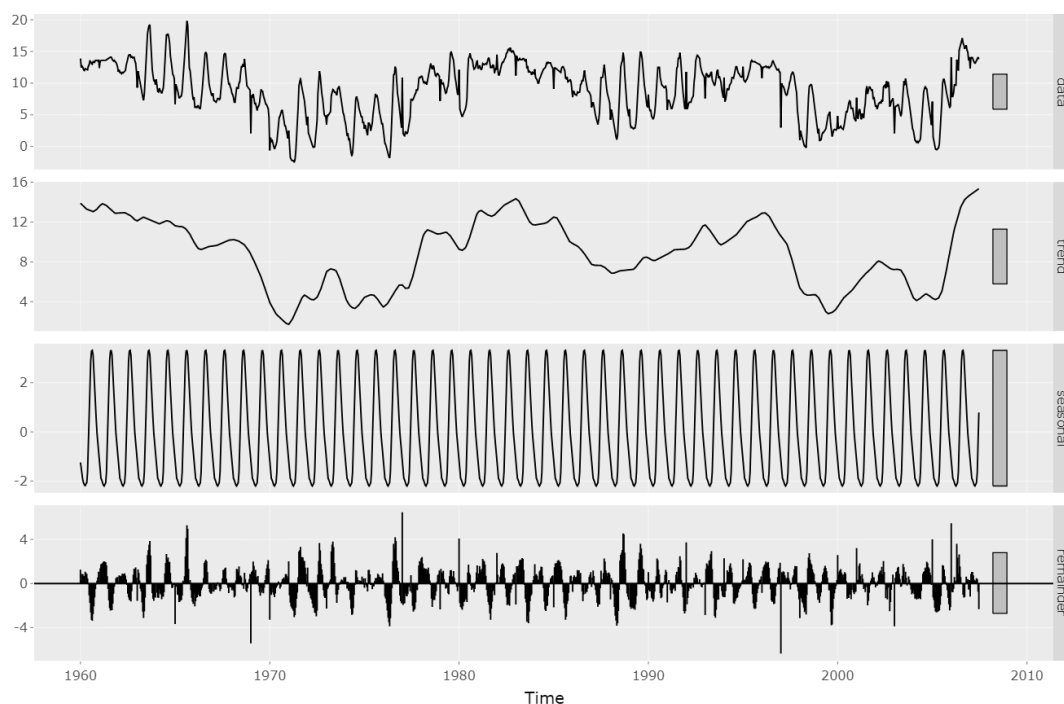


Figura 3.10: Temperatura hora a hora ($^{\circ}\text{C}$) com sazonalidade diária (frequência = 24).

3.3.2 Precipitação

Ao contrário das outras variáveis sob análise, a precipitação precisa de ser estudada de modo distinto face às suas particularidades. Especificamente, o tratamento dos *missing values* e dos *outliers*. Assim sendo, desconsiderou-se a hipótese de existirem *outliers*, à exceção de registos que ultrapassassem os limites descritos na Tabela 3.4, e admitiu-se que os valores desconhecidos correspondem a ausência de precipitação nessa determinada hora e, portanto, foram igualados a zero.

Na Figura 3.11 pode observar-se o comportamento dos valores horários de precipitação, compreendidos entre 1997 e 2006. Através desta figura é possível confirmar a ausência de registos durante alguns meses, nomeadamente de fevereiro a dezembro em 1997 e em 2006, de maio a dezembro de 2003, e a totalidade do ano de 2004.

Tendo em conta que tanto nos anos de 1997 como de 2006, apenas existem registos do mês de janeiro, estes anos foram desconsiderados das análises seguintes, para que os seus resultados não interferissem no estudo desta variável, no período de tempo considerado.

Através de uma breve análise foi possível verificar que a hora de maior precipitação registada neste período de tempo ocorreu no dia 22 de outubro de 2001 às 9 horas, com 25.50 mm. Já o dia onde o total de precipitação acumulada foi maior, corresponde a 21 de março também do mesmo ano.

Na Figura 3.12 está representada, através de um gráfico de barras, as precipitações anuais

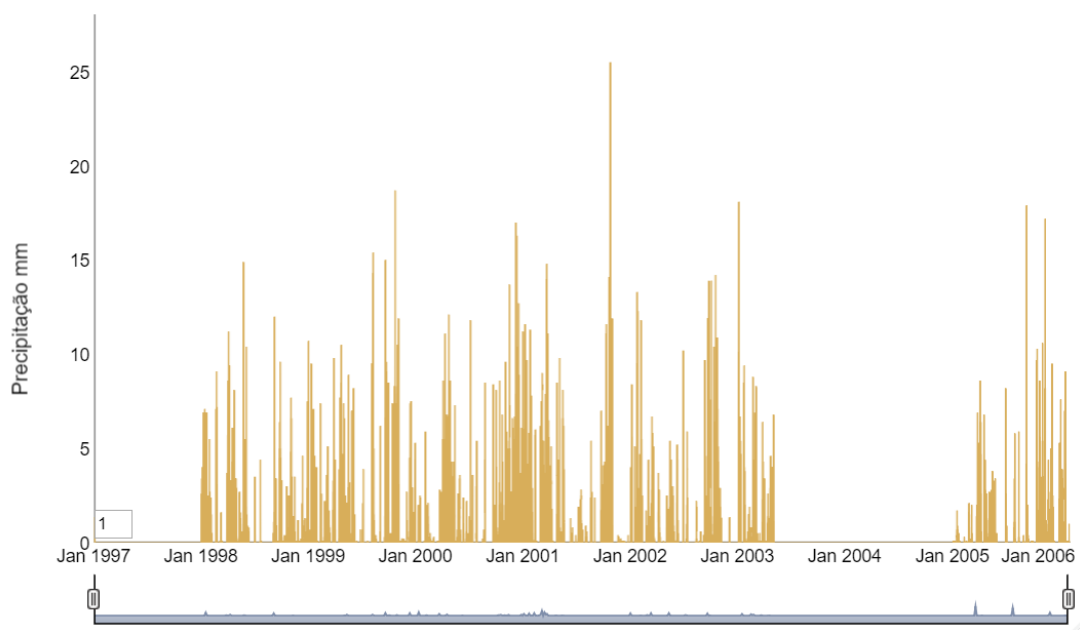


Figura 3.11: Gráfico da precipitação, registros de hora a hora (1997-2006).

no intervalo de tempo considerado (1998 a 2005). Tendo em conta a linha de tendência que acompanha o aumento ou decréscimo dos totais de precipitação, pode observar-se uma tendência crescente até o ano de 2000, à qual segue um decréscimo até 2003. Os anos de 2000 e 2001 foram aqueles cuja quantidade de precipitação foi maior, 1577.6 mm e 1561.7 mm respectivamente, e 2003 menor quantidade, 676.4 mm .

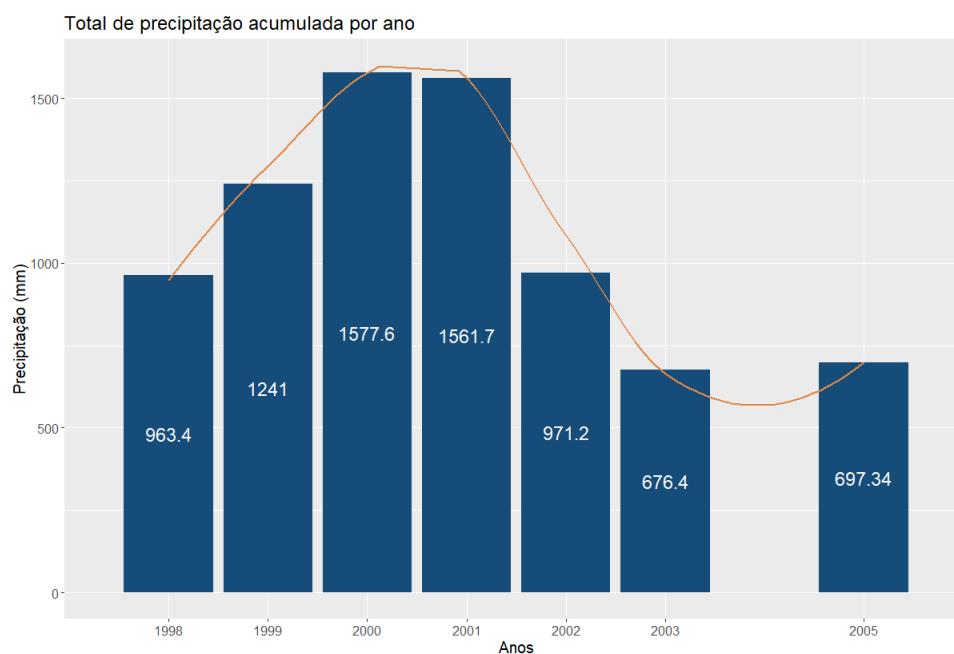


Figura 3.12: Total de precipitação por ano (1998-2005).

Também para a variável de precipitação foi produzido um *heatmap* (cf. Figura 3.13) representando o total de precipitação acumulada por mês e seu respectivo ano. No mesmo, podemos confirmar a

ausência dos meses de maio a dezembro de 2003, que registam um total de precipitação de 0 mm , e que o mês de março de 2001 foi o mais chuvoso, seguido de dezembro de 2000. É possível verificar que os meses de verão, especificamente junho, julho e agosto, são os meses onde ocorreu menor precipitação. Agosto de 1999 registou totais atípicos face aos meses de agosto dos outros anos da série, tendo em conta o gradiente de cores associado.

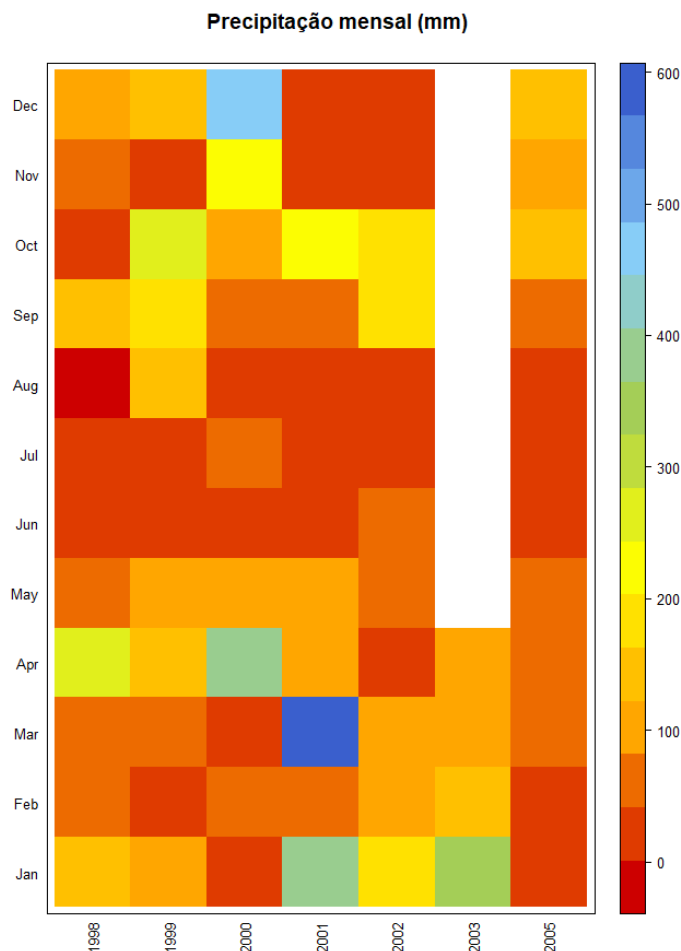


Figura 3.13: Heatmap da precipitação mensal (mm).

Observando depois, na Figura 3.14 o gráfico de barras com as precipitações acumuladas por mês, ao longo do período considerado anteriormente, e respetiva linha de tendência, puderam confirmar-se alguns dos resultados obtidos através da Figura 3.13. Respetivamente, os meses mais quentes de junho, julho e agosto mostram totais de precipitação mais baixos, contrariamente a janeiro, abril e março, que representam meses mais chuvosos.

Na Figura 3.15 estão representados os *boxplots* obtidos para a precipitação agrupada por acumulação mensal, de 1998 a 2005. Este gráfico permite não só comprovar a distribuição desta variável por mês, como identificar possíveis registos de precipitação que sejam incomuns. É exemplo disso o suposto *outlier* observado nos meses de março e agosto, que se afastam dos valores esperados para esses meses. Este tipo de valores pode influenciar os resultados obtidos,

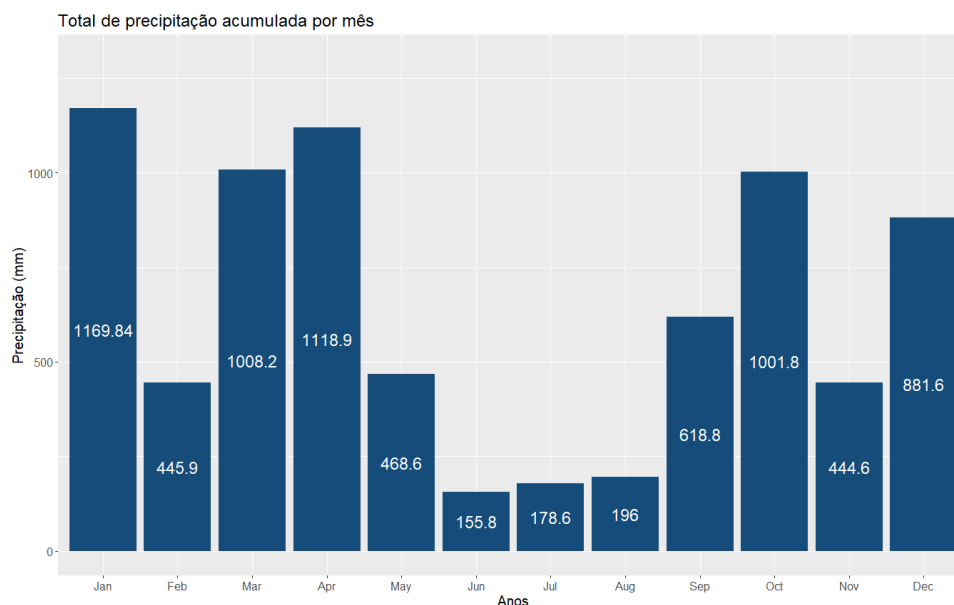


Figura 3.14: Total de precipitação por mês (1998-2005).

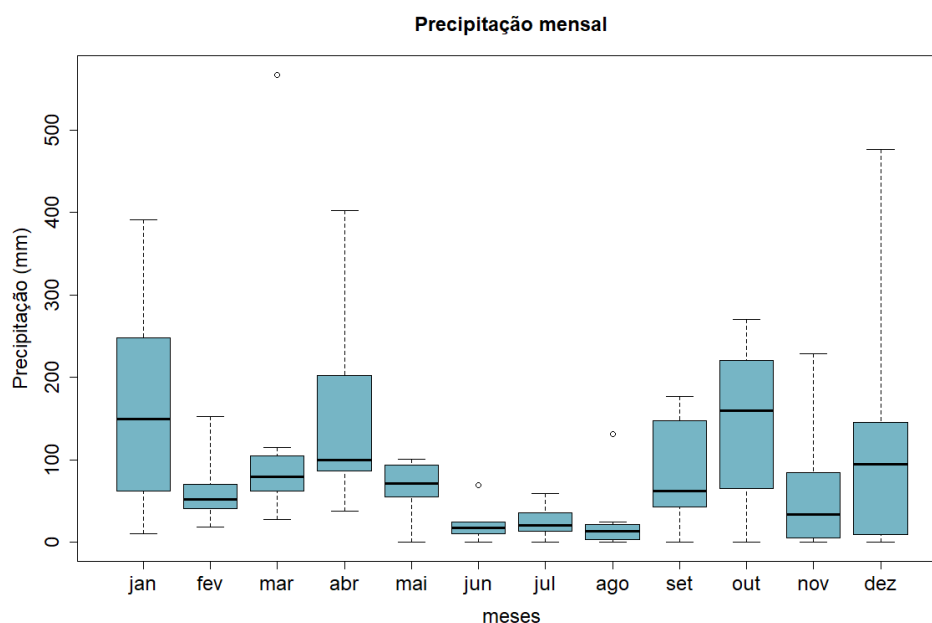


Figura 3.15: Boxplot das precipitações mensais (1998-2005).

contudo, não existem informações suficientes que comprovem que seja um *outlier*.

Ao consultar os ficheiros Excel com os registos de precipitação, reparou-se que algumas horas e dias de certos meses registaram avarias na instrumentação utilizada, podendo assim justificar alguns dos resultados obtidos. Essas avarias encontram-se sumariadas na Tabela A.10 no Apêndice A.

Tendo como base, a Figura 3.16, e a Tabela B.1, no Apêndice B, foi possível perceber o comportamento da precipitação no decorrer das estações do ano. A Figura 3.16 mostra a evolução da precipitação em cada estação do ano no período de 1998 a 2005, e ao lado, o

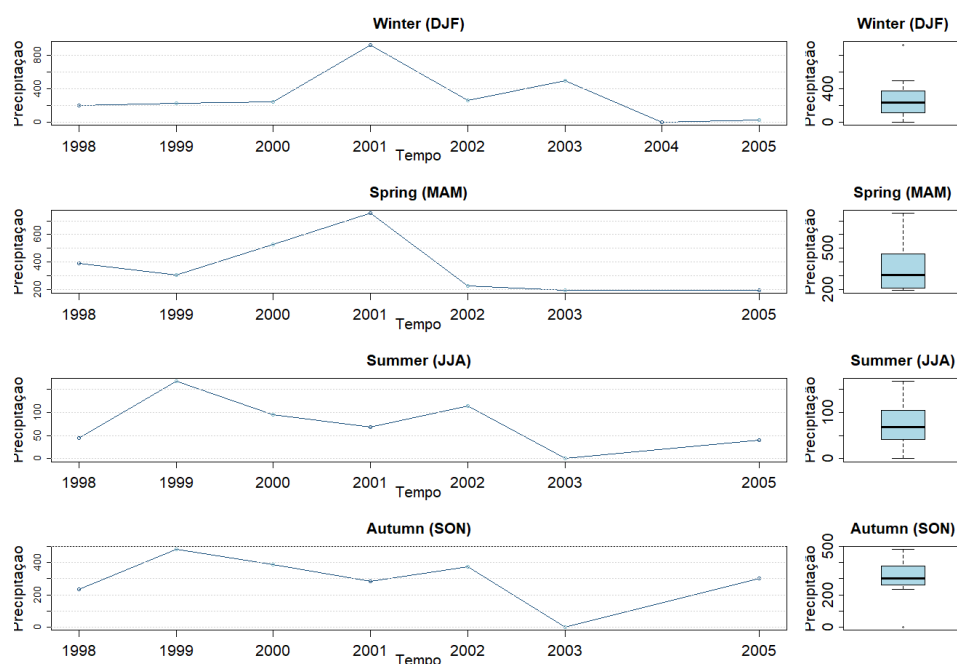


Figura 3.16: Evolução da precipitação anual acumulada nas estações do ano e seu *boxplot* (1998-2005).

boxplot respectivo a essa estação. Analisando estes gráficos e a tabela pode-se verificar que durante o verão ocorre menos precipitação (530.4 mm de precipitação total acumulada no período) face às demais e que, talvez contrariamente ao esperado, a primavera foi a estação com maior quantidade de precipitação (total de 2594.8 mm no período). Ainda assim, no período considerado, a estação/ano que registou a maior quantidade de precipitação foi o inverno de 2001. Estas informações encontram-se na Tabela B.1, no Apêndice B, que mostra, para cada ano, a quantidade de precipitação acumulada por estação e o seu total cumulado. A figura e a tabela indicam o verão e o outono de 2003 com precipitação nula, no entanto, tal como referido anteriormente, há ausência de registos de maio a dezembro desse ano, pelo que se pode considerar a estação com menos precipitação como sendo o verão de 2005.

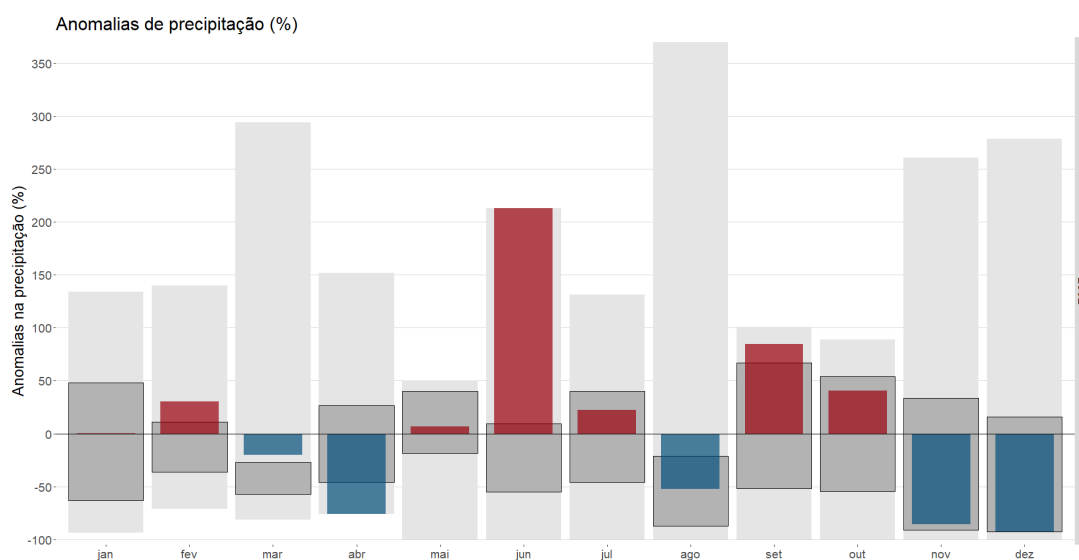
Ao analisar alguns índices de precipitação constatou-se a existência de 273 dias que se classificam como dias de precipitação forte, isto é, dias cuja precipitação foi igual ou superior a 10 mm (índice R10), e 114 dias com precipitação muito forte, igual ou superior a 20 mm (índice R20). O ano de 2001 foi aquele onde ocorreu precipitação forte e muito forte. Por outro lado, relativamente aos dias secos, encontrou-se um total de 1884 dias, ou seja, a precipitação foi inferior a 1 mm (índice DD), sendo o ano mais seco o de 2003. Estes valores podem encontrar-se descritos na Tabela 3.9, acrescida do total de precipitação acumulada por cada índice. O índice R25 corresponde ao número de dias de precipitação extremamente forte, com precipitação igual ou superior a 25 mm.

A Figura 3.17 mostra as anomalias da precipitação de 2002 e 2005, Figura 3.17a e Figura 3.17b, respetivamente. Estes foram os anos onde se verificaram um maior e menor número de anomalias positivas respetivamente. A anomalia positiva de junho de 2002 corresponde também ao valor máximo deste mês, e oposto verifica-se no caso de dezembro de 2002, e janeiro e fevereiro de 2005

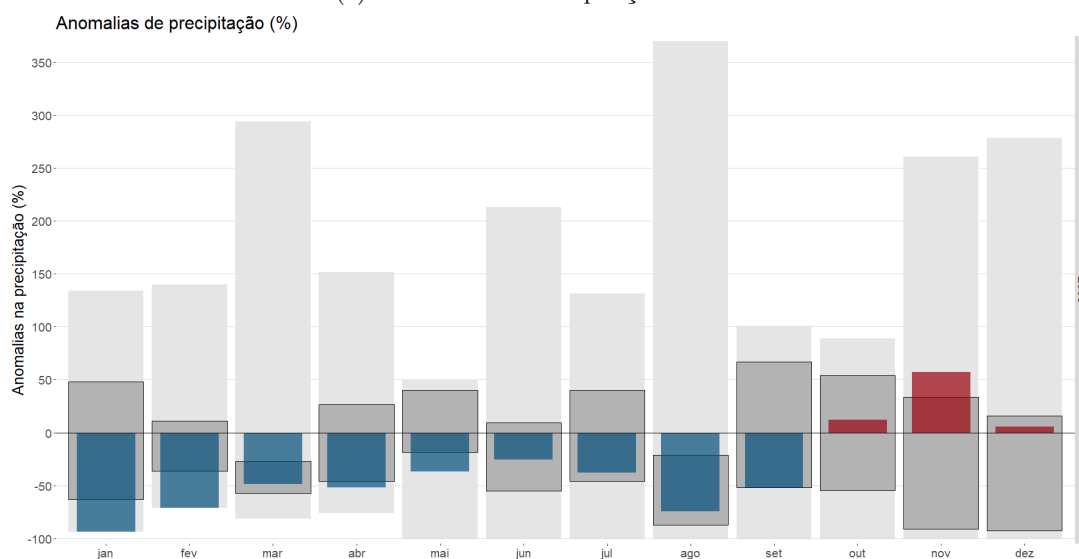
Tabela 3.9: Índices de precipitação e respectivos valores. (DD - Dias secos; R10 - Dias de forte precipitação; R20 - Dias de precipitação muito forte; R25 - Dias de precipitação extremamente forte)

Índice	Total de dias	Precipitação Acumulada (mm)	Ano de maior/menor precipitação
DD	1884	90.1	2000
R10	273	5810.3	2001
R20	114	3566.5	2001
R25	67	2518.6	2001

cujas anomalias negativas atingem o mínimo da série temporal. Também para esta variável foi aplicado o teste *Augmented Dickey-Fuller*, que se revelou positivo com um *p-value* inferior a 0.05.



(a) Anomalias de Precipitação em 2002.



(b) Anomalias de Precipitação em 2005.

Figura 3.17: Anomalias de precipitação dos anos 2002 e 2005.

Ao observar o comportamento das componentes das séries temporais de precipitação diárias e mensais, respetivamente, nas Figuras B.6 (presente no Apêndice B), 3.18, verifica-se um aumento da tendência até finais do ano 2000 seguido de uma diminuição que se reverte em 2002. Dada a ausência de dados em 2004 e de mais registos em 2006 não é possível concluir se posteriormente a tendência tende a aumentar ou não.

No que diz respeito às componentes da série temporal da precipitação, obtiveram-se as mesmas para a série composta por observações de hora a hora, e médias diárias, ambas com sazonalidade anual, e para séries de precipitação média mensal, entre 1998 e 2005. As três componentes destas séries, tendência, sazonalidade e restante, podem ser encontradas nas Figuras B.6, 3.18 e B.7, respetivamente. As três séries mostram igual tendência indicativa de um aumento da tendência até finais do ano 2001 seguido de uma diminuição. Atendendo à ausência de dados no ano de 2004, não é possível aferir se a tendência se manteve entre 2003 e 2005 ou se se observou outro comportamento no decorrer deste período de tempo.

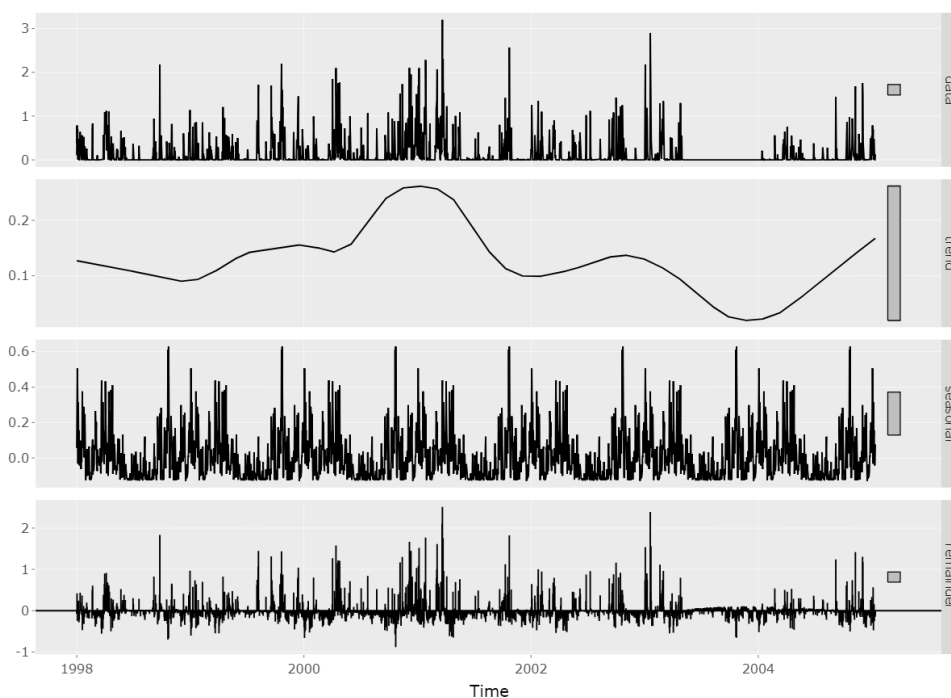


Figura 3.18: Decomposição da série temporal da precipitação: médias diárias (mm) com sazonalidade anual (frequência = 365).

3.3.3 Pressão atmosférica

Ao proceder à análise da pressão atmosférica verificou-se que, ainda que não seja a variável com uma maior percentagem de *missing values* (2.19%) foi a variável onde foram identificados mais *outliers*, registados na Tabela A.5 do Apêndice A, totalizando 181 casos. Na Figura 3.19 pode ver-se o comportamento desta variável no período de tempo registado, de 1999 a 2007, de hora a hora, onde também, alguns desses *outliers* podem ser identificados, nomeadamente nos meses de

outubro, novembro e dezembro de 2005, com valores que se afastam dos demais.

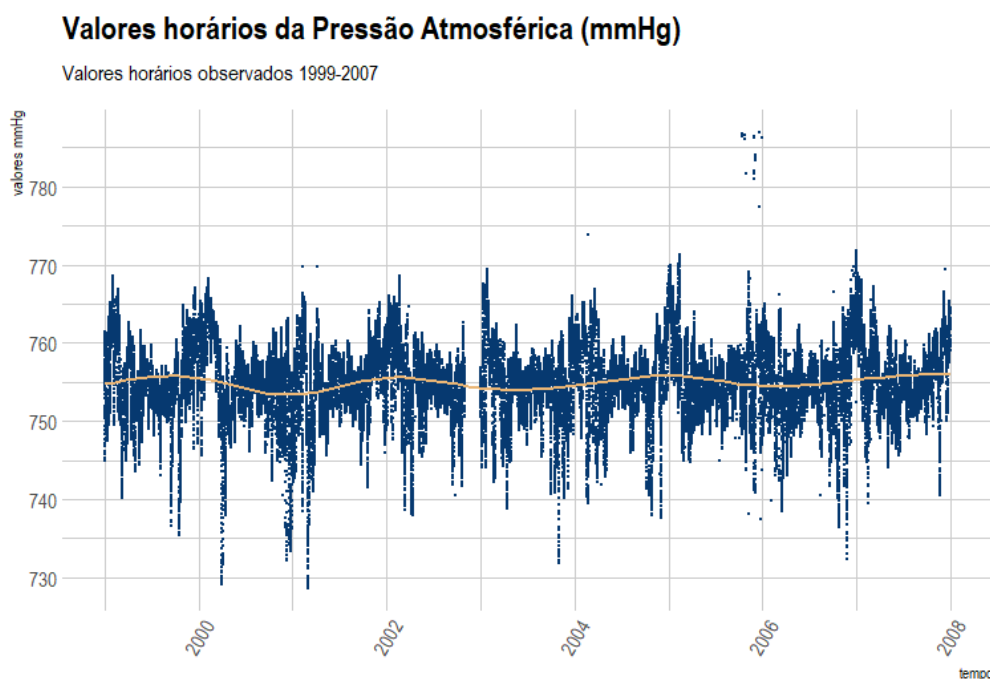


Figura 3.19: Série temporal da pressão atmosférica (*mmHg*) e linha da tendência (1999-2007).

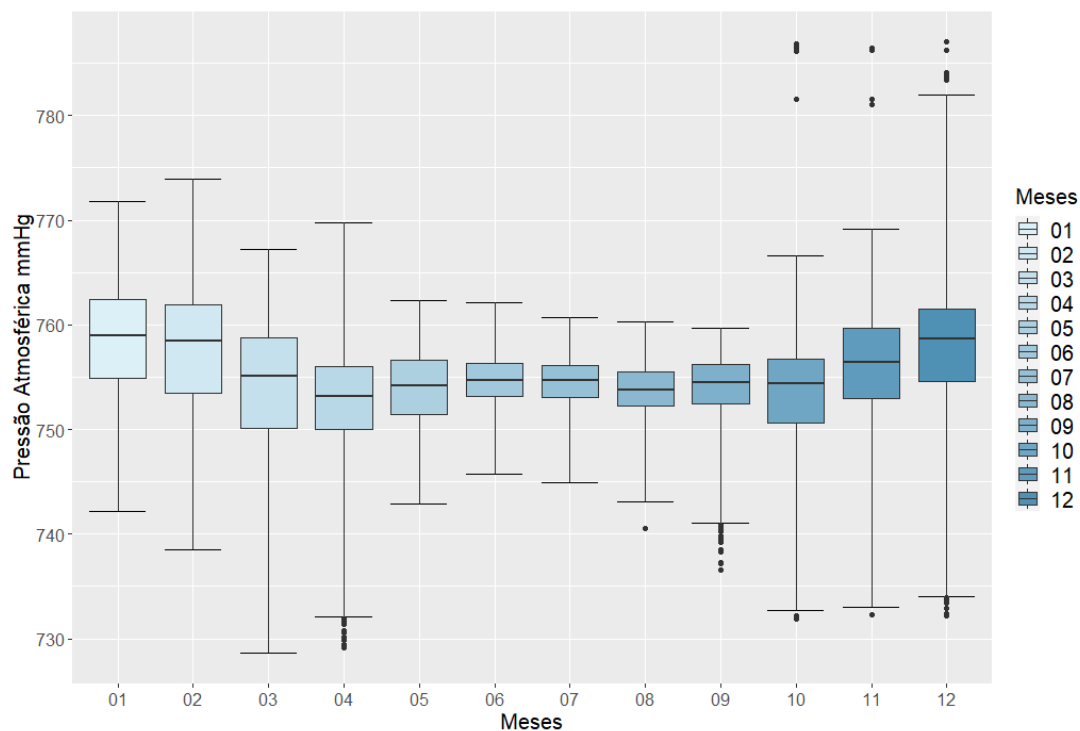


Figura 3.20: *Boxplot* da pressão atmosférica onde é possível identificar outliers extremos (1999-2007).

Para melhor observar a distribuição de *outliers* extremos, na Figura 3.20 encontra-se o *boxplot*

que apresenta esses resultados por mês, ao longo do período considerado (1999-2007). Através do mesmo pode reforçar-se a presença mais acentuada de *outliers* nos meses de outubro, novembro e dezembro, assim como outros, com valores mais baixos do que o esperado, nos meses de março, e de agosto a dezembro.

Na Figura B.8, no Apêndice B pode observar-se a distribuição de *missing values*, no intervalo de tempo registado, que foram corrigidos do mesmo modo que a temperatura. Esses *missing values* encontram-se destacados a vermelho e correspondem a 49 registos, 0.06% da totalidade dos valores observados.

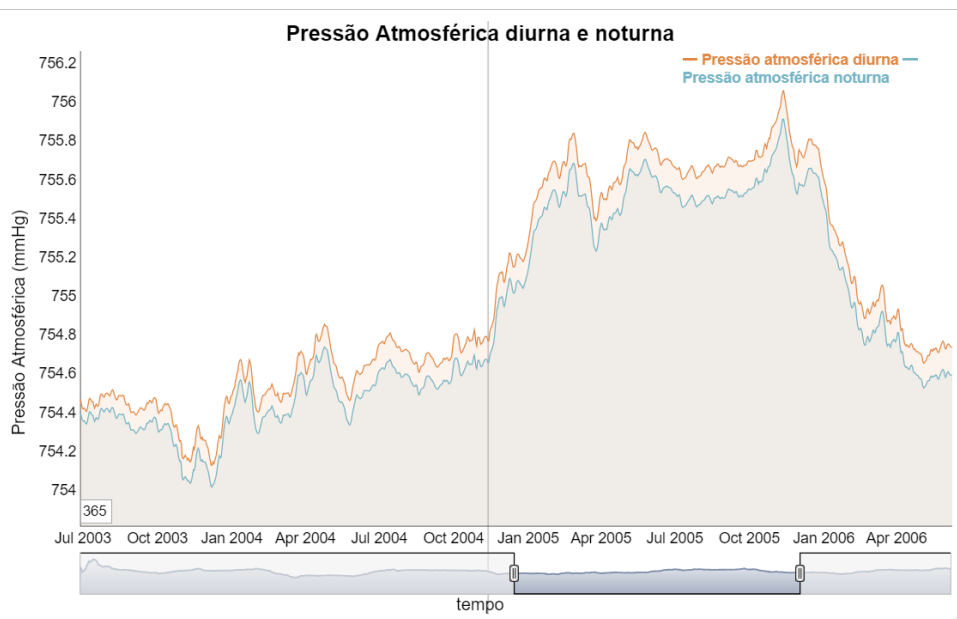


Figura 3.21: Pressão atmosférica diária diurna e noturna com *moving average* de ordem 365.

De seguida, recorreu-se à representação dos dados horários da pressão atmosférica através de um *heatmap*, Figura B.10, Apêndice B. Este gráfico revelou-se, novamente, muito útil a identificar períodos de ausência de registos e possíveis *outliers*. Na Figura B.10b encontramos um exemplo de ausência de registos, nas três primeiras horas do dia de abril e na Figura B.10a é possível observar a presença de muitos dos *outliers* anteriormente mencionados, dada a diferença abrupta do gradiente de cores, que sugerem valores mais elevados que o normal, por exemplo, no dia 5 de dezembro de 2007, às 7 horas, que regista um valor de 783.4 mmHg enquanto que as suas horas adjacentes registam 753.4 mmHg. Outras ausências de registos na pressão atmosférica de 1999-2007, encontram-se na Tabela A.9 no Apêndice A.

Comparou-se depois o comportamento da série com valores diários de pressão atmosférica diurnos e noturnos. Este pode ser observado na Figura 3.21 que representa uma janela, com uma *moving average* anual, de ordem 365 para melhor visualizar a distribuição dos dados, aplicada à série de pressão atmosférica com dados diários. Nesta janela podem identificar-se as séries compostas pelos valores diários diurnos e, a outra cor, os valores diários noturnos, e aferir que estes são sempre mais elevados durante o dia do que pela noite, ao longo de toda a série temporal (1999-2007).

Atendendo agora às médias mensais desta variável, na Figura 3.22 estão representadas a diferentes cores, os vários anos que compõem esta série, com as respetivas médias mensais da pressão atmosférica, ao longo dos seus doze meses. Analisando a mesma, verifica-se que é nos meses de inverno, especificamente janeiro e dezembro, que se registam valores mais elevados, contrariamente ao meses de março e abril que aparentam os valores mais baixos. Já nos meses de maio, junho e julho, nota-se que a pressão atmosférica regista uma menor variabilidade ao longo dos anos.

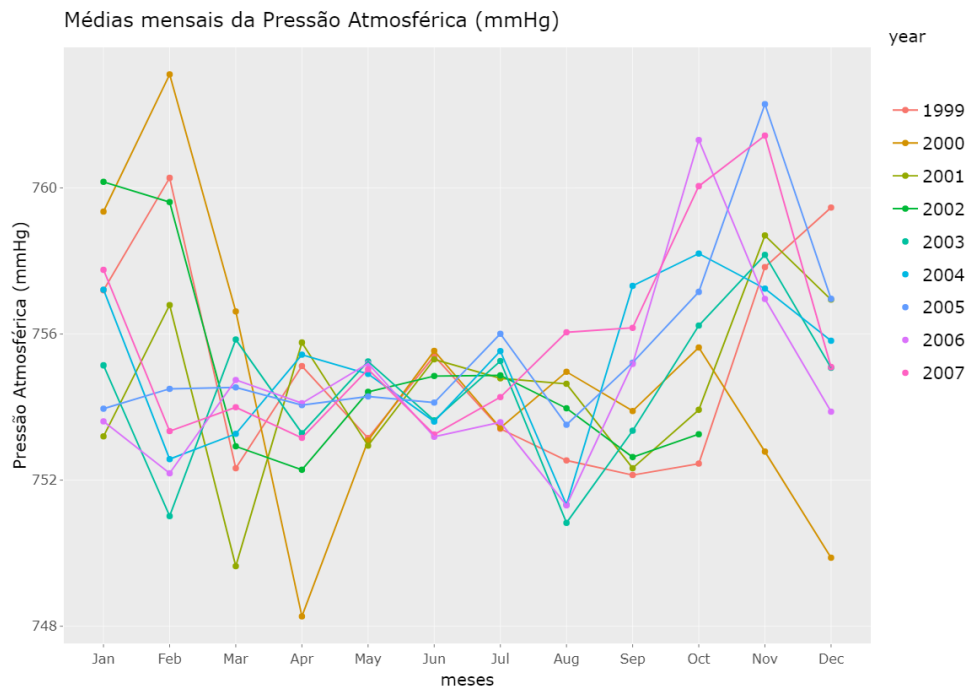


Figura 3.22: Gráfico das médias mensais da pressão atmosférica ao longo dos anos (1999-2007).

Na Figura 3.23 apresentam-se as distribuições mensais da pressão atmosférica média diária. Podemos observar que os meses de junho a agosto, são aqueles que apresentam uma distribuição mais concentrada.

Finalmente, procurou analisar-se o comportamento das componentes da série temporal da pressão atmosférica. Analogamente à temperatura, o teste *Augmented Dickey-Fuller* obteve um *p-value* igual a 0.01 e foram calculadas as séries temporais com médias diárias e mensais, para observar o comportamento das componentes das mesmas, e dos dados originais, de hora a hora, face a diferentes sazonalidades. As Figuras B.11, B.12, presentes no Apêndice B, e 3.24 mostram os resultados obtidos para as séries anteriormente referidas, entre 1999 e 2007.

É nas figuras com sazonalidade anual, nomeadamente Figuras B.12 e 3.24, que a barra lateral se mostra mais próxima da escala dos dados iniciais, querendo isto dizer que, se se encurtasse a escala da componente sazonal até se tornar do mesmo tamanho que a barra dos dados iniciais, o intervalo de variação da componente sazonal seria semelhante, apenas ligeiramente menor. Ainda assim, é possível identificar a existência de sazonalidade diária e anual nas séries de pressão atmosférica. Já no que diz respeito à evolução da tendência, a mesma mostrou-se crescente a partir do ano 2000, ainda que com algumas oscilações durante os anos que se seguiram.

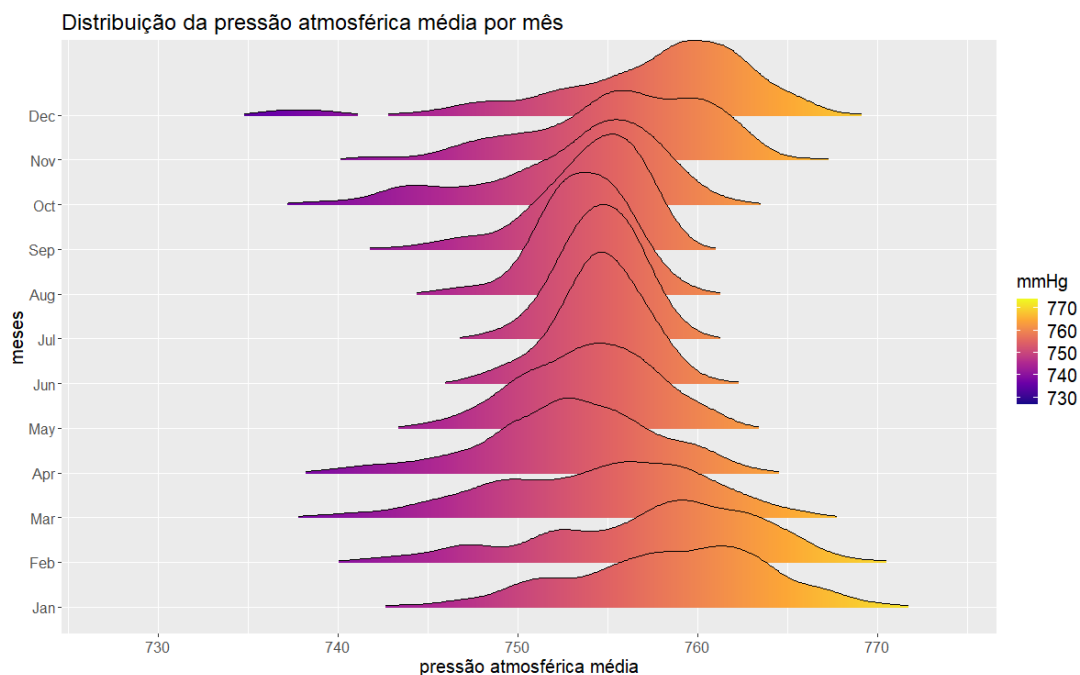


Figura 3.23: Distribuição da pressão atmosférica média diária, por mês.

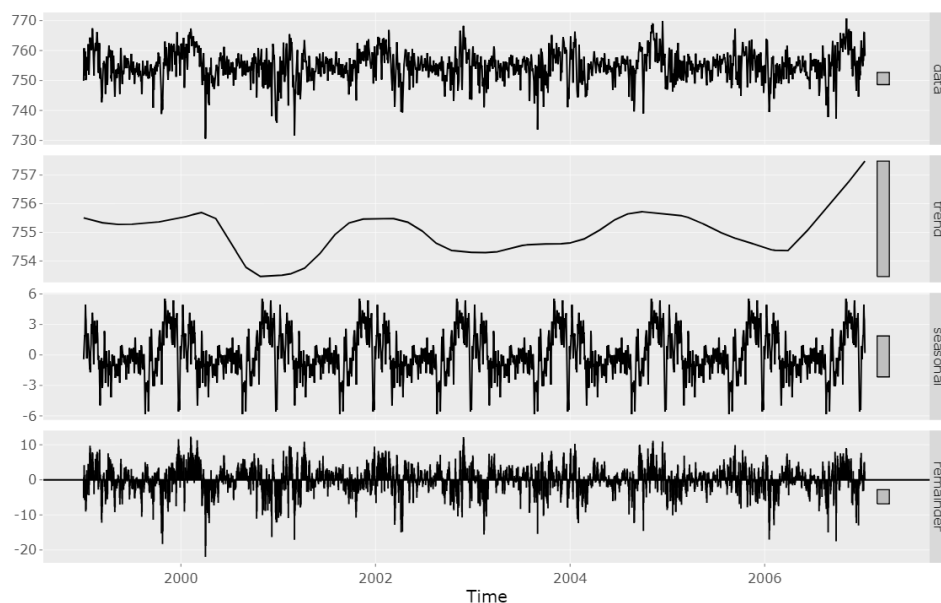


Figura 3.24: Decomposição da série temporal da pressão atmosférica: médias diárias (*mmHg*) com sazonalidade anual (frequência = 365).

3.3.4 Velocidade do vento

Relativamente à velocidade do vento deparamo-nos com a maior ausência de registos durante o período de tempo registado. Tendo em conta a série temporal regular no tempo, caracterizada pelos registos horários desta variável, de 1994 a 2007, 33.6% corresponde a *missing values*, os quais podem ser observados na Figura B.13 assinalados a vermelho, no Apêndice B.

No entanto a distribuição dos *missing values* submetidos a correção, encontra-se retratada na Figura 3.25, identificados a vermelho. Foram corrigidos apenas os casos onde faltassem apenas um ou dois registos consecutivos, tal como referido na secção 3.2.1. Esta correção correspondeu unicamente a 0.027% dos dados correspondes de 1994 a 2007, totalizando 21 *missing values*.

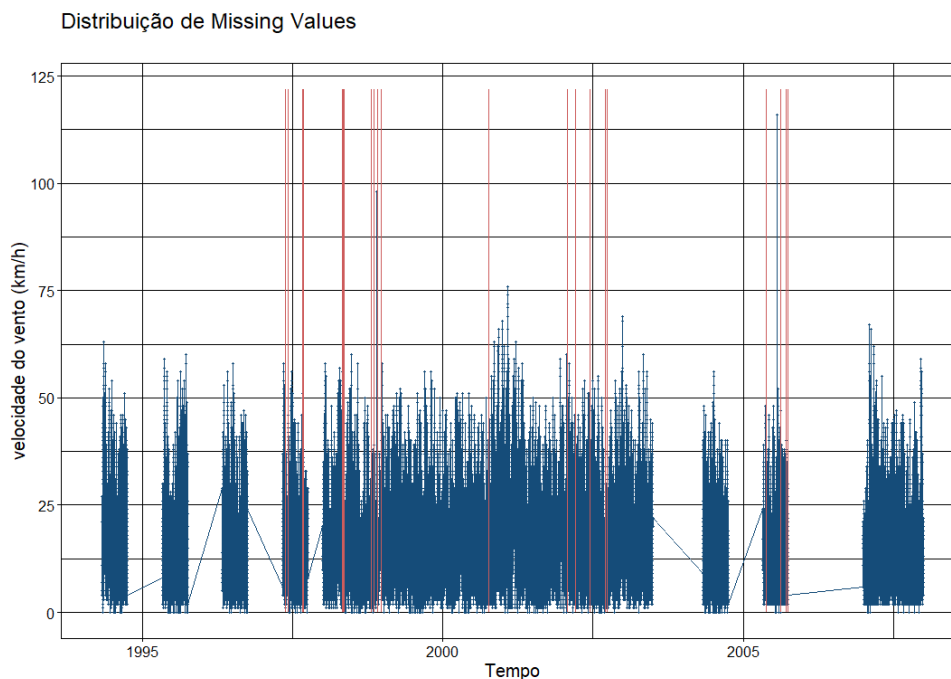


Figura 3.25: Distribuição de *missing values* na velocidade do vento (1994-2007).

Procurando entender o comportamento desta série temporal, na Figura B.15, Apêndice B, estão representados os valores horários da velocidade do vento e respetiva linha de tendência. Nesta figura é possível identificar alguns *outliers* como é o caso do dia 26 de novembro de 1998, que regista uma velocidade de 98 km/h valor este que difere dos demais em larga escala.

Para perceber que outros possíveis *outliers* podiam existir, na Figura 3.26 encontram-se representados os *boxplots* agrupados por mês ao longo do período de dados registados (1994-2007) e que permitiram a identificação de *outliers*. Através da figura verifica-se que de março a junho e agosto, foram os únicos meses que não registaram qualquer *outlier* e que em julho, foi encontrado o registo mais extremo destes dados, no dia 24 de 2005, que assinalou 116 km/h . Os restantes *outliers*, identificados para esta variável, e respetiva proposta de correção, podem ser consultados na Tabela A.6, no Apêndice A.

De modo análogo às restantes variáveis, a Figura B.17, no Apêndice B mostra o heatmap com os valores registados de hora a hora pela velocidade do vento. Na Figura B.17a podemos observar que existem algumas horas, ao longo dos meses, que não possuem qualquer registo. Também nas Figuras B.17b e Figuras B.17c, são apenas observados os meses de maio a setembro, uma vez que, os restantes são inexistentes. No caso da Figura B.17b também é possível observar alguns valores em falta, no dia 5 de maio, das 20 às 23 horas. As restantes ausências de dados, observadas através do heatmap, podem ser observadas na Tabela A.11, no Apêndice A. Através

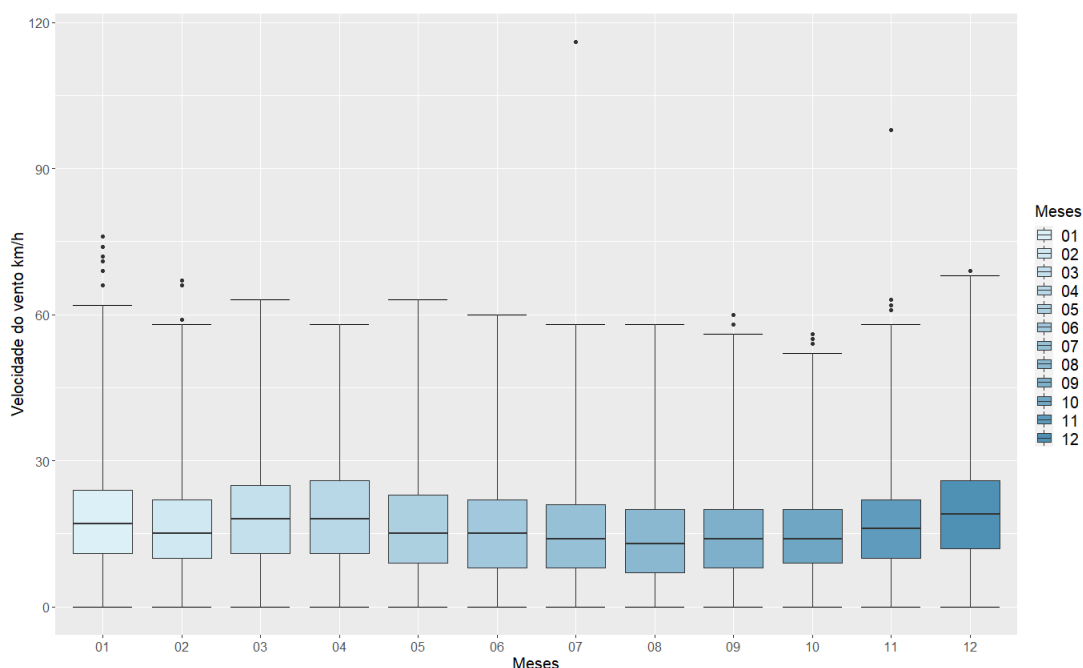


Figura 3.26: *Boxplot* da velocidade do vento onde é possível identificar *outliers*.

da Figura B.17c é possível ainda identificar um dos *outliers* desta variável, nomeadamente o do dia 24 de julho de 2005, já referido anteriormente.

Apesar da ausência de registos, na série temporal da velocidade do vento, poder influenciar o estudo da sua evolução ao longo do período de tempo considerado, 1994 a 2007, já nas Figuras B.17b e Figuras B.17c é notório o aumento da velocidade nos meses de maio a setembro do ano 2004 comparativamente ao de 2005. Ao observar o comportamento da velocidade do vento durante o dia e durante a noite conclui-se que não existe variabilidade significativa ao longo do período registado.

Na Figura 3.27, onde estão representadas as médias anuais da velocidade do vento ao longo do período de tempo registado (1994-2007). É de notar que apenas os anos de 1998 a 2002 e 2007 se encontram completos, pelo que, se observarmos a evolução da tendência nesse período de tempo, repara-se num aumento de 1998 a 2001.

A Figura 3.28 mostra as média mensais da velocidade do vento no período 1994 a 2007. Nela podemos identificar que os meses de Verão apresentam velocidades de vento bem mais reduzidas do que os meses de Inverno. Em particular, o mês de agosto apresenta as velocidades do vento mais baixas (média de 14.71 Km/h), enquanto que o mês de dezembro apresenta uma média mais elevada (19.85 Km/h).

Na Figura 3.29 apresentam-se as distribuições da média da velocidade do vento diária por mês. Podemos observar as distribuições dos vários meses são muito semelhantes, embora os meses de verão apresentem uma caudas mais curtas para velocidades de vento elevadas.

Procurou-se, também, observar o comportamento da tendência aplicando *moving average* aos

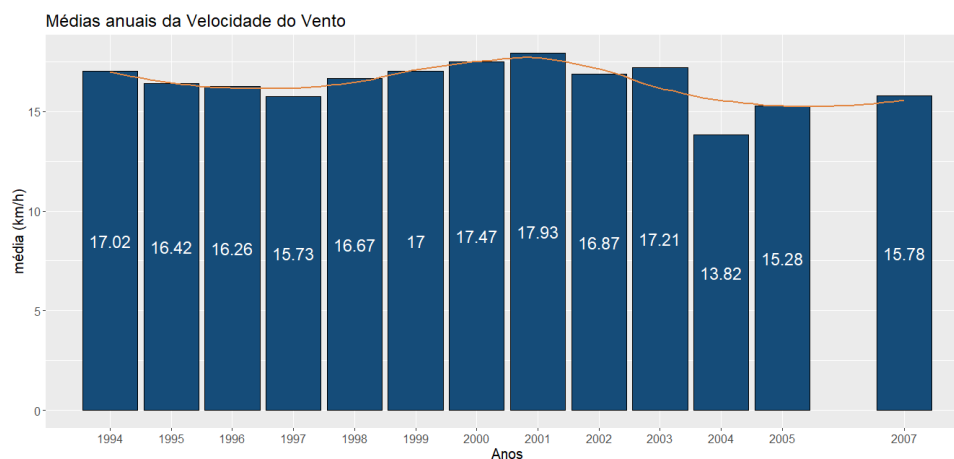


Figura 3.27: médias anuais da velocidade do vento (Km/h)

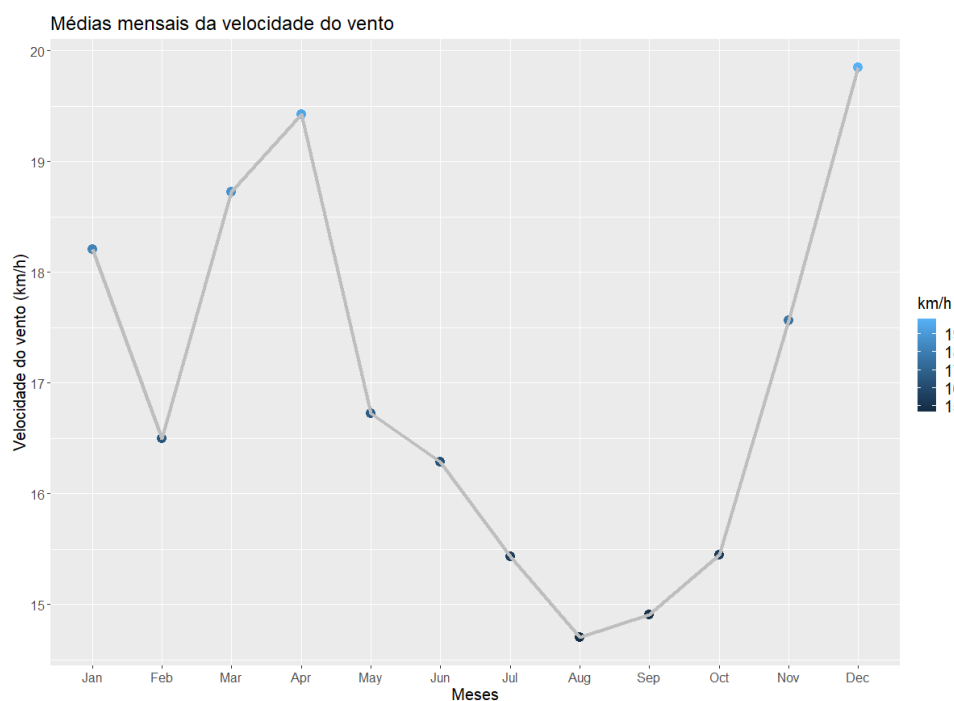


Figura 3.28: Médias mensais da velocidade do vento por mês.

dados relativos a médias diárias da velocidade do vento. A Figura 3.30 mostra a aplicação de três *moving averages*, de diferentes ordens, na mesma série temporal de velocidade do vento com médias diárias, aplicadas ao intervalo de tempo entre 2000 e 2001.

Mais uma vez, o *Augmented Dickey-Fuller* aplicado a esta série temporal obteve o mesmo resultado que as demais estudadas, com um *p-value* inferior a 0.05 identificando a série como estacionária.

Observou-se de seguida, nas Figuras B.18 e B.19 (presentes no Apêndice B), o comportamento dos dados da velocidade do vento, hora a hora, médias diárias, e anuais, nas diferentes sazonalidades possíveis. Verificou-se, na Figura B.19 uma sazonalidade quase impercetível, ainda que essa

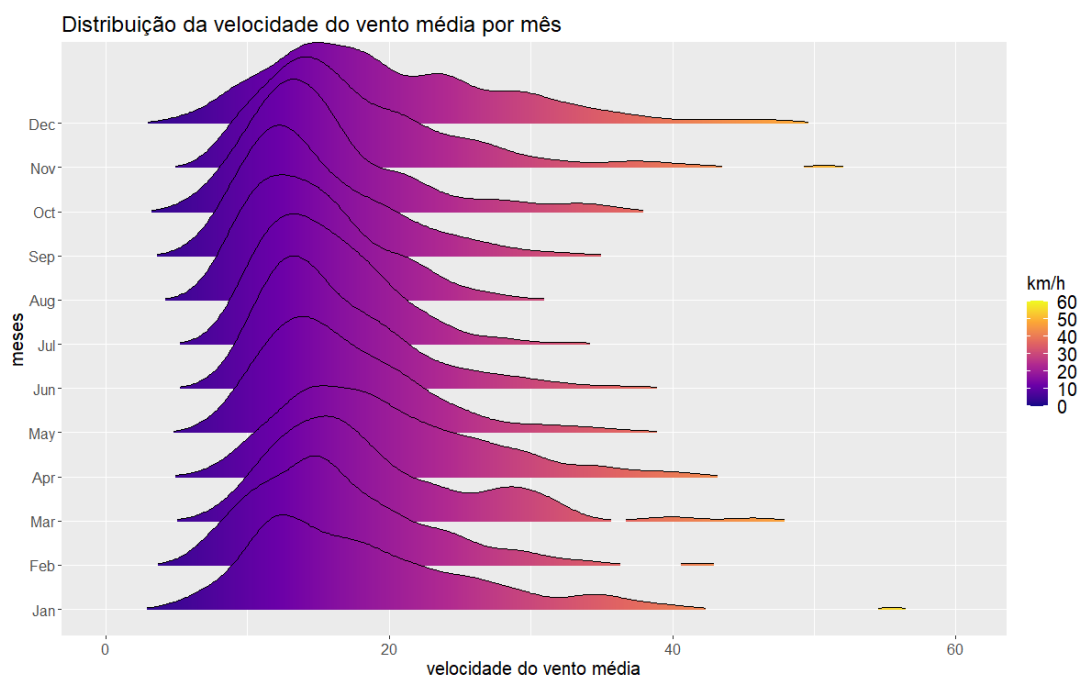


Figura 3.29: Distribuição da velocidade do vento média diária, por mês.

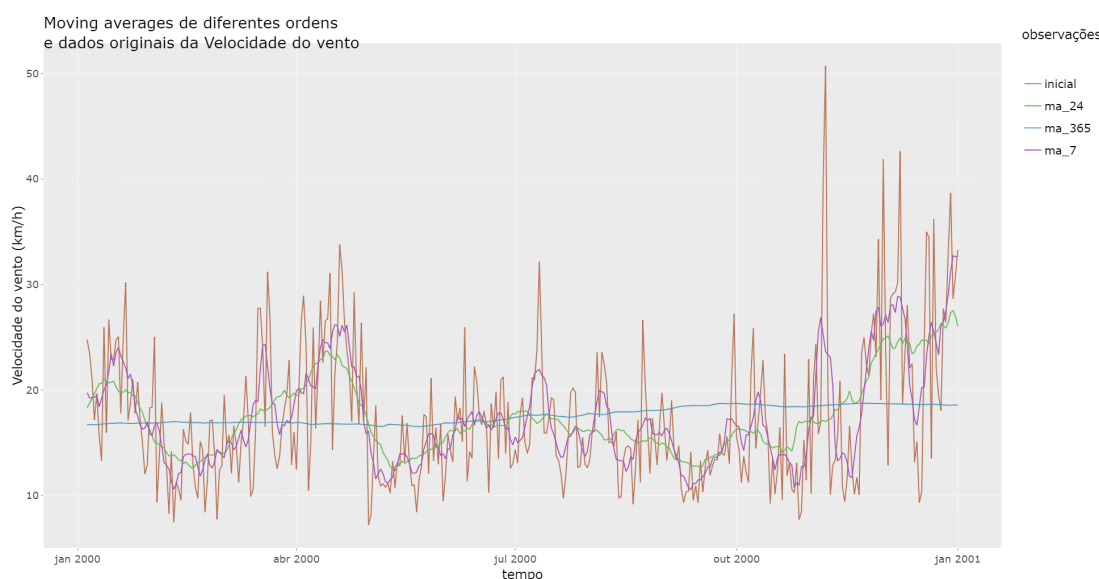


Figura 3.30: Diferentes *moving averages* aplicadas à série temporal da velocidade do vento diária (2000).

e a componente sazonal da Figura 3.31 tenham registado o sinal mais forte entre todas as componentes sazonais das figuras referidas anteriormente, relativas à decomposição das séries da velocidade do vento, presentes no Apêndice B. Ainda assim, é possível observar a presença de sazonalidade em todas as séries temporais de velocidade do vento em causa.

Relativamente à evolução da tendência nesta variável ao longo do tempo considerado (1994 a 2007), esta componente mostra-se relativamente estável ao longo dos anos, à excepção de 1999 caracterizado por um aumento na linha da tendência, opostamente ao ano de 2001 onde se verificou um decréscimo acentuado nos valores de velocidade do vento. Apesar da tendência

observada descrever este comportamento, é importante ter em conta a irregularidade dos dados de velocidade do vento registados, o que implica a fraca veracidade das conclusões obtidas relativamente a esta variável.

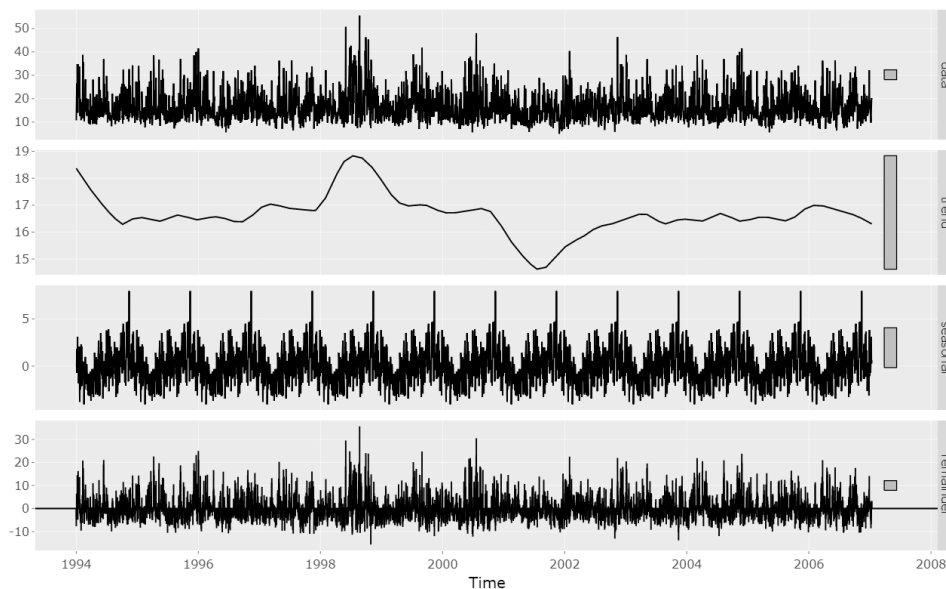


Figura 3.31: Decomposição da série temporal da velocidade do vento (1994-2007): médias diárias (km/h) com sazonalidade anual (frequência = 365).

3.4 Análise de dados multivariada

Procurando perceber de que modo as variáveis meteorológicas se correlacionam no período de tempo disponível, foi realizada uma análise que visa esse mesmo propósito. Para tal, observou-se o comportamento das quatro variáveis sob estudo: a temperatura do ar, a precipitação, a pressão atmosférica e a velocidade do vento, e de que modo as mesmas evoluem no tempo, e se correlacionam.

Atendendo ao facto de que as quatro variáveis em estudo foram registadas em períodos de tempo diferentes, foi criada uma nova série temporal que teve em conta apenas as datas comuns às quatro, procurando evitar pesos desmedidos entre as mesmas. Posto isto, começou por se aplicar o teste de correlação de Pearson a estas variáveis, obtendo os resultados descritos na Figura 3.32. É necessário ter em conta que estes coeficientes são muito sensíveis a valores extremos de dados pelo que não se devem tomar estes resultados como determinantes. Ainda aassim, observando a figura, verifica-se que nenhuma das variáveis se mostra estritamente correlacionada apesar dos coeficientes mais altos corresponderem aos grupos de precipitação e velocidade do vento (com um coeficiente positivo) e pressão atmosférica com precipitação e velocidade do vento (ambas com um coeficiente igual e negativo). Não se sabe se estes valores podem ou não ser consequência da escala reduzida de tempo comum às quatro variáveis.

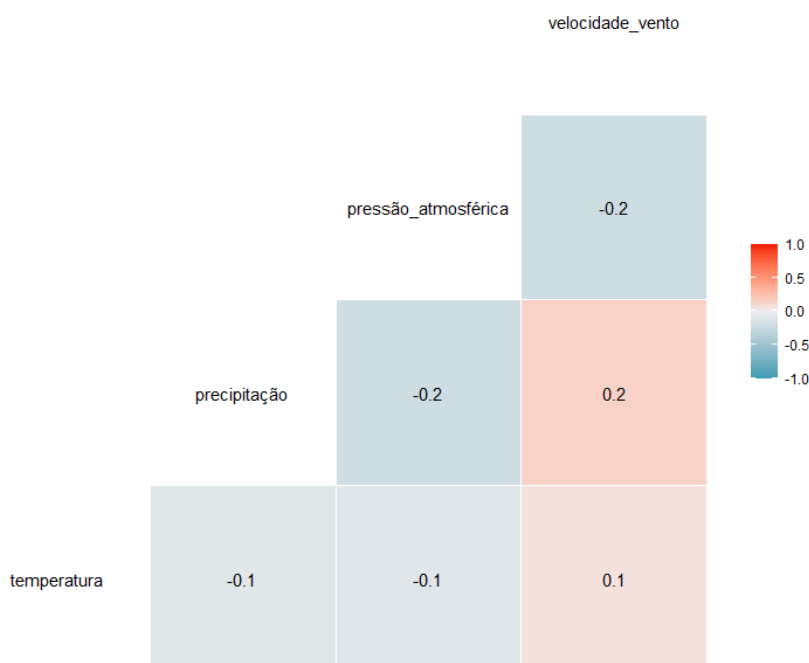


Figura 3.32: Matriz de correlação entre as variáveis meteorológicas (2000-2005).

Foram criados três gráficos de dispersão correspondentes às variáveis que mostraram resultados mais altos no teste de correlação de *Pearson*. A Figura 3.33 representa a única correlação positiva, das três observadas, entre a precipitação e a velocidade do vento cujo coeficiente de correlação é de 0.2. Por outro lado, as Figuras 3.34 e 3.35 mostram a correlação entre a pressão atmosférica com a precipitação e com a velocidade do vento. Nestes, ambos os coeficientes de *Pearson* são negativos, (-0.2) o que implica que para valores mais altos de pressão atmosférica os de precipitação e de velocidade do vento tenderão a ser mais baixos.

Posto isto, procurou analisar-se a correlação entre a temperatura e a precipitação. Na Figura 3.36 encontramos o climograma que relaciona o total de precipitação mensal acumulada com a respetiva temperatura média. Através do mesmo, é possível confirmar o coeficiente negativo que a matriz encontrou entre estas duas variáveis, uma vez que, à medida que as temperaturas aumentam, a quantidade de precipitação diminui. Estas conclusões são mais evidentes durante os meses de verão, de junho a setembro, que registaram as médias de temperatura mais alta.

Ainda observando as possíveis correlações entre a temperatura e a precipitação, foram observados, para cada estação do ano, o respetivo *Quadrant Type Scatter Plot* das anomalias da temperatura e precipitação. Através da análise do mesmo é possível caracterizar um determinado ano como sendo húmido ou seco, e quente ou frio, tendo em conta as anomalias da temperatura e da precipitação. Neste gráfico, as anomalias de precipitação foram expressas em percentagem com a média fixada a zero. A Figura 3.37 mostra os resultados obtidos, para o período de tempo comum às duas variáveis, durante o inverno. Os gráficos das restantes estações podem ser consultados no Apêndice B.

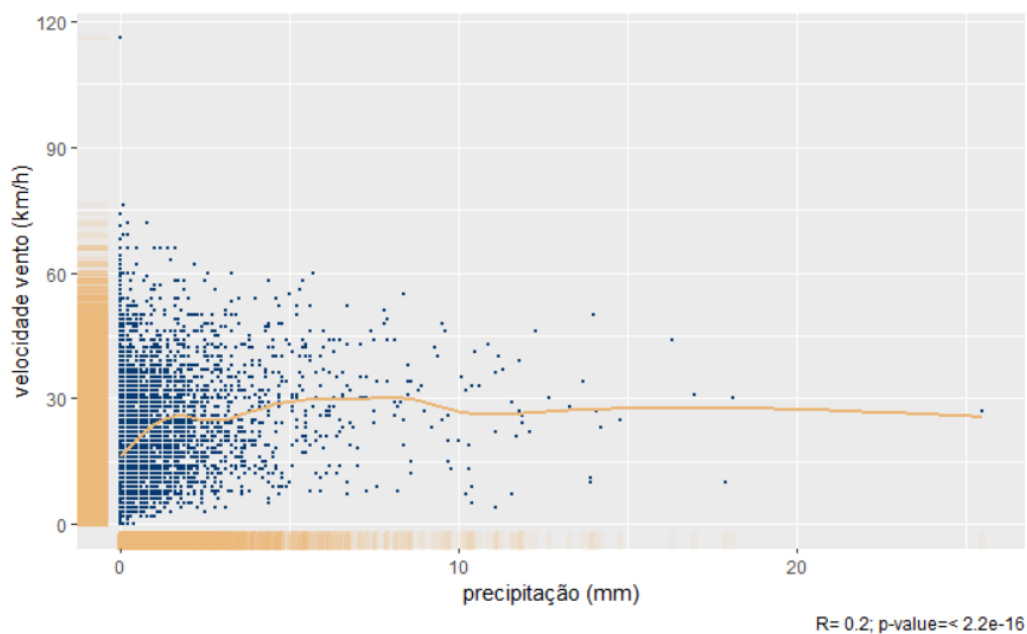


Figura 3.33: Gráfico de dispersão entre a precipitação e a velocidade do vento, linha de tendência e respectivo valor R e $p\text{-value}$ (2000-2005).

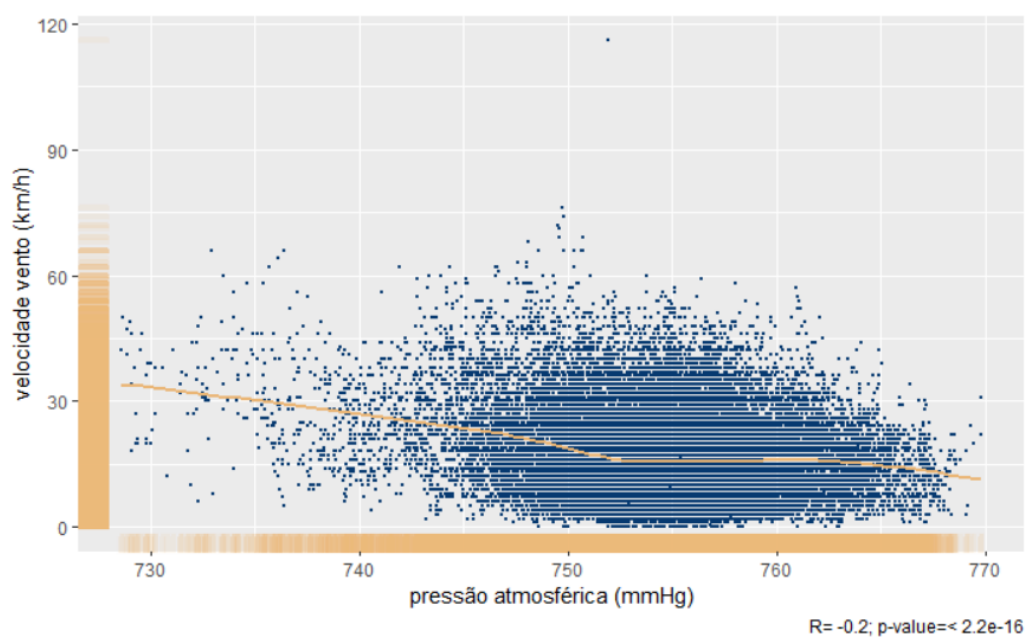


Figura 3.34: Gráfico de dispersão entre a pressão atmosférica e a velocidade do vento, linha de tendência e respectivo valor R e $p\text{-value}$ (2000-2005).

Para além do teste de *Pearson* foram ainda calculados os valores de *normalized cross-correlation* entre as séries temporais atrás mencionadas. Os resultados obtidos foram iguais aos dos obtidos no teste de *Pearson* e encontra-se na Figura 3.38 o exemplo da correlação entre a precipitação e a velocidade do vento, onde as variáveis se correlacionaram no *lag* 0 com um valor de 0.2.

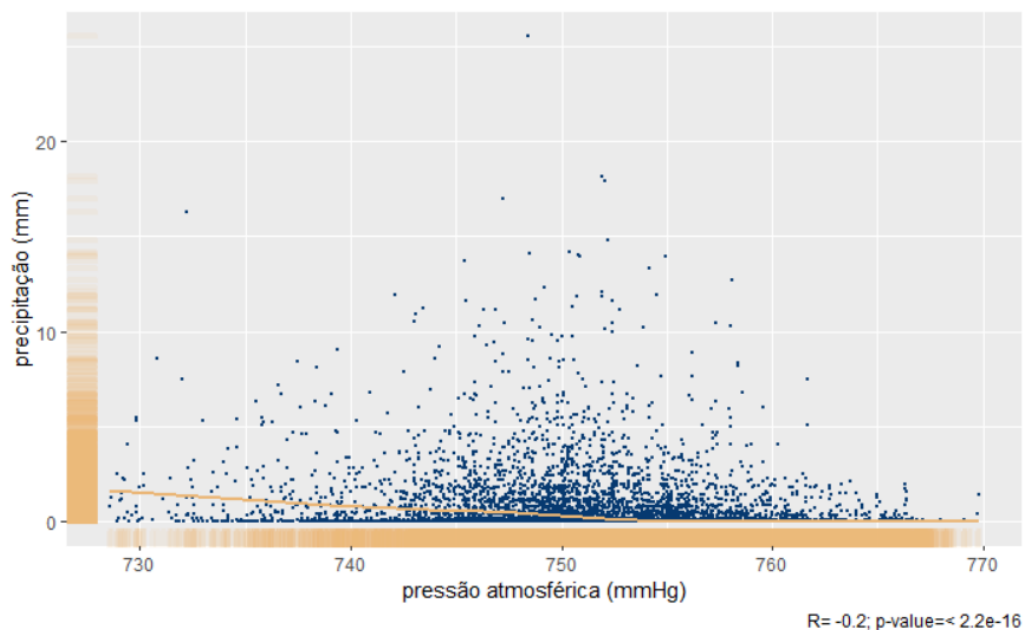


Figura 3.35: Gráfico de dispersão entre a pressão atmosférica e a precipitação, e linha de tendência e respectivo valor R e *p-value* (2000-2005).

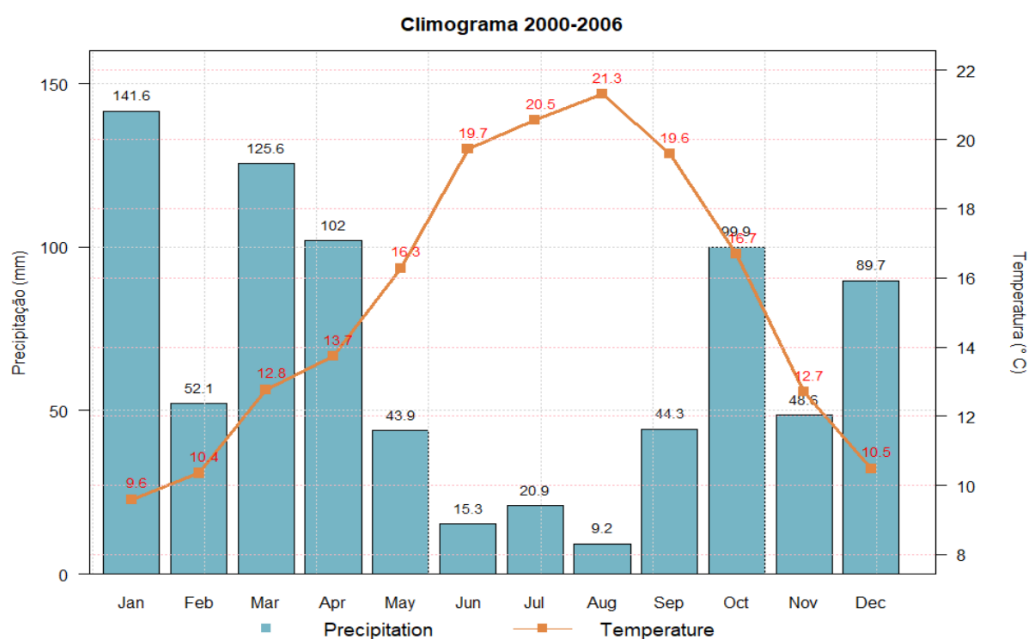


Figura 3.36: Climograma (2000-2006).

3.5 Discussão

Com base na análise exploratória realizada sobre dados meteorológicos registados no período de tempo de 1960 a 2007 pelo IGUP na cidade do Porto, podemos resumir o conjunto das observações principais que foi possível extrair dos dados.

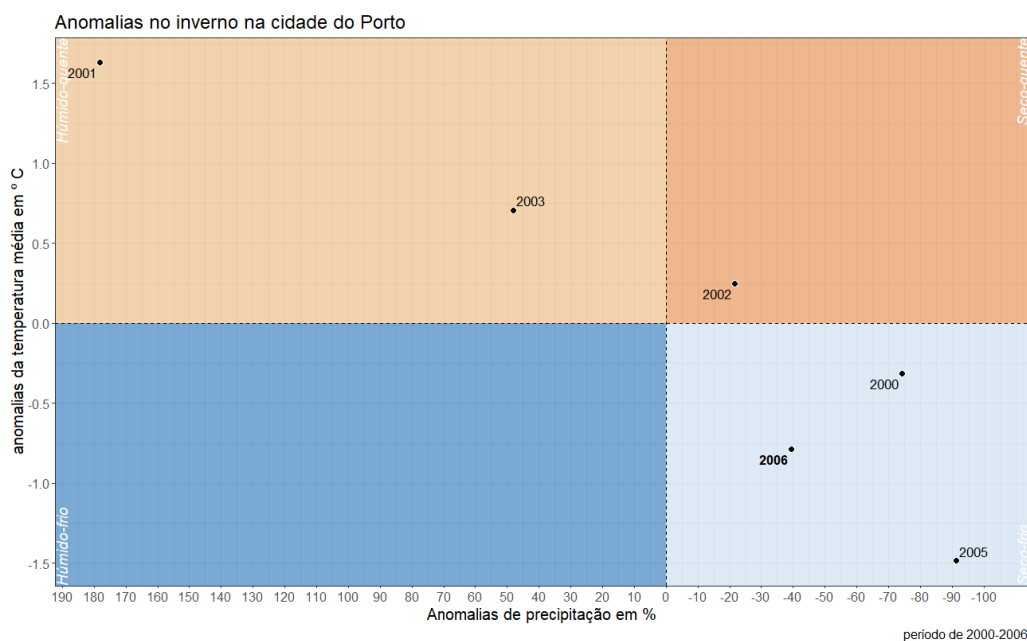


Figura 3.37: *Quadrant Type Scatter Plot* das anomalias de temperatura e precipitação, no inverno (2000-2006).

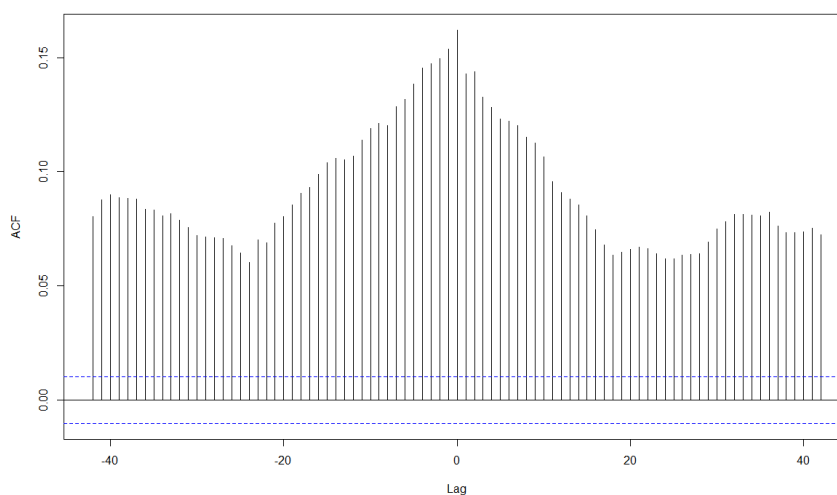


Figura 3.38: *Normalized cross-correlation* entre a precipitação e a velocidade do vento.

Relativamente à temperatura, ainda que com a ausência de alguns registos (entre 1976-1980 e 1995-2000), foi possível observar um aumento dos valores observados, com uma maior incidência nos últimos 5 anos. Analogamente, o mesmo comportamento verificou-se quando se analisaram os valores diários diurnos e noturnos, por ano, ou inclusive observando o período de tempo considerado, no decorrer das estações de ano. Atendendo às temperaturas mínimas, a década de 1960 foi a que registou as temperaturas mais baixas, sendo que, os dias de frio intenso diminuíram e as noites tropicais aumentaram. No caso das temperaturas máximas, foi na década de 2000 que se registaram as temperaturas mais elevadas.

No caso da precipitação, e no período considerado de 1998 a 2005, os anos de 2000 e 2001

registaram uma precipitação elevada. Em 2001 ocorreu uma precipitação forte e muito forte, observando-se depois um decréscimo da mesma, com o ano de 2003 a registar a menor precipitação acumulada. Face a este período de tempo, 2002 foi o ano onde se verificou o maior número de anomalias positivas, isto é, precipitação superior à média de precipitação durante o período considerado, contrariamente ao ano de 2005 que obteve o maior número de anomalias negativas. Face às estações do ano, é no verão, durante os meses de junho, julho e agosto que a precipitação é mais baixa, enquanto que o inverno regista a precipitação mais alta com um total acumulado de 2344.14mm. O mês de janeiro é aquele que regista também um total mais elevado.

No caso da pressão atmosférica, o valor mais baixo foi registado em 2001 com 728.60mmHg e o mais alto em 2005 com 787.00mmHg. Posto isto, numa primeira análise, pode verificar-se que os valores diurnos desta variável foram superiores aos noturnos, ao longo de todo o período de tempo observado, entre 1999 e 2007. Considerando depois os vários meses do ano, janeiro e dezembro registaram os valores mais altos, contrariamente a março e abril com os mais baixos, sendo que maio, junho e julho foram aqueles que registaram menor variabilidade.

Finalmente, e atendendo à análise realizada aos valores observados da velocidade do vento, é de notar a ausência de registo de grande parte das observações no período de 1994 a 2007. Analisando o comportamento desta variável no período em análise, é nos meses de verão que se verifica velocidades do vento mais baixas, e são os meses de inverno que apresentam velocidades mais elevadas.

Considerando o período de observações comuns às quatro variáveis analisadas, foi difícil estabelecer uma correlação entre as mesmas cujos resultados fossem evidentes. Ainda assim, tomando apenas em consideração os registos de temperatura e de precipitação, durante o período de 2000 a 2006, podemos afirmar que, durante o inverno, o ano de 2001 foi o mais quente e húmido, com uma anomalia de temperatura e de precipitação positiva, face à média do período de tempo considerado. O ano de 2005 apresentou as anomalias de temperatura e de precipitação mais negativas, caracterizando-se como um ano seco e frio. Já no verão (período 2000-2005), o ano de 2005 apresentou a maior anomalia de temperatura, classificando-se como seco-quente.

Capítulo 4

Testes de Homogeneidade

Neste capítulo é apresentada a aplicação dos testes de homogeneidade, descritos no Capítulo 2, às séries temporais em estudo. São discutidos os seus resultados, procurando classificar as séries qualitativamente de acordo com os mesmos.

4.1 Aplicação de testes de homogeneidade

Os testes seleccionados para estudar a homogeneidade das séries foram: o *Buishand range test*, *Standard normal homogeneity test* (SNHT) para deteção única de *breaks*, *Pettitt test* e *Von Neumann ratio test*. Estes testes de homogeneidade são, tipicamente, os mais utilizados quando aplicados a séries temporais climáticas. Os quatro testes assumem uma hipótese nula de que os valores anuais Y_i da variável testada Y são independentes e identicamente distribuídos. Já na hipótese alternativa, os testes SNHT, *Buishand range* e *Pettitt test* assumem a presença de uma mudança gradual na média (*break*), e são capazes de localizar o ano onde este ocorreu. O teste de *Von Neumann*, sob a hipótese alternativa, assume que os dados não se encontram distribuídos de forma aleatória, e não dá informações acerca do ano onde ocorreu o *break*, em séries de frequência mensal ou anual. Embora os três primeiros testes sejam semelhantes em muitas características, o SNHT deteta *breaks* mais facilmente, se se encontrarem perto do início ou do fim de uma série, enquanto o *Buishand range test* e *Pettitt test* são mais sensíveis no caso dos *breaks* se encontrarem no meio da série.

Conjuntos de dados climáticos anuais conseguem melhores resultados de homogenização em relação a conjuntos de dados mensais, devido ao aumento de variabilidade das séries temporais [49]. Por esse motivo, para aplicação dos testes e obtenção de resultados, foram consideradas as médias anuais mínimas e máximas, o intervalo (diferença entre máximo e mínimo) e valores médios observados por ano, para as séries temporais da temperatura, precipitação, pressão atmosférica e velocidade do vento. As observações máximas e mínimas podem responder de modo diferente a possíveis variações na localização dos instrumentos de medição, daí a importância de utilizar o intervalo entre estas duas medidas. A deteção de possíveis *breaks* pode ser fraca nas séries dos

valores máximos e mínimos, mas visível quando observado o comportamento do intervalo [39].

Após a aplicação dos testes de homogeneidade, e caso a classificação das séries de uma determinada variável o justifique, foi aplicado o teste *Mann-Kendall* procurando analisar a tendência monotónica dessa variável. De acordo com [55], depois de aplicados os três testes que detetam o ano de ocorrência de um específico *break*, se dois deles identificarem um mesmo ano, com um determinado nível de confiança, podemos assumi-lo como um ano de não-homogeneidade. Para os quatro testes em estudo, a hipótese nula será rejeitada caso os valores obtidos nos testes excedam um determinado nível de confiança, ou valor crítico.

De seguida, serão apresentados os resultados da aplicação destes testes às séries temporais em estudo: temperatura, precipitação, pressão atmosférica e velocidade do vento.

4.1.1 Temperatura

Foi analisada a homogeneidade da variável de temperatura no período de tempo registado pela mesma, isto é de 1960 a 2007.

A Figura 4.1 apresenta o comportamento das séries temporais anuais de temperatura, mínimas, máximas, intervalo (DTR) e médias, para os três testes aplicados, capazes de identificar o ano do *break*, e respetivos valores críticos (de 5% e 1%) para níveis de confiança de 95% e 99%, respetivamente. Os valores críticos utilizados como referência nos gráficos podem ser consultados na Tabela C.1 do Apêndice C.

Conforme se pode verificar em todos os testes, e para todas as séries temporais em estudo, as hipóteses nulas foram rejeitadas uma vez que excederam os dois níveis de confiança atrás referidos. Isto significa que a hipótese nula pode ser rejeitada e, portanto, a hipótese alternativa que sugere a presença de não-homogeneidade é provável nestas séries.

Os resultados destes três testes podem ser observados na Tabela 4.1 que regista o resultado de cada teste e o respetivo ano identificado como *break*, na série temporal, assim como o resultado do teste de *Vonn Neumann*. Os testes aplicados identificam um possível *break* por cada iteração, pelo que, o ano identificado como *break* nos gráficos correspondentes é identificado pelo valor máximo obtido no teste. Todos os resultados obtidos registaram um *p-value* inferior a 0.05 e 0.5 no *Pettitt Test* pelo que podemos afirmar que são confiáveis. De acordo com os dados da tabela verifica-se que os anos de 1985, 1982 e 1986 podem ser considerados *breaks* na série de temperatura, uma vez que dois ou três, dos três testes aplicados os identificaram como tal. No caso do teste de *Von Neumann*, todos os resultados obtidos (>2), indicaram a presença de não homogeneidades nas séries em estudo, rejeitando a hipótese nula para níveis de confiança de 95%.

Atendendo ao facto de que todos os testes aplicados, isto é, mais de dois testes, rejeitaram a hipótese nula, a série temporal de dados de temperatura pode classificar-se como classe 3 ou *suspect*, o que significa que possíveis análises de tendência e respetiva variabilidade podem ser consideradas como falsas, e requerem muita atenção e cuidado.

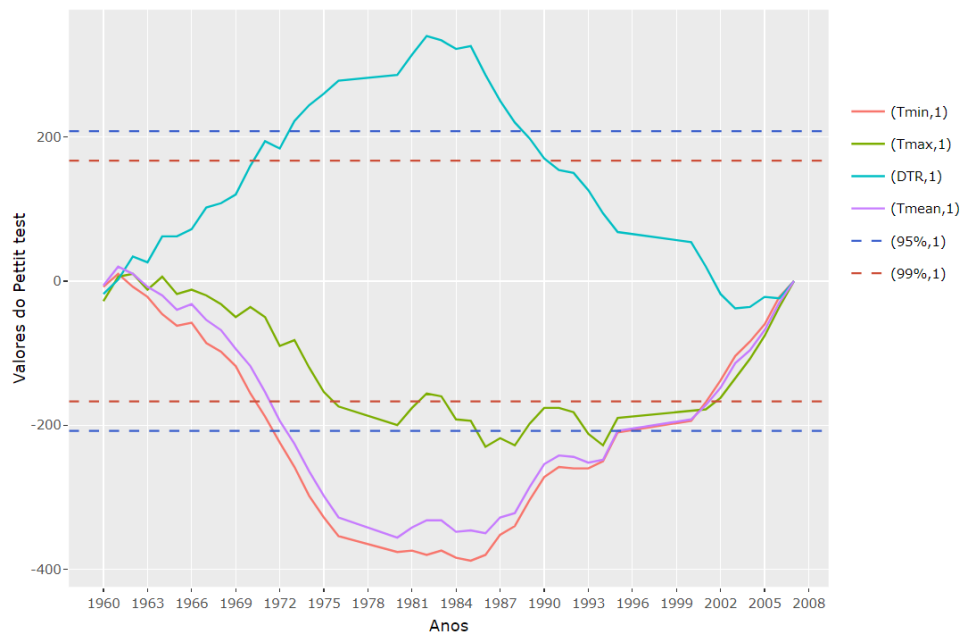
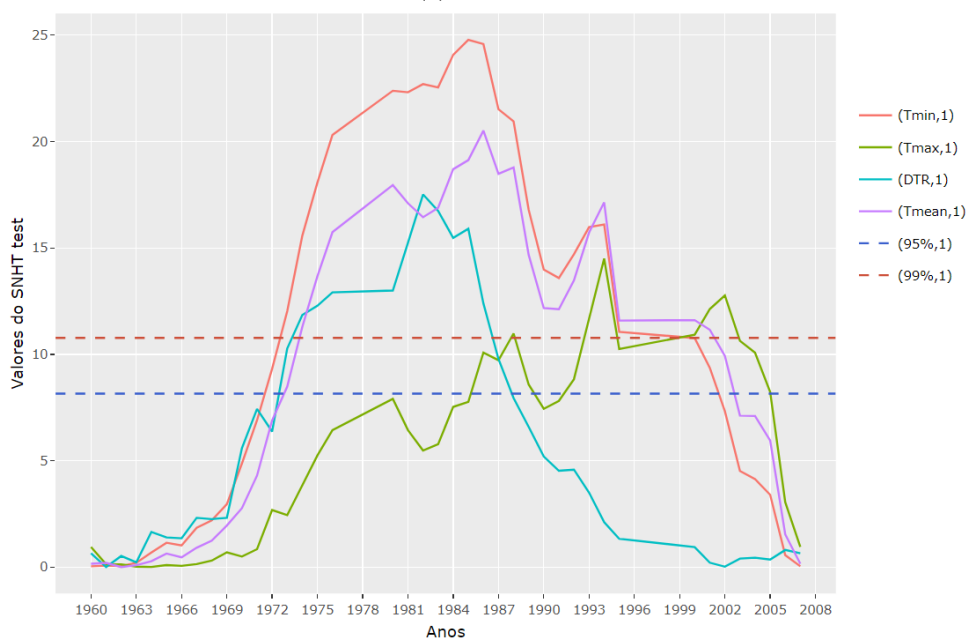
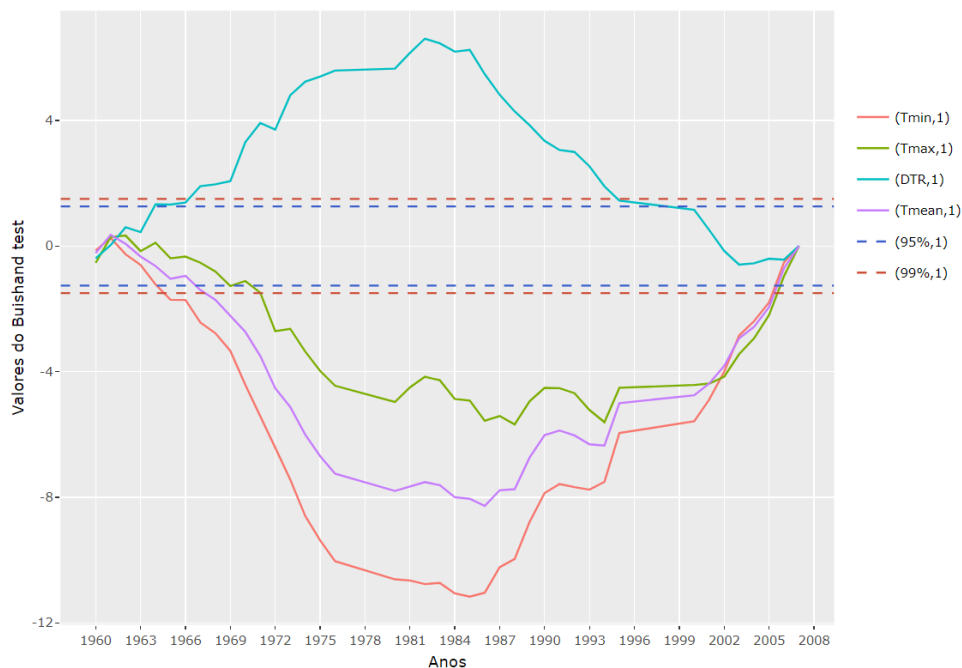
(a) *Pettit test*(b) *SNHT test*

Figura 4.1: Gráficos dos testes de homogeneidade às séries temporais de temperatura anual, valores médios (T_{mean}), máximos (T_{max}), mínimos (T_{min}), intervalo (DTR) e níveis de confiança de 95 e 99%.



(c) Buishand range test

Figura 4.1: Gráficos dos testes de homogeneidade às séries temporais de temperatura anual, valores médios (T_{mean}), máximos (T_{max}), mínimos (T_{min}), intervalo (DTR) e níveis de confiança de 95 e 99%.

Tabela 4.1: Comparação dos resultados obtidos para as séries anuais de temperatura, mínimas (T_{min}), máximas (T_{max}), intervalo (DTR) e médias (T_{mean}).

série temporal	<i>Pettitt Test</i> resultado (ano)	<i>SNHT test</i> resultado (ano)	<i>Buishand range Test</i> resultado (ano)	<i>Von Neuman</i> (resultado)
T_{min}	388 (1985)	24.78 (1985)	2.53 (1985)	0.43
T_{max}	230 (1986)	14.50 (1994)	1.69 (1988)	1.44
DTR	340 (1982)	17.51 (1982)	2.28 (1982)	0.87
T_{mean}	356 (1980)	20.51 (1986)	2.33 (1986)	0.54

4.1.2 Precipitação

Do ponto de vista da precipitação, o estudo da homogeneidade desta série temporal não se presta à aplicação dos testes anteriores de igual modo. Isto porque, a precipitação é uma variável que apresenta grande variabilidade entre os registos e a mesma pode resultar em incertezas quando se trata de estudar a sua homogeneização. Neste caso, flutuações climáticas verdadeiras, na precipitação diária, podem ser interpretadas como pontos de mudança (*breaks*) e removidas, erroneamente, das séries temporais, como não-homogeneidades, aquando da correção das mesmas. Estas não-homogeneidades podem ainda apresentar uma magnitude elevada causada por essas situações meteorológicas variáveis.

Apesar destas séries serem muito importantes para caracterizar o clima e respetiva evolução no tempo, para ultrapassar as questões anteriormente referidas são requeridas técnicas complexas

pelo que é necessário melhorar alguns métodos de homogeneização, antes da sua utilização em séries climáticas mensais e anuais de precipitação. Por estes motivos, optou-se por não aplicar os testes de homogeneidade referidos a esta série temporal.

Ainda assim, o teste não paramétrico *Mann-Kendall* (MK) é popularmente utilizado para determinar a significância estatística da tendência de séries temporais hidrológicas, como é o caso da precipitação. Foi aplicado este teste à série temporal anual da precipitação, composta por observações entre o intervalo de tempo 1998-2005, uma vez que o ano de 1997 e de 2006 apenas inclui registos do mês de janeiro, podendo influenciar os resultados obtidos. Antes de aplicar este teste, é necessário perceber, se os dados se encontram ou não correlacionados no período de tempo registado. Para esse efeito, foram gerados os gráficos da Figura 4.2a, com os valores da *autocorrelation*, e da Figura 4.2b com os da *partial autocorrelation*.

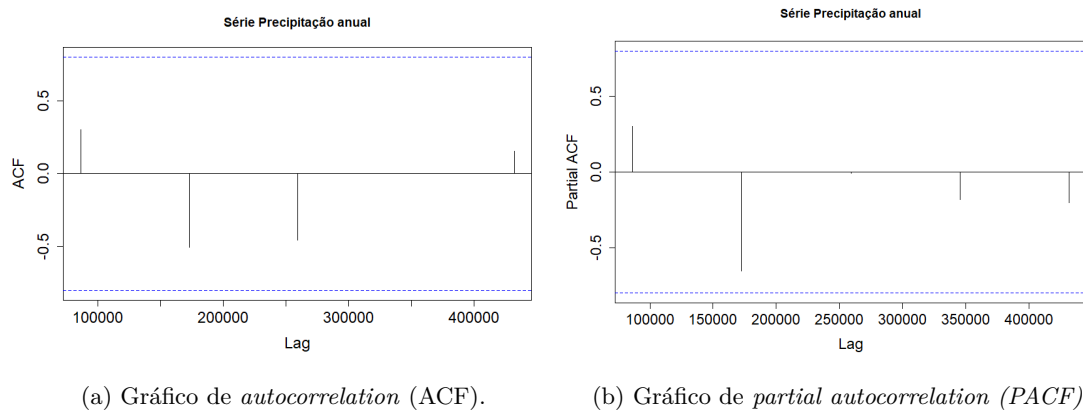


Figura 4.2: Gráficos de ACF e PACF da série temporal da precipitação anual (1998-2005).

De acordo com o observado nos gráficos da Figura 4.2, a *autocorrelation* e *partial autocorrelation* da série anual não são significativas. Apenas se os “traços” verticais atingissem as linhas azuis do gráfico. Posto isto, pode proceder-se à aplicação do teste *Mann-Kendall*. Apesar de um valor de S negativo indicar uma tendência decrescente, os resultados deste teste revelaram a ausência de tendência nesta série temporal. Isto porque, ainda que $S = -3$, combinado com um p -value de $0.7071 > 0.05$ é indicativo de inexistência de tendência. No entanto, sabendo à priori que se trata de uma limitação deste teste, o tamanho da série temporal utilizada ($n = 6$) poderá ter sido pequeno demais para conseguir detetar a tendência, podendo isto ser mitigado num conjunto de dados maior.

Posto isto, testou-se a mesma série temporal de precipitação mas atendendo agora a uma janela temporal de médias mensais, procurando verificar se os resultados seriam diferentes dos obtidos anteriormente. De modo análogo, e atendendo aos gráficos das Figuras C.1, no Apêndice C, com uma *autocorrelation* e *partial autocorrelation* não significativas, procedeu-se ao teste *Mann-Kendall*. Considerando os valores obtidos no teste, com um p -value de $0.003764 < 0.05$, ou seja, significativo, e $S = -466$, podemos concluir que a tendência é decrescente, (um valor positivo de S apontaria para uma tendência crescente).

4.1.3 Pressão atmosférica

Sobre a pressão atmosférica foram aplicados os testes de homogeneidade referidos, no período de dados registados, de 1999 a 2007.

Na Figura 4.3 podemos encontrar o comportamento das séries anuais de pressão atmosférica sob teste, respetivamente, mínimas, máximas, intervalo e médias, aplicadas ao *Standard Normal Homogeneity Test* e respetivos valores críticos de 5 e 1%. Nenhuma das séries atinge o limite destes níveis de confiança o que, mais uma vez, descarta a hipótese alternativa de que existam *breaks* na série. Os testes *Pettitt* e *Buishand range* mostram um comportamento semelhante ao observado na Figura 4.3, e podem ser consultados no Apêndice C.

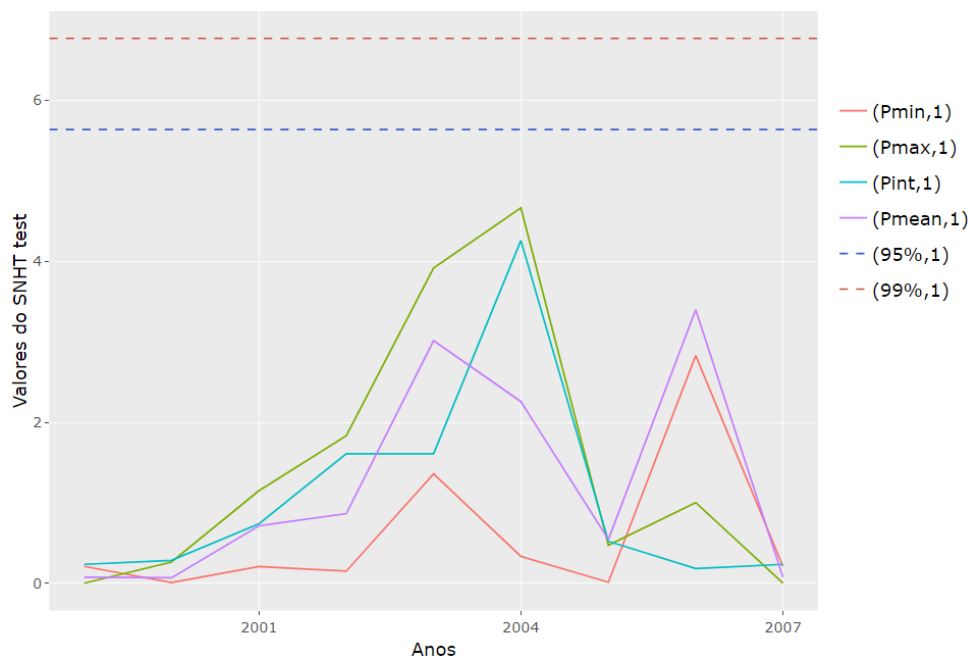


Figura 4.3: *Standard Normal Homogeneity Test* aplicados às series anuais de pressão atmosférica mínimas (P_{min}), máximas (P_{max}), intervalo (P_{int}) e médias (P_{mean}) e níveis de confiança de 95 e 99%.

Para esta variável, o tamanho do conjunto de dados ($n < 10$) não permitiu a aplicação do teste de *Vonn Neumann*. No entanto, os resultados obtidos, pelos restantes testes, encontram-se sumariados na Tabela 4.2. A negrito encontram-se os valores cujo *p-value* indica um resultado não significativo do ponto de vista estatístico, isto é, superior a 0.5 no caso do *Pettitt test* e superior a 0.05 nos restantes. Apenas as séries com valores anuais de pressão atmosférica máximos, intervalo e médias, obtiveram resultados estatisticamente significativos e, para níveis de confiança de 95% e de 99% nenhum atingiu a hipótese alternativa do respetivo teste.

Tendo em conta os testes que rejeitaram a sua hipótese nula, nas séries estudadas, podemos inserir os dados da pressão atmosférica na classe 1 ou *useful*, considerando a série suficientemente apta para análise de tendência.

Dada esta classificação, procurou analisar se o comportamento da tendência através do teste, não

Tabela 4.2: Comparação dos resultados obtidos para as séries anuais de pressão atmosférica mínimas (P_{min}), máximas (P_{max}), intervalo ($P_{intervalo}$) e médias (P_{mean}).

série temporal	<i>Pettitt test</i> resultado (ano)	<i>SNHT test</i> resultado (ano)	<i>Buishand range test</i> resultado (ano)
P_{min}	10 (2003)	2.83 (2006)	0.72 (2003)
P_{max}	16 (2003/4)	4.66 (2004)	1.01 (2003)
$P_{intervalo}$	14 (2004)	4.26 (1994)	1.11 (1996)
P_{mean}	14 (2003)	3.40 (2006)	0.95 (2003)

paramétrico, *Mann-Kendall*. Começou-se por verificar se estas séries se correlacionavam, através dos gráficos de *autocorrelation* e *partial autocorrelation* e concluiu-se que não, isto é, as séries poderiam ser submetidas ao teste *Mann-Kendall*. No entanto, os resultados deste teste, ainda que com valores $S > 0$, mostraram *p-values* superiores a 0.05, concluindo assim a ausência de tendência nas séries em questão. Tal como na precipitação, estes resultados não garantem que não exista tendência, uma vez que os mesmos podem ser influenciados pelo tamanho reduzido do conjunto de dados, neste caso $n = 9$, podendo em conjuntos maiores, detetar possíveis tendências.

4.1.4 Velocidade do vento

Quando se trata de aplicar testes de homogeneidade a séries temporais é imperativo que as mesmas sejam regulares na sua frequência temporal para que os possíveis *breaks* sejam devidamente identificados. Por esse motivo, ainda que existam registos da velocidade do vento de 1994 a 2007 apenas os anos de 1998 a 2002 e 2007 se encontram completos, pelo que foram aplicados os testes de homogeneidade apenas a este intervalo de tempo.

Na Figura 4.4 podemos observar o comportamento do teste *Buishand range* aplicado às séries anuais de velocidade do vento mínimas, máximas, intervalo e médias. Este foi o único teste que registou a hipótese alternativa para as séries anuais máximas e médias, para níveis de confiança de 95 e 99%. Os resultados dos restantes testes podem ser consultados no Apêndice C.

De modo idêntico à pressão atmosférica, o tamanho do conjunto de dados ($n = 6$) é inferior a 10, pelo que impossibilita a aplicação do teste de *Vonn Neuman*. Os resultados dos restantes testes encontram-se sumariados na Tabela 4.3 para as respetivas séries com os valores anuais da velocidade do vento. A negrito encontram-se os resultados estatísticos dos testes cujo *p-value* foi superior a 0.5 no caso do *Pettitt test* e superior a 0.05 nos restantes.

Observando os valores da Tabela 4.3, nenhum dos testes obteve resultados estatisticamente significativos, e tendo em conta que, à exceção dos casos observados na Figura 4.4, os outros testes obtiveram resultados que assumem a hipótese nula, a série temporal de dados de velocidade do vento pode classificar-se homogénea, pertencendo à classe 1 ou *useful*.

Dada a classificação da série procurou analisar-se o comportamento da sua tendência através do teste *Mann Kendall*. Aplicaram-se os testes de *autocorrelation* e *partial autocorrelation* às quatro

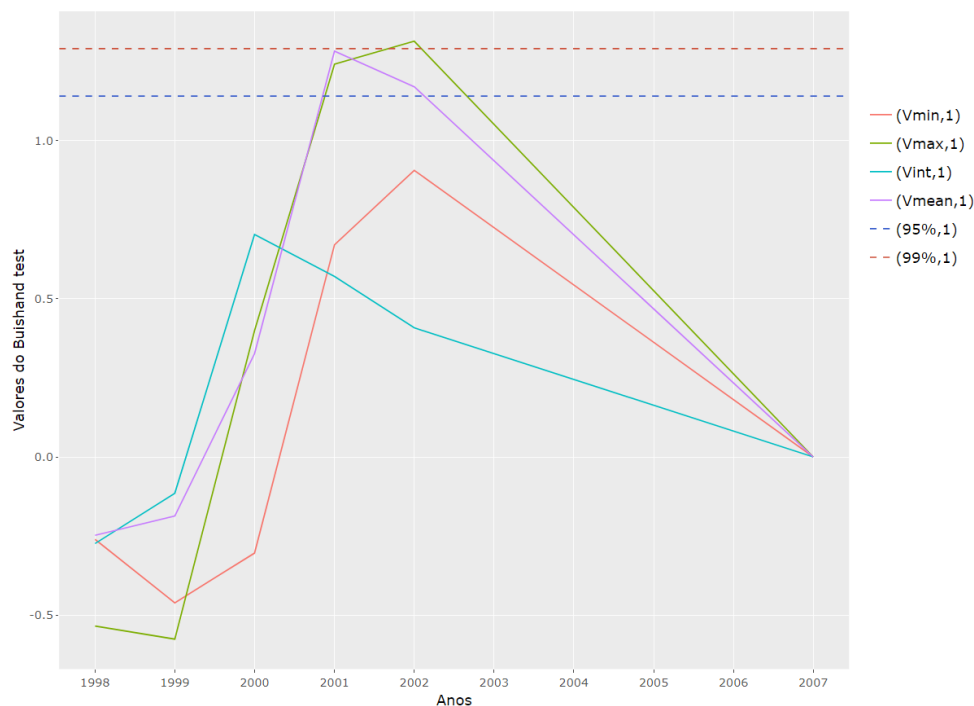


Figura 4.4: *Buishand range test* aplicado às séries anuais de velocidade do vento mínimas (V_{min}), máximas (V_{max}), intervalo (V_{int}), médias (V_{mean}) e níveis de confiança de 95 e 99%.

Tabela 4.3: Comparação dos resultados obtidos para as séries anuais de velocidade do vento mínimas (V_{min}), máximas (V_{max}), intervalo ($V_{intervalo}$) e médias (V_{mean}).

série temporal	<i>Pettitt test</i> resultado (ano)	<i>SNHT test</i> resultado (ano)	<i>Buishand range test</i> resultado (ano)
V_{min}	5 (2002)	2.51 (2002)	0.89 (2002)
V_{max}	5 (2002)	2.82 (2002)	0.89 (2002)
$V_{intervalo}$	6 (2001)	1.68 (2000)	0.90 (2000)
V_{mean}	6 (2001)	3.13 (2002)	0.86 (2001)

séries de velocidade do vento em estudo, procurando saber se se correlacionavam entre si e os resultados foram negativos. No entanto, ao aplicar o teste *Mann Kendall* em todos os casos se obtiveram *p-values* inferiores a 0.05 o que implica que, independentemente do resultado estatístico deste teste, conclui-se a ausência de tendência. Este resultado pode dever-se ao facto do conjunto de dados ter um tamanho reduzido ($n = 6$) e não deve ser considerado como absoluto.

4.2 Discussão

Depois de realizados os diferentes testes de homogeneidade às variáveis climáticas em estudo aferiu-se que apenas a série com dados de temperatura se classificou como não homogénea e de classe 3 ou *suspect*.

Tanto os dados de pressão atmosférica como os da velocidade do vento foram classificados como

useful, na classe 1 e foram submetidos ao teste *Mann Kendall*. Este teste avalia o comportamento da tendência. Contudo, o tamanho de ambos os conjuntos de dados mostrou-se demasiado pequeno no intervalo de tempo considerado, para que se reconhecesse uma possível tendência de forma significativa.

No caso da precipitação, dada a sua variabilidade mais complexa e tendo em conta também o tamanho reduzido do conjunto de dados ($n = 6$) optou-se por não se testar a homogeneidade da série, apesar de também ter sido submetida ao teste *Mann Kendall*. Pelos mesmos motivos atrás referidos, neste teste, os resultados foram inconclusivos ao testar a série anual, devido ao pequeno tamanho do conjunto de dados, mas indicativos de uma tendência decrescente, quando analisados do ponto de vista mensal.

Séries temporais com presença de não-homogeneidades podem causar tendências enviesadas nos dados e interferir com a avaliação das mesmas ao longo do tempo. Por esse motivo é muito importante ter registos climáticos precisos e daí advém o interesse em corrigir séries não homogêneas. No entanto, o processo de correção é muito delicado e deve ser complementado sempre com informações históricas das estações onde foram feitos os registos. Ainda assim, o número de procedimentos para corrigir descontinuidades nas séries é limitado e por vezes são excluídas análises a séries que não possuam metadados disponíveis, considerando apenas para estudo os períodos homogêneos mais longos [49]. Por estes motivos, nesta dissertação optou-se por não se proceder ao processo de correção da série temporal da temperatura, a única para a qual foram identificados *breaks* de forma significativa.

Capítulo 5

Conclusões

Neste capítulo, com base no tema desta dissertação, *Análise de Séries Temporais de Dados Meteorológicos da Cidade do Porto*, são apresentadas as conclusões obtidas face ao desenvolvimento deste projeto, assim como as principais limitações encontradas durante o mesmo. São também apresentadas algumas propostas de trabalho futuro, com o propósito de mitigar alguns dos obstáculos encontrados.

5.1 Contribuições

Este estudo compreendeu a análise das variáveis de temperatura, precipitação, pressão atmosférica e velocidade do vento, registados pelo IGUP, no período de 1960 a 2007 na cidade do Porto.

Tendo em conta os objetivos inicialmente propostos, e considerando o período de tempo registado por cada variável, foi possível:

- identificar valores em falta, erros de consistência e *outliers* globais, contextuais e extremos;
- visualizar e interpretar, graficamente, o comportamento das variáveis em estudo e em diferentes intervalos de tempo, nomeadamente, hora a hora, dias, meses e anos;
- analisar as principais componentes das séries temporais dos dados em estudo, especificamente, tendência, sazonalidade e componente restante;
- estudar alguns índices característicos da temperatura e precipitação;
- observar as relações entre as variáveis meteorológicas de 2000 e 2005 (anos comuns às quatro variáveis em estudo);
- classificar as séries temporais face à sua homogeneidade, interpretando os resultados dos testes aplicados.

Numa primeira fase procedeu-se ao pré-processamento dos dados, visando a identificação e remoção de *missing values*, erros de consistência e de diferentes tipos de *outliers*. Posto isto, procedeu-se à análise exploratória das séries temporais em causa. Através da mesma, obtiveram-se as principais conclusões acerca do comportamento de cada variável, no respetivo período de tempo considerado, assim como, do ponto de vista de outros intervalos de tempo, nomeadamente, diários, mensais e anuais.

De um modo geral, podemos destacar o valor elevado da precipitação em 2001, que se evidenciou face aos restantes. Também é de realçar o comportamento da temperatura, que registou valores mais baixos nos anos iniciais, apresentando uma tendência crescente nos últimos anos da série.

Foram ainda observadas as possíveis relações entre as quatro variáveis, durante o intervalo de tempo comum a todas, compreendido de 2000 a 2005.

Finalmente, foram aplicados diferentes testes de homogeneização a cada uma das variáveis, com vista a classificá-las qualitativamente, discutindo os resultados obtidos. A temperatura foi a única que se revelou não-homogênea classificando-se assim como classe 3 ou *suspect*.

5.2 Limitações do estudo e trabalho futuro

Ainda que os principais objetivos tenham sido cumpridos, ao longo do presente estudo foram encontradas algumas limitações importantes. Em particular, os registos horários das diferentes variáveis apresentavam muitos valores em falta. Sendo o intervalo de tempo de observações tão reduzido, torna-se difícil preencher lacunas de dados mensais ou mesmo anuais, de modo a tornar as séries temporais regulares no tempo. Nesse sentido, o estudo e análise do comportamento destas variáveis poderá não ser totalmente fidedigno, para a realidade observada entre os primeiro e último anos de registo. Também a presença de erros nos ficheiros iniciais e de *outliers* nos dados registados, poderiam ser mitigadas através de uma pré-confirmação por parte de peritos desta área, com vista a obter resultados o mais próximos possíveis da realidade.

Finalmente, em investigações futuras seria, igualmente interessante, reunir metadados relativos às variáveis em estudo, ou mesmo séries temporais de estações mais próximas, que facilitassem a correção de não-homogeneidades encontradas, e comparar os seus resultados aos dados iniciais. Deste modo, seria possível fazer uma análise mais detalhada da homogeneidade das séries temporais.

Resultados da análise da qualidade dos dados

A Figura A.1 apresenta o o *printscreen* de um ficheiro Excel, com dados horários relativos à velocidade do vento, do mês de julho de 2007.



A.1 Erros de consistência

Inconsistências encontradas na importação dos dados dos ficheiros excel.

A.1.1 Temperatura

Tabela A.1: Erros de consistência encontrados ao importar os dados da Temperatura.

Ficheiro (ano)	Sheet	Tipo de erro
1983	6	o mês da sheet não coincide
1983	7	o mês da sheet não coincide
1984	-	meses a mais nas sheets
1985	6	o ano da sheet não coincide com o ano do ficheiro
1985	7	o ano da sheet não coincide com o ano do ficheiro
1985	8	o ano da sheet não coincide com o ano do ficheiro
1985	9	o ano da sheet não coincide com o ano do ficheiro
1985	10	o ano da sheet não coincide com o ano do ficheiro
1985	11	o ano da sheet não coincide com o ano do ficheiro
1986	12	o mês da sheet não coincide
1986	12	o ano da sheet não coincide com o ano do ficheiro
1988	4	o ano da sheet não coincide com o ano do ficheiro
1988	12	o ano da sheet não coincide com o ano do ficheiro
1960	8	ausência de dia/s no mês
1975	4	ausência de dia/s no mês
1975	8	ausência de dia/s no mês
1985	5	ausência de dia/s no mês
2001	4	excesso de dia/s no mês

A.1.2 Precipitação

Tabela A.2: Erros de consistência encontrados ao importar os dados da Precipitação.

Ficheiro (ano)	Sheet	Tipo de erro
2003	5	o ano da sheet não coincide com o ano do ficheiro
2003	6	o ano da sheet não coincide com o ano do ficheiro
2003	7	o ano da sheet não coincide com o ano do ficheiro
2003	8	o ano da sheet não coincide com o ano do ficheiro
2003	9	o ano da sheet não coincide com o ano do ficheiro
2003	10	o ano da sheet não coincide com o ano do ficheiro
2003	11	o ano da sheet não coincide como ano do ficheiro
2003	12	o ano da sheet não coincide com o ano do ficheiro
2006	2	o mês da sheet não coincide
2006	2	o ano da sheet não coincide com o ano do ficheiro
2006	3	o mês da sheet não coincide
2006	3	o ano da sheet não coincide com o ano do ficheiro
2006	4	o mês da sheet não coincide
2006	4	o ano da sheet não coincide com o ano do ficheiro
2006	5	o mês da sheet não coincide
2006	5	o ano da sheet não coincide com o ano do ficheiro
2006	6	o mês da sheet não coincide
2006	6	o ano da sheet não coincide com o ano do ficheiro
2006	7	o ano da sheet não coincide com o ano do ficheiro
2006	8	o mês da sheet não coincide
2006	8	o ano da sheet não coincide com o ano do ficheiro
2006	9	o ano da sheet não coincide com o ano do ficheiro
2006	10	o ano da sheet não coincide com o ano do ficheiro
2006	11	o ano da sheet não coincide com o ano do ficheiro
2006	12	o ano da sheet não coincide com o ano do ficheiro

A.2 Outliers

A.2.1 *Outliers* globais

Tabela A.3: *Outliers* globais

Parâmetro	Data	Hora	Valor Suspeito
Temperatura	2001/03/01	4h	74 °C
Velocidade do Vento	2000/10/04	16h	156 Km/h
Pressão Atmosférica	204/02/19	3h	1207 mmHg
	2004/07/21	2h	1222.6 mmHg
	2004/07/31	12h	705.31 mmHg
	2004/10/10	4h	1149.8 mmHg
	2005/07/02	6h	705.7 mmHg
	2005/12/15	15h	787.68 mmHg
	2005/12/16	15h	789.33 mmHg
	2005/12/17	4h	788.6 mmHg
	2005/12/17	15h	787.6 mmHg

Os *outliers* identificados nas seguintes tabelas foram encontrados através da função *tsoutliers* [26], assim como as suas respectivas sugestões de correção.

A.2.2 Temperatura

Tabela A.4: Outliers encontrados por mês do ano, e respetivo valor sugerido, para valores horários de temperatura (1960-2007).

Data	Hora	Valor	Esperado	Data	Hora	Valor	Esperado
1960/02/28	13h	28.4	27.1	1974/05/22	15h	32.4	31.3
1960/02/28	14h	28.7	27.6	1974/05/22	16h	32.4	31.3
1964/05/16	15h	33.2	32.4	2006/05/27	14h	33.5	25.0
1965/05/10	13h	33.4	31.9	2006/05/27	15h	33.4	27.6
1965/05/10	14h	34.2	32.0	2006/05/27	16h	33.6	30.1
1965/05/10	15h	33.3	32.0	1981/06/14	13h	36.6	35.9
1965/05/10	16h	33	32.1	1981/06/14	14h	37.7	35.9
1968/05/16	14h	31.6	30.5	1981/06/14	15h	38.6	35.9
1970/05/26	14h	31.4	30.6	1981/06/14	16h	36.5	35.9
1970/05/26	15h	31.4	30.8	1981/06/15	13h	36.4	35.3
1970/05/26	16h	31.6	30.9	1981/06/15	14h	36.5	35.4
1974/05/21	11h	31.5	30.0	1981/06/15	15h	36.5	35.6
1974/05/21	12h	32.4	29.8	2001/08/22	13h	44.9	22.8
1974/05/21	16h	32.3	31.1	1970/09/04	14h	36.3	35.2
1974/05/22	11h	31.8	30.2	1970/09/05	12h	36.8	34.7

Tabela A.4 (continuação): Outliers encontrados por mês do ano, e respetivo valor sugerido, para valores horários de temperatura (1960-2007).

Data	Hora	Valor	Esperado	Data	Hora	Valor	Esperado
1974/05/22	12h	32.4	30.7	1980/10/02	14h	31.6	30.7
1974/05/22	14h	31.6	31.3	1980/10/02	15h	31.8	30.8

A.2.3 Pressão atmosférica

Tabela A.5: Outliers encontrados por mês do ano, e respetivo valor sugerido, para valores horários de pressão atmosférica (1999-2007).

Data	Hora	Valor	Esperado	Data	Hora	Valor	Esperado
2004/02/22	16h	733.9	743.9	2005/10/13	24h	786.4	754.8
2000/04/02	9h	729.81	730.0	2005/10/13	23h	786.8	755.8
2000/04/02	10h	729.8	730.0	2005/10/23	07h	786.1	756.3
2000/04/02	11h	729.8	729.9	2005/10/25	07h	786.6	756.6
2000/04/02	12h	729.39	729.9	2005/10/26	07h	781.6	751.7
2000/04/02	13h	729.4	729.9	2005/11/09	18h	738.06	768.3
2000/04/02	14h	729.3	729.9	2005/11/29	09h	781.05	750.6
2000/04/02	15h	729.3	729.9	2005/11/29	10h	781.6	751.2
2000/04/02	16h	729.1	729.9	2005/11/30	09h	786.27	756.1
2000/04/02	17h	729.1	729.8	2005/11/30	10h	786.4	756.5
2000/04/02	18h	729.2	729.8	2006/11/24	13h	735.6	736.7
2005/07/15	5h	744.9	754.9	2006/11/24	14h	734	736.7
2006/08/16	24h	740.5	751.6	2006/11/24	15h	732.99	736.7
1999/09/18	24h	740.2	752.1	2006/11/24	16h	732.3	736.8
1999/09/18	17h	741.8	742.9	2006/11/24	17h	733.9	736.8
1999/09/18	18h	741.47	742.6	2006/11/24	18h	734.8	736.8
1999/09/18	19h	741.5	742.4	2006/11/24	19h	735	736.8
1999/09/18	20h	741.6	742.1	2006/11/24	20h	735.1	736.8
1999/09/18	22h	741.5	742.7	2006/11/24	21h	735.8	736.8
1999/09/18	23h	741	742.6	2000/12/06	24h	736.4	749.1
1999/09/19	01h	739.8	744.9	2000/12/06	21h	737.6	741.6
1999/09/19	02h	739.3	745.4	2000/12/06	22h	737	743.4
1999/09/19	03h	738.3	745.9	2000/12/06	23h	736.7	745.2
1999/09/19	04h	737.3	746.3	2000/12/07	01h	736.4	746.4
1999/09/19	05h	736.6	746.8	2000/12/07	02h	736.1	745.8
1999/09/19	06h	737.2	747.3	2000/12/07	03h	735.7	745.3
1999/09/19	07h	737.3	747.8	2000/12/07	04h	735.4	744.7
1999/09/19	08h	738.3	748.3	2000/12/07	05h	734.6	744.1
1999/09/19	09h	738.5	748.8	2000/12/07	06h	734	743.5
1999/09/19	10h	739.2	749.2	2000/12/07	07h	733.4	743.0
1999/09/19	11h	739.6	749.7	2000/12/07	08h	732.3	742.4
1999/09/19	13h	740.5	748.5	2000/12/07	9h	732.2	741.8
1999/09/19	14h	741	746.9	2000/12/07	10h	732.4	741.2
1999/09/19	15h	741.3	745.2	2000/12/07	11h	732.9	740.7

Tabela A.5 (continuação): Outliers encontrados por mês do ano, e respetivo valor sugerido, para valores horários de pressão atmosférica (1999-2007)

Data	Hora	Valor	Esperado	Data	Hora	Valor	Esperado
1999/09/19	16h	741.5	743.6	2000/12/07	12h	733.47	740.1
2002/09/24	16h	740.6	749.5	2000/12/07	13h	734	739.5
2003/09/30	24h	741.6	748.6	2000/12/07	14h	736.7	738.9
2003/09/30	09h	744.21	744.8	2004/12/01	03h	739.8	740.0
2003/09/30	10h	744.1	745.3	2004/12/01	04h	739.2	740.0
2003/09/30	11h	743.8	745.7	2004/12/01	05h	738.7	740.0
2003/09/30	12h	743.31	746.2	2004/12/01	06h	738.5	740.0
2003/09/30	13h	742.9	746.7	2004/12/01	07h	738.3	740.0
2003/09/30	14h	742	747.2	2004/12/01	08h	738.2	740.0
2003/09/30	15h	741.28	747.7	2004/12/01	09h	738.11	740.0
2003/09/30	16h	741	748.2	2004/12/01	10h	738.1	740.0
2003/09/30	17h	741	748.6	2004/12/01	11h	738.1	740.0
2003/09/30	18h	740.62	749.1	2004/12/01	12h	738.11	740.0
2003/09/30	19h	740.8	749.6	2004/12/01	13h	738	740.0
2003/09/30	20h	741	750.1	2004/12/01	14h	737.8	740.0
2003/09/30	21h	741.3	750.5	2004/12/01	15h	737.7	740.0
2003/09/30	22h	741.5	751.0	2004/12/01	16h	737.9	740.0
2003/09/30	23h	741.6	751.5	2004/12/01	17h	738.3	740.0
2006/09/21	04h	742	743.4	2004/12/01	18h	738.7	740.0
2006/09/21	05h	741.7	743.3	2004/12/01	19h	739.1	740.0
2006/09/21	06h	741.8	743.2	2004/12/01	20h	739.7	740.0
2003/10/31	02h	733.5	734.6	2005/12/01	24h	741.2	759.5
2003/10/31	03h	733.3	734.6	2005/12/01	04h	782	752.0
2003/10/31	04h	733.2	734.7	2005/12/01	22h	741.9	744.3
2003/10/31	05h	733	734.7	2005/12/01	23h	741.6	746.3
2003/10/31	06h	733	734.7	2005/12/02	01h	740.9	747.7
2003/10/31	07h	732.7	734.7	2005/12/02	02h	740.9	746.9
2003/10/31	08h	732.2	734.7	2005/12/02	03h	740.9	746.2
2003/10/31	09h	732.6	734.8	2005/12/02	04h	741	745.5
2003/10/31	10h	732	734.8	2005/12/02	05h	741.2	744.7
2003/10/31	11h	732.1	734.8	2005/12/02	06h	741.5	744.0
2003/10/31	12h	731.95	734.8	2005/12/02	07h	741.8	743.2
2003/10/31	13h	731.9	734.9	2005/12/04	01h	784.1	753.8
2003/10/31	14h	731.9	734.9	2005/12/04	07h	783.5	753.6
2003/10/31	15h	731.92	734.9	2005/12/05	01h	783.8	754.5
2003/10/31	16h	732	734.9	2005/12/05	07h	783.4	753.4
2003/10/31	17h	732.7	734.9	2005/12/18	04h	787	754.0
2003/10/31	18h	733.41	735.0	2005/12/18	10h	777.4	757.4
2003/10/31	19h	734.1	735.0	2005/12/24	03h	737.5	757.6
				2005/12/28	21h	786.2	756.1

A.2.4 Velocidade do vento

Tabela A.6: Outliers encontrados por mês do ano, e respetivo valor sugerido, para valores horários de velocidade do vento (1994-2007).

Data	Hora	Valor	Esperado	Data	Hora	Valor	Esperado
2001/01/28	05h	69	66	2005/07/24	19h	116	16.5
2001/01/28	06h	72	66	1995/09/19	17h	60	41.5
2001/01/28	07h	72	66	2000/10/20	13h	55	39.33
2001/01/28	08h	74	66	2000/10/20	14h	52	42.67
2001/01/28	09h	71	66	1998/11/26	03h	98	12.5
2001/01/28	10h	76	66	2000/11/06	14h	63	58.0
2001/01/28	14h	69	63	2000/11/06	16h	62	57
2002/01/23	02h	60	46	2000/12/30	03h	68	62
2007/02/28	23h	67	36	2002/12/27	15h	68	62
2007/02/16	16h	66	40.5	2002/12/27	16h	69	62

A.3 Ausência de registos

A.3.1 Temperatura

Tabela A.7: Ausência de registos de temperatura identificados através do heatmap.

Ano	Meses	Tipo de ausência
1982	setembro	não-justificada
	outubro	
	novembro	
	dezembro	
1986	novembro	não-justificada
	dezembro	
1992	novembro	avaria
1993	janeiro	avaria
2006	dezembro	avaria
2007	janeiro	avaria

A.3.2 Pressão atmosférica

Tabela A.8: Ausência de registos de pressão atmosférica identificados através do heatmap.

Ano	Meses	Tipo de ausência
2002	novembro	não-justificada
	dezembro	(meses completos)
2007	abril	não-justificada
		(algumas horas)

A.3.3 Precipitação

Na precipitação foram identificadas as ausências correspondentes a avarias na instrumentação utilizada, levando desse modo a ausência de registos, nesse período considerado

Tabela A.10: Avarias registadas nos instrumentos de medição da precipitação.

Ano	Mês	Período de avaria registado
1998	junho	10 horas do dia 10 até 24 horas do dia 15 19 horas do dia 1 até 9 horas do dia 3
	agosto	19 horas do dia 8 até 9 horas do dia 10 19 horas do dia 14 até 9 horas do dia 17
	setembro	2 horas do dia 26 até 24 horas do dia 29 dias 4, 5 e 6
	outubro	10 horas do dia 18 até 9 horas do dia 19
	novembro	10 horas do dia 1 até 9 horas do dia 2 10 horas do dia 6 até 9 horas do dia 8
2001	fevereiro	10 horas do dia 3 até 9 horas do dia 7
2003	abril	10 horas do dia 13 até 9 horas do dia 14

A.3.4 Velocidade do vento

Tabela A.11: Ausência de registos de velocidade do vento identificados através do heatmap.

Ano	Meses	Tipo de ausência
1994	janeiro a abril	não-justificada
	outubro a dezembro	(meses completos)
1995	janeiro a abril	não-justificada
	outubro a dezembro	(meses completos)
1996	janeiro a abril	não-justificada
	outubro a dezembro	(meses completos)
1997	janeiro a abril	não-justificada
	outubro a dezembro	(meses completos)
1998	maio	
	junho	não-justificada
	agosto	(algumas horas)
	outubro a dezembro	
2002	janeiro	não-justificada
	março	(algumas horas)
	julho	
2003	julho a dezembro	não-justificada (meses completos)
2004	maio	não-justificada (algumas horas)
2005	janeiro a abril	não-justificada
	outubro a dezembro	(meses completos)

Apêndice B

Resultados da análise exploratória de dados

Neste apêndice são apresentados gráficos e tabelas que complementam a análise exploratória das quatro variáveis meteorológicas em estudo, realizada no Capítulo 3.

B.1 Análise univariada

B.1.1 Temperatura

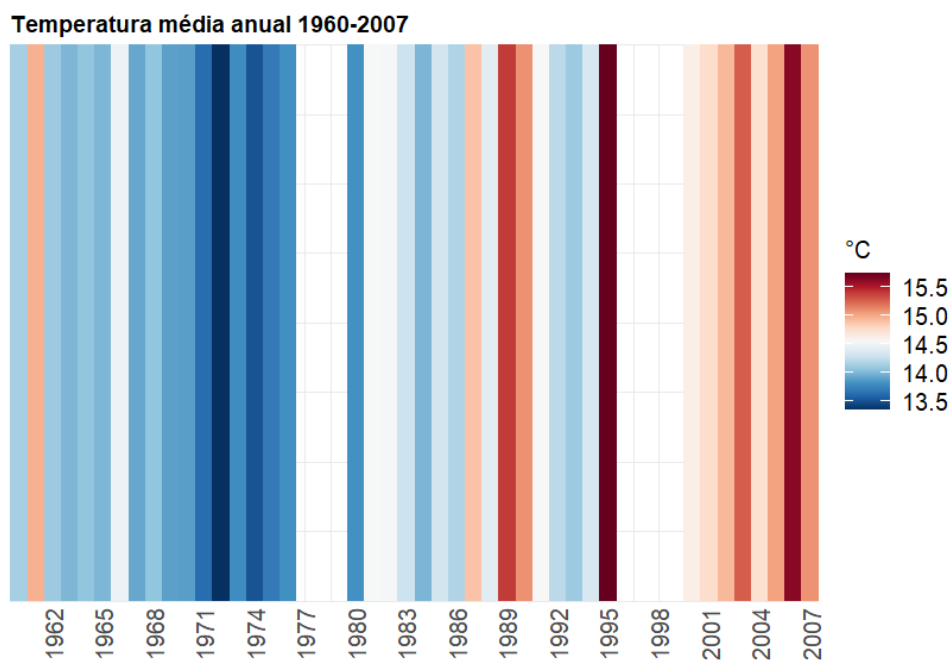


Figura B.1: *Warming stripes* das temperaturas médias anuais (°C) (1960-2007).

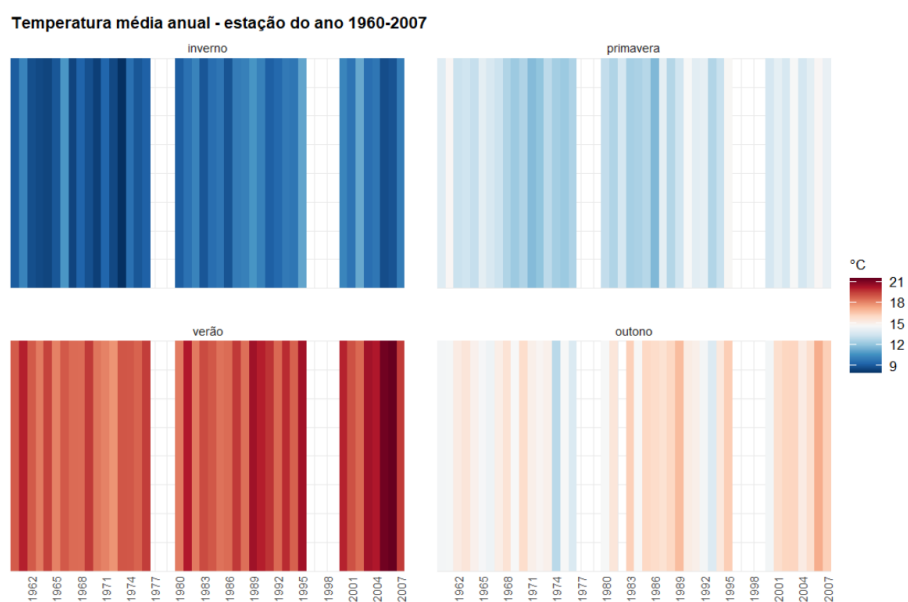
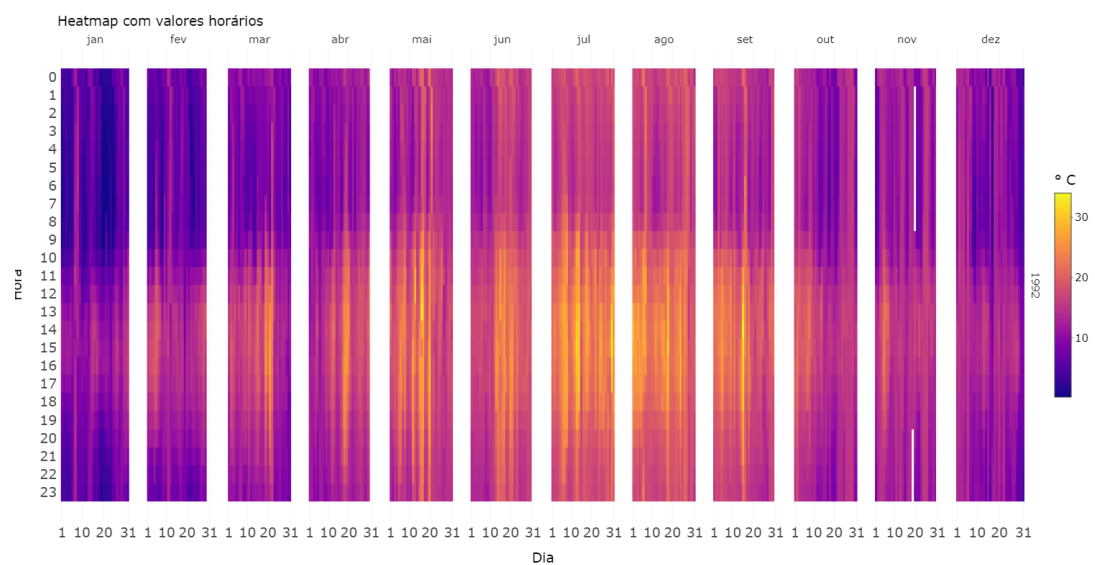
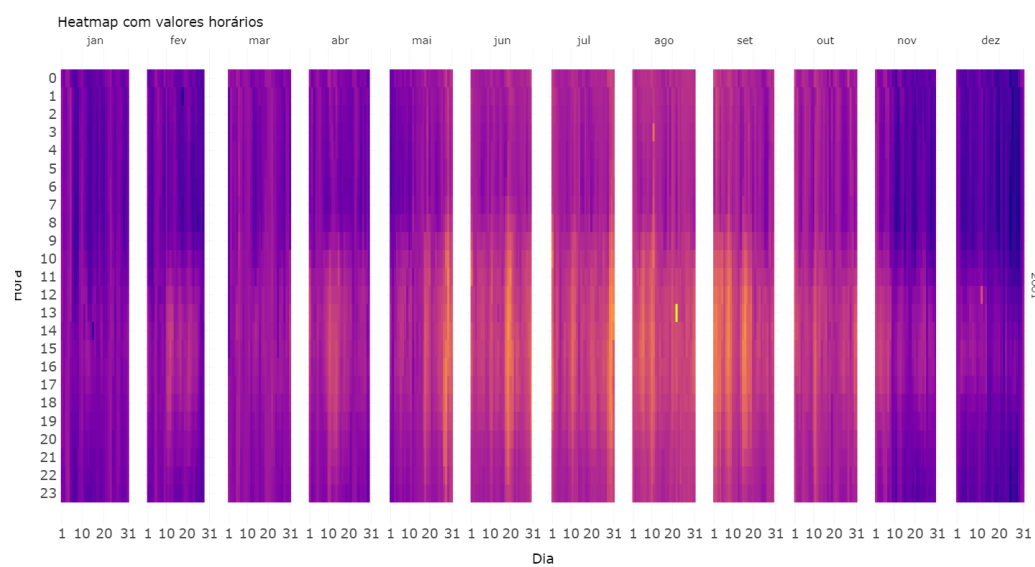


Figura B.2: *Warming stripes* das temperaturas médias anuais por estação do ano (°C) (1960-2007).



(a) Heatmap das temperaturas de hora a hora de 1992.



(b) Heatmap das temperaturas de hora a hora de 2001.

Figura B.3: Heatmap das temperaturas de hora a hora de 1992 e 2001(°C).

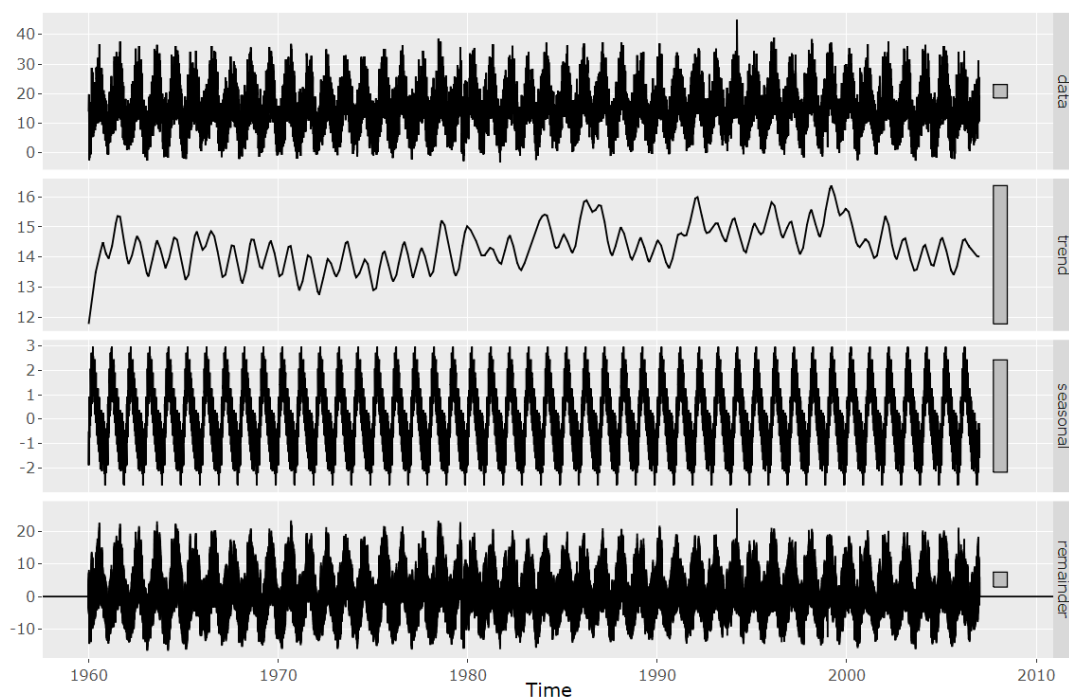


Figura B.4: Decomposição da série temporal da temperatura (1960-2007): hora a hora com sazonalidade anual (frequência = 8766).

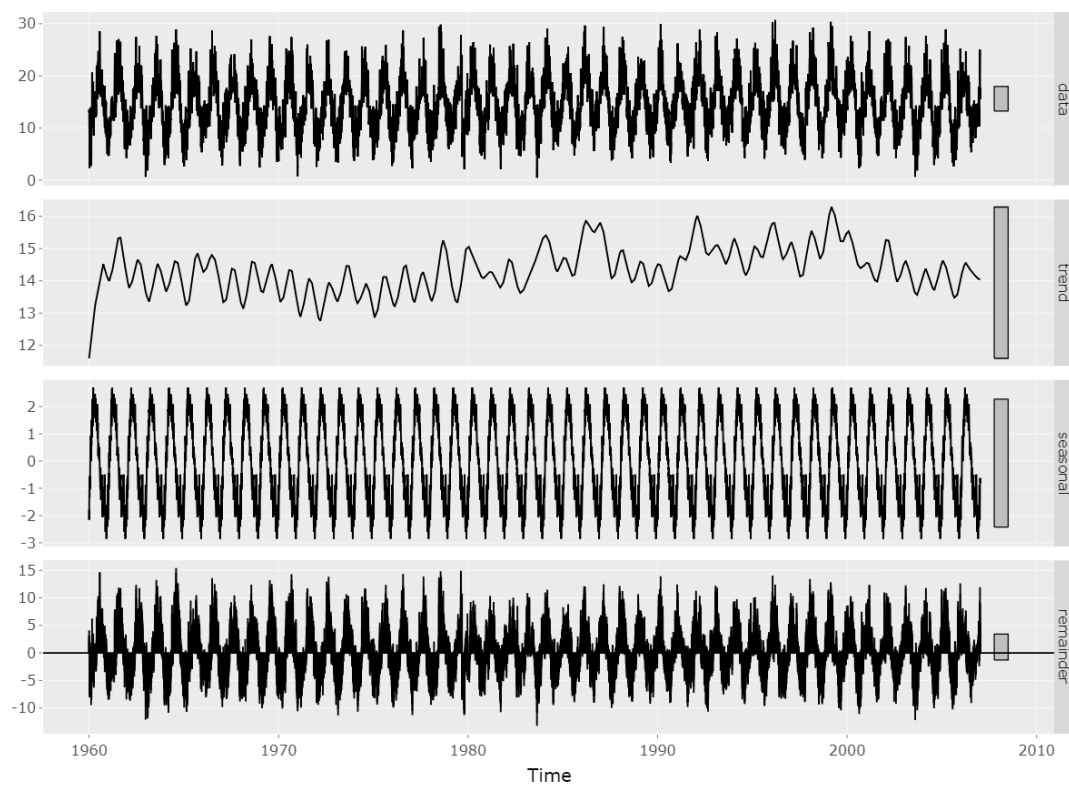
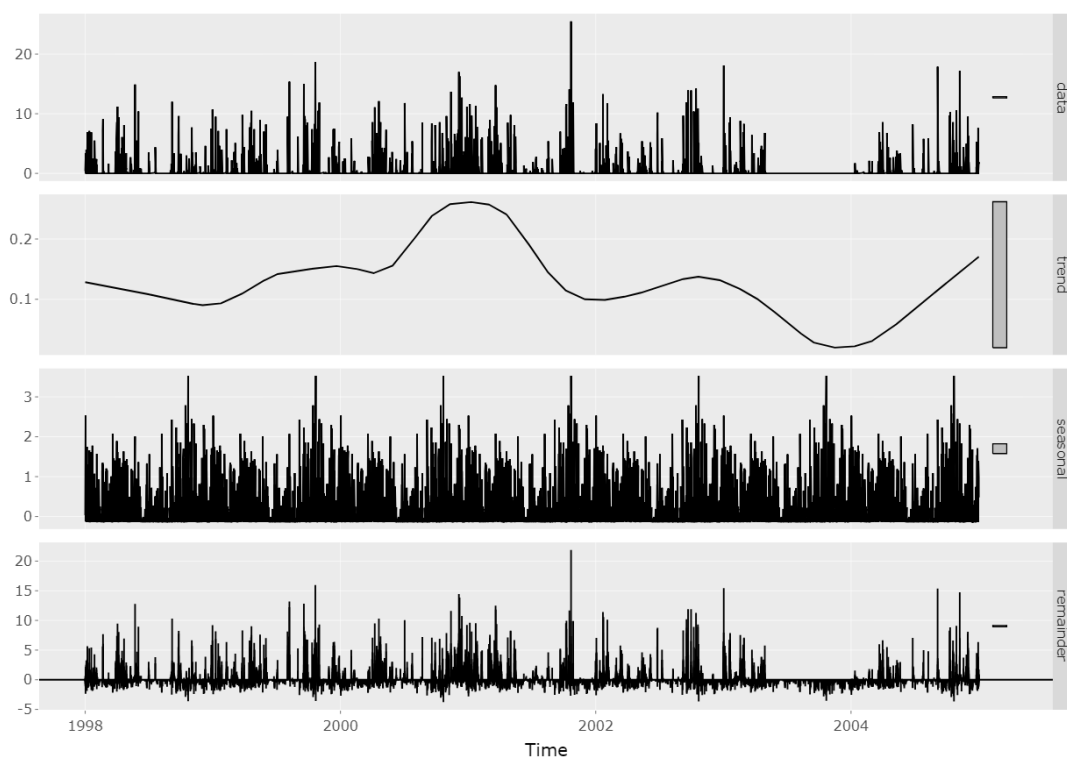


Figura B.5: Decomposição da série temporal da temperatura (1960-2007): Médias diárias com sazonalidade anual (frequência = 365).

B.1.2 Precipitação

Tabela B.1: Precipitação acumulada (mm) por estação do ano e total (1998-2005).

Estação	Ano							Total
	1998	1999	2000	2001	2002	2003	2005	
Inverno	197.90	224.10	244.10	920.20	259.40	489.60	28.84	2344.14
Primavera	392.0	303.0	530.2	756.1	224.4	195.9	193.2	2594.8
Verão	44.7	167.7	95.9	68.3	114.2	0.0	39.6	530.4
Outono	233.9	481.5	389.4	284.8	373.1	0.0	302.5	2065.2

Figura B.6: Decomposição da série temporal da precipitação (1998-2005): hora a hora (mm) com sazonalidade anual (frequência = 8766).

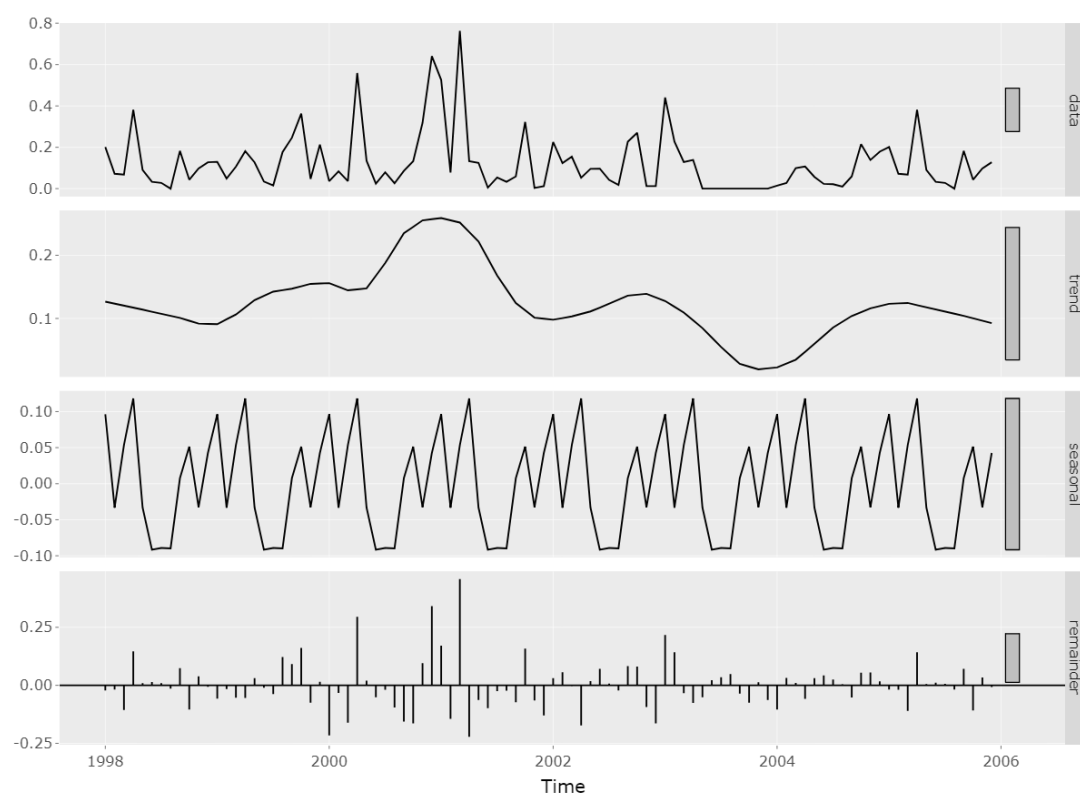


Figura B.7: Decomposição da série temporal da precipitação (1998-2005): médias mensais de precipitação (mm) (frequência = 12).

B.1.3 Pressão atmosférica

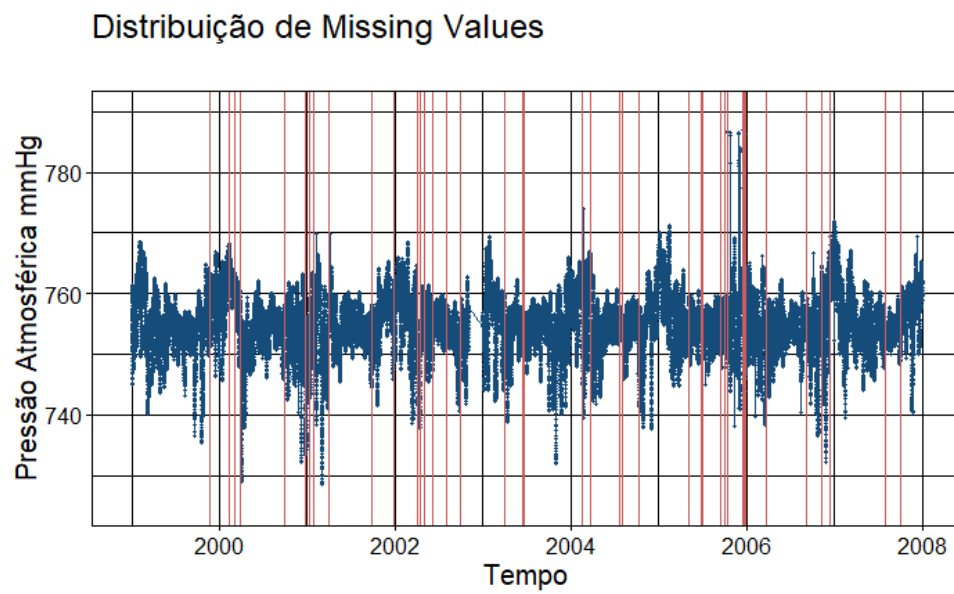
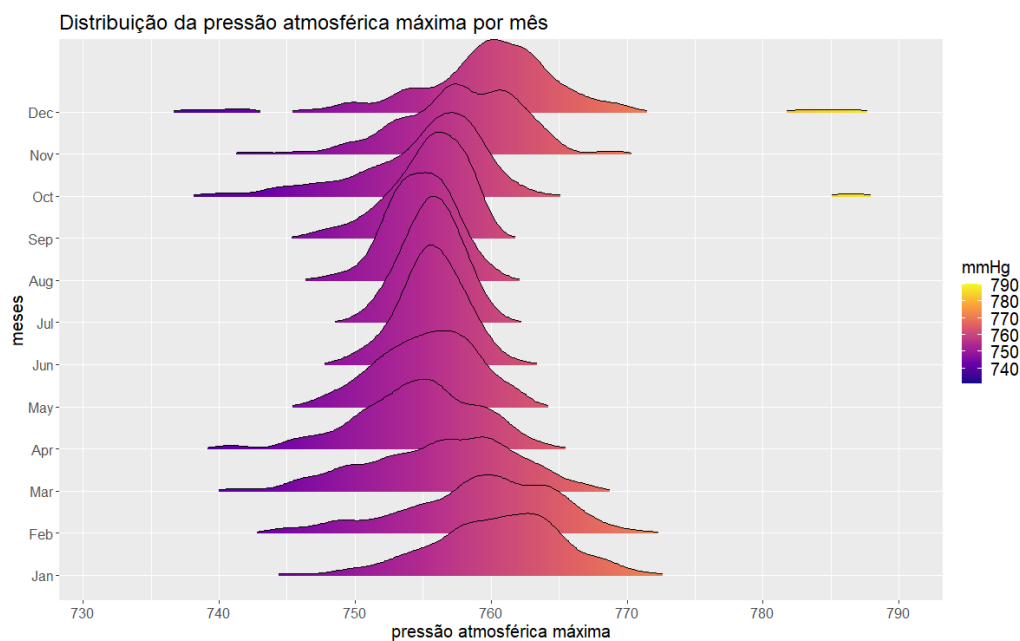
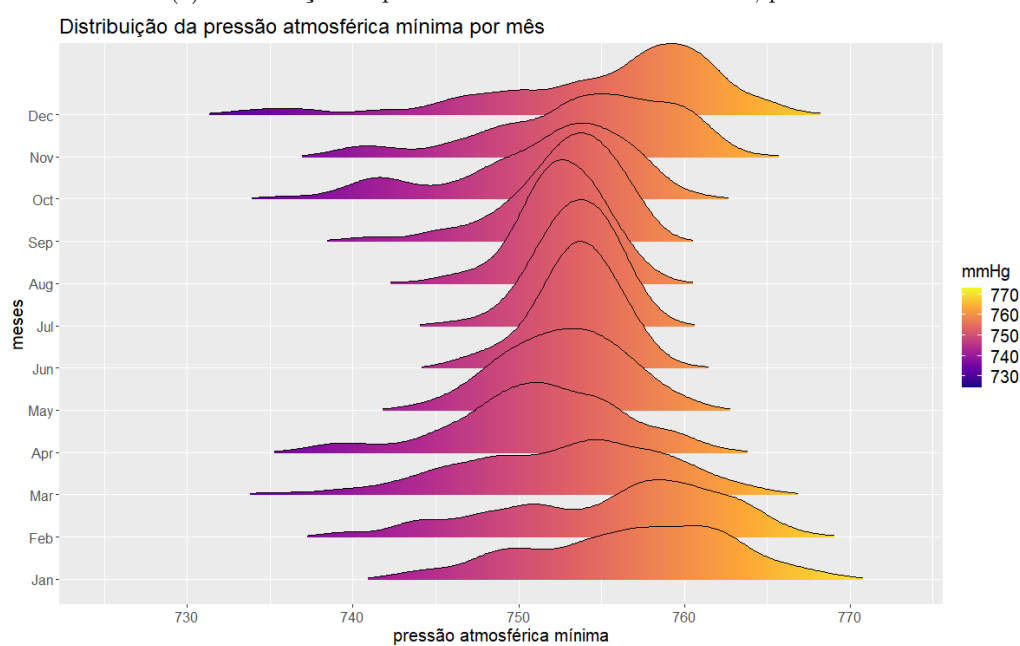


Figura B.8: Distribuição de *missing values* na pressão atmosférica e submetidos a correção (1999-2007).

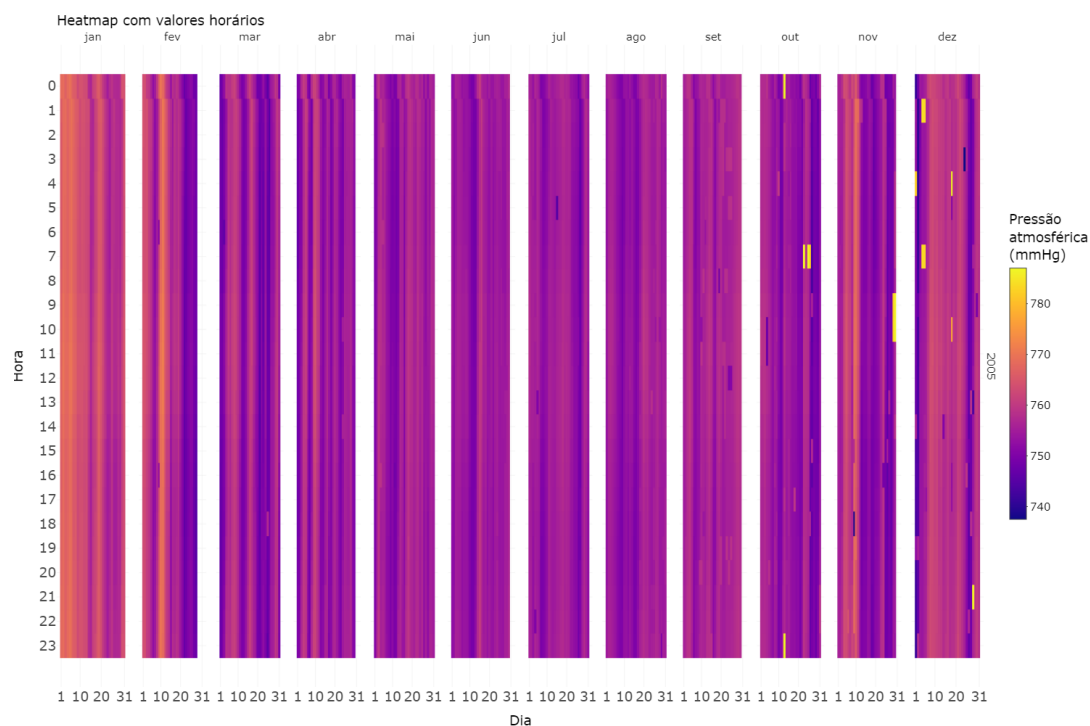


(a) Distribuição da pressão atmosférica máxima diária, por mês.

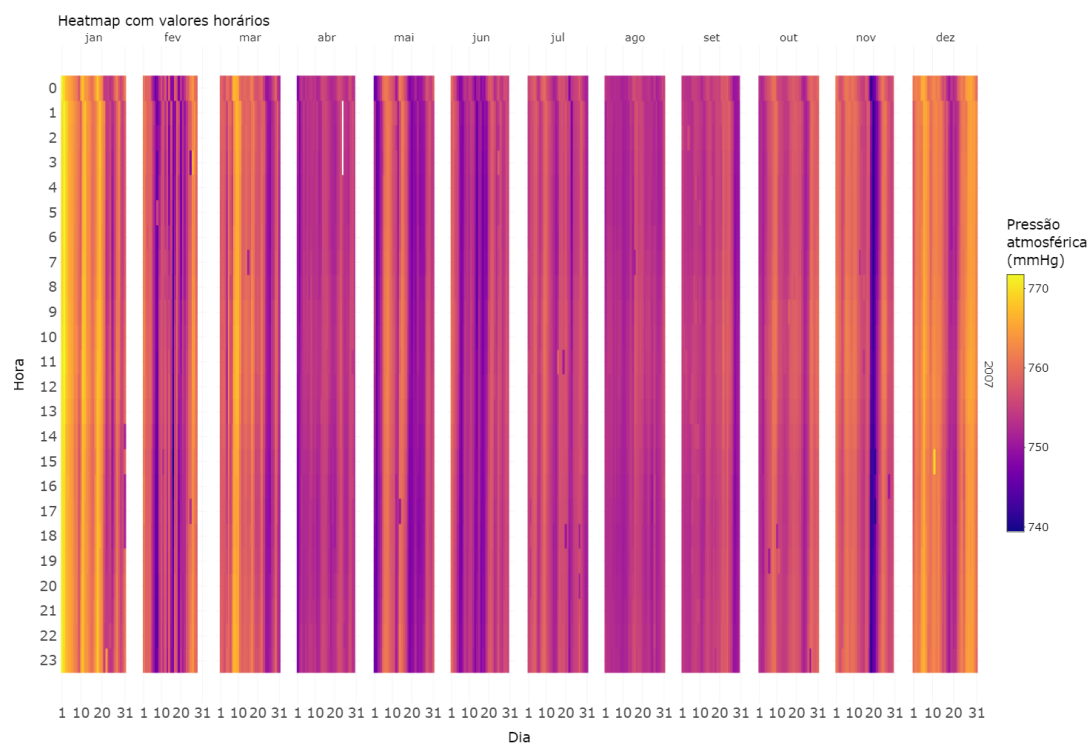


(b) Distribuição da pressão atmosférica mínima diária, por mês.

Figura B.9: Distribuição da pressão atmosférica máxima e mínima diária, por mês.



(a) Ano 2005.



(b) Ano 2007.

Figura B.10: *Heatmap* da Pressão Atmosférica horária dos anos 2005 e 2007

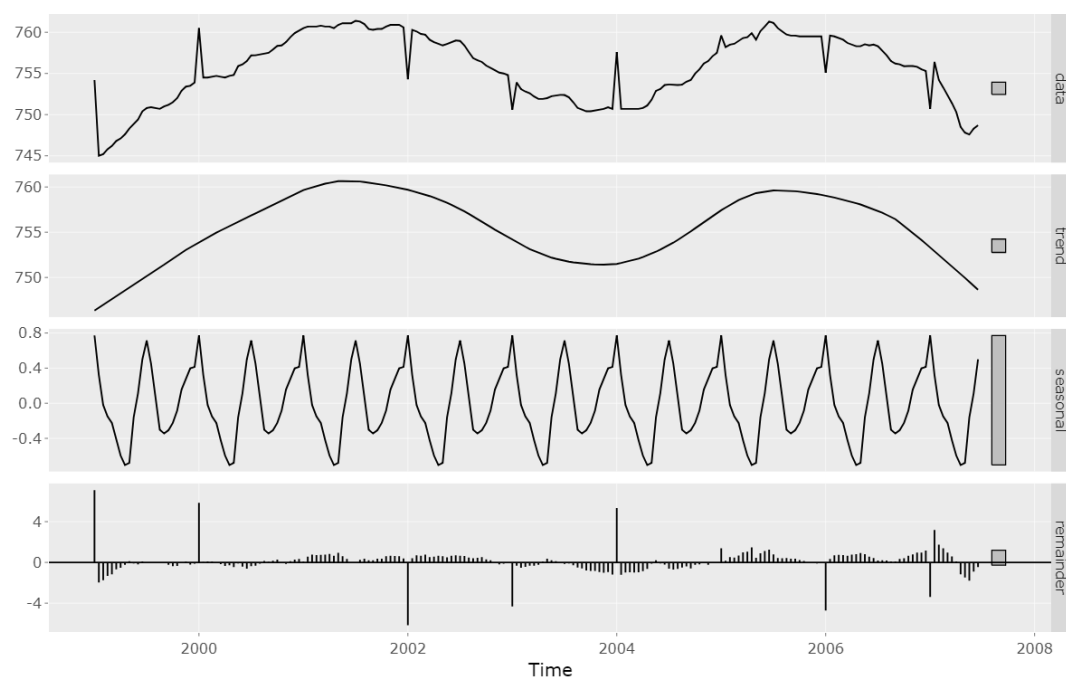


Figura B.11: Decomposição das série temporal da pressão atmosférica: hora a hora ($mmHg$) com sazonalidade diária (frequência = 24).

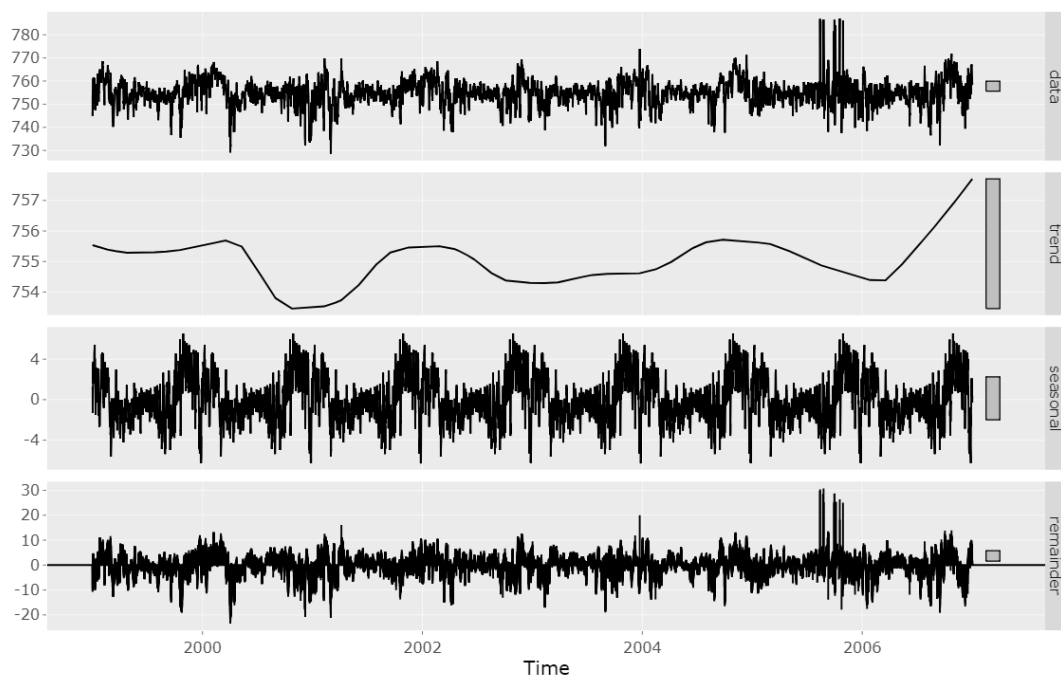


Figura B.12: Decomposição das série temporal da pressão atmosférica: hora a hora ($mmHg$) com sazonalidade anual (frequência = 8766).

B.1.4 Velocidade do Vento

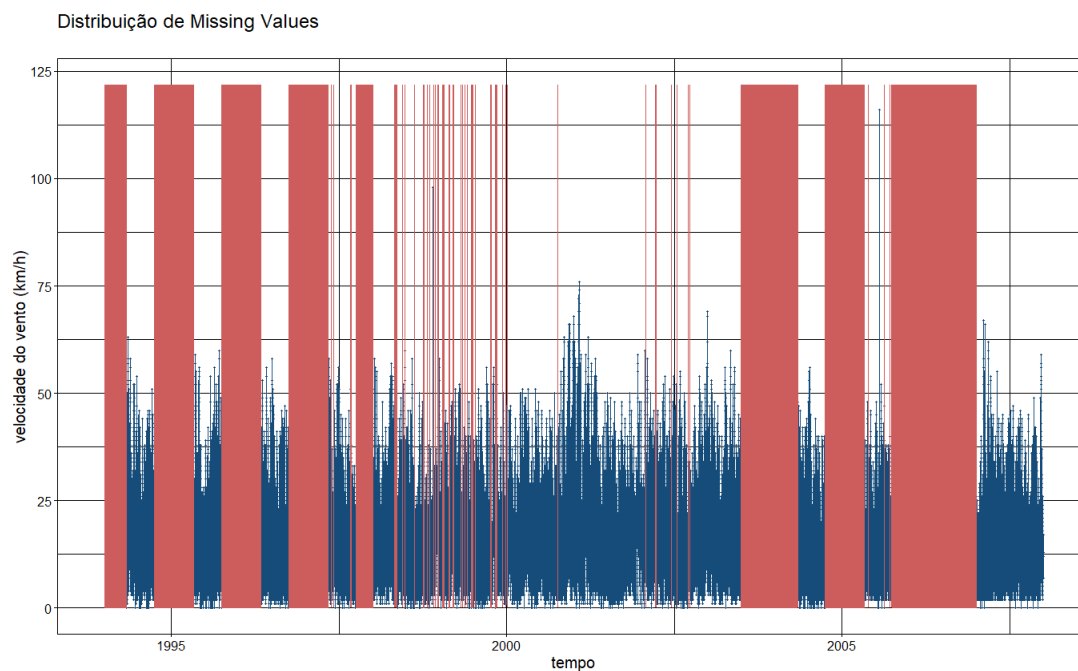


Figura B.13: Distribuição de *missing values* na velocidade do vento antes de qualquer correção (1994-2007).

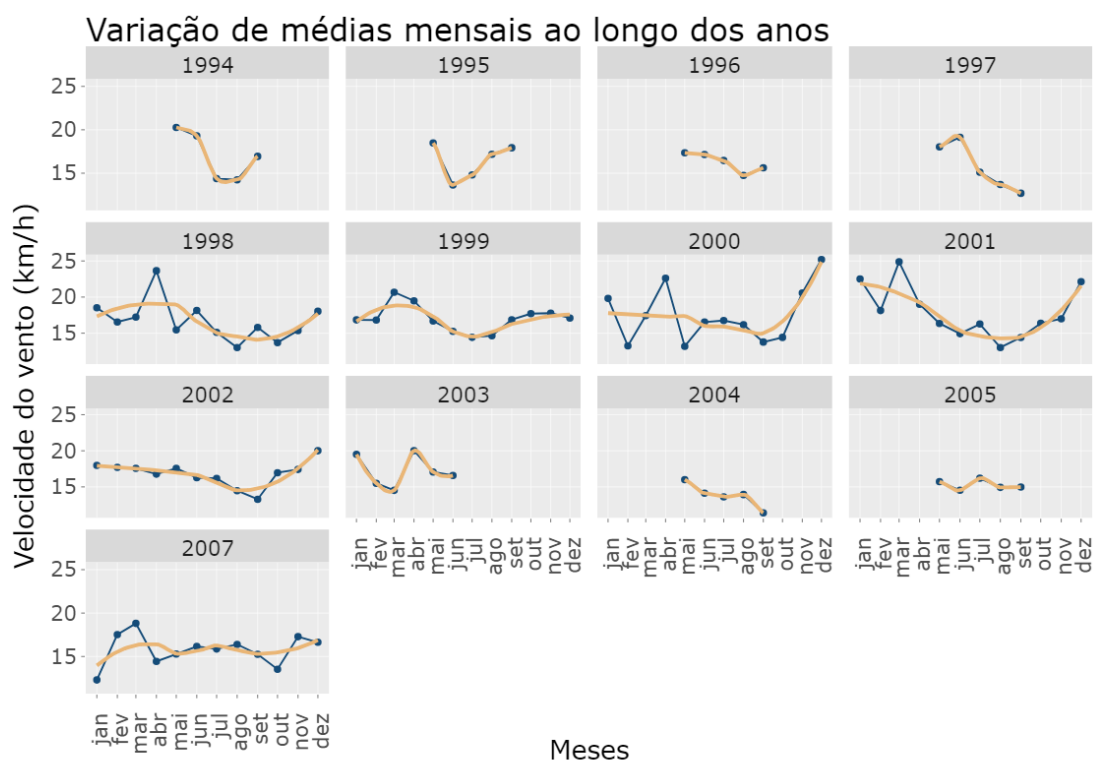


Figura B.14: Médias mensais da velocidade do vento ao longo dos anos e linha de tendência.

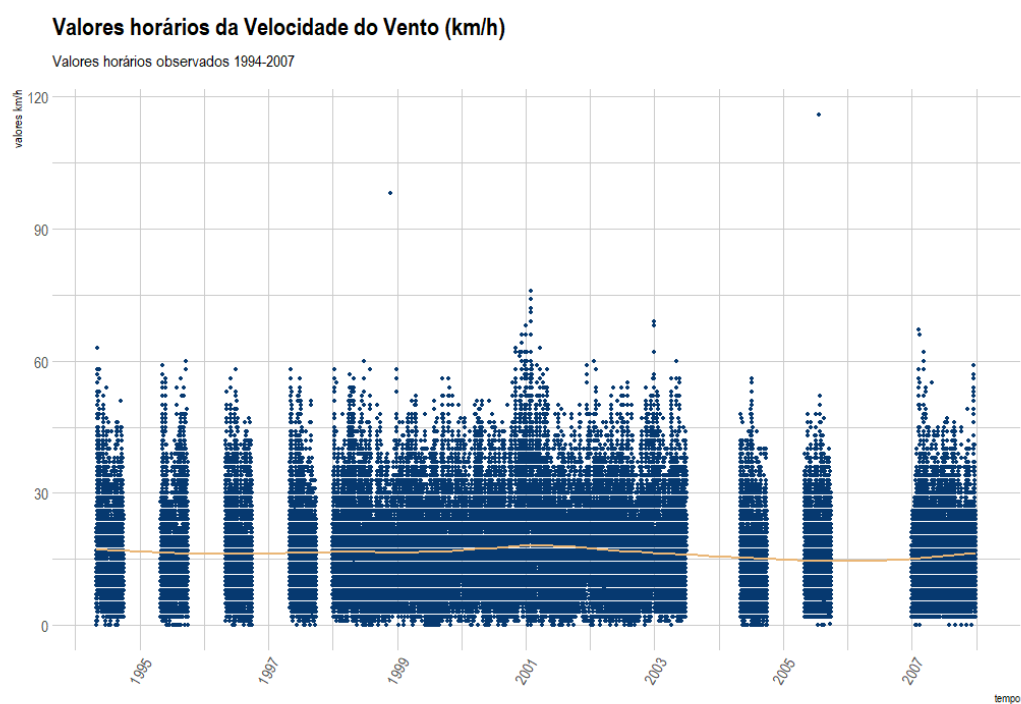
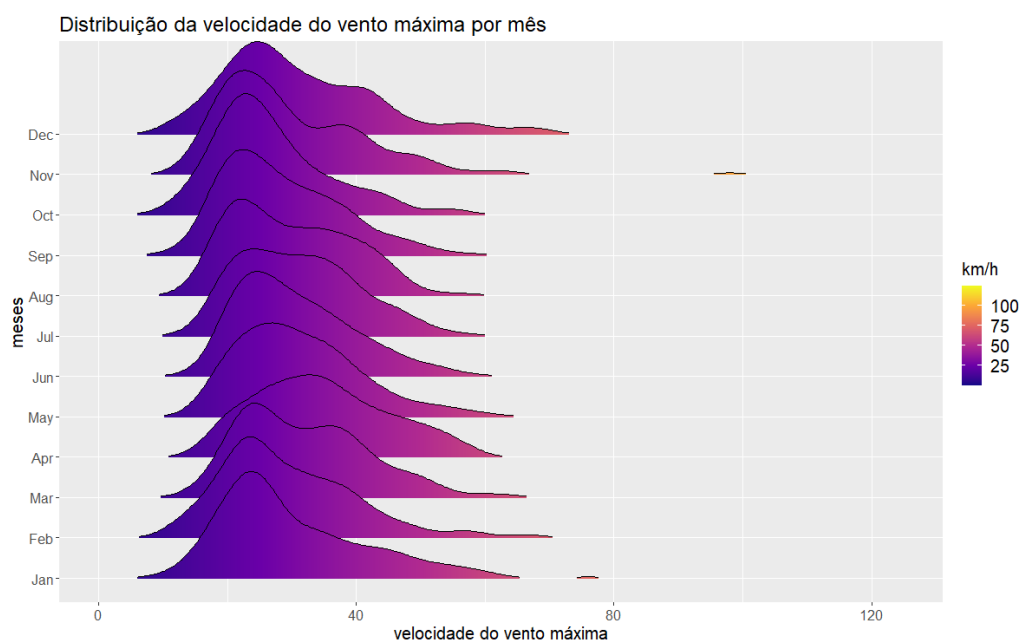
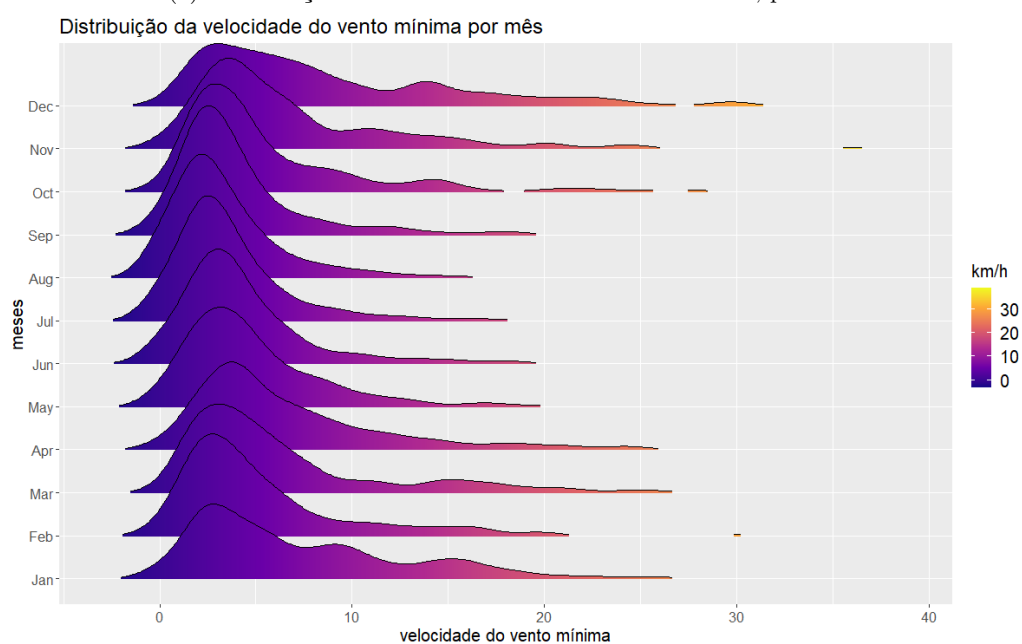


Figura B.15: Série temporal da velocidade do vento horária e linha da tendência (1994-2007).

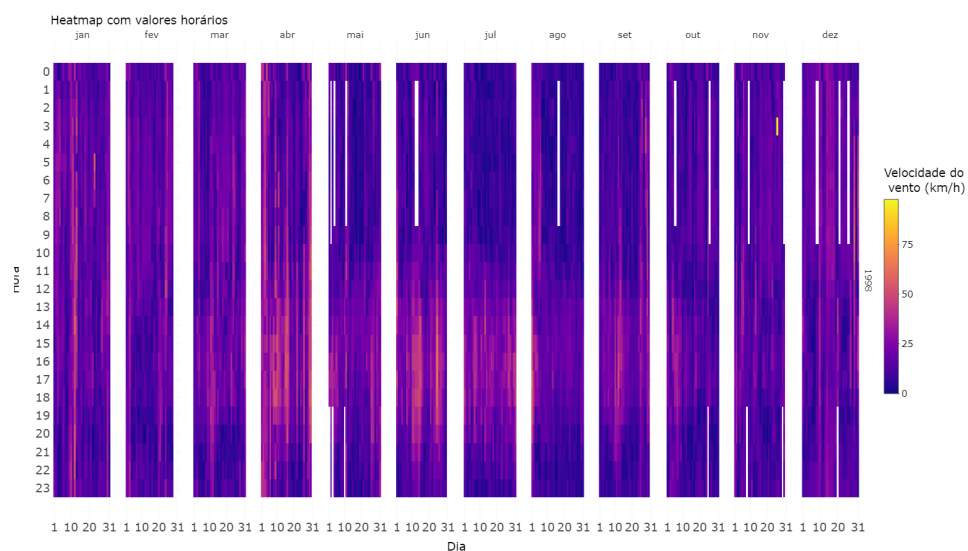


(a) Distribuição da velocidade do vento máxima diária, por mês.

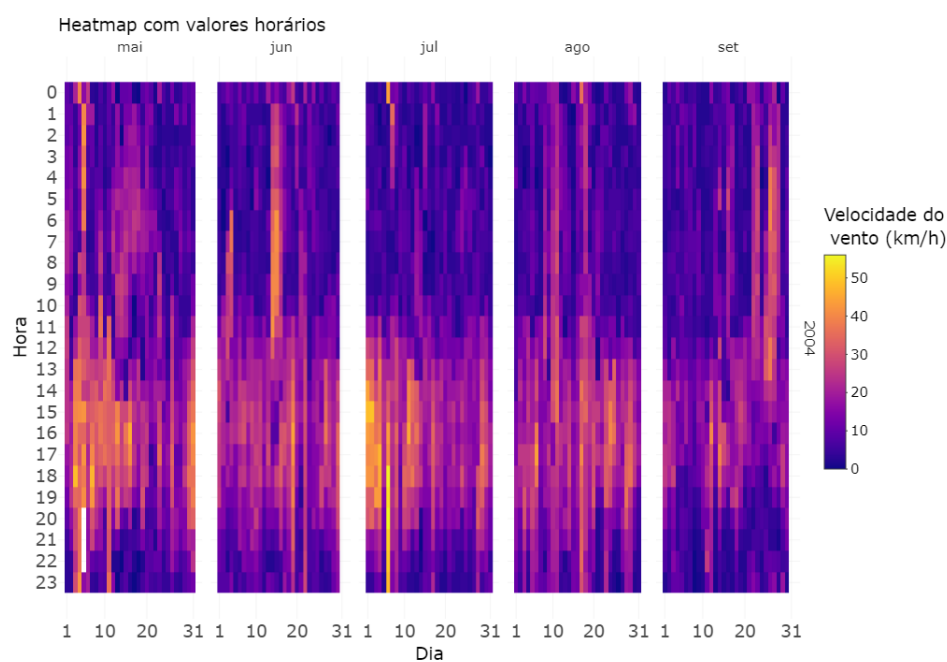


(b) Distribuição da temperatura mínima diária, por mês.

Figura B.16: Distribuição da velocidade do vento máxima e mínima diária, por mês.

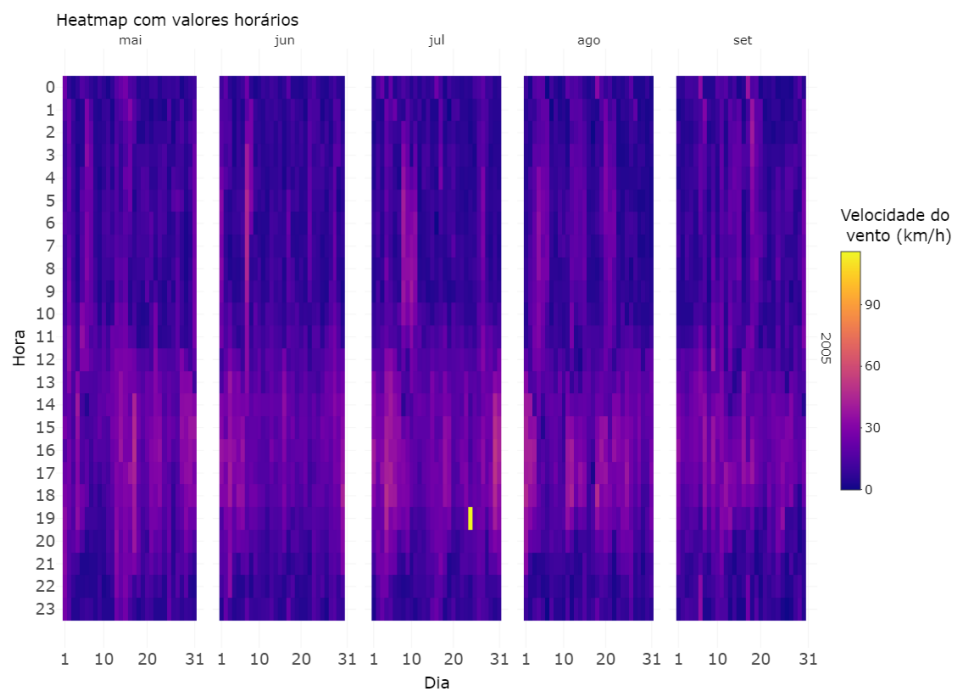


(a) Heatmap das velocidade do vento de hora a hora em 1992.



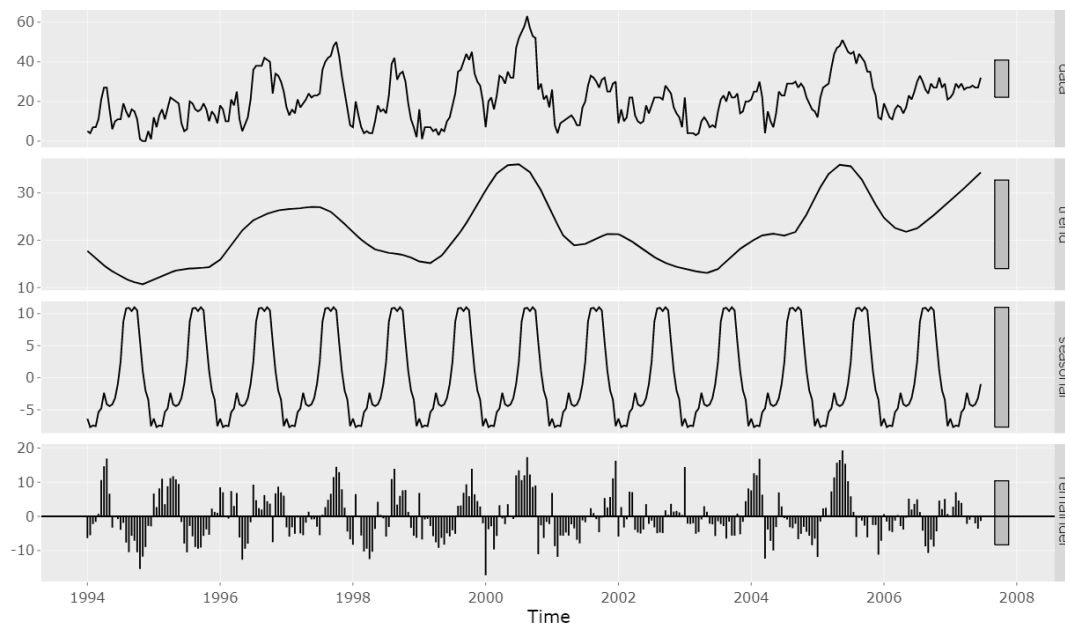
(b) Heatmap das velocidade do vento de hora a hora em 2004.

Figura B.17: Heatmap das velocidade do vento de hora a hora.



(c) Heatmap das velocidades do vento de hora a hora em 2005.

Figura B.17: Heatmap das velocidades do vento de hora a hora.

Figura B.18: Decomposição da série temporal da velocidade do vento (1994-2007): hora a hora (km/h) com sazonalidade diária (frequência = 24).

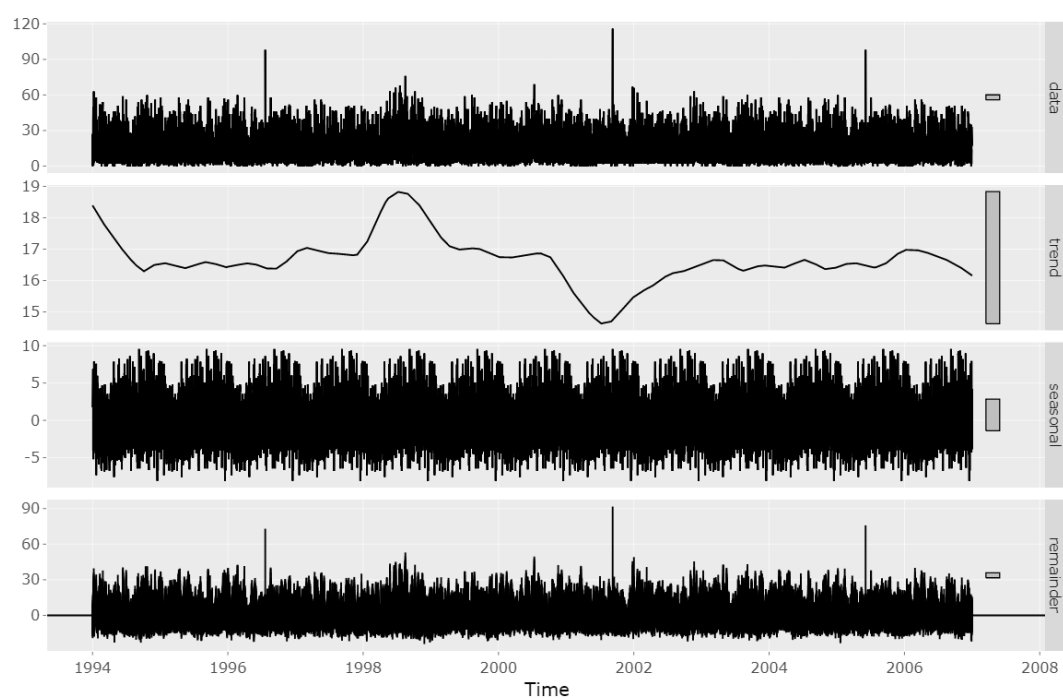


Figura B.19: Decomposição da série temporal da velocidade do vento (1994-2007): hora a hora (km/h) com sazonalidade anual (frequência = 8766).

B.2 Análise multivariada

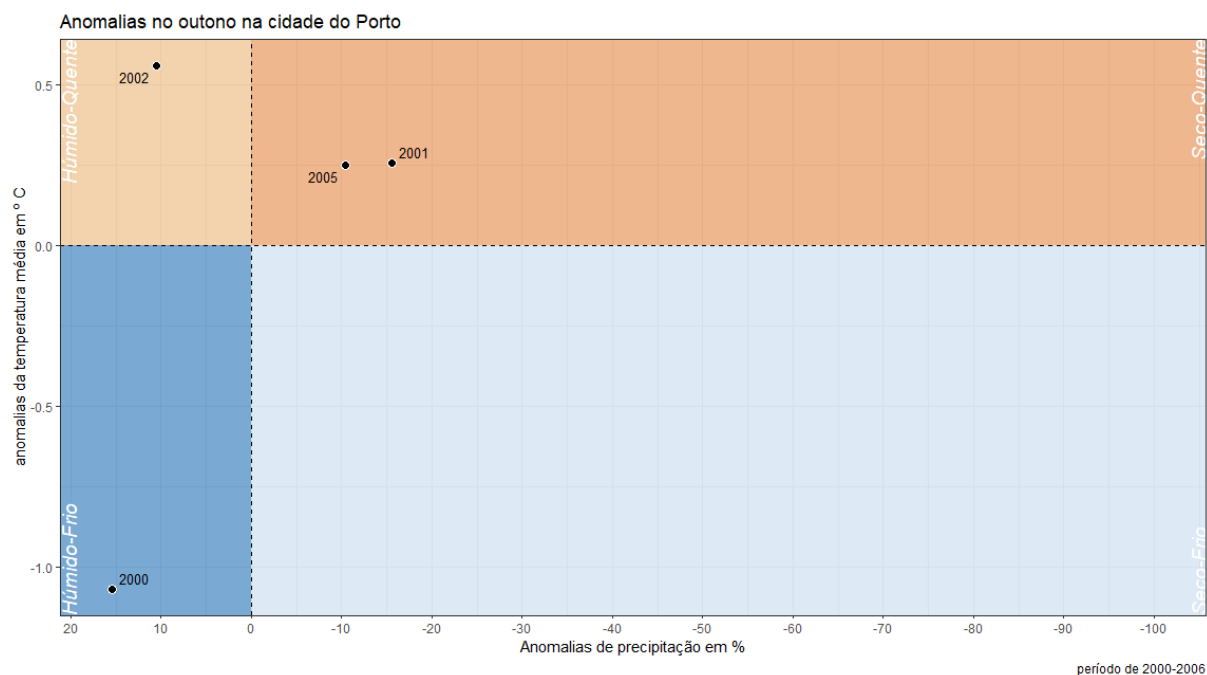


Figura B.20: *Quadrant Type Scatter Plot* das anomalias de temperatura e precipitação, no outono (2000-2006).

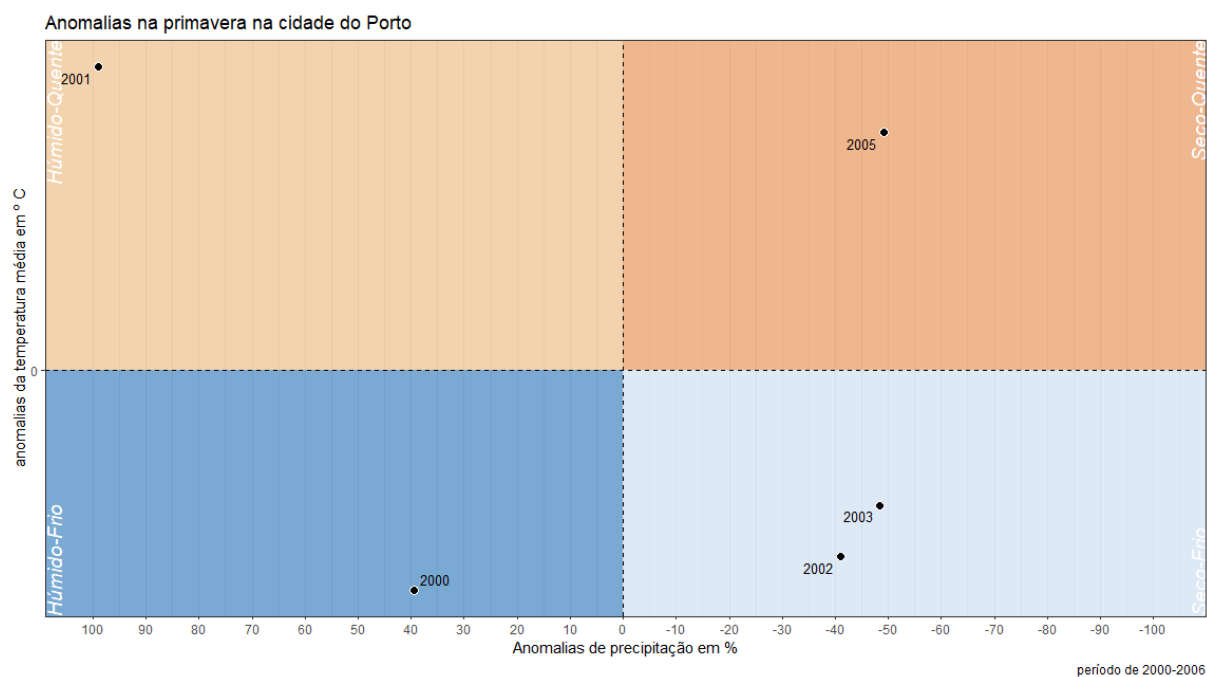


Figura B.21: *Quadrant Type Scatter Plot* das anomalias de temperatura e precipitação, na primavera (2000-2006).

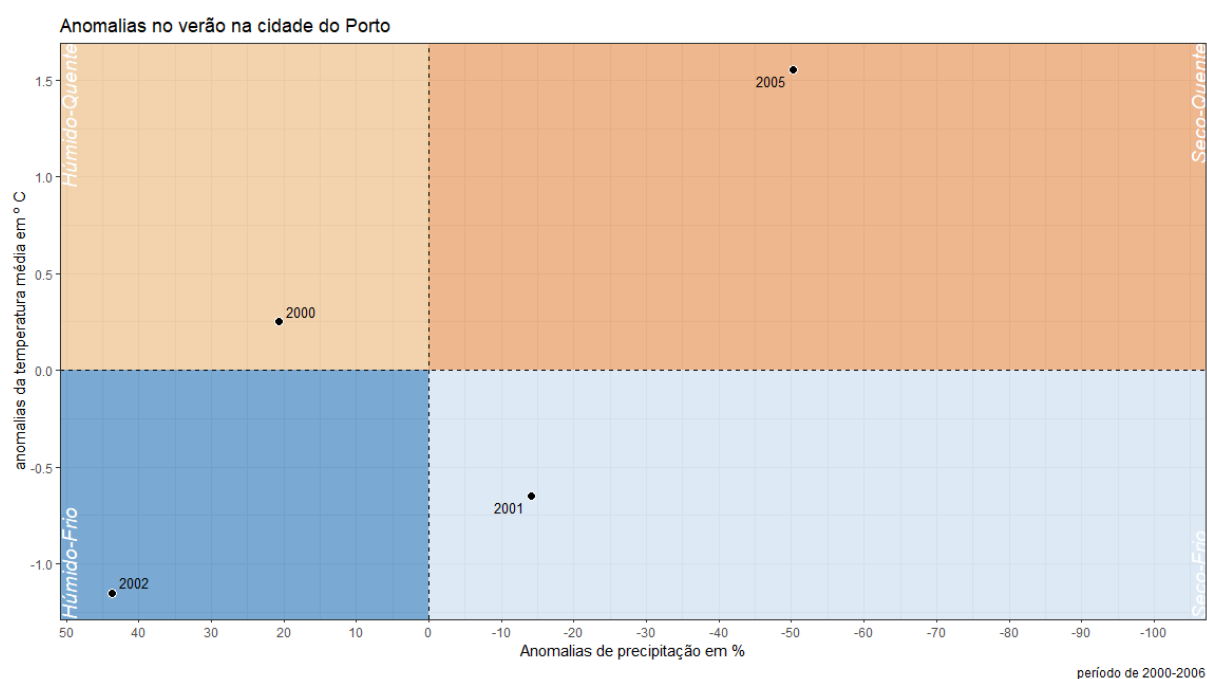


Figura B.22: *Quadrant Type Scatter Plot* das anomalias de temperatura e precipitação, no verão (2000-2006).

Apêndice C

Resultados da aplicação de testes de homogeneização

Neste apêndice são apresentadas as tabelas com os valores críticos de apresentados de 1% e 5%, para os testes de homogeneização abordados nesta dissertação. São também apresentados alguns resultados dos testes de homogeneização aplicados no Capítulo 4.

C.1 Tabelas de valores críticos para testes de homogeneização

Tabela C.1: Valores críticos de 5% e 1% para os testes de homogeneização Buishand Range Test, Pettit Test e Standard Normal Homogeneity Test (SNHT), para conjuntos de dados de diferentes tamanhos [21].

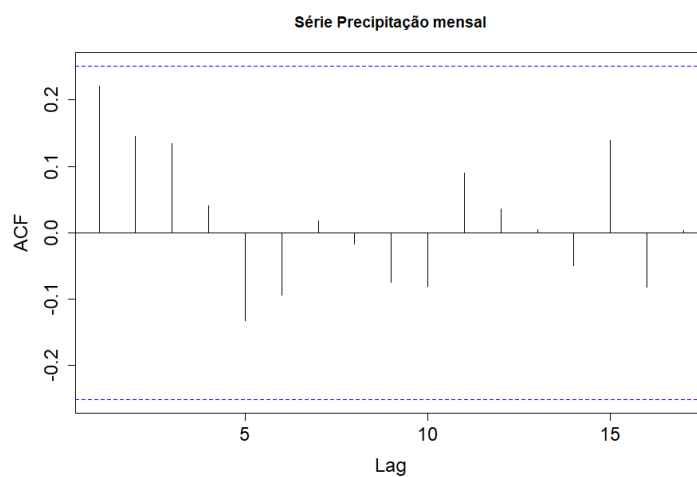
Tamanho do conjunto de dados	Critical Values											
	20		30		40		50		70		100	
	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
Buishand	1.22	1.60	1.24	1.70	1.26	1.74	1.27	1.78			1.29	1.86
Pettit	71	57	133	107	208	167	293	235	488	393	841	677
SNHT	7.089	9.113	7.747	10.153	8.151	10.771	8.432	11.193	8.814	11.737	9.167	12.228

Tabela C.2: Valores críticos de 1% e 5% para o teste de homogeneização Von Neumann [21].

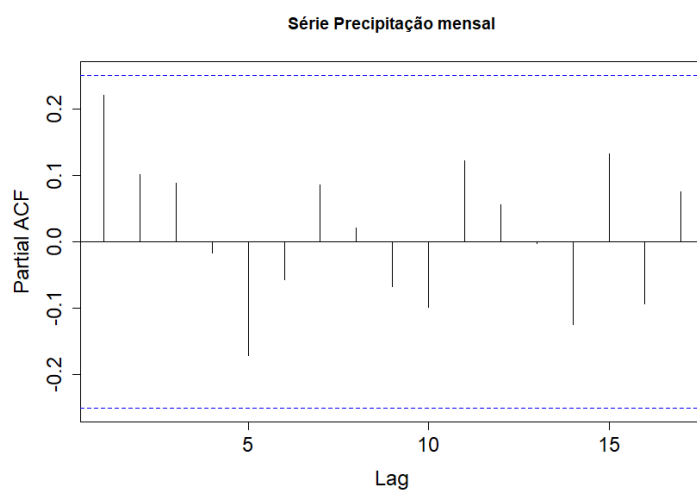
Critical Values						
Tamanho do conjunto de dados	20	30	40	50	70	100
1%	1.04	1.20	1.29	1.36	1.45	1.54
5%	1.30	1.42	1.49	1.54	1.61	1.67

C.2 Resultados dos testes de homogeneização

C.2.1 Precipitação



(a) Gráfico de *autocorrelation* (ACF).



(b) Gráfico de *partial autocorrelation* (PACF)

Figura C.1: Gráficos de ACF e PAFC da série temporal da precipitação mensal (1998-2005).

C.2.2 Pressão atmosférica

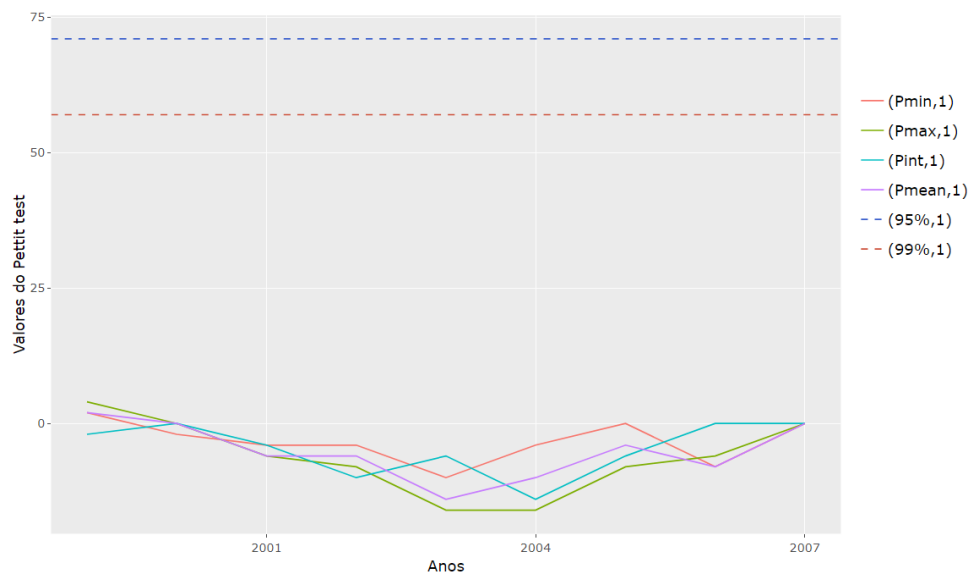


Figura C.2: *Pettitt test* aplicados às series anuais de pressão atmosférica mínimas (P_{min}), máximas (P_{max}), intervalo (P_{int}), médias (P_{mean}) e níveis de confiança de 95 e 99%.

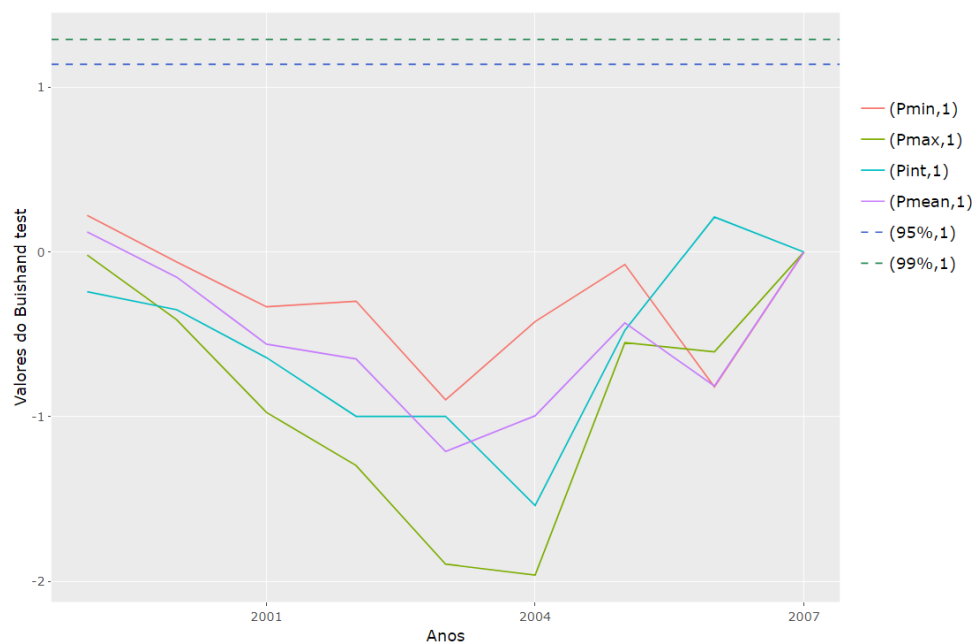


Figura C.3: *Buishand range test* aplicados às series anuais de pressão atmosférica mínimas (P_{min}), máximas (P_{max}), intervalo (P_{int}), médias (P_{mean}) e níveis de confiança de 95 e 99%.

C.2.3 Velocidade do vento

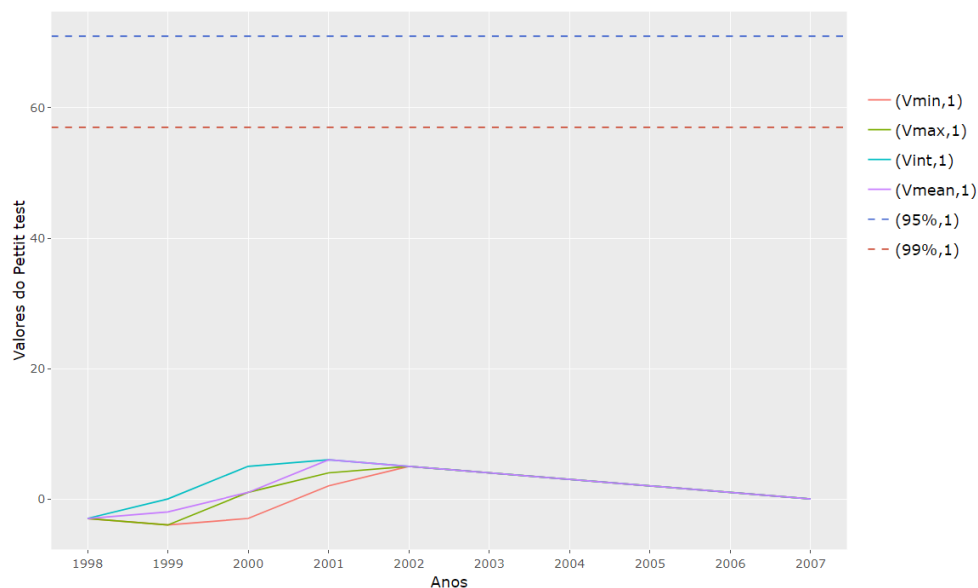


Figura C.4: *Pettitt test* aplicado às series anuais de velocidade do vento mínimas (V_{min}), máximas (V_{max}), intervalo (V_{int}), médias (V_{mean}) e níveis de confiança de 95 e 99%.

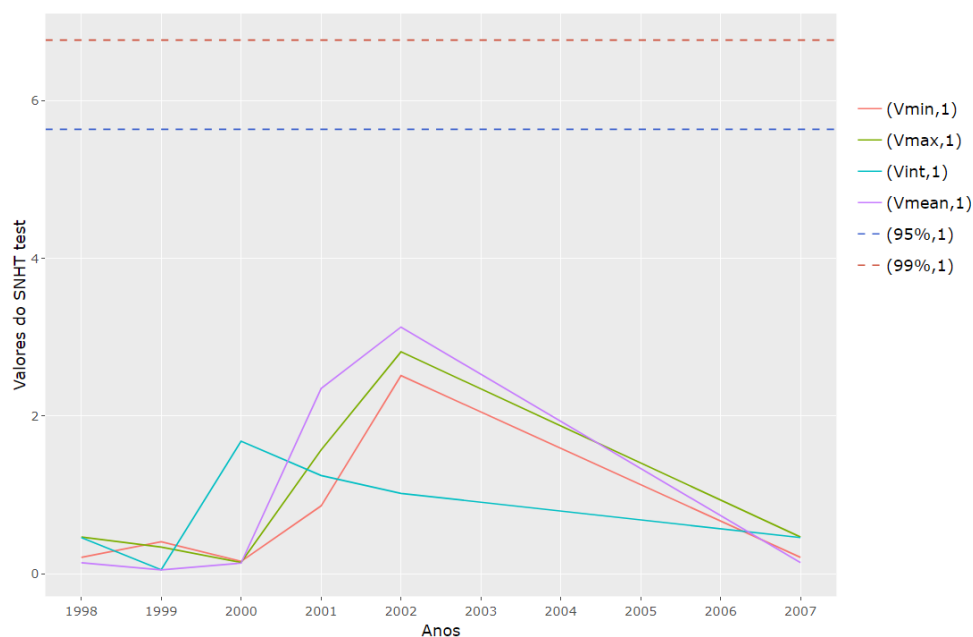


Figura C.5: *Standard Normal Homogeneity Test* aplicado às series anuais de velocidade do vento mínimas (V_{min}), máximas (V_{max}), intervalo (V_{int}), médias (V_{mean}) e níveis de confiança de 95 e 99%.

Apêndice D

Ferramentas

O estudo apresentado nesta dissertação foi desenvolvido na linguagem R [46]. Aqui são apresentados os principais *packages* e funções utilizadas ao longo deste estudo.

- **tidyverse** [69] – o tidyverse é uma coleção de 8 *packages* cujo propósito visa a possibilitar uma melhor e mais fácil manipulação, exploração e visualização de dados. Seguem-se alguns dos mais utilizados/úteis para o propósito deste projeto:
 - **dplyr** [70] – consiste num conjunto de funções que ajudam o utilizador a lidar com os desafios mais comuns na manipulação de dados: desde adicionar novas variáveis, selecioná-las com base no seu valor a filtrar ou agrupar diferentes casos, de entre muitas outras funções que facilitam estas operações;
 - **tidyr** [68] – este *package* tem como objetivo a criação de *Tidy data*, sendo estes dados onde cada coluna representa uma variável, cada linha uma observação e cada célula um único valor. Dados neste formato facilitam a sua análise e manipulação nos restantes *packages* do *tidyverse*;
 - **tibble** [40] – permite a criação de *tibbles*, que proporcionam uma melhor formatação face aos tradicionais *data frames*, facilitando a sua manipulação;
 - **stringr** [67] – fornece um conjunto de funções concebidas para facilitar a manipulação de *strings*;
 - **lubridate** [22] – é um *package* concebido para tornar a manipulação de datas e intervalos de tempo, o mais fácil e intuitiva possível. As funções deste *package* são tolerantes a vários formatos de dados (*date*, *POSIXct* ou séries temporais *zoo/xts*), isto é, têm um comportamento consistente, e são úteis para extrair e manipular componentes dos mesmos, tais como, o ano, mês, dia da semana, horas ou minutos. Além disso, apresenta ainda funções capazes de lidar com diferentes fusos horários, anos bissextos, horários de verão ou outras peculiaridades relacionadas com o tempo;
 - **ggplot2** [66] – é um *package* dedicado à visualização de dados. Permite a construção de quase qualquer tipo de gráfico e pode melhorar muito a sua qualidade estética tornando-o mais eficiente na sua criação. Para cada tipo de gráfico deste *package* podem ser aplicadas imensas opções como personalizar títulos ou legendas, adicionar anotações, usar *facets*, de entre muitas outras. O ggplot2 constrói os gráficos através de camadas utilizando funções `geom_` podendo, por vezes, combinar diferentes tipos de gráficos num só;

- **GGally** [57] – trata-se de uma extensão do ggplot2 com a adição de algumas funções nomeadamente *ggcorr* (Correlation value plot) que calcula as correlações entre os dados fornecidos;
- **dygraphs** [63] – *package* de interfaces gráficas que permite representar automaticamente objetos *xts* com eixos configuráveis e recursos interativos (como por exemplo zoom da série temporal);
- **plotly** [60] – focado na criação de gráficos interativos através do ggplot2 e/ou interface personalizada;
- **ggribes** [72] – *package* para criar gráficos *Ridgeline*, úteis para observar mudanças nas distribuições de uma dada variável ao longo do tempo ou do espaço;
- **stats** [47] – *package* formado por um conjunto de funções que auxiliam cálculos estatísticos e ferramentas que facilitam a análise de dados;
- **timetk** [12] – este *package* disponibiliza um conjunto de ferramentas para visualizar, manipular e transformar dados de séries temporais;
- **imputeTs** [38] – é um *package* focado na imputação de *missing values* em séries temporais univariadas. Dispõe de diversos algoritmos de imputação assim como funções para imprimir e visualizar estatísticas relacionadas com esses valores em falta;
- **hydroTSM** [35] – focado na análise, interpolação e representação gráfica de séries temporais orientadas a dados hidrológicos;
- **climindex.pcic** [19] – este *package* é focado exclusivamente no cálculo dos índices que caracterizam a temperatura e a precipitação;
- **tseries** [62] – *package* com funções para analisar séries temporais;
- **forecast** [26] – fornece métodos e ferramentas necessárias para representar e analisar previsões de séries temporais univariadas;
- **xts** [54] – *eXtensible Time Series* (xts) é um *package* que fornece uma classe de séries temporais, cujo index é baseado no tempo da mesma. Este *package* permite a criação e análise de objetos *xts*, assim como ferramentas para facilitar a sua manipulação, nomeadamente substituir e atualizar valores, selecionar, indexar e criar subconjuntos deste objeto, lidar com valores em falta, ou realizar operações aritméticas, entre outros. O index deste tipo de objetos suporta classes de tempo ou datas (como *Date*, *POSIXct*, *chron*, *yearmon*, *yearqtr*, e *dateTime*). Sendo o *xts* uma extensão do *package zoo*, o primeiro consegue aplicar as mesmas funções que o *zoo* e é otimizado para lidar com grandes volumes de dados;
- **trend** [44] – *package* com diversos testes de detecção de tendência e pontos de não-homegeidades/*breaks*, nomeadamente: teste de tendência de *Mann-Kendall*, e testes de homogeneidades como os testes de *Pettitt*, *Buishand*, ou *Von Neumann*, entre outros;
- **climtrends** [21] – análogo ao *package trend*, no que diz respeito à presença de funções que testes absolutos e relativos de homogeneidade, como os testes de *Pettitt*, *Buishand*, ou *Von Neumann*, entre outros, assim como os respetivos valores críticos.

Nas Tabelas D.1, D.2 e D.3 são elencadas algumas das funções da linguagem R que foram utilizadas neste estudo, respetivos *packages*, e uma breve descrição das mesmas.

A Tabela D.1 apresenta algumas funções úteis para manipular dados em R, enquanto que as Tabelas D.2 e D.3, mostram, exclusivamente, algumas das principais funções, focadas na análise de séries temporais e especializadas em estudos de homogeneização de séries temporais, por esta ordem, assim como o respetivo *package* a que correspondem e uma breve descrição do *output* dessas funções.

Tabela D.1: *Packages* e funções da linguagem R utilizadas e respectivas descrições.

<i>package</i>	Função (nome)	Descrição
tibble [40]	tibble(<i>Build a data frame</i>)	criação de <i>tibble</i>
	as_tibble (<i>Coerce lists, matrices, and more to data frames</i>)	conversão de objetos (listas, matrizes, ou <i>data frame</i>) em <i>tibble</i>
dplyr [70]	bind (<i>Efficiently bind multiple data frames by row and column</i>)	'bind_rows' ou 'bind_cols', combina linhas ou colunas de diferentes <i>data frames</i> num só
	mutate_joins (<i>Mutating joins</i>)	adiciona colunas de uma tabela (y) a outra (x): são incluídas todas as linhas de x em 'left_join'; todas as de y em 'right_join'; todas as de x e y em 'inner_join' e todas as de x ou y em 'full_join'
	leap_year (<i>Is a year a leap year?</i>)	testa se é ou não ano bisexto
lubridate [22]	month (<i>Get/set months component of a date-time</i>)	obtem a componente da data conforme a função; (outras variações: 'year', 'hour', 'week', 'semester')
	force_tz (<i>Replace time zone to create new date-time</i>)	mesma data e tempo (de relógio), mas com uma nova <i>time-zone</i>
	ymd_hms (<i>Parse date-times with year, month, and day, hour, minute, and second components</i>)	transforação de datas em formato <i>string</i> ou numérico para objetos <i>POSIXct</i>
ggplot2 [66]	geom_bar (<i>Bar charts</i>)	gráfico de barras
	geom_boxplot (<i>A box and whiskers plot</i>)	<i>boxplot</i>
	geom_point (<i>Points</i>)	<i>scatter plot</i>
	geom_smooth	tendência de um conjunto de dados, através de suavização; o método de suavização aplicado pode variar entre <i>loess</i> , <i>gam</i> ou <i>linear model</i> para tendências lineares
plotly [60]	Ggplotly (<i>Convert ggplot2 to plotly</i>)	conversão de objeto <i>ggplot2</i> num outro: <i>ggplotly</i>

Tabela D.2: *Packages* e funções da linguagem R utilizadas em séries temporais, e respectivas descrições.

<i>package</i>	Função (nome)	Descrição
stats [47]	decompose (<i>Classical Seasonal Decomposition by Moving Averages</i>)	decomposição de série temporal nas componentes: tendência, sazonalidade e restante utilizando <i>moving averages</i> ; lida com modelos aditivos e multiplicativos
	stl (<i>Seasonal Decomposition of Time Series by Loess</i>)	decomposição de série temporal nas componentes: tendência, sazonalidade e restante utilizando <i>loess</i> ; lida com modelos aditivos e multiplicativos
dygrpahs [63]	dygraph (<i>dygraph interactive plot for time series data</i>)	gráfico <i>dygraph</i> , interativo, de objeto <i>xts</i>
timetk [12]	tk_summary_diagnostics (<i>Group-wise Time Series Summary</i>)	sumário de uma ou mais séries temporais em formato tibble (início, fim, unidades, escala, assim como mínimo, 1º quartil, mediana, média, 3º quartil, e frequência máxima)
	ts_impute_vec (<i>Missing Value Imputation for TimeSeries</i>)	correção de <i>missing values</i> através de interpolação linear; na presença de sazonalidade é aplicada primeiro uma decomposição <i>stl</i> e adicionada de volta, depois de corrigidos
tseries [62]	adf.test (<i>Augmented Dickey-Fuller Test</i>)	teste de estacionaridade de séries temporais
	kpss.test (<i>KPSS Test for Stationarity</i>)	
imputeTs [38]	statNa (<i>Print Statistics about Missing Values</i>)	sumário sobre <i>missing values</i> com as informações: n° de observações (incluindo NAs), n° e percentagem de NA's, n° de um ou mais NAs consecutivos (<i>gaps</i>) e média do tamanho de <i>gaps</i> , <i>gap</i> mais longo e mais frequente; tamanho do maior gap na série temporal; visão geral da frequência com que cada série de NAs consecutivos ocorre
	na_interpolation (<i>Missing Value Imputation by Interpolation</i>)	correção de <i>missing values</i> através de interpolação linear, interpolação <i>Stineman</i> ou interpolação por <i>spline</i>
	na_remove (<i>Remove Missing Values</i>)	remoção de <i>missing values</i> da série temporal
	ggplot_na_distribution (<i>Visualize the Distribution of Missing Values</i>)	visualização da distribuição de <i>missing values</i> na série temporal
forecast [26]	ma (<i>Moving-average smoothing</i>)	série temporal com os valores da <i>moving average</i>
	tsoutliers (<i>Identify and replace outliers in a time series</i>)	identificação de <i>outliers</i> e proposta de correção; na presença de sazonalidade é aplicada primeiro uma decomposição <i>stl</i> caso contrário é utilizado o método <i>supsmu</i>
	tsclean (<i>Identify and replace outliers and missing values in a time series</i>)	correção de <i>outliers</i> (com as estimativas de 'tsoutliers') e de <i>missing values</i>
	Acf (<i>(Partial) Autocorrelation and Cross-Correlation Function Estimation</i>)	cálculo das funções de <i>autocorrelation</i> , <i>partial auto correlation</i> ou <i>cross correlation</i>
xts [54]	xts (<i>Create Or Test For An xts Time-Series Object</i>)	criação de objeto <i>xts</i>
	apply.monthly; apply.daily; apply.weekly; (<i>Apply Function over Calendar Periods</i>)	subdivisão da série temporal em diferentes períodos de tempo (frequência diária, semanal, mensal,...)
	coredata.xts (<i>Extract/Replace CoreData of an xts Object</i>)	extração dos dados de um objeto <i>xts</i> numa matriz (index do <i>xts</i> convertido nos nomes das linhas)
	periodicity (<i>Approximate Series Periodicity</i>)	periodicidade do objeto <i>xts</i> (diários, semanais, anuais,...)
climdex.pcic [19]	climdex.su (<i>Frost Days</i>)	calcula o índice de temperatura TN0
	climdex.fd (<i>Summer Days</i>)	calcula o índice de temperatura TX25
	climdex.rx5day (<i>Monthly Maximum Consecutive 5 day Precipitation</i>)	calcula o índice de precipitação RX5D
hydroTSM [35]	climdex.rx1day (<i>Monthly Maximum Consecutive 1 day Precipitation</i>)	calcula o índice de precipitação RX5D
	Climograph (<i>Climograph</i>)	climograma baseado em dados de precipitação e temperatura
	matrixplot (<i>Matrixplot</i>)	<i>heatmap</i> de uma matriz (de duas dimensões) de dados
	hydroplot (<i>Hydrological time series plotting and extraction</i>)	gráficos de linhas, <i>boxplots</i> , e/ou histogramas, de séries temporais diárias, mensais, anuais ou sazonais
	seasonalfunction (<i>Seasonal Function</i>)	valores representativos das estações do ano (somas, médias,...)

Tabela D.3: *Packages* e funções da linguagem R utilizadas na homogeneização de séries temporais e respetiva descrição.

<i>package</i>	Função (nome)	Descrição
trend [44]	br.test (<i>Buishand Range Test for Change-Point Detection</i>)	testes de <i>Buishand</i> , <i>Pettitt</i> , <i>Standard Normal Homogeneity</i> e <i>Von Neumann</i> , respetivamente; (nome da série utilizada, valor estimado, <i>p-value</i> , tamanho dos dados, hipótese alternativa, e, caso se aplique, <i>index</i> do break na série temporal
	pettitt.test (<i>Pettitt's Test for Change-Point Detection</i>)	
	snh.test (<i>Standard Normal Homogeneity Test (SNHT) for Change-Point Detection</i>)	
	bartels.test (<i>Bartels Test for Randomness</i>)	
	mk.test (<i>Mann-Kendall Trend Test</i>)	
climrends [21]	Buishand.Critical.Values (<i>Critical values for the Buishand test</i>)	valores críticos dos testes <i>Buishand</i> , <i>Pettitt</i> , <i>Standard Normal Homogeneity</i> e <i>Von Neumann</i> , respetivamente
	Pettitt.Critical.Values (<i>Critical values for the Pettitt test</i>)	
	SNHT.Critical.Values (<i>Critical values for the standard normal homogeneity test (SNHT)</i>)	
	Von.Neumann.Ratio.Test.Critical.Values (<i>Critical values for the Von Neumann ratio test</i>)	

Bibliografia

- [1] [Interpret all statistics and graphs for cross correlation](#). Accessed: 2021-09-01.
- [2] [Exploratory data analysis](#). Accessed: 2021-08-10.
- [3] Visualising temperatures in amsterdam as a heatmap in r — part i | by deepak gulati | towards data science. <https://towardsdatascience.com/visualising-temperatures-in-amsterdam-as-a-heatmap-in-r-part-i-7e1be1e251c0>. (Accessed on 07/09/2021).
- [4] Rakesh Aggarwal and Priya Ranganathan. [Common pitfalls in statistical analysis: The use of correlation techniques](#). *Perspectives in Clinical Research*, 7:187, 10 2016. doi:10.4103/2229-3485.192046.
- [5] C Donald Ahrens. *Essentials of meteorology: an invitation to the atmosphere*. Nelson Education, 2011.
- [6] Lonneke Boels, Arthur Bakker, Wim Van Dooren, and Paul Drijvers. Conceptual difficulties when interpreting histograms: A review. *Educational Research Review*, 28:100291, 2019.
- [7] Artem Cheprasov. [Take online courses. earn college credit. research schools, degrees careers](#). Accessed: 2021-08-07.
- [8] Robert B Cleveland, William S Cleveland, Jean E McRae, and Irma Terpenning. Stl: A seasonal-trend decomposition. *J. Off. Stat*, 6(1):3–73, 1990.
- [9] M. Fátima Coelho, Isabel De Lima, Alexandre Ramos, and Ricardo Trigo. [Trends in seasonal surface air temperature in mainland portugal, since 1941](#). *International Journal of Climatology*, 34, 05 2014. doi:10.1002/joc.3803.
- [10] Ana Cristina Costa and Amílcar Soares. Homogenization of climate data: review and new perspectives using geostatistics. *Mathematical geosciences*, 41(3):291–305, 2009.
- [11] Paul S.P. Cowpertwait and Andrew V. Metcalfe. *Introductory time series with R*. Use R! Springer, Dordrecht, Heidelberg, London, cop. 2009. Fundo de Investigação MagIC.
- [12] Matt Dancho and Davis Vaughan. [timetk: A Tool Kit for Working with Time Series in R](#), 2021. R package version 2.6.1.
- [13] Robert Dawson. How significant is a boxplot outlier? *Journal of Statistics Education*, 19(2), 2011.
- [14] M. Isabel P. de Lima, Fátima Espírito Santo, Alexandre M. Ramos, and João L.M.P. de Lima. [Recent changes in daily precipitation and surface air temperature extremes in mainland](#)

- portugal, in the period 1941–2007. *Atmospheric Research*, 127:195 – 209, 2013. ISSN: 0169-8095. doi:<https://doi.org/10.1016/j.atmosres.2012.10.001>.
- [15] M. Isabel P. de Lima, Fátima Espírito Santo, Alexandre M. Ramos, and João L.M.P. de Lima. [Recent changes in daily precipitation and surface air temperature extremes in mainland portugal, in the period 1941–2007](#). *Atmospheric Research*, 127:195 – 209, 2013. ISSN: 0169-8095. doi:<https://doi.org/10.1016/j.atmosres.2012.10.001>.
- [16] ATLAS CLIMÁTICO DOS ARQUIPÉLAGOS, DAS CANÁRIAS, and DA MADEIRA E DOS AÇORES. Atlas climático de los archipiélagos de canarias, madeira y azores.
- [17] Stephen Few and Perceptual Edge. Criteria for evaluating visual eda tools. *Visual Business Intelligence Newsletter*. Perceptual Edge, 2012.
- [18] Manish Karkera (Institute for Environmental Analytics). [Show your stripes](#). Accessed: 2021-08-10.
- [19] David Bronaugh for the Pacific Climate Impacts Consortium. *climindex.pcic: PCIC Implementation of Climindex Routines*, 2020. R package version 1.1-11.
- [20] Michael Galarnyk. [Understanding boxplots](#), Jul 2020. (Accessed on 07/09/2021).
- [21] Jose Gama. *climtrends: Statistical Methods for Climate Sciences*, 2016. R package version 1.0.6.
- [22] Garrett Grolemund and Hadley Wickham. [Dates and times made easy with lubridate](#). *Journal of Statistical Software*, 40(3):1–25, 2011.
- [23] Giles Harrison. *Meteorological measurements and instrumentation*. John Wiley & Sons, 2015.
- [24] NM Holden and AJ Brereton. Definition of agroclimatic regions in ireland using hydro-thermal and crop yield data. *Agricultural and Forest Meteorology*, 122(3-4):175–191, 2004.
- [25] Mia Hubert and Ellen Vandervieren. An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, 52(12):5186–5201, 2008.
- [26] Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O’Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *forecast: Forecasting functions for time series and linear models*, 2020. R package version 8.13.
- [27] Robin John Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Australia, 2nd edition, 2018.
- [28] Stephanie Hänsel, Deusdedit M. Medeiros, Jörg Matschullat, Reinaldo A. Petta, and Isamara de Mendonça Silva. [Assessing homogeneity and climate variability of temperature and precipitation series in the capitals of north-eastern brazil](#). *Frontiers in Earth Science*, 4:29, 2016. ISSN: 2296-6463. doi:10.3389/feart.2016.00029.
- [29] Daniel A Keim, Ming C Hao, and Umeshwar Dayal. Hierarchical pixel bar charts. *IEEE Transactions on Visualization and Computer Graphics*, 8(3):255–269, 2002.
- [30] Nina Kirchner. Warming stripes—visualisations of annual temperature series from the nordic countries. dataset version 3.0. bolin centre database, 2021. Accessed: 2021-08-05.
- [31] Matthieu Komorowski, Dominic Marshall, Justin Saliccioli, and Yves Crutain. *Exploratory Data Analysis*, pages 185–203. 09 2016. ISBN: 978-3-319-43740-8 (Print) 978-3-319-43742-2 (Online).

- [32] Matthew Lane, Alberto Maiocco, Sanjiv K. Bhatia, and Sharlee Climer. [Chapter four - eyeing the patterns: Data visualization using doubly-seriated color heatmaps](#). volume 119 of *Advances in Computers*, pages 121–156. Elsevier, 2020. doi:<https://doi.org/10.1016/bs.adcom.2020.03.003>.
- [33] Pedro M A Miranda. *Introdução à Meteorologia*. 2012.
- [34] João Maroco. *Análise Estatística – Com utilização do SPSS*. EdiçõesSílabo, 3^a edição edition, 2003.
- [35] Mauricio Zambrano-Bigiarini. [hydroTSM: Time Series Management, Analysis and Interpolation for Hydrological Modelling](#), 2020. R package version 0.6-0 . doi: <https://doi.org/10.5281/zenodo.83964>.
- [36] Karen A Monsen. Exploratory data analysis. In *Intervention Effectiveness Research: Quality Improvement and Program Evaluation*, pages 77–85. Springer, 2018.
- [37] Stephan Morgenthaler. Exploratory data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):33–44, 2009.
- [38] Steffen Moritz and Thomas Bartz-Beielstein. [imputeTS: Time Series Missing Value Imputation in R](#). *The R Journal*, 9(1):207–218, 2017. doi:10.32614/RJ-2017-009.
- [39] Anna L Morozova and Maria Antónia Valente. Homogenization of portuguese long-term temperature data series: Lisbon, coimbra and porto. *Earth System Science Data*, 4(1):187–213, 2012.
- [40] Kirill Müller and Hadley Wickham. [tibble: Simple Data Frames](#), 2021. R package version 3.1.1.
- [41] Mohamed N. Nounou and Bhavik R. Bakshi. [Chapter 5 - multiscale methods for denoising and compression](#). In Beata Walczak, editor, *Wavelets in Chemistry*, volume 22 of *Data Handling in Science and Technology*, pages 119 – 150. Elsevier, 2000. doi:[https://doi.org/10.1016/S0922-3487\(00\)80030-1](https://doi.org/10.1016/S0922-3487(00)80030-1).
- [42] Shay Palachy. [Detecting stationarity in time series data](#), Nov 2019. Accessed: 2020-10-25.
- [43] Thomas C Peterson, David R Easterling, Thomas R Karl, Pavel Groisman, Neville Nicholls, Neil Plummer, Simon Torok, Ingeborg Auer, Reinhard Boehm, Donald Gullett, et al. Homogeneity adjustments of in situ atmospheric climate data: a review. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 18(13):1493–1517, 1998.
- [44] Thorsten Pohlert. [trend: Non-Parametric Trend Tests and Change-Point Detection](#), 2020. R package version 1.1.4.
- [45] Universidade do Porto. [Memria u.porto](#). Accessed: 2020-11-10.
- [46] R Core Team. [R: A Language and Environment for Statistical Computing](#). R Foundation for Statistical Computing, Vienna, Austria, 2019.
- [47] R Core Team. [R: A Language and Environment for Statistical Computing](#). R Foundation for Statistical Computing, Vienna, Austria, 2019.
- [48] Alexandre M Ramos, Ricardo M Trigo, and Fátima E Santo. Evolution of extreme temperatures over portugal: recent changes and future scenarios. *Climate Research*, 48(2-3):177–192, 2011.
- [49] Sara Ribeiro, Júlio Caineta, and Ana Cristina Costa. Review and discussion of homogenisation methods for climate data. *Physics and Chemistry of the Earth, Parts A/B/C*, 94:167–179, 2016.
- [50] Rosa Maria Rodriguez, Agueda Benito, and A Portela. *Meteorología y Climatología (FECYT). España*. Fecyt (fundación española para la ciencia y la tecnología) edition, 2004. ISBN: 84-688-8535-5.

- [51] Dominic Royé. [Visualize monthly precipitation anomalies](#), Jul 2019. Accessed: 2021-08-10.
- [52] Dominic Royé. [Visualize climate anomalies](#), Mar 2020. Accessed: 2021-08-10.
- [53] Kim Rutledge, Tara Ramroop, Diane Boudreau, and Erin Sprout Hilary Costa Hilary Hall Jeff Hunt Melissa McDaniel, Santani Teng. [Atmospheric pressure](#). Online, March 2011. Accessed: 2020-12-11.
- [54] Jeffrey A. Ryan and Joshua M. Ulrich. [xts: eXtensible Time Series](#), 2020. R package version 0.12.1.
- [55] A.K.M. Saifuddin and A. W. Jayawardena. *Homogeneity and trend analysis of temperature for urban and rural areas*. PhD thesis.
- [56] Fatima Espirito Santo, Alexandre M Ramos, M Isabel P de Lima, and Ricardo M Trigo. Seasonal changes in daily precipitation extremes in mainland portugal from 1941 to 2007. *Regional environmental change*, 14(5):1765–1788, 2014.
- [57] Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. [GGally: Extension to 'ggplot2'](#), 2021. R package version 2.1.2.
- [58] Howard J Seltman. *Experimental design and analysis*. Carnegie Mellon University, 01 2012.
- [59] R.H. Shumway and D.S. Stoffer. [Time Series Analysis and Its Applications: With R Examples](#). Springer Texts in Statistics. Springer New York, 2010. ISBN: 9781441978646.
- [60] Carson Sievert. [Interactive Web-Based Data Visualization with R, plotly, and shiny](#). Chapman and Hall/CRC, 2020. ISBN: 9781138331457.
- [61] Chiawa H Sim, Francis F Gan, and Tang C Chang. Outlier labeling with boxplot procedures. *Journal of the American Statistical Association*, 100(470):642–652, 2005.
- [62] Adrian Trapletti and Kurt Hornik. [tseries: Time Series Analysis and Computational Finance](#), 2020. R package version 0.10-48.
- [63] Dan Vanderkam, JJ Allaire, Jonathan Owen, Daniel Gromer, and Benoit Thieurmél. [dygraphs: Interface to 'Dygraphs' Interactive Time Series Charting Library](#), 2018. R package version 1.1.1.6.
- [64] David Vetter. From dresses to electric cars, why are these stripes all over social media? *Forbes*, June 2021.
- [65] Michael F. Wehner. [Characterization of long period return values of extreme daily temperature and precipitation in the cmip6 models: Part 2, projections of future change](#). *Weather and Climate Extremes*, 30:100284, 2020. ISSN: 2212-0947. doi:<https://doi.org/10.1016/j.wace.2020.100284>.
- [66] Hadley Wickham. [ggplot2: Elegant Graphics for Data Analysis](#). Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4.
- [67] Hadley Wickham. [stringr: Simple, Consistent Wrappers for Common String Operations](#), 2019. R package version 1.4.0.
- [68] Hadley Wickham. [tidyr: Tidy Messy Data](#), 2020. R package version 1.1.2.
- [69] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemond, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohnske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. [Welcome to the tidyverse](#). *Journal of Open Source Software*, 4(43):1686, 2019. doi:10.21105/joss.01686.

-
- [70] Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. R package version 1.0.3.
- [71] JB Wijngaard, AMG Klein Tank, and GP Können. Homogeneity of 20th century european daily temperature and precipitation series. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 23(6):679–692, 2003.
- [72] Claus O. Wilke. *ggridges: Ridgeline Plots in 'ggplot2'*, 2021. R package version 0.5.3.
- [73] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.
- [74] G WMO. *WMO No. 8. Guide to meteorological instruments and methods of observation*. WMO, 2014.
- [75] Sheng Yue and ChunYuan Wang. The mann-kendall test modified by effective sample size to detect trend in serially correlated hydrological series. *Water resources management*, 18(3):201–218, 2004.
- [76] Zhu Zhongming, Lu Linong, Zhang Wangqiang, Liu Wei, et al. Show your stripes. 2019.