Clustering de instalações, com (re)definição de segmentos em função do comportamento energético

Joana Martins

Mestrado em Engenharia Matemática

Departamento de Matemática da Faculdade de Ciências da Universidade do Porto

2014

Orientador

Prof Dr João Nuno Tavares, Professor Associado, FCUP

Profª Dra Ana Paula Rocha, Professora Auxiliar, FCUP

Profª Dra Margarida Silva, Professora Associada, FCUP

Profª Dra Maria Eduarda Silva, Professora Associada, FEP

Dra Susana Magalhães, DGE, EDP Distribuição – Energia S.A.

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

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

O Presidente do Júri,

Porto, _____/_____/_____

Agradecimentos

Gostaria de agradecer aos meus orientadores, Prof Dr João Nuno Tavares, Profª Dra Ana Paula Rocha, Profª Dra Margarida Brito, Profª Dra Maria Eduarda Silva, Dra Susana Magalhães e Eng Duarte Duarte, não só pela disponibilidade e acompanhamento, mas também pelas suas sugestões e críticas que tiveram um contributo fundamental para o desenvolvimento deste estágio.

A todos os elementos da DGE/EDP Distribuição - Energia, S.A. que me acolheram durante o período de estágio, obrigado pelo carinho e boa disposição que todos os dias transmitiram.

Aos meus colegas de mestrado, em especial à minha colega de estágio Elena Selaru, obrigado pelos momentos de entusiasmo e apoio partilhados em conjunto.

Agradeço ainda às pessoas mais especiais da minha vida pelo incentivo, compreensão e encorajamento, durante todo este período.

Resumo

Este relatório relata todo o trabalho realizado ao longo do estágio curricular do Mestrado em Engenharia Matemática, na empresa EDP Distribuição - Energia, SA.

A EDP Distribuição tem instalados contadores de leitura inteligente (telecontagem) em várias empresas (instalações) situadas em território nacional. O presente trabalho teve como objetivo o *agrupamento de instalações através do seu consumo energético diário*.

Para alcançar o objetivo foi necessário, inicialmente, perceber os consumos energéticos, definindo variáveis explicativas do consumo, como por exemplo o horário de funcionamento de cada instalação, dias de feriado e variáveis climáticas (temperatura, humidade, etc).

De seguida realizou-se um estudo sobre os consumos numa perspetiva de séries temporais, utilizando métodos de decomposição, *suavização* e *análise espectral singular*, e *regressão linear múltipla*.

Observou-se que apenas algumas variáveis são explicativas do consumo e por este motivo utilizaram-se os métodos *backward*, *random forest* e *correlação* para encontrar as variáveis mais significativas dos consumos energéticos.

Após selecionar as instalações a utilizar no trabalho, procedeu-se à pesquisa de métodos de clustering de séries temporais para dados de grande dimensão. Utilizaram-se os métodos *u-shapelets* e *fatores de similaridade*.

Abstract

This report describes all the work realized throughout the Mathematical Engineering Master Degree's internship at EDP Distribuição - Energia, S.A. EDP Distribuição has installed smart meter reading (telemetry) in various companies located in the national territory. The present study's aim was to *cluster companies considering their daily energy consumption*. Initially, to achieve this goal, it was necessary to understand energy consumption by setting explanatory variables of consumption, such as the opening hours of each company, holidays and climatic variables (temperature, humidity, etc). Then, a study on consumption in a time series perspective was conducted, using decomposition methods, *smoothing* and *singular spectral analysis*, and *multiple linear regression*. It was observed that only a few variables influence the consumption and, for this reason, *backward*, *random forest* and *correlation* methods were employed to find the most significant variables in energy consumption. After selecting the companies for this work, methods of clustering time series for big data were researched and *u-shapelets* and *similarity factors* methods were used.

Conteúdo

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas

ComprDia Variável explicativa que indica o Comprimento do Dia (nº de horas de luz solar)

DiaSemana Variável explicativa que indica o dia da semana (1 - Segunda, 2 - Terça, ...)

HorTrab Variável explicativa que indica o horário de trabalho de uma instalação por dia da semana

HumMax Variável explicativa que indica a Humidade Máxima por dia, medida em percentagem (%)

HumMedia Variável explicativa que indica a Humidade Média por dia, medida em percentagem (%)

HumMin Variável explicativa que indica a Humidade Mínima por dia, medida em percentagem (%)

PC Componentes Principais (do inglês Principal Components)

PCA Análise de Componentes Principais (do inglês Principal Components Analysis)

PresMax Variável explicativa que indica a Pressão ao nível do Mar Máxima por dia, medida em hetopascal (hPa)

PresMedia Variável explicativa que indica a Pressão ao nível do Mar Média por dia, medida em hetopascal (hPa)

PresMin Variável explicativa que indica a Pressão ao nível do Mar Mínima por dia, medida em hetopascal (hPa)

PtOrvMax Variável explicativa que indica o Ponto de Orvalho Máximo por dia, medida em graus Celsius (ºC)

PtOrvMedio Variável explicativa que indica o Ponto de Orvalho Médio por dia, medida em graus Celsius (ºC)

PtOrvMin Variável explicativa que indica o Ponto de Orvalho Mínimo por dia, medida em graus Celsius (ºC)

SSA Análise Espectral Singular (do inglês Singular Spectrum Analysis)

STM Série temporal multivariada

SVD Decomposição de Valor Singular (do inglês Singular Value Decomposition)

TempMax Variável explicativa que indica a Temperatura Máxima por dia, medida em graus Celsius (ºC)

TempMedia Variável explicativa que indica a Temperatura Média por dia, medida em graus Celsius (ºC)

TempMin Variável explicativa que indica a Temperatura Mínima por dia, medida em graus Celsius (ºC)

VelRajVentoMax Variável explicativa que indica a Velocidade Máxima de Rajada de Vento por dia, medida em quilómetros por hora (km/h)

VelVentoMax Variável explicativa que indica a Velocidade do Vento Máxima por dia, medida em quilometros por hora (km/h)

VelVentoMedia Variável explicativa que indica a Velocidade do Vento Média por dia, medida em quilómetros por hora (km/h)

Capítulo 1

Introdução

A EDP Distribuição tem instalados contadores de leitura inteligente (telecontagem) em várias empresas (instalações) situadas em território nacional. A implementação destas redes inteligentes permite automatizar a gestão da rede, melhorar a qualidade de serviço, fornecendo ao consumidor meios que permitem gerir e otimizar o seu consumo diário, minimizando custos e impactos ambientais.
Os sistemas inteligentes registam o consumo de energia elétrica em intervalos de 15 minutos (96 leituras diárias). A estes registos dá-se o nome de *diagrama de carga* da instalação.
Estes equipamentos extraem uma quantidade enorme de dados que necessitam ser analisados. Surgem vários desafios:

- Caracterizar tendências base de consumo;
- Detetar e caracterizar padrões de consumo;
- Analisar correlação com fatores externos (temperatura, humidade, iluminação, aquecimento, etc.);
- Detetar comportamentos anómalos;
- Analisar fraudes.

A propósito destes problemas, a EDP Distribuição proporcionou dois estágios para alunos do Mestrado em Engenharia Matemática, subordinados aos temas seguintes:

Estágio 1 Desagregação do consumo energético em sub-conjuntos;

Estágio 2 Clustering de instalações, com (re)definição de segmentos em função do comportamento energético.

O estágio 1 foi desenvolvido pela aluna Elena Selaru e este relatório descreve o desenvolvimento do estágio 2. Ambos os estágios avançaram em paralelo, tendo uma parte inicial comum.
Os objetivos comuns a ambos estágios foram:

1. Compreender os consumos energéticos;
2. Extrair os fatores externos correlacionados com o consumo;
3. Selecionar as instalações a utilizar;

Os objetivos referentes ao tema *Clustering de instalações* foram:

5. Pesquisa e implementação de métodos de clustering e medidas de semelhança;
6. Avaliação e validação dos métodos.

Este relatório de estágio está organizado em sete capítulos, designadamente:
No primeiro capítulo, contextualiza-se e apresenta-se o estágio realizado. No segundo capítulo, descreve-se a análise gráfica efetuada a um diagrama de carga e no terceiro capítulo, a análise numa perspetiva de séries temporais, construindo um conjunto de medidas e variáveis explicativas do consumo. No quarto capítulo, apresenta-se o estudo realizado aos conjuntos anteriores, concluindo quais as medidas a utilizar e quais as variáveis que são significativas. No quinto capítulo, enumera-se o conjunto de critérios necessários na seleção de instalações a utilizar. Os primeiros 5 capítulos dizem respeito à parte comum dos estágios.
No sexto capítulo expõe-se a metodologia utilizada para alcançar o objetivo final do estágio e apresentam-se os resultados obtidos. No sétimo capítulo, são apresentadas as conclusões do estágio e propõem-se algumas sugestões para futuros trabalhos no mesmo contexto. O trabalho termina com as referências bibliográficas, seguidas dos anexos que incluem os instrumentos essenciais para o trabalho realizado.

Capítulo 2

Análise preliminar de uma instalação

2.1 Introdução

Como já foi referido no capítulo 1, o objetivo principal do estágio foi agrupar um conjunto de instalações segundo o seu consumo energético. Para tal, foi necessário perceber como é um consumo energético de uma instalação e compreender como se comporta.
Inicialmente foi disponibilizado o diagrama de carga de uma instalação. O diagrama de carga de uma instalação contém a informação da potência, energia consumida a cada intervalo de tempo (neste caso definido em períodos de 15 minutos, ou seja, 96 leituras diárias). Neste caso, o período temporal máximo foi de 1 de Janeiro de 2011 a 31 de Agosto de 2013, o que corresponde a 93 504 observações.
Na próxima secção será apresentada a análise gráfica realizada a este diagrama de carga. Esta análise foi desenvolvida no Excel.

2.2 Análise gráfica

O objetivo desta análise foi perceber o consumo energético da instalação disponibilizada, sendo assim foi necessário converter a potência em energia. Na figura 2.1 está representado o diagrama de carga no ano 2013. Observando a figura constatou-se uma enorme

dificuldade em perceber o comportamento da curva pois existia uma grande quantidade de observações, sendo apenas possível concluir que o consumo energético varia entre 0 e 130 kWh.

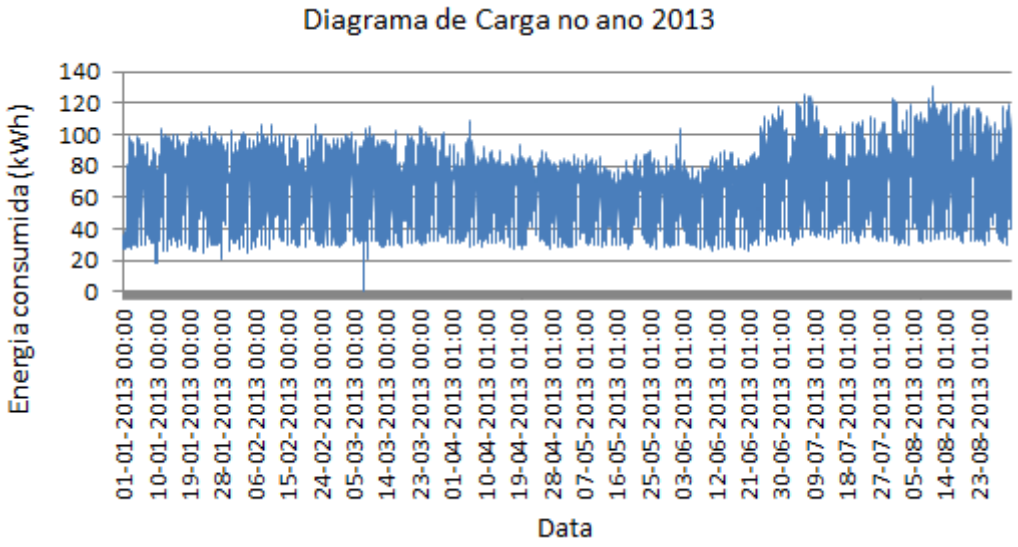

Figura 2.1: Diagrama de Carga no ano 2013 da instalação inicial

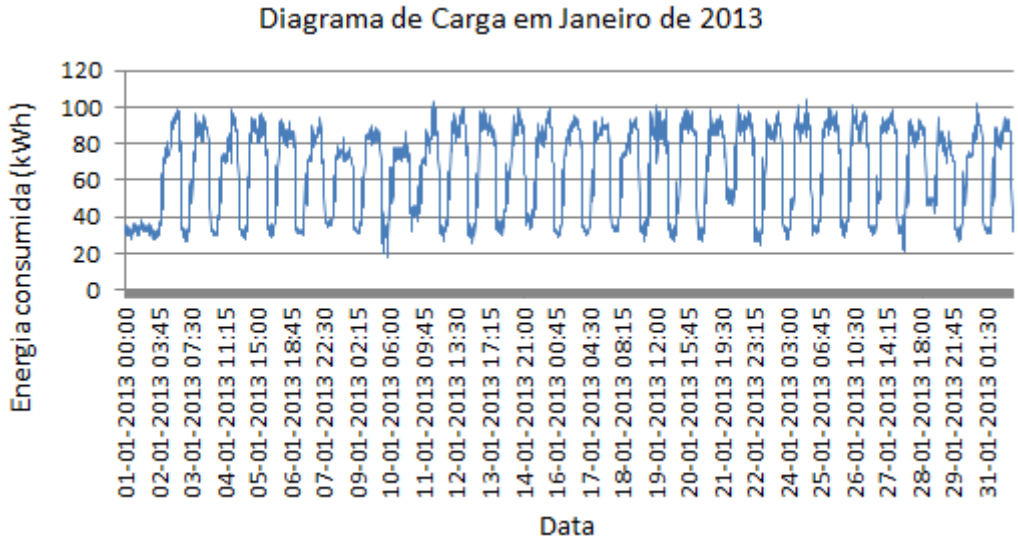

Figura 2.2: Diagrama de Carga em Janeiro de 2013 da instalação inicial

Para uma melhor compreensão dos dados, na figura 2.2 está representado o diagrama de carga no mês de Janeiro de 2013, podendo observar-se que existe um ciclo diário, ou seja, concluiu-se que nesta instalação o consumo energético não é constante ao longo do dia e é semelhante em todos os dias do mês de Janeiro com exceção do primeiro dia. Observou-

se também que o consumo energético baixo (entre 20 a 40 KWH) acontece durante a noite e o consumo energético alto (entre 80 a 100 KWH) durante o dia, o que sugere que o horário de trabalho desta instalação é durante o dia.

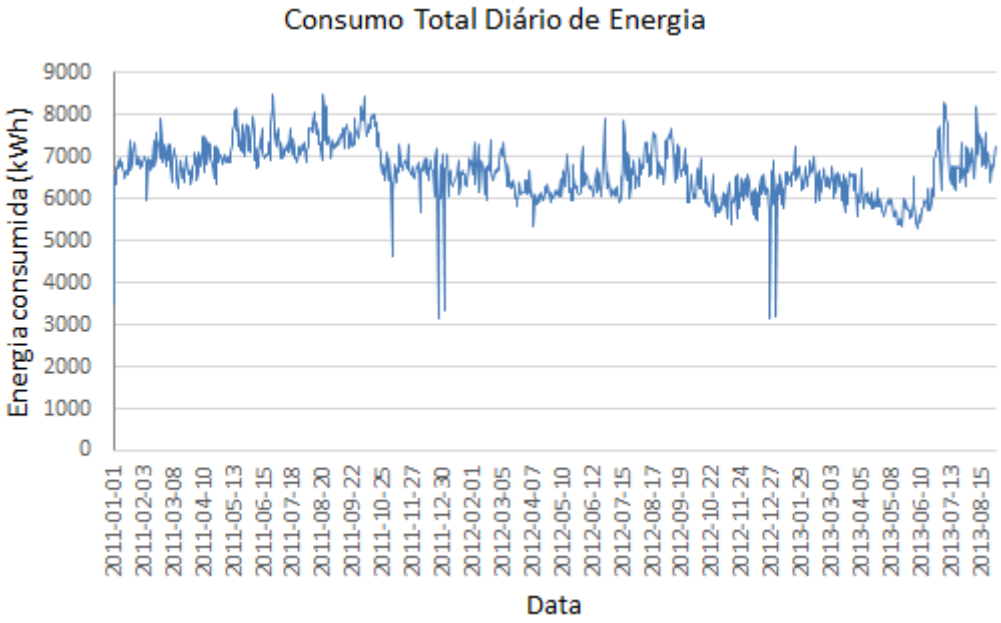

Figura 2.3: Energia total diária consumida na instalação entre 1 de Janeiro de 2011 e 31 de Agosto de 2013

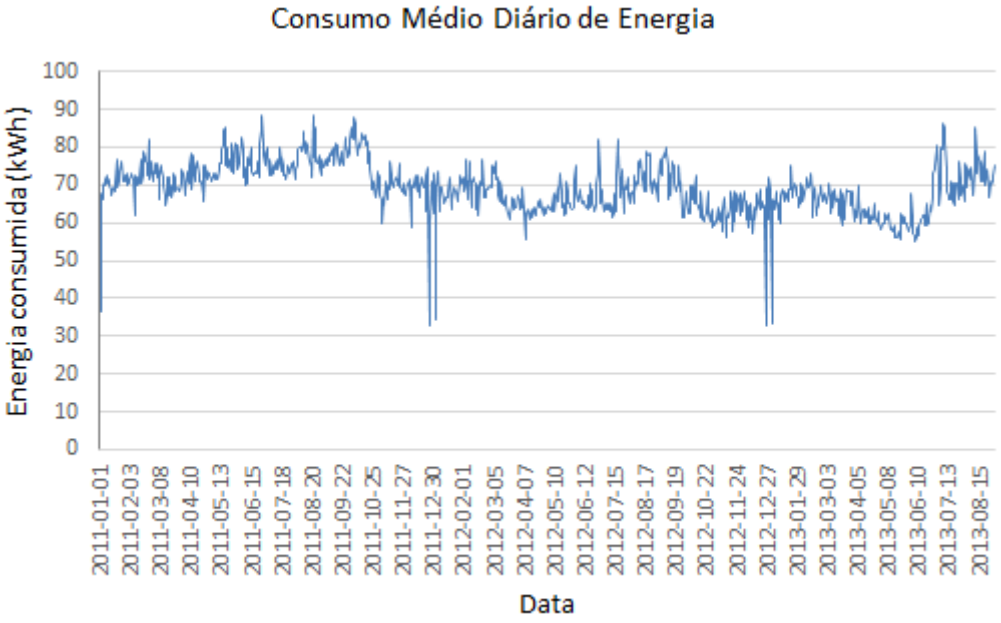

Figura 2.4: Energia média diária consumida na instalação entre 1 de Janeiro de 2011 e 31 de Agosto de 2013

Uma vez que existe demasiado detalhe, foi decidido agregar os dados por dia. Nas figuras

2.3 e 2.4 estão apresentados os consumos diários desta instalação agregados por soma e média, respetivamente, no período de 1 de Janeiro de 2011 a 31 de Agosto de 2013.
Analisando ambos os gráficos observou-se que as curvas de energia eram bastante semelhantes, a menos da escala, e que existia alguns dias em que a energia consumida era bastante inferior aos restantes. Esses dias dizem respeito aos dias de Natal (25 de Dezembro) e Ano Novo (1 de Janeiro) e o consumo sugere que esta instalação esteve fechada nesses dias. Para esta instalação, os únicos feriados que se destacaram foram os de Natal e Ano Novo, mas foi necessário ter em atenção que o consumo de outras instalações pode ser influenciado por outros feriados. Os feriados nacionais utilizados neste trabalho foram Ano Novo, Carnaval, Sexta-feira Santa, Páscoa, 25 de Abril, Dia do Trabalhador, Corpo de Deus[1], Dia de Portugal, Assunção de Nossa Senhora, Implantação da República[1], Dia de Todos os Santos[1], Restauração da Independência[1], Imaculada Conceição e Natal.
No gráfico 2.3, ao contrário do gráfico 2.4, observou-se um dia (4 de Novembro de 2011) onde a energia era também significativamente baixa em relação aos restantes dias. O dia 4 de Novembro não é um feriado e apenas no ano 2011 se verifica este baixo consumo de energia. Assim, foi analisado as observações de 15 em 15 minutos do dia 4 de Novembro de 2011 e observou-se que existiam alguns períodos em falta (20 períodos consecutivos) e por este motivo, a energia agregada pela soma nesse dia era obviamente mais baixa que a energia agregada num dia em que não existia falta de períodos.
Esta situação alertou para o facto de ser necessário verificar a incompletude dos dados, definindo critérios para exclusão de dias que não tivessem os dados completos e/ou para ajustar o consumo diário para 96 períodos (24h).

Variação Anual

Após analisar os dados em períodos de 15 minutos e diários foi analisada a variação do consumo energético médio ao longo dos 3 anos. Nas figuras 2.5 e 2.6 estão apresentados 2 gráficos referentes ao consumo energético médio anual. O gráfico 2.5 é o gráfico de barras da energia média consumida nos diferentes anos e, uma vez que no ano 2013 apenas se tinha 8 meses de observações, foi calculada a média da energia. Em 2.6 tem-se a energia média diária consumida para os diferentes anos. Observou-se que existe uma diminuição de consumo energético de ano para ano. Esta observação pode

[1] feriado suspenso no ano de 2013

dever-se à situação económica do país, ao acesso aos diagramas de carga, condições meteorológicas, etc.

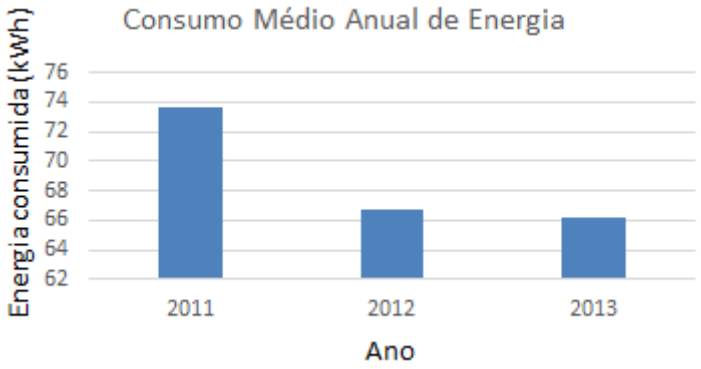

Figura 2.5: Gráfico de barras do consumo médio energético da instalação inicial por ano

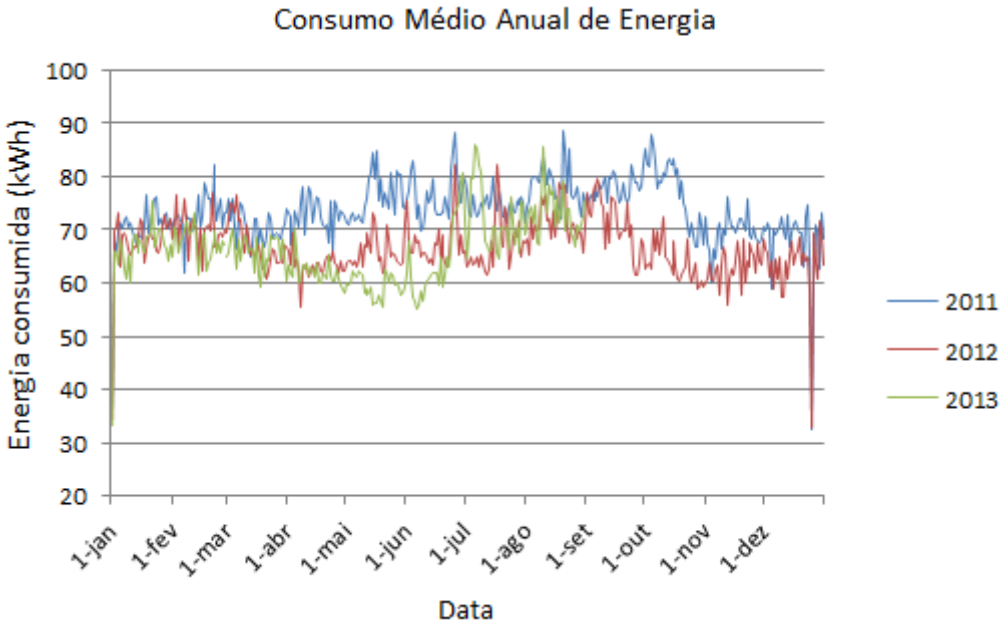

Figura 2.6: Energia média diária consumida na instalação por ano

Resta salientar que, se no gráfico 2.6 fosse usado a medida de agregação soma, a conclusão seria a mesma.

Variação Mensal

A análise que surgiu imediatamente a seguir à anual foi a análise mensal, ou seja, observar o comportamento do consumo energético médio mensal. Sendo assim, na figura 2.7 tem-se a energia média mensal consumida nos diferentes anos. Observou-se que o consumo

mensal é diferente para os 3 anos, como já foi visto, e que existe variação no consumo ao longo dos meses. Nos meses de Agosto e Setembro verificou-se um consumo maior que nos restantes e nos meses de Janeiro, Março, Abril, Maio e Dezembro (para os anos 2011 e 2012) observou-se um baixo consumo de energia. Estas observações sugeriram que as estações do ano podem influenciar o consumo de energia e, sendo assim, a próxima análise gráfica foi ao consumo energético nas diferentes estações do ano.

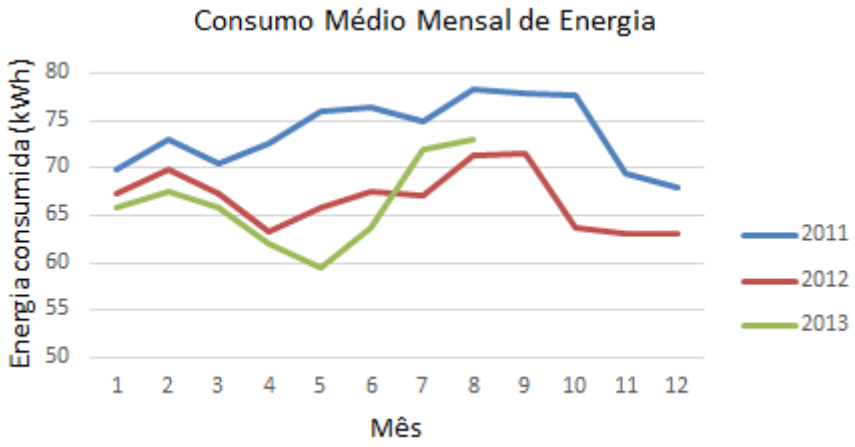

Figura 2.7: Energia média mensal consumida na instalação por ano

Variação nas Estações do Ano

Para verificar se as estações do ano influenciavam o consumo energético desta instalação, foi construído o gráfico de barras da energia média consumida nas 4 estações, podendo ser visualizado na figura 2.8. O ano foi dividido nas seguintes estações: *Inverno* - meses de Dezembro, Janeiro e Fevereiro; *Primavera* - meses de Março, Abril e Maio; *Verão* - meses de Junho, Julho e Agosto; *Outono* - meses de Setembro, Outubro e Novembro.

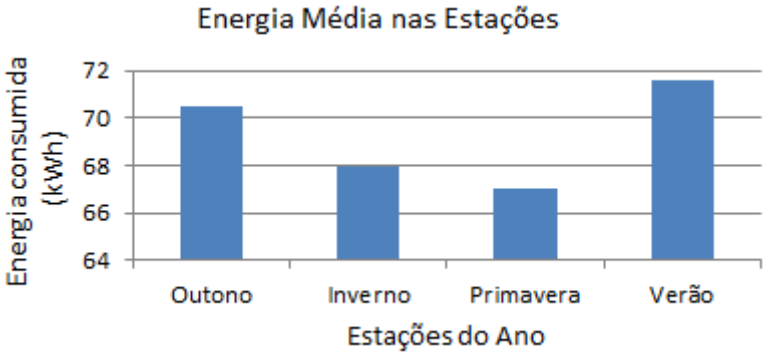

Figura 2.8: Gráfico de barras da energia média consumida na instalação inicial por estações do ano

Observou-se que existiam grandes diferenças no consumo energético segundo as estações do ano - nas estações Outono e Verão existia um maior consumo de energia do que no Inverno e Primavera. Nesta análise foi utilizada a medida de agregação média uma vez que não existia observações da estação Outono no ano 2013.

Variação Semanal e Diária

Através da figura 2.2 observou-se que, em Janeiro de 2013, o consumo de energia variava ao longo do dia, sendo mais baixo durante a noite, e que para os diferentes dias o consumo era semelhante. Para analisar a variação semanal e diária, foi necessário observar se o que foi dito anteriormente para o mês de Janeiro de 2013 acontece nos 3 anos.

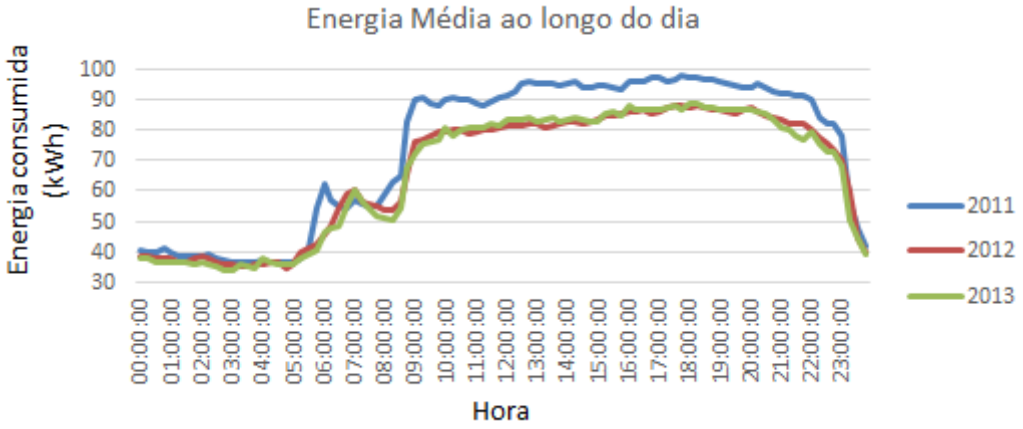

Figura 2.9: Energia média horária por ano da instalação inicial

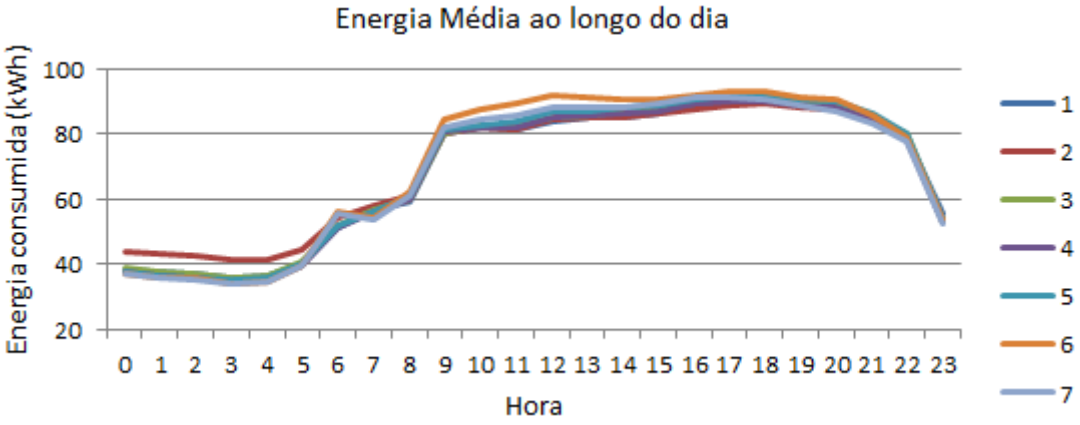

Figura 2.10: Energia média horária por dia de semana (1 - Segunda-feira) da instalação inicial

Na figura 2.9 está representado o consumo médio de energia ao longo do dia, para os 3 anos, e observou-se que para todos a variação da energia era semelhante, ou seja, o

horário de trabalho da instalação era entre as 8h e as 23h (caso onde o consumo de energia era maior). Restava verificar se este horário era semelhante para os diferentes dias da semana. Na figura 2.10 está apresentado o gráfico da energia média consumida ao longo do dia para os 7 dias da semana. Observou-se que, para todos os dias, o consumo de energia era muito semelhante, assim como o horário de trabalho da instalação. O mesmo se concluía se a medida de agregação fosse a soma.

Através desta análise foi possível perceber o horário de funcionamento da instalação nos diferentes dias de trabalho e saber se o trabalho foi semelhante ou não nos diferentes dias da semana. Sendo assim, na ferramenta R (`http://cran.r-project.org/`) criou-se uma função que devolve o horário da instalação para os 7 dias da semana e para os feriados. Para esta função os dados do consumo energético tinham que estar agregados por hora e por dia da semana (incluindo Feriado), ou seja, uma tabela de 8 linhas referentes aos 7 dias da semana mais feriados e 24 colunas referentes às horas do dia, para cada instalação. O algoritmo consistia em calcular as diferenças de consumo entre horas consecutivas de um mesmo dia da semana. Caso a diferença fosse significativamente grande (após vários testes, o valor mais adequado foi 13% da amplitude do consumo) obtinha-se o horário da instalação do dia da semana correspondente. A primeira diferença positiva corresponderia à hora inicial de trabalho e a última negativa corresponderia à hora final de trabalho. Caso não existisse diferenças significativas então o horário de funcionamento da instalação seria o dia inteiro ou a instalação não funcionava nesse dia, ou seja, a hora inicial de trabalho era igual à hora final.

2.3 Conclusão

Na secção anterior analisou-se o consumo energético de uma instalação agregando o consumo através de várias medidas e variáveis. A necessidade de agrupar o consumo surgiu pela quantidade elevada de observações, que tornou impossível a análise visual, e também pela morosidade do processamento computacional dos dados.

As medidas de agregação mais utilizadas foram a média e a soma, uma vez que o mínimo e o máximo são medidas de extremos, não sendo as mais adequadas para análise do

comportamento do consumo energético. No entanto, ao longo desta análise, observou-se que a soma não era a melhor medida a utilizar quando existia falta de observações.
Ao longo da análise gráfica foram construídas várias variáveis que podiam influenciar o consumo de energia de uma instalação:

- Ser feriado ou não (Feriado)
- Ano
- Estações do Ano (Estação)
- Dia da Semana (DiaSemana)
- Ser Fim-de-Semana ou não (FimSemana)
- Horário de trabalho (HorTrab)

A análise anterior foi realizada a mais 5 instalações e obtiveram-se as mesmas conclusões. Para algumas delas verificou-se que o horário de funcionamento era semelhante nos dias úteis, mas diferente no fim de semana e em feriados, o que aumentou a relevância das variáveis anteriores.

Capítulo 3

Análise de séries temporais

3.1 Introdução

O consumo energético diário e o diagrama de carga de uma instalação eram variáveis medidas sequencialmente ao longo do tempo. Quando uma variável é medida sequencialmente ao longo ou num intervalo do tempo, os dados resultantes formam uma *série temporal* [4][23][25]. Uma série temporal diz-se *contínua* quando as observações são feitas continuamente no tempo. Quando as observações são feitas em tempos específicos, geralmente equiespaçados, diz-se que a série temporal é *discreta* [23]. Neste contexto, como o consumo energético é medido em períodos de 15 min, o estudo deste trabalho é sobre séries temporais discretas.

As principais medidas de muitas séries temporais são a tendência e as variações sazonais que podem ser modeladas deterministicamente com funções do tempo. Em geral, uma variação sistemática no tempo que não aparenta ser periódica é conhecida como *tendência* [4][23][25]. Um padrão que tende a repetir-se a cada ano é conhecido como *variação sazonal*, embora o termo seja aplicado mais geralmente a padrões repetidos num período fixo [4].

Ao analisar-se uma ou mais séries temporais, a representação gráfica dos dados sequencialmente ao longo do tempo é fundamental. O gráfico não só revela padrões e medidas de comportamento importantes como também valores anormais ou falsos [4][23][25].

Notação: Uma série temporal discreta (observada) de tamanho n é representada por

$$\{x_t : t = 1, ..., n\} = \{x_1, x_2, ..., x_n\}$$

A notação será abreviada para $\{x_t\}$ quando o tamanho n da série não necessita de ser especificado.

Pode-se consultar preliminares e modelos básicos de séries temporais em [4] e [7].

Neste capítulo será apresentada a análise ao consumo energético analisado no capítulo anterior, mas numa perspetiva de séries temporais.

3.2 Análise gráfica

O diagrama de carga tinha registos de 1 de Janeiro de 2011 a 31 de Agosto de 2013 de 15 em 15 minutos. Uma vez que eram muitos instantes, foram analisados os dados médios diários.

Na secção 2.2 observou-se na figura 2.4 a existência de valores anormais, nomeadamente os dias de Natal e Ano Novo. No entanto, observaram-se também a existência de uma tendência, que crescia nas estações Primavera e Verão e decrescia nas restantes estações, e ainda a existência de componentes sazonais. Por exemplo, os dias de Natal e Ano Novo formavam um padrão anual, e, para além deste, observaram-se outros de período inferior.

Para este consumo, verificou-se que a variação ao longo do dia era uma variação sazonal uma vez que se repete em todos os dias (ver figura 2.10).

As séries temporais frequentemente podem ser caraterizadas pelos seus momentos de primeira ordem (média) e segunda ordem (variância), incluindo o momento cruzado (autocovariância) [7], quando estes momentos são invariantes por translações temporais, ou seja, quando x_t e $x_{t+\tau}$ têm os mesmos momentos qualquer que seja τ. Neste caso, a série diz-se *fracamente estacionária*.

A média foi calculada por segmentos, considerando-se intervalos de tempo de duração 1 mês, obtendo-se a figura 3.1. A média varia ao longo do tempo, sendo maior no Verão, estação do ano em que se gasta mais energia, eventualmente devido às condições ambientais (figura 3.1).

Para $\tau = 1$ mês, observando o gráfico da média mensal do consumo, concluiu-se que a média não se mantém constante ao longo do tempo, logo a série em causa não era estacionária.

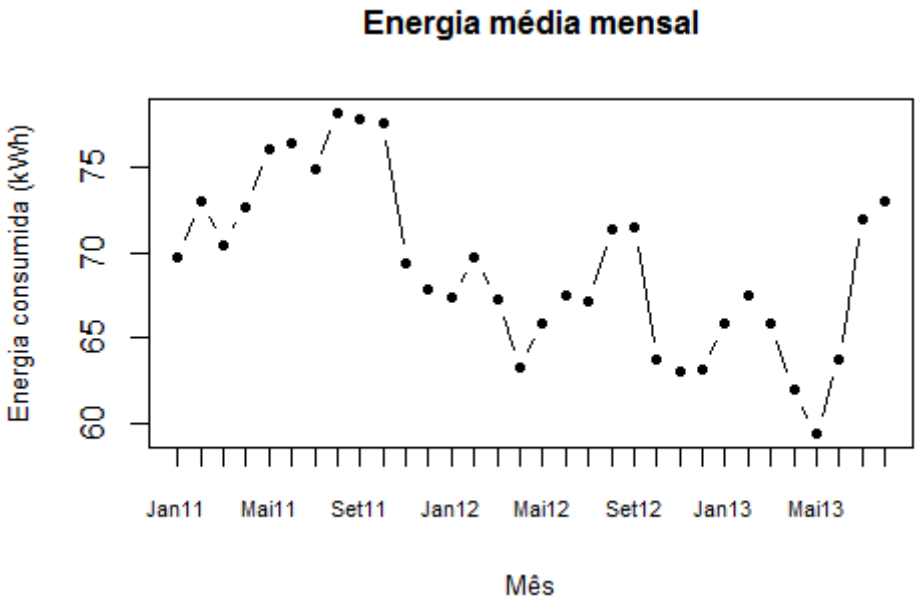

Figura 3.1: Energia média mensal consumida na instalação entre Janeiro de 2011 e Agosto de 2013

3.3 Decomposição

Alguns métodos aplicáveis no estudo de séries temporais baseiam-se na decomposição em movimentos ou forças componentes. As componentes reconhecidas são:

- Tendência;
- Sazonalidade;

a que se tem de juntar também um resíduo designado por componente errática, irregular ou aleatória.
Um modelo simples de decomposição aditiva é dado por

$$x_t = m_t + s_t + z_t$$

onde x_t é a série observada, m_t é a tendência, s_t é o efeito sazonal e z_t é o resíduo que é, em geral, uma sequência de variáveis aleatórias correlacionadas com média zero [4].
Se o efeito sazonal tende a aumentar com o aumento da tendência, um modelo multiplicativo pode ser mais apropriado [4]:

$$x_t = m_t \times s_t \times z_t$$

O modelo que melhor se ajusta varia, é claro, de série para série e, quando se aceita a ideia de decomposição, o melhor será fazer vários ensaios até chegar ao modelo que reduz ao máximo a componente residual, sem prejuízo da respetiva aleatoriedade.

3.3.1 Suavização (smoothing)

Existem vários procedimentos para estimar a tendência m_t e os efeitos sazonais s_t, no tempo t. Um método relativamente simples, disponível no `R` na função `decompose`, e que não assume nenhuma forma específica, consiste em calcular uma *média móvel* centrada em x_t. Um segundo algoritmo disponível também no `R` é `stl`. Este utiliza uma técnica de regressão ponderada, conhecida como *loess* [4]. Os detalhes de cada um dos métodos podem ser consultados em [4].
O método `decompose` estima a tendência e as variações sazonais tanto para o modelo aditivo como para o modelo multiplicativo. No entanto, o método `stl` apenas estima para o modelo aditivo.
Para o consumo energético médio diário, analisado na secção anterior, foi difícil perceber qual o modelo mais adequado, se o aditivo ou se o multiplicativo. Assim, foram estimadas a tendência e as variações sazonais pelo `stl` e `decompose`, assumindo-se que esta série era um modelo aditivo, e foram estimadas as mesmas características pelo `decompose`, assumindo-se que a série era um modelo multiplicativo.
Nas figuras 3.2 e 3.3 apresentam-se os resultados de decomposição da série, assumindo-se que esta é um modelo aditivo. Observou-se que, para ambos métodos, a tendência é decrescente e a componente sazonal contém padrões anuais e de período inferior. Porém, observou-se que no resíduo existe tendência e componentes sazonais, o que indica que a extração da tendência e da sazonalidade da série não foi bem sucedida.

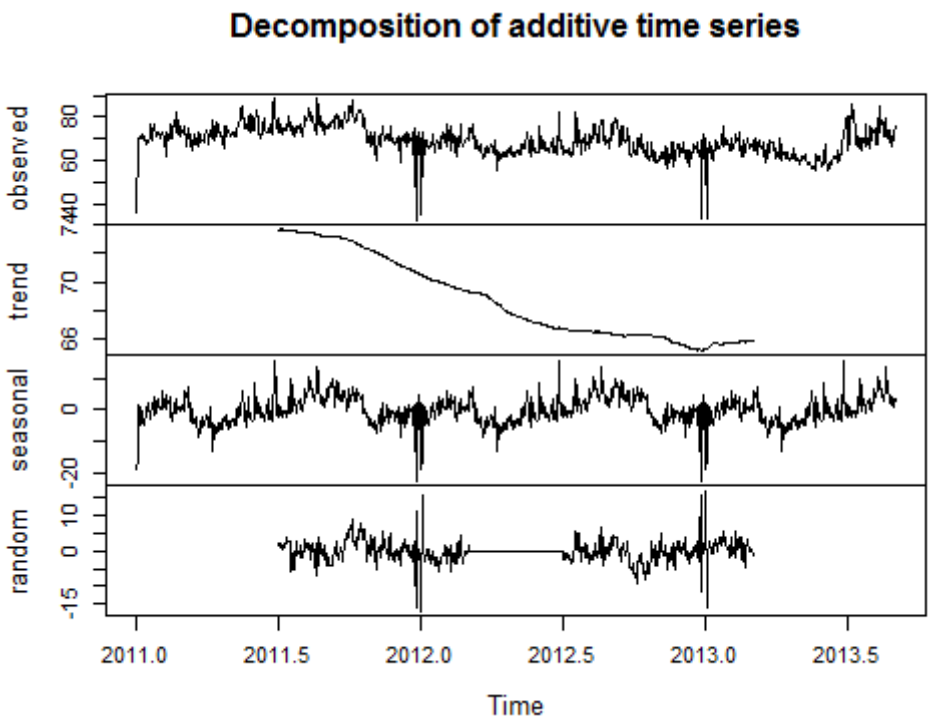

Figura 3.2: Decomposição da série, segundo um modelo aditivo, usando o método *decompose*

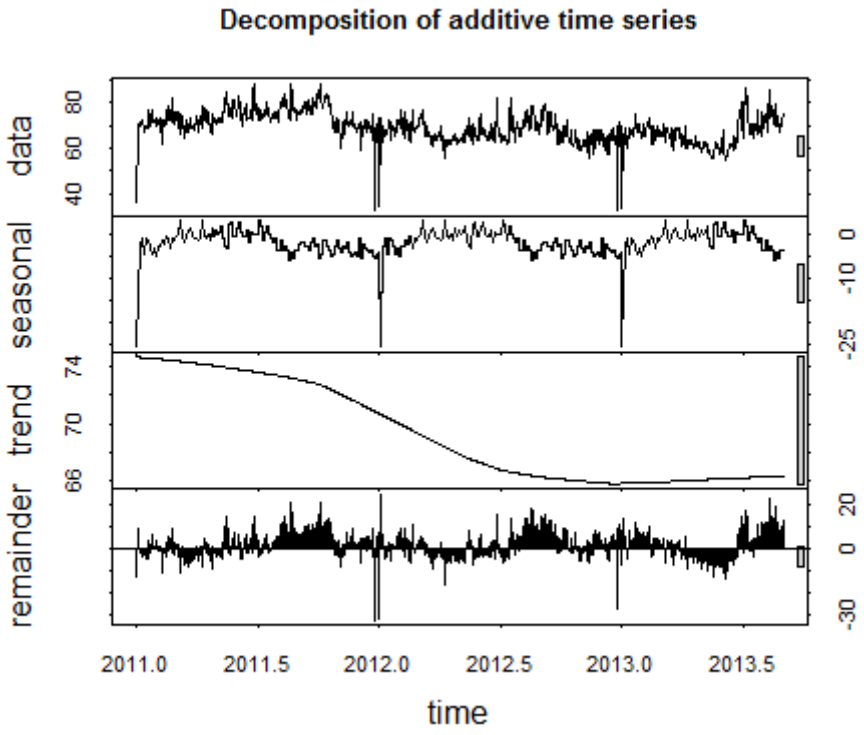

Figura 3.3: Decomposição da série, segundo um modelo aditivo, usando o método *stl*

Na figura 3.4 apresenta-se o resultado de decomposição da série assumindo-se que esta é um modelo multiplicativo. Mais uma vez, observou-se que as componentes não foram corretamente estimadas, uma vez que o resíduo apresenta sazonalidade.

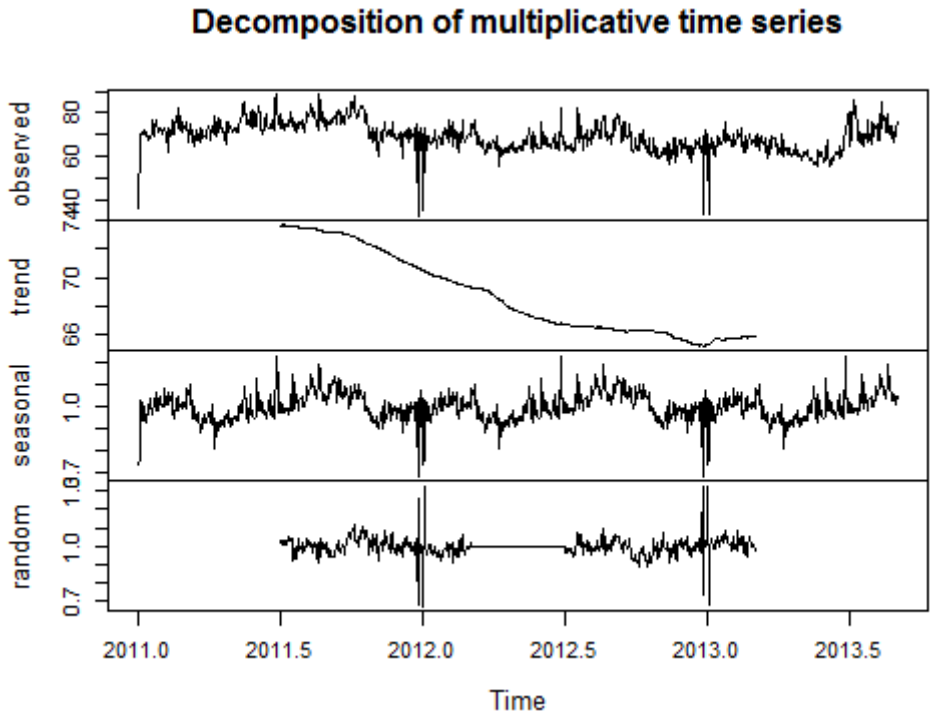

Figura 3.4: Decomposição da série, segundo um modelo multiplicativo, usando o método *decompose*

Deste modo, decidiu-se aplicar outra técnica de decomposição de séries temporais, apresentada a seguir.

3.3.2 Análise Espectral Singular (SSA)

Uma outra técnica de decomposição é a *Análise Espectral Singular* (SSA, do inglês Singular Spectrum Analysis). SSA é mencionada como uma técnica moderna e poderosa para análise de séries temporais, que dispensa o conhecimento sobre o modelo paramétrico da série e pode ser aplicada a qualquer série, simples ou complexa, pois não exige suposições estatísticas relativas à série.

De uma forma muito sumária, na técnica de SSA básica, a série temporal inicial é decomposta numa soma de poucas subséries, de modo a que cada subsérie constitui uma componente principal (padrão temporal). De seguida selecionam-se grupos de subséries

com os quais se procede à reconstrução da série temporal. Estes passos estão descritos detalhadamente no anexo A.

A dificuldade da técnica está na escolha dos parâmetros, que apenas são dois: o comprimento da janela L e o agrupamento dos triplos próprios (*eigentriple grouping*). Na subsecção A.3 do anexo A pode ser consultado o procedimento para a escolha de cada um dos parâmetros.

Como mencionado na secção 3.1, a série em análise continha variações sazonais (anuais) e tendência. Visualizou-se a existência de outros períodos, contudo, apesar de se observarem variações, não foi possível determinar-los. Para a escolha do parâmetro L é necessário saber os períodos da série.

Assim, sendo L menor que $N/2 = 487$, teve-se que escolher $L = 365$ dias $= 1$ ano. Como a série tem um aspeto complexo, foi usado $L = 7$ dias $= 1$ semana para extrair a componente reconstruída de tendência, uma vez que esta é uma curva mais suave e posteriormente foi aplicada a técnica SSA à série sem tendência com $L = 365$ (consultar secção *Informações Adicionais* do anexo A).

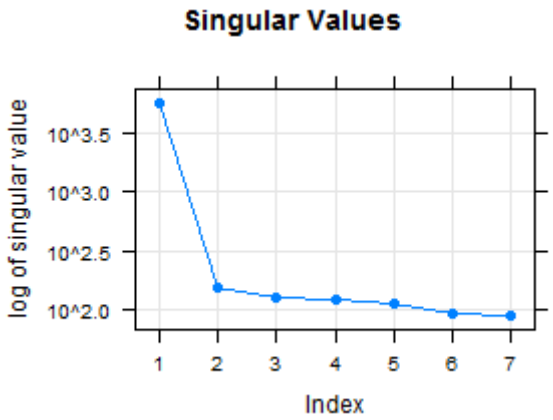

Figura 3.5: Gráfico dos valores singulares da decomposição da série com $L = 7$

Segundo [12], o primeiro valor singular normalmente está associado à componente de tendência. Observando a figura 3.5, concluiu-se que o primeiro valor estava associado à tendência, uma vez que a partir do segundo observou-se um decaimento lento e com pouca variação nos valores singulares. Assim, a partir do segundo, considerou-se que tais valores singulares estavam associados ao ruído do sinal (A.3).

Na figura 3.6 está apresentada a tendência obtida pela técnica SSA com $L = 7$ (a vermelho, em cima), sobre a série original (a preto, em cima), obtendo a série sem tendência (a verde,

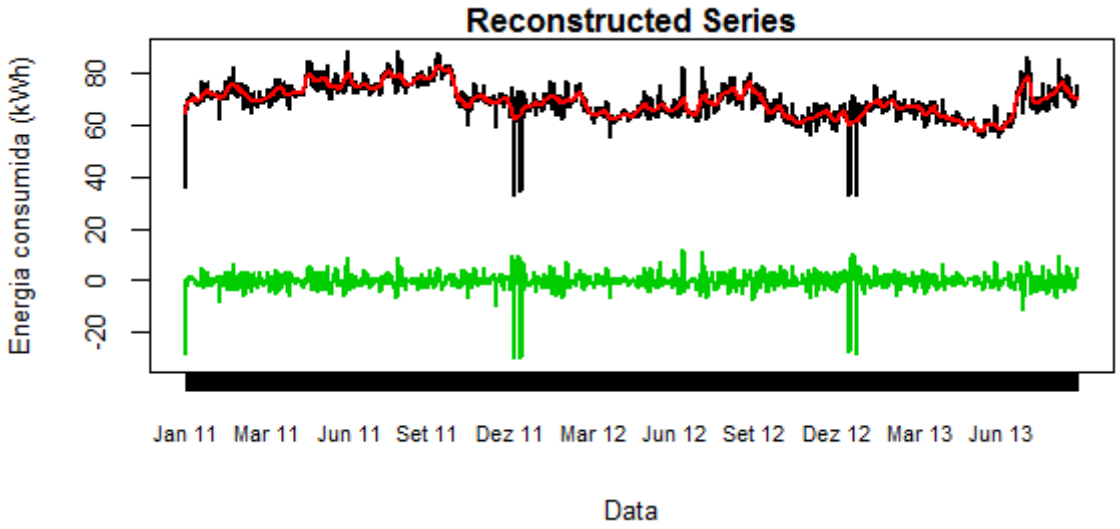

Figura 3.6: Extração da tendência da série em estudo. Em cima: Série original a preto, curva da tendência a vermelho. Em baixo: Série original sem tendência.

em baixo). Observando a série a verde, concluiu-se que a tendência foi bem calculada uma vez que os valores da série aparentam ter média 0 (ao longo do tempo).

Após extrair a tendência, decompôs-se a série resultante em componentes sazonais e ruído. Utilizando $L = 365$, obteve-se os gráficos das figuras 3.7, 3.8 e 3.9 para auxiliar na escolha dos triplos próprios (consultar A.3).

Observando a figura 3.7, concluiu-se que a partir do valor singular 17 estava-se na presença de valores associados ao ruído da série (sem tendência), uma vez que os valores inferiores ao 16 decrescem lentamente.

Dois valores próprios são denominados como um par quando estes estão suficientemente próximos e são associados a uma componente sazonal. Os primeiros 3 pares de valores singulares, observados na figura 3.7, estavam associados (cada par) a uma componente sazonal. Os próximos 2 pares levantaram algumas dúvidas: se cada par estava associado a uma componente sazonal ou se os 2 pares pertenciam a um mesmo grupo, estando associados à mesma componente. Por fim, os restantes 6 valores singulares tinham valores muito próximos, o que sugeriu que os 6 estavam num mesmo grupo.

Observando o gráfico dos vetores próprios obtidos na decomposição (à esquerda) e os "scartterplot's" de pares de vetores próprios (à direita) na figura 3.8, concluiu-se que os pares (1, 2) (3, 4) (5, 6) (7, 8) e (9, 10) estavam associados a componentes sazonais. De facto, os vetores próprios associados a cada par aparentavam ser uma componente

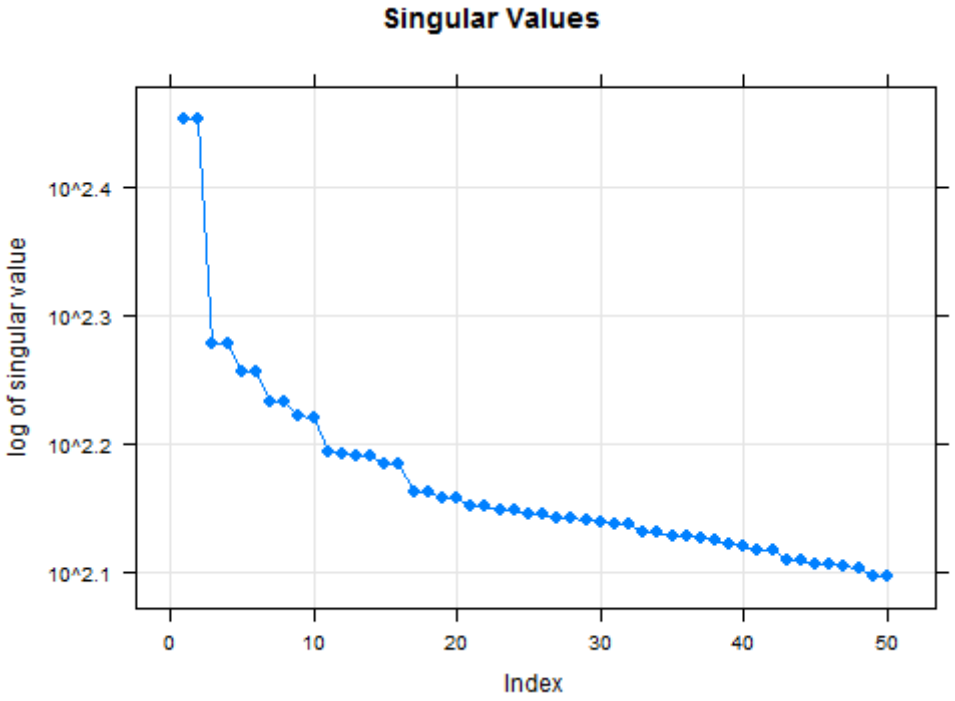

Figura 3.7: Valores singulares obtidos na decomposição da nova série (sem tendência) com $L = 365$

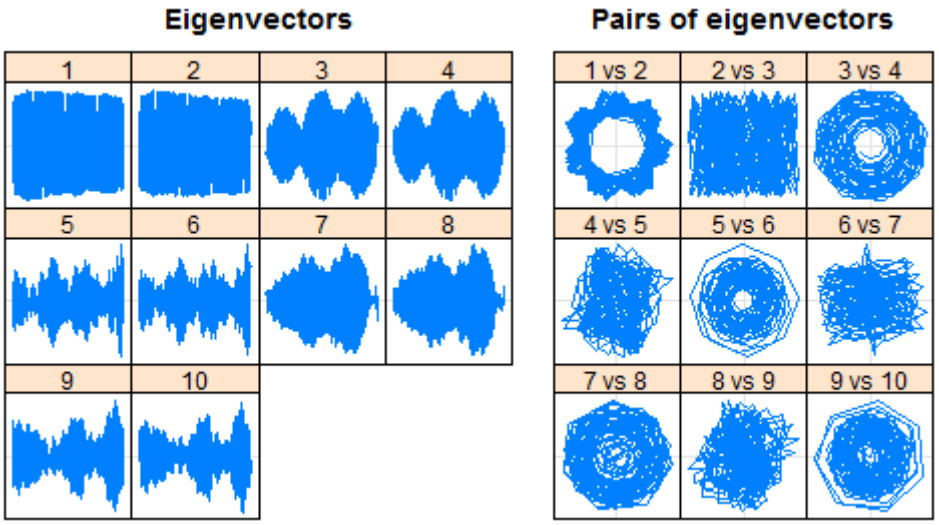

Figura 3.8: À esquerda: Gráfico dos 10 primeiros vetores próprios obtidos na decomposição da nova série (sem tendência) com $L = 365$. À direita: scatterplot's dos 9 primeiros pares de vetores singulares

sazonal e eram semelhantes entre si, e os scatterplot's aparentavam ser polígonos regulares (ver A.3). Este agrupamento coincidiu com um dos observados no gráfico dos valores singulares, contudo neste caso só foi possível analisar até ao vetor próprio 10 (característica da função `ssa()` do `R`).

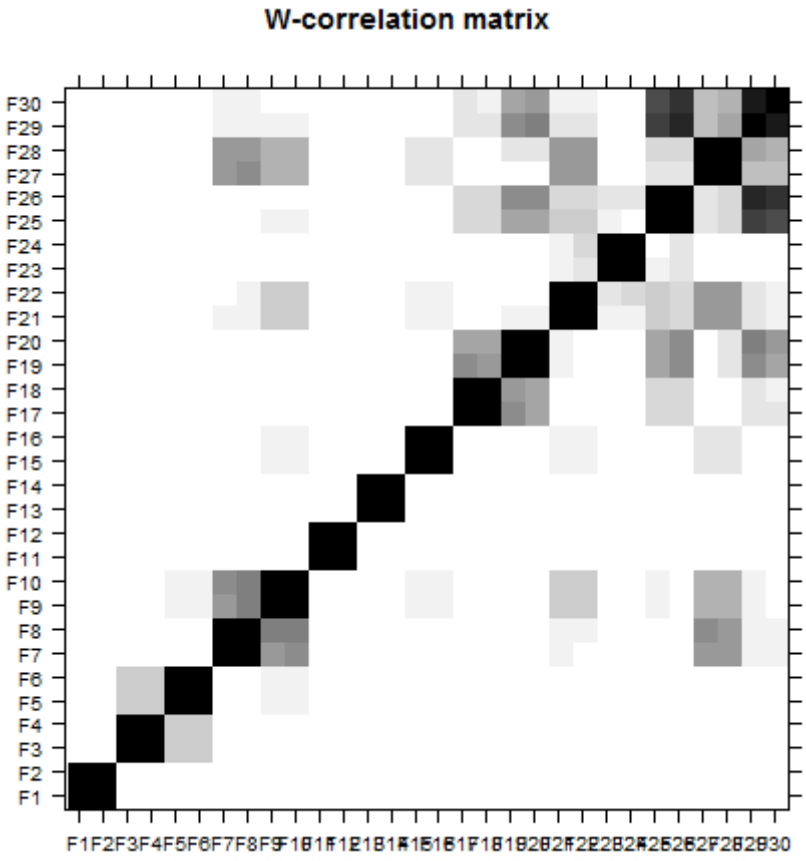

Figura 3.9: Matriz w-correlação das primeiras 30 componentes SVD resultantes da etapa decomposição da técnica SSA aplicada à nova série (sem tendência) com $L = 365$

Através da matriz w-correlação (figura 3.9) observou-se que o sinal da série (série original sem tendência) poderia ser formado pelos 16 triplos próprios principais, indicando que estes eram suficientes para reconstruir a série. Com efeito, a partir da 17ª componente observou-se a existência de mais componentes com correlações de tonalidade cinza. Observou-se que existia um grupo com 4 triplos próprios (7-10), pois tinham uma correlação forte entre eles, e as restantes componentes associadas ao sinal da série eram agrupadas por pares.

A análise anterior conduziu a alguns agrupamentos diferentes:

- (1, 2) (3, 4) (5, 6) (7, 8) (9, 10) (11, 14);
- (1, 2) (3, 4) (5, 6) (7, 8) (9, 10);
- (1, 2) (3, 4) (5, 6) (7-10) (11-14);

- (1, 2) (3, 4) (5, 6) (7-10) (11, 12) (13, 14) (15, 16).

Para perceber qual dos agrupamentos anteriores é mais adequado, reconstruiu-se as componentes para cada um e de seguida analisou-se as componentes sazonais e os resíduos. As componentes de sazonalidade obtidas não suscitaram qualquer interpretabilidade no contexto do problema, não auxiliando na escolha do melhor agrupamento. Os resíduos deviam apresentar uma variação constante e média 0. Ora, analisando os resíduos, figura 3.10, observou-se que existia ainda a sazonalidade anual observada inicialmente na série original.

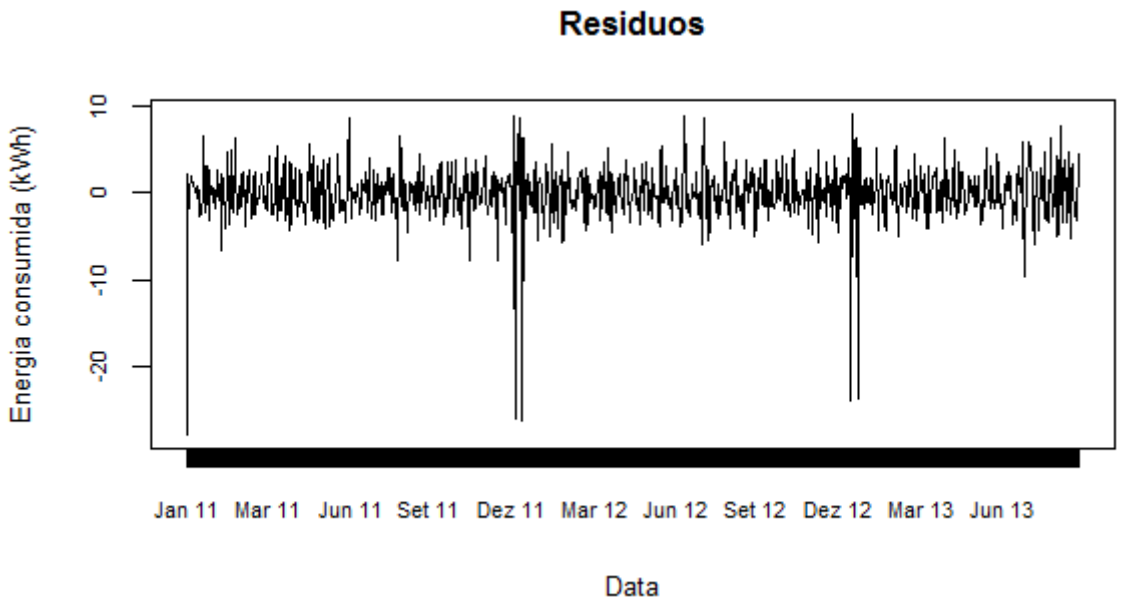

Figura 3.10: Resíduos obtidos após decompor a série em tendência e componentes sazonais usando o primeiro agrupamento: (1, 2) (3, 4) (5, 6) (7, 8) (9, 10) (11, 14)

Obteve-se a mesma conclusão para as restantes opções de agrupamento.

Concluiu-se que a tendência da série em questão foi bem estimada, contudo a sazonalidade principal (a anual, visualizada graficamente) não. Estes resultados podem ter sido obtidos por uma má escolha dos parâmetros do método ou pela falta de dados. Uma vez que se pretendeu estimar uma componente sazonal anual, talvez fosse necessário dados de maior quantidade (mais anos).

O procedimento de agrupamento de triplos próprios pode ser considerado um ponto crítico da análise via SSA, pois por ser um parâmetro estrutura, a sua definição baseia-se em poucos critérios objetivos.

Com o objetivo de resolver este problema e uma vez que se conhecia os dias em que o consumo de energia era anormal, referentes à sazonalidade anual, decidiu-se substituir tais consumos pela média e aplicar a técnica SSA. Observou-se mais uma vez que os resíduos apresentam sazonalidade para todas as opções de agrupamento mais adequadas e que a tendência era influenciada pelos valores anormais de consumo.

3.4 Regressão linear múltipla

Um modelo de regressão linear modela a relação linear entre uma variável continua Y e um conjunto de variáveis (de qualquer natureza) $X_1, ..., X_p$. A Y dá-se o nome de *variável resposta ou dependente* e às variáveis $X_1, ..., X_p$ dá-se o nome de *variáveis explicativas ou independentes*.

Os objetivos desta modelação são descrever a estrutura geral dos dados, avaliar o efeito das variáveis $X_1, ..., X_p$ sobre Y e prever observações futuras.

Neste trabalho o interesse foi utilizar a regressão linear para descrever a estrutura geral dos dados e avaliar o efeito das variáveis explicativas sobre o consumo energético diário.

Para mais detalhes sobre regressão linear múltipla consultar [31][6]. Do capítulo 2 chegou-se à conclusão que as variáveis Ano, Estação, Dia da Semana, Fim de Semana, Feriado e Horário de Trabalho explicavam parte do consumo energético. Uma vez que as estações do ano influenciavam a energia média consumida, a temperatura e outras variáveis climáticas podiam ser variáveis explicativas significativas do consumo. Assim, utilizaram-se diversas variáveis climáticas, como por exemplo, temperatura máxima, temperatura mínima, humidade máxima, etc (extraídas do site *Weather Underground* [34] para a cidade de Lisboa, localização da instalação). A análise detalhada deste conjunto de variáveis pode ser consultada no anexo B e no capítulo 4.

Os coeficientes da regressão linear múltipla, utilizando todas as variáveis anteriores como explicativas, foram estimados através do comando `lm()` do `R`. Posteriormente, foi utilizado o comando `step()` (com método *stepwise*) para obter as variáveis que melhor explicam o consumo energético, ou seja, as variáveis estatisticamente significativas e o comando `summary()` para obter as estatísticas associadas [31].

Foram usadas várias medidas de agregação do diagrama de carga por dia: soma, média e

mediana. No Excel não foi possível agregar os dados por mediana, em tabelas dinâmicas, então foi necessário construir em R uma função para agregar os diagramas de carga através da mediana. Para estes consumos, não foi possível utilizar a variável Horário de Trabalho, uma vez que esta variável não era diária.

Através de testes de hipóteses sobre os coeficientes de regressão foi possível encontrar as variáveis significativas ($valor - p$ menor que 0.05) para explicar o consumo. Foi também possível quantificar o ajustamento do modelo aos dados, através do coeficiente de determinação ajustado, R_a^2, uma medida de qualidade do ajustamento do modelo aos dados [31].

```
> summary(modelo)

Call:
lm(formula = Media.Energia ~ Ano + Feriado + Estacao + DiaSemana +
    Compr_Dia + Temp_Max + Temp_Min + Pt_Orvalho_Medio + Pt_Orvalho_Min +
    Humidade_Media + Vel_Vento_Max + Vel_Vento_Media + Dir_Vento_Graus,
    data = data[, -1])

Residuals:
    Min      1Q  Median      3Q     Max
-35.281  -2.463   0.015   2.560  15.009

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      69.323858   5.084551  13.634  < 2e-16 ***
Ano2012          -6.009127   0.350745 -17.132  < 2e-16 ***
Ano2013          -6.519265   0.406325 -16.044  < 2e-16 ***
Feriadoutil       4.683880   0.765621   6.118 1.38e-09 ***
EstacaoOutono    -3.277679   0.553699  -5.920 4.50e-09 ***
EstacaoPrimavera -2.864102   0.689945  -4.151 3.60e-05 ***
EstacaoVerÆo     -2.047702   0.880500  -2.326 0.020249 *
DiaSemana2        1.235847   0.538407   2.295 0.021928 *
DiaSemana3        0.495609   0.537870   0.921 0.357060
DiaSemana4       -0.038304   0.538646  -0.071 0.943324
DiaSemana5        0.936401   0.537740   1.741 0.081943 .
DiaSemana6        2.224427   0.536180   4.149 3.64e-05 ***
DiaSemana7        0.685810   0.537575   1.276 0.202356
Compr_Dia        -0.493742   0.209449  -2.357 0.018608 *
Temp_Max          0.297754   0.094277   3.158 0.001637 **
Temp_Min          0.229952   0.111300   2.066 0.039093 *
Pt_Orvalho_Medio  0.400070   0.207175   1.931 0.053771 .
Pt_Orvalho_Min   -0.227451   0.101530  -2.240 0.025306 *
Humidade_Media   -0.082288   0.047639  -1.727 0.084434 .
Vel_Vento_Max     0.173695   0.033337   5.210 2.31e-07 ***
Vel_Vento_Media  -0.199013   0.046581  -4.272 2.13e-05 ***
Dir_Vento_Graus  -0.004579   0.001371  -3.339 0.000873 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.472 on 952 degrees of freedom
Multiple R-squared:  0.5607, Adjusted R-squared:  0.551
F-statistic: 57.85 on 21 and 952 DF,  p-value: < 2.2e-16
```

Figura 3.11: Sumário com as estatísticas associadas da regressão linear múltipla aplicada ao consumo médio diário

Os resultados da regressão linear múltipla para as medidas anteriores foram muito se-

melhantes. Na figura 3.11 podem ser visualizados os resultados para a medida média. As variáveis estatisticamente significativas foram: *Ano*, *Feriado*, *Estação*, *Dia da Semana*, *Comprimento do Dia*, *Temperatura Máx.*, *Temperatura Mín.*, *Ponto de Orvalho Mín.*, *Velocidade do Vento Máx.*, *Velocidade do Vento Média* e *Direção do Vento* e o coeficiente de determinação ajustado, R_a^2, foi, aproximadamente, 0.55, ou seja, o modelo explicava cerca de 55% da variância dos dados.

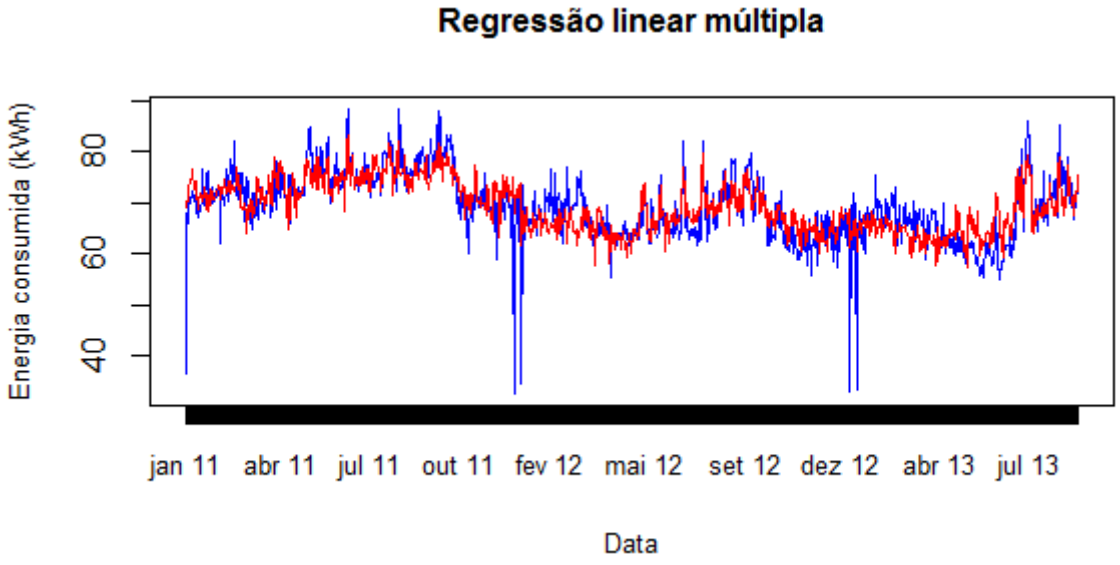

Figura 3.12: Resultados da regressão linear múltipla aplicada ao consumo médio diário. A azul: Consumo energético médio diário da serie original; A vermelho: Consumo energético estimado pela regressão

Para uma melhor compreensão do ajustamento do modelo aos dados, na figura 3.12 pode-se observar o consumo energético da série temporal (a azul) e o consumo energético estimado pela regressão linear múltipla (a vermelho). Observou-se que a regressão não estima os valores anormais da série temporal.

Utilizou-se também o diagrama de carga como variável resposta (neste caso, acrescenta-se a variável Horário de Trabalho nas variáveis explicativas). Quase todas as variáveis foram significativas e o R_a^2 foi 61%. O coeficiente foi melhor do que usando os dados diários, contudo este processo demorou bastante tempo, não sendo viável, computacionalmente, aplicar a todas as instalações (usando os dados agregados demorou cerca de 3 segundos enquanto que usando os diagramas de carga demorou 5 minutos).

3.5 Conclusão

Nas secções anteriores concluiu-se:

- A série temporal era não estacionária;
- Através dos métodos *Smoothing*, a tendência e a sazonalidade não foram estimadas corretamente;
- Através da técnica *SSA*, a tendência foi estimada corretamente, contudo a sazonalidade não;
- Usando *Regressão Linear Múltipla* observou-se quais as variáveis explicativas significativas e quanto o modelo se ajustava aos dados.

As metodologias anteriores foram realizadas a mais 5 instalações e obtiveram-se as mesmas conclusões, excepto na regressão linear múltipla. Observou-se que as variáveis explicativas significativas variavam de série para série, assim como o modelo se ajustava aos dados. No entanto, as variáveis categóricas *Ano*, *Feriado*, *Dia da Semana* e *Estação* foram significativas em todas as instalações.
Ao longo do trabalho, deparou-se com a existência de duas divisões de estações diferentes [5], *estações meteorológicas* e *estações astronómicas*. No capítulo 2, as estações usadas foram as meteorológicas. As estações astronómicas foram definidas como: *Inverno* - de 21 de Dezembro até 20 de Março; *Primavera* - de 21 de Março até 20 de Junho; *Verão* - de 21 de Junho até 20 de Setembro; *Outono* - de 21 de Setembro até 20 de Dezembro.
Decidiu-se comparar as duas divisões através do coeficiente de determinação ajustado. Obtiveram-se melhores resultados, para as 5 instalações, para as estações astronómicas. Desta forma, nos próximos capítulos quando se refere à variável *Estação*, é a astronómica. Resta salientar, que para cada uma das 5 instalações utilizaram-se as observações climáticas da localização da instalação.

Capítulo 4

Seleção de medidas, variáveis e características

4.1 Introdução

No capítulo 2 foi discutido que os dados teriam que ser agregados, uma vez que o problema foi constituído por uma grande quantidade de instalações e, para cada uma delas, existia uma grande quantidade de observações. À medida que a análise gráfica foi realizada, foi considerado um conjunto de variáveis explicativas do consumo energético e foi observada a existência de valores de consumo energético anormais (os dias de Natal e Ano Novo). Na secção 3.4 foi decidido utilizar um conjunto de variáveis climáticas para explicar o consumo energético, mas observou-se que várias não foram estatisticamente significativas.
Deste modo, nas próximas secções apresentam-se as discussões dos seguintes problemas: qual a medida de agregação a utilizar; quais as variáveis explicativas mais importantes para este trabalho; como detetar os valores anormais de consumo energético.
A análise realizada nos capítulos 2 e 3 incidiu apenas sobre uma amostra de 6 instalações. Nesta fase utilizou-se uma amostra de 97 instalações (chamada de Lote 1) onde a variabilidade dos consumos era garantida.

4.2 Medidas de agregação

A escolha da medida de agregação mais adequada foi uma das questões mais importantes deste trabalho. Caso a escolha não fosse a mais adequada, podia-se estar a fazer um estudo sobre dados enviesados. Sendo assim, foi necessário ter atenção às observações em falta e aos valores extremos, uma vez que podiam afetar o valor real da soma e da média dos dados, respetivamente.

As medidas de agregação testadas nos capítulos anteriores foram: média, soma, máximo e mínimo.

Como foi referido em 2.3, os dados diários deviam ser agregados pela soma ou média, uma vez que o objetivo foi analisar o comportamento tipico do consumo energético. Sendo assim, para toda a análise que necessitou dos dados diários utilizou-se a média ou a soma como medidas de agregação.

Para a construção da variável *Horário de Trabalho* foi necessário agregar os dados por hora e dia da semana. Testando com as medidas enumeradas anteriormente chegou-se à conclusão que os resultados obtidos pela média e soma foram bastante semelhantes e visualizou-se facilmente o horário de trabalho; agregando por máximo não se visualizaram variações significativas no consumo energético sendo mais difícil detetar o horário de trabalho; agregando por mínimo foi impossível detetar o horário de trabalho uma vez que bastava haver um dia ou horas em que a instalação não funcionou, ou houve falha de energia, ou falha nas comunicações para o horário da instalação ser admitido como constante (hora inicial igual à hora final).

Por estas razões, foram usadas as medidas de agregação *soma* e *média*, tendo em conta que a soma é sensível às falhas de dados e a média à existência de outliers. Quando a medida soma foi utilizada, caso existissem falhas de períodos de 15 minutos ao longo de algum dia, usou-se o critério de seleção de dias, e caso o dia fosse selecionado, o consumo energético diário era ajustado para 96 períodos (24h).

4.3 Variáveis Externas

4.3.1 Análise dos dados

O período de tempo máximo de observações foi de 1 de Janeiro de 2010 a 31 de Agosto de 2013. Numa análise inicial às observações climáticas para a cidade de Lisboa, extraídas do site *Weather Underground*, no período temporal acima referido, deparou-se com a existência de falhas (ou seja, valores desconhecidos) e algumas variáveis incoerentes. A análise descritiva de cada variável pode ser consultada no anexo B.

A variável *Precipitação* contém informação da percentagem de precipitação diária, enquanto que a variável *Eventos* regista a ocorrência de Chuva, Trovoada, Nevoeiro, etc. No entanto, em alguns casos quando *Eventos* toma valor Chuva a *Precipitação* é 0. Como o valor da *Precipitação* não foi compreensível, decidiu-se retirar esta variável e utilizar apenas a variável *Eventos*, uma vez que esta contém informação sobre outras ocorrências climáticas.

A variável *Visibilidade Máxima* tem valor constante igual a 10 (à exceção dos valores desconhecidos) e, deste modo, não contribuiu com qualquer informação adicional, tendo sido retirada do conjunto de dados.

A existência de valores desconhecidos no conjunto de variáveis climáticas é discutido a seguir:

Tratamento de falhas

Segundo [29], [28], [14] e [18] existem várias estratégias para lidar com os valores desconhecidos (`NA`), as mais comuns são:

- Remover os casos ou variáveis com valores desconhecidos;
- Preencher os valores desconhecidos com o valor mais frequente da variável em questão;
- Preencher os valores desconhecidos explorando semelhanças entre casos (vizinhos mais próximos).

Neste contexto, a primeira opção apenas foi aplicada a variáveis com demasiados valores em falta. Em relação aos casos (dias) com algumas falhas, para não perder informações

climáticas desses dias, foi necessário preencher as falhas. Segundo [14], preencher os valores desconhecidos pode enviesar os dados, mas, no entanto, esta última abordagem é uma estratégia comum. Em comparação com outros métodos, este usa o máximo de informação a partir dos dados para prever valores em falta [29][14]. Uma vez que as variáveis climáticas dependem da estação do ano, a opção mais adequada foi preencher os valores desconhecidos explorando semelhanças entre casos. Para detalhes sobre este método, consultar anexo B.

Ao analisar os dados das variáveis climáticas observou-se que a variável *Velocidade Máxima de Rajada de Vento* tem uma grande quantidade de `NA's` (cerca de 88%). Sendo assim, usou-se a primeira abordagem eliminando esta variável do conjunto. As restantes variáveis com `NA's` foram tratadas no `R` através da função `knnImputation()`.

Jönsson e Wohlin [18] sugerem que um valor adequado para o número k de vizinhos mais próximos é, aproximadamente, a raiz quadrada do número de casos completos nos dados. Neste caso existiam 1176 casos completos, logo $k \approx \sqrt{1176} \approx 34$.

Uma vez que existiam variáveis climáticas com valores inteiros, usou-se a mediana, medida mais robusta, com valor de k ímpar. Deste modo, foi usada a mediana dos 33 vizinhos mais próximos ao caso com `NA`, para lhe atribuir um valor.

4.3.2 Seleção de regiões de Portugal Continental

No capítulo 3 foram usadas as variáveis climáticas para explicar o consumo de energia média diária de um pequeno conjunto de instalações (em Regressão Linear Múltipla). Para todo o conjunto foram usadas as observações climáticas de Lisboa, embora nem todas as instalações sejam de Lisboa ou arredores - existe uma instalação de Beja e outra do Algarve. Surgiu então as questões: *É necessário utilizar observações climáticas de regiões diferentes do País, de acordo com a localização da instalação? Ou seja, a influência das observações climáticas sobre o consumo de energia diferem por região do País? Que regiões utilizar?*

Na maior parte de Portugal continental o clima é temperado [17]. No entanto, verifica-se que no interior do País o clima é mais seco e quente que no litoral e que o norte e centro do País são mais frios do que o sul. Deste modo, esperou-se que a influência sobre o consumo energético, usando observações climáticas de regiões distantes, fosse diferente.

Nesta fase apenas se tinha uma medida para quantificar a influência de um conjunto de variáveis sobre a variável resposta, o coeficiente de determinação ajustado R_a^2 (quanto maior, melhor explica a variação da variável resposta). Utilizou-se o coeficiente de determinação ajustado com o objetivo de verificar influências diferentes no consumo de energia médio diário para observações climáticas de regiões diferentes. Para o conjunto de instalações inicial, realizou-se a regressão linear múltipla com diferentes observações climáticas para se obter o R_a^2 (usaram-se observações climáticas da localização real da instalação e de Lisboa e Porto). Os resultados podem ser observados na tabela 4.1.

Instalação	Localização	Local real	Lisboa	Porto
A	Lisboa	0.60	0.60	0.57
B	Lisboa	0.77	0.77	0.76
C	Lisboa	0.33	0.33	0.38
D	Lisboa	0.72	0.72	0.71
E	Beja	0.87	0.85	0.83
F	Algarve	0.92	0.93	0.91

Tabela 4.1: Coeficiente de determinação ajustado para o conjunto inicial de instalações com diferentes observações climáticas

Para as 4 primeiras instalações na tabela 4.1 a localização real coincidiu com Lisboa logo os valores do coeficiente de determinação ajustado foram comuns nas duas colunas (Local real e Lisboa). Comparando os valores obtidos utilizando observações climáticas da Localização real, Lisboa e Porto em todas as instalações, observou-se que são bastante semelhantes, sendo superiores na Localização real à exceção das instalações C e F.

O facto de existirem instalações em que o melhor coeficiente de determinação ajustado não se obteve usando as observações climáticas da localização real da instalação, indica que esta medida não foi a mais adequada para esta análise. Esperava-se, pelo senso comum, que as observações climáticas da localização real se ajustassem melhor ao consumo energético, que as observações de regiões bastante distantes da real. Esperava-se também, pelo clima de Portugal continental, que existisse uma maior diferença entre os valores do R_a^2, quando usadas observações climáticas de regiões bastante distantes no mapa.

Sendo assim, a análise anterior não foi a mais apropriada para analisar a influência de

observações climáticas de regiões distantes.

Decidiu-se pesquisar a existência de divisões climatéricas fundamentadas de Portugal continental para se poder utilizar neste contexto. Apenas se encontrou uma divisão, a Classificação de Köppen-Geiger [17], figura 4.1, baseada na temperatura do ar e na precipitação [36]. Segundo esta classificação, Portugal continental tem 2 regiões climatéricas [17]. Como neste estágio se usaram mais variáveis climáticas, tentou-se confirmar a classificação de Köppen usando todas as variáveis, testando se observações de locais distantes, mas na mesma região, são semelhantes (através do teste de hipóteses t [32]), no entanto obtiveram-se resultados inconclusivos, ou seja, para algumas variáveis obteve-se que as curvas são semelhantes podendo pertencer a uma mesma região e para outras obteve-se a conclusão contrária.

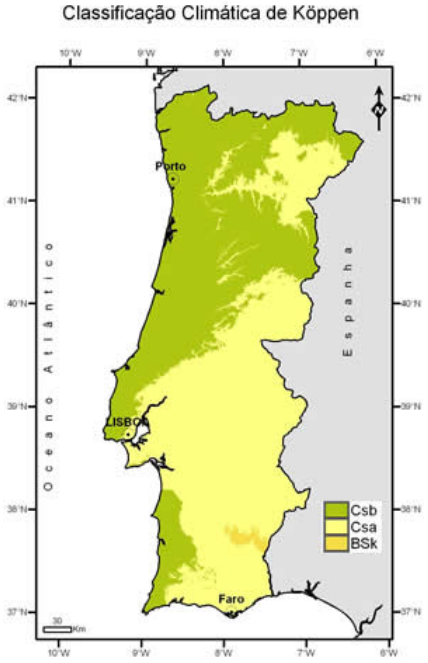

Figura 4.1: Classificação de Köppen-Geiger para Portugal continental [17]

Decidiu-se também verificar para quais distritos de Portugal existem informações climáticas no site de meteorologia *Weather Underground*, caso fosse necessário utilizar os distritos de Portugal continental como divisão em regiões. Verificou-se que existem observações para Bragança, Ovar/Aveiro, Faro, Monte Real/Leiria, Beja, Castelo Branco, Penhas Douradas/Guarda, Montijo/Setúbal, Lisboa, Sintra/Lisboa e Alverca/Lisboa.

Assim, com base numa análise crítica subjetiva e existindo um conhecimento generalizado

das diferenças climatéricas entre as regiões Norte, Centro, Sul e as respetivas regiões litorais e interiores, num debate de grupo decidiu-se a divisão do país em 6 regiões:

- Região do Porto, que contém os distritos de Viana do Castelo, Porto e Aveiro;
- Região de Bragança, que contém os distritos de Vila Real, Bragança e Guarda;
- Região de Lisboa, que contém os distritos de Coimbra, Leiria e Lisboa;
- Região de Castelo Branco, que contém os distritos de Castelo Branco, Santarém e Portalegre
- Região de Faro, que contém os distritos de Setúbal, Algarve e Litoral de Beja (concelho Odemira);
- Região de Beja, que contém os distritos de Évora e Interior de Beja (todos os concelhos de Beja exceto Odemira).

Concluindo, foram construídas tabelas de variáveis climáticas para cada uma das regiões acima. A variável comprimento do dia foi extraída do site *Sun or Moon Rise/Set Table for One Year* [27], sendo necessário colocar as coordenadas geográficas das regiões, consultadas em [1].

4.3.3 Seleção de variáveis

Após a definição de regiões de Portugal continental a usar, foi necessário proceder ao tratamento de falhas das tabelas, como em 4.3.1.
Para a variável *Direção do Vento em graus*, as observações de Bragança e Castelo Branco são constantes, então decidiu-se retirar esta variável do conjunto de dados por não fornecer informação relevante. As observações da variável *Cobertura de Nuvens* na região de Bragança contêm 88% de falhas. Mais uma vez, devido à grande quantidade de valores desconhecidos, foi decidido retirar esta variável do conjunto de dados.
Para as restantes falhas utilizaram-se os vizinhos mais próximos a um caso com falhas para preencher esse valor. Decidiu-se usar o mesmo número k de vizinhos mais próximos ao caso com falha para todas as regiões, sendo necessário verificar qual o número mínimo

de casos completos. As observações climáticas de Bragança são as que contêm o menor número de casos completos, 869, ou seja, $k \approx \sqrt{869} \approx 29$. Deste modo, os valores desconhecidos nas observações climáticas das seis regiões foram preenchidos usando a mediana dos seus 29 vizinhos mais próximos.

Como referido, as variáveis construídas no capítulo 2 e as variáveis climáticas explicavam parte do consumo energético. Contudo, tinha-se 23 variáveis explicativas e viu-se no capítulo 3 que algumas das variáveis climáticas não são estatisticamente significativas para explicar o consumo energético. Uma vez que as variáveis estatisticamente significativas obtidas pela regressão linear múltipla variam consoante a instalação é necessário realizar outras abordagens:

4.3.3.1 Backward - Regressão Linear Múltipla

No capítulo 3 usou-se a Regressão Linear Múltipla para analisar a influência das variáveis externas sobre o consumo energético. Observou-se que, para uma instalação, nem todas as variáveis são estatisticamente significativas ($valor - p \leq 0.05$) para explicar o consumo energético. Foi pois necessário restringir o conjunto de variáveis externas às variáveis que mais explicam o consumo. No entanto, utilizando uma outra instalação, as variáveis mais significativas foram diferentes das obtidas anteriormente.

Na regressão linear múltipla existem métodos de seleção de variáveis: *Forward*, *Backward* e *Stepwise* (ver [6]). Estes métodos selecionam as variáveis que evidenciam os efeitos mais fortes de diferentes formas. Neste trabalho usou-se o método Backward, uma vez que lida com o problema da multicolinearidade de uma forma melhor [6].

Assim, decidiu-se utilizar o Lote 1 como amostra e obter as variáveis selecionadas pelo *backward* para cada instalação. As variáveis mais significativas para o conjunto de instalações foram as que ocorrem mais vezes. Deste modo, na tabela 4.2 estão apresentadas as frequências relativas para cada variável selecionada. Observando a tabela 4.2 concluiu-se que as variáveis mais significativas ($\geq 50\%$) foram: *Estação*, *Ano*, *Feriado*, *Humidade Mínima*, *Comprimento do Dia*, *Temperatura Máxima*, *Visibilidade Média Ponto de Orvalho Médio*, *Pressão Máxima*, *Pressão Média* e *Ponto de Orvalho Máximo*. Por exemplo, a variável *Ano* ocorreu como significativa em 95% das instalações, ou seja, em 92 instalações de 97.

Variável	Estação	Ano	Feriado	HumMin	DiaSemana
Frequência	98%	95%	77%	77%	75%
Variável	ComprDia	TempMax	VisibMedia	PtOrvMedio	PresMax
Frequência	71%	67%	58%	56%	53%
Variável	PresMedia	PtOrvMax	PresMin	HumMedia	TempMedia
Frequência	52%	51%	40%	37%	37%
Variável	PtOrvMin	VelVentpMedia	Eventos	TempMin	VisibMin
Frequência	34%	34%	33%	32%	29%
Variável	HumMax	VelVentoMax	FimSemana		
Frequência	14%	10%	4%		

Tabela 4.2: Frequências relativas das variáveis selecionadas através do método *backward*

4.3.3.2 Random Forests

Random forests é um algoritmo popular e bastante eficiente, baseado num conjunto de árvores de decisão, para ambos problemas de regressão e classificação [29][14][10].
O algoritmo random forests é poderoso em muitas aplicações diferentes [10], incluindo a seleção de variáveis importantes. A função `randomForest()` está disponível na package `randomForest` do `R` e, caso se queira avaliar a importância das variáveis, é necessário colocar o parâmetro `importance=T` (os restantes parâmetros da função podem ser consultados e compreendidos em [3]). A medida de importância das variáveis é dada por uma nova taxa de erro interno. A quantidade pela qual este erro excede o erro do conjunto original de teste é definida como a importância da variável [3].
Assim, decidiu-se utilizar o Lote 1 como amostra e obter as variáveis mais importantes para cada instalação (o critério que mostrou ser o mais adequado foi considerar as 10 variáveis com maior erro). As variáveis mais significativas para o conjunto de instalações foram as que ocorrem mais vezes. Deste modo, na tabela 4.3 estão apresentadas as frequências relativas para cada variável selecionada. Observando a tabela 4.3 concluiu-se que as variáveis mais significativas ($\geq 50\%$) foram: *Comprimento do Dia*, *Estação*, *Temperatura Média*, *Temperatura Máxima*, *Temperatura Mínima*, *Ano*, *Ponto de Orvalho Máximo*, *Dia da Semana*, *Humidade Media* e *Feriado*.

Variável	ComprDia	Estação	TempMedia	TempMax	*TempMin*
Frequência	99%	99%	96%	95%	91%
Variável	Ano	PtOrvMax	DiaSemana	HumMedia	Feriado
Frequência	84%	78%	64%	62%	52%
Variável	HumMin	PtOrvMedio	PtOrvMin	VisibMedia	FimSemana
Frequência	49%	39%	23%	14%	13%
Variável	PresMin	PresMax	VelVentoMedia	*HumMax*	PresMedia
Frequência	9%	8%	8%	5%	5%
Variável	Eventos	VelVentoMax	VisibMin		
Frequência	4%	1%	1%		

Tabela 4.3: Frequências relativas das variáveis selecionadas através do método *random forests*

O problema deste método foi que seleciona variáveis muito correlacionadas, como é o caso das *Temperaturas*. Mas, neste contexto, não existia interesse em selecionar variáveis correlacionadas uma vez que se queria diminuir o número de variáveis. Testou-se a função `cforest()` da package `party` onde o problema das variáveis correlacionadas está resolvido, contudo este demorou cerca de 12 minutos para uma instalação com 1 ano de observações, não sendo viável utilizá-la para o Lote 1 completo.

4.3.3.3 Correlação parcial e cruzada

Nas subsecções anteriores as variáveis "significativas" selecionadas eram correlacionadas. No entanto, a existência de variáveis significativas correlacionadas significa que ambas fornecem a mesma informação e, portanto, pode-se eliminar uma delas.

No anexo B pode-se verificar que o máximo, a média e o mínimo de uma variável (por exemplo a temperatura) são bastante correlacionadas.

Assim, nesta subsecção foram selecionadas as variáveis mais significativas não correlacionadas através da correlação cruzada e parcial [16][13].

Não foi possível quantificar a correlação entre uma variável numérica e uma categórica, por isso consideraram-se como significativas as variáveis categóricas devolvidas pelos métodos anteriores: *Estação*, *Ano*, *Feriado* e *Dia da Semana*.

Inicialmente, utilizou-se a correlação parcial entre as variáveis explicativas e a variável resposta (consumo energético diário) para determinar as variáveis estatisticamente correlacionadas com o consumo, excluindo o efeito das restantes. De seguida, usou-se a correlação cruzada para eliminar das variáveis anteriores as correlacionadas entre si, daí resultando as variáveis mais correlacionadas com o consumo energético diário e não correlacionadas entre si. A correlação cruzada está disponível na função `rcorr` da package `Hmisc` e a correlação parcial na função `pcor` da package `ppcor` do `R`.
O resultado foi *Comprimento do Dia*, *Ponto de Orvalho Máximo* e *Humidade Mínima*. Segundo o meteorologista *Jeff Haby* [8] o conforto humano é definido usando o *Ponto de Orvalho* e a *Humidade*. Num contexto empresarial, o conforto humano é relevante nas horas de trabalho. Uma vez que a maior parte das instalações trabalham durante o dia, as observações das variáveis *Ponto de Orvalho* e *Humidade* registadas nesse período foram a máxima e o mínimo, respetivamente. Esta informação destacou este método como o mais adequado para a escolha das variáveis significativas do consumo energético.

Em suma:

- Os métodos *Backward* e *Random Forests* retornaram como variáveis significativas, variáveis correlacionadas;
- O método que utiliza as *correlações cruzada e parcial* não selecionou variáveis categóricas;
- A seleção das variáveis categóricas mais significativas foi realizada através dos dois primeiros métodos, selecionando: *Estação*, *Dia da Semana*, *Ano* e *Feriado*;
- As variáveis numéricas foram selecionadas através do último método, selecionou-se: *Comprimento do Dia*, *Ponto de Orvalho Máximo* e *Humidade Mínima*. Resta salientar que estas variáveis também foram significativas através do método Backward.

4.4 Deteção de valores anormais

No capítulo 2 observou-se que a instalação inicial continha dois valores de consumo energético diário bastante mais baixo que os restantes valores, por ano. Esses valores dizem respeito aos dias de Natal e Ano Novo e são chamados *valores anormais*. Ao analisar o Lote 1 de instalações, verificou-se que existem outras instalações com esses valores (nos mesmos dias), enquanto que outras não contêm valores anormais visíveis. A deteção destes valores anormais é útil para caracterizar uma instalação, servindo para identificar as instalações que são sensíveis a feriados e as que não são. Deste modo, nesta secção mostra-se o desenvolvimento de um algoritmo para detetar valores anormais neste contexto, utilizando o Lote 1 de instalações como amostra.

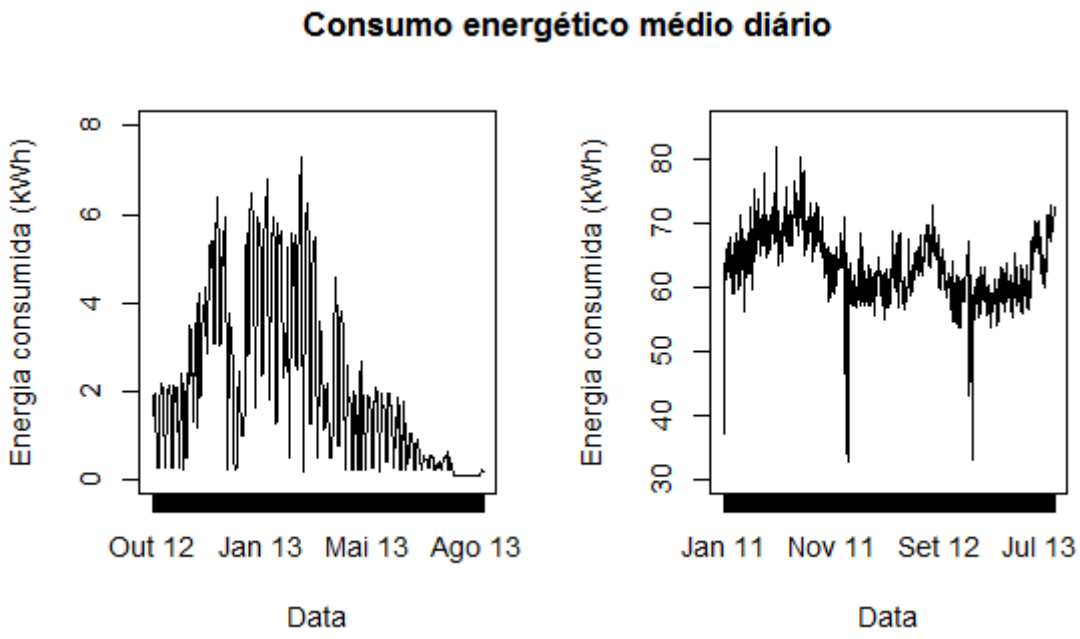

Figura 4.2: Consumo de energia média diária de duas instalações típicas de cada grupo: Esquerda - Instalação típica do grupo sem valores anormais; Direita - Instalação típica do grupo com valores anormais

No Lote 1 foram distinguidos visualmente dois grupos de instalações. Na figura 4.2 estão representados os consumos de energia médios diários de duas instalações típicas de cada grupo. Observou-se que a instalação à esquerda não contém nenhum valor anormal, enquanto que a instalação à direita contém 5 valores extremos.

Para detetar os valores anormais pensou-se em utilizar uma abordagem simples com a utilização da média, desvio padrão ou variância do consumo. Na figura 4.3 estão no-

vamente representados os consumos das instalações acima, mas com limites, que definem valor normal ou não, determinados usando a média mais/menos o desvio padrão ou variância. Analisando a figura observou-se que os limites usando a média mais/menos o desvio padrão detetaram demasiados valores anormais em ambos os casos, ou seja, foi necessário utilizar um fator de escala superior a 1 no desvio padrão. Os limites usando a média mais/menos a variância detetaram valores anormais na instalação que não os contém e não detetaram nenhum valor anormal na instalação que contém 5 valores anormais. Consequentemente, estes limites não puderam ser utilizados para detetar estes valores.

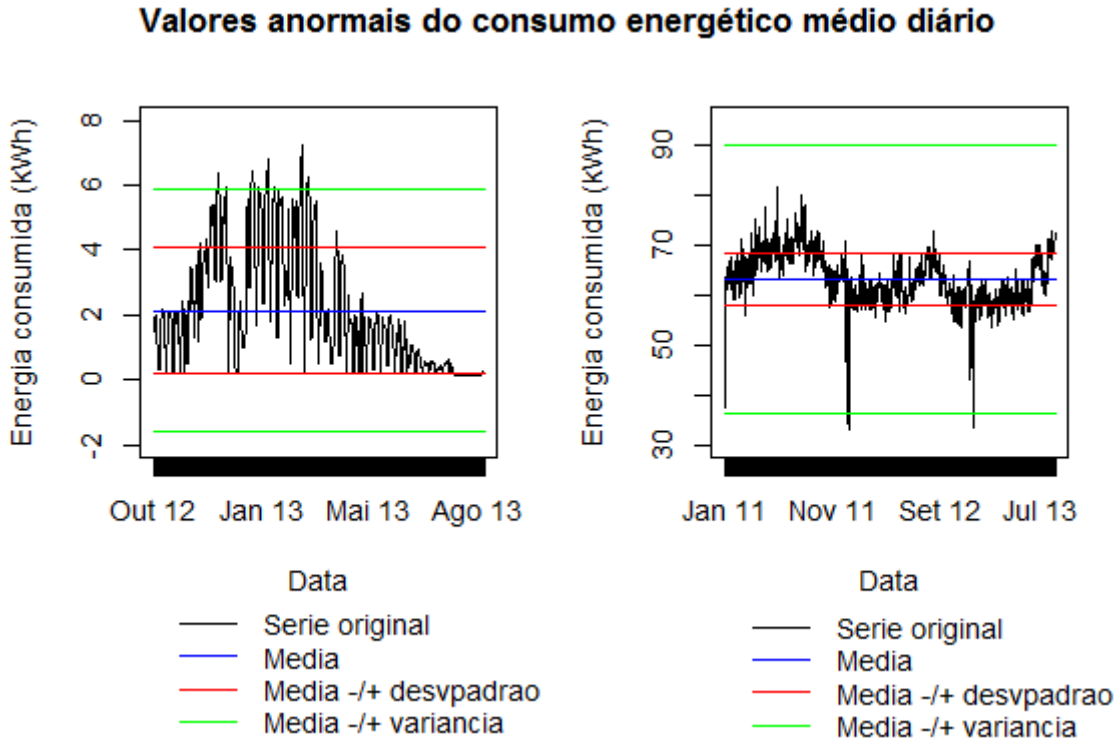

Figura 4.3: Limites de ser um valor de consumo energético normal usando a média, desvio padrão e variância em duas instalações típicas de cada grupo: Esquerda - Instalação típica do grupo sem valores anormais; Direita - Instalação típica do grupo com valores anormais

Testando vários fatores de escala no desvio padrão, chegou-se à conclusão que um fator igual a 3.5 foi suficiente para detetar corretamente os valores anormais de ambas instalações, como se pode verificar na figura 4.4. Contudo, testando esta abordagem para as restantes instalações do Lote 1 verificou-se que em 21 instalações de 97 não se detetou corretamente os valores anormais de consumo energético.

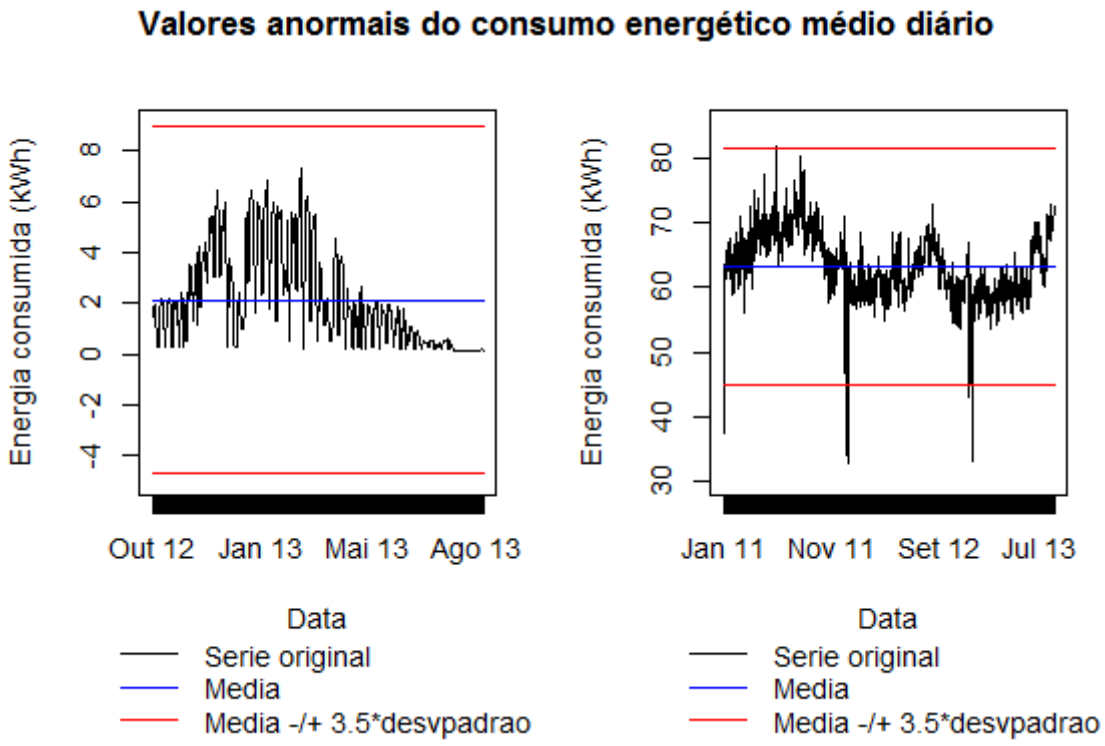

Figura 4.4: Limites de ser um valor de consumo energético normal usando a média e 3.5 do desvio padrão em duas instalações típicas de cada grupo: Esquerda - Instalação típica do grupo sem valores anormais; Direita - Instalação típica do grupo com valores anormais

Decidiu-se utilizar a abordagem anterior, mas usando a tendência do consumo em vez da média, com o objetivo de diminuir o número de instalações onde não se detetaram corretamente os valores anormais. Na figura 4.5 pode-se observar os consumos das instalações com os limites, que definem valor normal ou não, determinados usando a tendência mais/menos a média, tendência mais/menos o desvio de padrão e tendência mais/menos a variância. Observou-se que para detetar corretamente os valores anormais na instalação à esquerda foi necessário aplicar um fator de escala superior a 1 à média, ao desvio padrão ou à variância. Quanto à instalação à direita foi necessário aplicar um fator de escala entre 0 e 1 à média ou à variância ou um fator de escala superior a 1 ao desvio padrão. Sendo assim, a única opção comum a ambos os casos foi aplicar um fator de escala superior a 1 ao desvio padrão.

Experimentaram-se vários fatores de escala no desvio padrão do consumo e chegou-se à conclusão que esta abordagem não foi suficiente para detetar corretamente os valores anormais de todas as instalações do Lote 1.

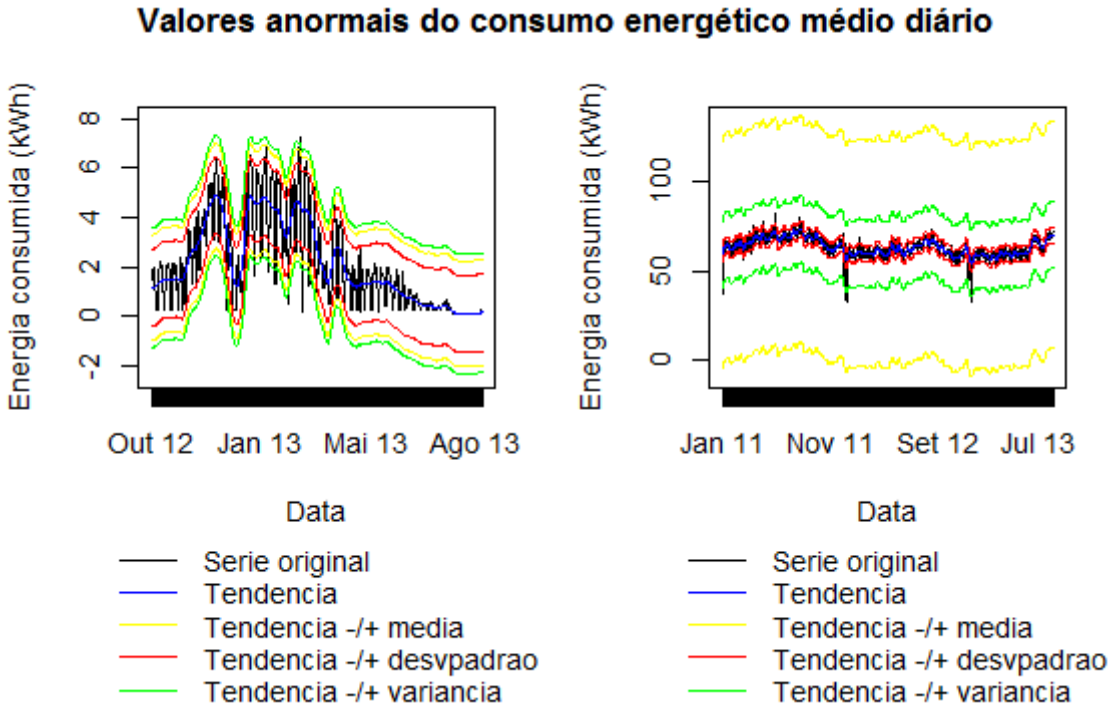

Figura 4.5: Limites de ser um valor de consumo energético normal usando a tendência, média, desvio padrão e variância da tendência em duas instalações típicas de cada grupo: Esquerda - Instalação típica do grupo sem valores anormais; Direita - Instalação típica do grupo com valores anormais

Decidiu-se utilizar a abordagem anterior em simultâneo com outro método. Utilizou-se um método de clustering de Data Mining, *DBSCAN*, que agrupa as observações criando também um grupo de outliers [14][28].

Este método, *DBSCAN*, necessita de 2 parâmetros e, após vários testes, detetou-se corretamente os valores anormais em 82 de 97 instalações. Ao exigir que um valor anormal satisfaça ambos os métodos, *tendência do consumo mais/menos desvio padrão* e *DBSCAN*, os resultados foram mais satisfatórios, detetando-se corretamente os valores anormais em 93 de 97 instalações.

Concluindo, para detetar os valores anormais neste contexto foram usados os métodos *tendência do consumo mais/menos desvio padrão*, com fator de escala 2 sobre o desvio padrão, e *DBSCAN* com `eps=1.9` e `MinPts=5`, em simultâneo. As conclusões foram as mesmas para a medida de agregação soma.

Resta salientar que inicialmente foi testado o método tradicional de deteção de outliers através do *Box-Plot* para o Lote 1 e um algoritmo de deteção de consumos energéticos anormais segundo [24]. Contudo, não se obtiveram os resultados pretendidos para ambos.

Como referido anteriormente, a deteção de valores anormais é uma forma de caraterizar uma instalação. Assim, decidiu-se construir uma série com valor zero em todos os instantes menos nos que contêm valores anormais, sendo o valor desses instantes igual ao original, com o objetivo de construir uma nova variável explicativa do consumo energético. Uma outra variável importante para o clustering foi a *Tendência*. Foi visto no capítulo 3 que os valores anormais influenciam a tendência. Deste modo, a variável *Tendência* foi definida como sendo a tendência da série sem os valores anormais (colocando nesses instantes o valor da tendência inicial da série).

4.5 Conclusão

Este capítulo apresentou a escolha das medidas de agregação, das variáveis significativas, das regiões de Portugal continental e de outras características como os valores anormais de consumo energético.

- Medidas de agregação: *soma* e *média*;
- Regiões de Portugal continental: *Região do Porto*, *Região de Bragança*, *Região de Lisboa*, *Região de Castelo Branco*, *Região de Beja*, *Região de Faro*;
- Variáveis significativas: *Ano*, *Estação*, *Dia da Semana*, *Feriado*, *Comprimento do dia*, *Ponto de Orvalho Máximo*, *Humidade Mínima*;
- Novas variáveis: *Valores Anormais* e *Tendência*.

As ferramentas computacionais utilizadas foram o `R` e o `SQL Server`. Os dados de consumos energéticos das instalações estavam registados em bases de dados no `SQL Server`. Contudo, foi necessário processar os dados, agregar segundo as medidas de agregação que foram usadas, processar as observações das variáveis externas, etc. Testou-se agregar os dados diariamente, utilizando a média, no `R` e demorou 12 segundos para apenas uma instalação, enquanto que no `SQL Server` apenas demora 1 segundo. Desta forma,

foi decidido processar todos os dados em SQL Server, incluindo a junção de variáveis climáticas segundo a região da instalação.

As restantes metodologias que necessitaram de comandos específicos do R foram executadas em R.

Nas próximas tabelas, 4.4 e 4.5, podem-se visualizar os tempos de execução de cada metodologia e ferramenta para o conjunto de instalações do Lote 1. Em SQL Server, agregaram-se os dados diariamente utilizando a soma e a média e agregaram-se os dados por dia da semana (incluindo Feriados) e hora utilizando a média para ser possível calcular o horário de trabalho (ver capítulo 2). Em R, foi necessário ler as tabelas construídas em SQL Server (através da package RODBC), construir o Horário de Trabalho para cada instalação, tratar os valores desconhecidos existentes nas tabelas de variáveis climáticas, selecionar as variáveis significativas através dos métodos Backward, Random Forests e Correlação e detetar os valores anormais e a tendência do consumo sem esses valores.

Tempos de Execução para o Lote 1

SQL Server	Tempos de Execução
Agregar diariamente	2 min.
Agregar para Horário de trabalho	1.5 min.

Tabela 4.4: Tempo de processamento dos dados das instalações do Lote 1 em SQL Server

R	Tempos de Execução
Ler dados agregados diariamente	26 seg.
Ler dados agregados para Horário de trabalho	1.2 seg.
Calcular o Horário de Trabalho	3 seg.
Tratamento de falhas (6 regiões)	8 seg.
Backward	50 seg.
Random Forests	19 min.
Correlação	16 seg.
Valores anormais e tendência	30 seg.

Tabela 4.5: Tempo de execução de metodologias aplicadas aos dados das instalações do Lote 1 em R

Capítulo 5

Seleção de instalações

5.1 Introdução

Neste capítulo serão expostos os critérios usados para selecionar as instalações que foram utilizadas no estudo. Por exemplo, uma instalação que continha demasiadas observações em falta não fornecia dados suficientes para ser possível comparar com outras instalações. Salienta-se que, ao longo do trabalho, os critérios foram ajustados consoante as necessidades.

5.2 Critérios de seleção de instalações

Ao analisar os dados, constatou-se que, por vezes, existiam várias observações para um mesmo instante de uma mesma instalação. Por vezes foi necessário substituir um contador, ou seja, este passa a estar inativo e o substituto será o contador ativo. Nestes casos, o nome do contador inativo passa a conter "!" no início e o ativo não contém nenhum "!".
Uma possibilidade para a existência de mais do que um registo, é que o contador substituído continue ativo, em simultâneo com o substituto, devido a algum erro.
Uma outra justificação é a existência de mais do que um contador. Por exemplo, uma empresa pode ter vários contadores, um para a cantina, outro para o armazém, outro para a produção e o consumo total dessa empresa é a soma dos registos de todos os

contadores. Neste caso, nenhum dos nomes dos contadores contém "!".
Consequentemente, foi necessário criar critérios de seleção de registos para estes casos. As possibilidades foram:

1. Caso um dos nomes dos contadores contenha no início "!" e exista um outro nome que não contenha, então escolhe-se os contadores que não contêm "!" no início do nome;
2. Caso não exista um contador com "!", somam-se os registos de um mesmo instante;

Após os critérios anteriores, foram criados outros devido às metodologias usadas.
Não existia qualquer interesse em analisar instalações que já estivessem inativas ou indicassem que deixaram de produzir. Assim, definiu-se:

3. Para uma instalação ser selecionada é necessário que contenha registos recentes, ou seja, se o último registo for antes de Março de 2013 não se seleciona a instalação;
4. Caso as observações recentes (1 Set 2012 a 31 Ago 2013) de consumo de uma instalação contenham demasiados zeros (80 %), considera-se que a instalação deixou de estar ativa e elimina-se do estudo.

Os próximos critérios dizem respeito à quantidade de falhas que pode existir ao longo dos registos.
Primeiro, uma vez que as observações foram agregadas diariamente, foi necessário garantir que o valor não era enviesado do real. Por exemplo, se se considerar uma instalação que trabalhe durante o dia, e para um dia apenas existem observações durante a noite, ao agregar, o consumo diário será muito baixo em relação aos restantes dias de trabalho.
Segundo, para uma análise viável aos dados foi necessário que estes não contivessem demasiados dias em falta. Assim foi necessário um critério que elimine os casos em que tal acontece.

5. Para o primeiro caso dividiu-se o dia em 4 períodos:

- Das 7h às 13h;
- Das 13h às 19h;
- Das 19h à 1h;
- Da 1h às 7h;

e para cada um destes, pode ter-se no máximo 1h de falhas (ou seja, 4 registos de 15 minutos). Consequentemente, no máximo por dia podem-se ter 4h de falhas. Caso contrário, o dia é eliminado;

6. Para o segundo caso, consideram-se as instalações que não contêm mais que 10% de dias em falha, no total;

Por fim, para o clustering foi necessário que para cada instalação existissem observações suficientes para se comparar com outras. Sendo assim, foi decidido considerar as instalações que continham 9 meses de observações consecutivas (sem qualquer falha). Ao conter 9 meses de observações garantia-se que existem pelo menos três estações do ano para comparar.
No entanto, no capítulo 6 será discutido que os dados deviam ser restringidos ao ano e meio mais recente. Assim, o método anterior só pôde ser aplicado após restringir os dados, uma vez que se queria ter os 9 meses de observações, sem falhas, nas observações que foram usadas no capítulo *Clustering*.

7. A instalação é selecionada caso contenha pelo menos 9 meses de observações consecutivas (sem falhas) no ano e meio mais recente de observações. Caso contrário, elimina-se do estudo.

5.3 Resultados

Como referido anteriormente, os dados dos consumos energéticos estavam em bases de dados em `SQL Server`. O objetivo de definir critérios de seleção de instalações foi eliminar da base de dados registos que não eram adequados ao estudo. Assim, esta

etapa foi realizada no `SQL Server` por ser mais simples e mais rápida a seleção de registos. Em `SQL Server` existem comandos como `SELECT ... FROM ...`, `DELETE ... FROM ...`, que permitem a manipulação de registos de uma tabela.

Testaram-se os critérios para o primeiro Lote de instalações que continham 98 empresas. Inicialmente testou-se a existência de mais que um registo num mesmo período mas, contudo, para este Lote tal não acontece. A necessidade de criar esses critérios surgiu quando se testaram metodologias nos Lotes 2 e 3.

Assim, nesta secção apresentam-se os resultados para os Lotes 1, 2 e 3. As observações das instalações dos Lotes 2 e 3 estavam na mesma tabela, deste modo foram tratadas como um só Lote. Neste Lote, Lotes 2 e 3, existiam 199 instalações.

Os critérios 3 e 4 foram executados ao mesmo tempo, depois eliminaram-se os dias que não eram viáveis (critério 5) e verificou-se se a quantidade de falhas era demasiada (critério 6). Uma vez que no critério 5 podiam ser eliminados registos recentes, aplicou-se novamente o critério 3. Por fim, selecionaram-se as instalações que continham 9 meses de observações consecutivas e recentes, critério 7.

Os resultados da aplicação destes critérios nos Lotes 1, 2 e 3 podem ser visualizados na tabela 5.1. Os critérios estão apresentados pela ordem que foram aplicados.

Observando os resultados para os Lotes 1, 2 e 3 verifica-se que, com estes critérios, poucas instalações são eliminadas.

No capítulo 4, como mencionado, o Lote 1 continha 97 instalações. Estas instalações foram sujeitas aos critérios de 1 a 6, uma vez que o critério 7 apenas foi aplicado para a metodologia *Clustering*.

Critério	Lote 1	Lotes 2 e 3
1	Sem alterações	Elimina-se 0.6% dos registos
2	Sem alterações	Soma-se 1.1% dos registos obtendo 93 625 novos registos
3 e 4	Elimina-se 1 instalação	Elimina-se 13 instalações
5	Eliminam-se 0.12% dos dias	Eliminam-se 0.23% dos dias
6	Sem alterações	Elimina-se 2 instalações
3	Sem alterações	Sem alterações
7	Eliminam-se 11 instalações	Eliminam-se 46 instalações
Total	Selecionam-se 86 instalações de 98	Selecionam-se 138 instalações de 199
Tempo	4 minutos	20 minutos

Tabela 5.1: Seleção de instalações e registos por critério para os Lotes 1, 2 e 3

Capítulo 6

Agrupamento (Clustering)

6.1 Introdução

Um problema que se coloca com alguma frequência é o de, dado um conjunto de n objetos, agrupá-los em classes, ou subgrupos, de tal forma a que (i) cada subgrupo seja internamente homogéneo (isto é, constituído por objetos "similares"), e a que (ii) os vários subgrupos sejam heterogéneos entre si (isto é, os indivíduos de subgrupos diferentes sejam "dissimilares"). A este processo dá-se o nome de *Agrupamento (Clustering)* (não supervisionado).

Nos capítulos anteriores foi possível perceber as variáveis que explicavam parte do consumo energético de uma instalação. Claro que, para além das variáveis encontradas, foi necessário acrescentar a variável *Trabalho*, que diz respeito à energia consumida pela instalação para produzir.

O objetivo principal deste estágio foi agrupar as instalações segundo o seu *trabalho*, ou seja, colocar num mesmo grupo instalações que continham curvas de consumo energético semelhantes (os objetos foram as instalações). No entanto, foi necessário ter atenção às variáveis externas, pois duas instalações podiam ter curvas semelhantes de energia gasta com a produção mas, devido às condições externas, terem curvas do consumo energético diferentes. Por exemplo, uma instalação situada num local de clima quente e seco com certeza gastará mais energia no ar condicionado do que uma instalação análoga situada num local de clima temperado. Deste modo, foi necessário comparar o consumo

relacionado com o trabalho de cada instalação.
Estimar o consumo energético relacionado com as variáveis externas foi um dos objetivos principais do estágio *Desagregação do consumo energético*, desenvolvido pela aluna Elena Selaru, como já foi referido na introdução. Assim, neste estágio não se desenvolveu esse assunto, uma vez que seria necessário utilizar os resultados do estágio anterior. De facto, como os estágios decorreram ao mesmo tempo, não foi possível obter os resultados a tempo de realizar o clustering baseado em consumo energético desagregado. Desta forma, foi usado o consumo energético completo (com o efeito das variáveis externas) no clustering.
Nas próximas secções serão descritos métodos que tenham atenção às situações anteriores. Existem duas possibilidades:

- Aplicar um método de clustering utilizando os consumos energéticos diários ou, quando for possível utilizar as estimativas do consumo relacionado com as variáveis externas, utilizar o consumo relacionado com o trabalho;
- Utilizar um método de clustering de séries temporais multivariadas, para se considerarem as observações das variáveis externas significativas.

Após encontrar os métodos de clustering mais adequados ao problema foi necessário avaliá-los e definir o número adequado de grupos, secção 6.4. Posteriormente apresentam-se os resultados.
Salienta-se que foram usados os dados agregados por dia através da média e da soma.

6.2 Métodos de transformação dos dados

Clustering de séries temporais é um problema que surge numa grande variedade de áreas e que recentemente tem atraído o interesse dos investigadores, particularmente das áreas de estatística, processamento de sinal, data mining, finança e outras [38][2].
Infelizmente, os métodos existentes para o clustering de séries temporais que se baseiam nos valores pontuais das séries podem tornar-se impraticáveis, uma vez que perdem eficiência computacional quando aplicados a bases de dados com uma quantidade enorme

de séries temporais, que podem ser muito longas, e muitos algoritmos de clustering não lidam facilmente com dados de grande dimensão [2]. Frequentemente, os métodos existentes baseiam-se em distâncias calculadas na série completa, usando a distância *Euclidiana* ou a distância *Dynamic Time Warping* [9]. Contudo, a presença de ruído significativo e de valores estranhos limita a precisão do clustering neste domínio. Além disso, não se pode esperar que as séries tenham tamanhos iguais [38].

Devido a estes problemas, decidiu-se optar por dois métodos que primeiro transformam a estrutura da base de dados, para que seja possível aplicar algoritmos de clustering usuais (secção 6.3).

6.2.1 Shapelets

Zakaria, Mueen e Keogh [38] estenderam um conceito recente de data mining, *shapelets*, e propuseram um novo método para clustering de séries temporais baseado nas chamadas *unsupervised shapelets*, ou apenas *u-shapelets*.

Ideia geral: Ignorar parte da série temporal e usar apenas algum padrão local.

Como um exemplo concreto, consideram-se as séries temporais apresentadas na figura 6.1.

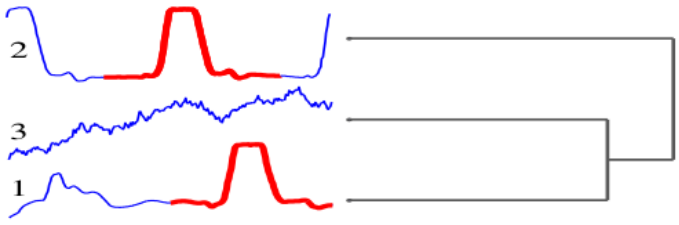

Figura 6.1: Clustering usando a distância Euclidiana entre séries inteiras [38]

Observando esta figura, conclui-se que se obtêm resultados de clustering fracos quando se utiliza a distância Euclidiana entre séries completas, pois esperava-se que as séries 1 e 2 pertencessem ao mesmo grupo.

Na figura 6.2 pode-se observar o resultado do clustering das séries ignorando parte delas e conclui-se que a situação melhora drasticamente. A questão é: *que partes da série ignorar?* Zakaria, Mueen e Keogh [38] apresentam um algoritmo onde esta questão é resolvida sem intervenção humana.

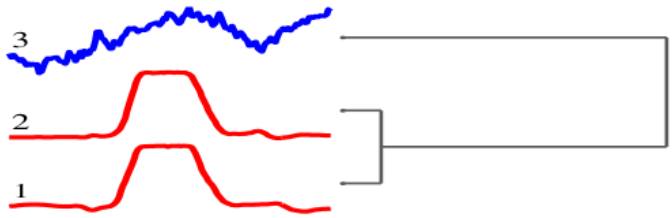

Figura 6.2: Clustering usando a distância Euclidiana ignorando parte das séries [38]

O conceito de shapelet foi introduzido por Ye [37] como sendo um padrão pequeno e local de uma série temporal que é altamente preditivo de uma classe. Segundo [38], as shapelets também podem ser altamente competitivas em clustering de séries temporais e foi proposto um novo conceito, *unsupervised-shapelets* (ou *u-shapelets*), onde não é necessário o conhecimento das classes reais das séries.

Definições

1. Distância Euclidiana normalizada pelo comprimento (length-normalized Euclidean distance): $dist(X,Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2}$, X e Y séries temporais;
2. Distância entre uma subsequência S de tamanho m e uma série temporal T de tamanho n ($m \leq n$): $sdist(S,T) = \min_{1\leq i\leq n-m} dist(S, T_{i,m})$, onde $T_{i,m}$ é a subsequência de T de tamanho m com inicio na i-ésima observação;
3. Uma u-shapelet $\acute{S}$ é uma subsequência de uma série temporal T que:
   - Dado um conjunto de séries temporais D,
   - $\acute{S}$ separa efetivamente D em dois grupos D_A e D_B tais que

     $$sdist(\acute{S}, D_A) << sdist(\acute{S}, D_B)$$

     ou seja, $sdist$ entre $\acute{S}$ e uma série temporal do grupo D_A é "muito menor" que $sdist$ entre $\acute{S}$ e qualquer série temporal do grupo D_B;
4. Matriz distância (distance map): matriz que contém as distâncias entre cada u-shapelet e cada série temporal de D. Supondo que existem m u-shapelets e N séries temporais, então a matriz distância tem dimensão $[N \times m]$.

Após obter a matriz distância, podem-se agrupar as séries fazendo clustering sobre a matriz, como será visto.

Cada série temporal é normalizada antes de calcular as distâncias, uma vez que é bem compreendido que não faz sentido comparar séries temporais com diferentes fases e amplitudes [38][19].

Algoritmo

A principal ideia do algoritmo [38] é procurar uma u-shapelet que separe de maneira ótima e remova um subconjunto de séries temporais das restantes. O algoritmo é então aplicado iterativamente ao conjunto restante de séries, até que não existam séries para separar.

Uma u-shapelet $\acute{S}$ ótima tem a capacidade de dividir o conjunto D em dois grupos (figura 6.3), D_A e D_B. D_A contém as séries temporais que têm subsequências similares a $\acute{S}$ enquanto que D_B contém as restantes séries de D. Assim, espera-se que a média de $sdist(\acute{S}, D_A)$ seja "muito menor" do que a média de $sdist(\acute{S}, D_B)$ e, portanto, quanto maior a distância entre essas duas médias, melhor será a u-shapelet.

Figura 6.3: Ilustração da separação de D em D_A e D_B [38]

Desta forma, para extrair u-shapelets (algoritmo 6.1) é necessário encontrar a subsequência que maximiza a distância (*gap*) entre dois subconjuntos de D. A distância é definida por [38]

$$gap = \mu_B - \sigma_B - (\mu_A + \sigma_A)$$

onde $\mu_A = mean(sdist(\acute{S}, D_A))$ e $\sigma_A = std(sdist(\acute{S}, D_B))$ (análogo para D_B).

Observa-se na figura 6.3 um ponto chamado de dt. Tal ponto maximiza gap, ou seja, pontos à esquerda de dt representam $sdist(\acute{S}, D_A)$, enquanto que pontos à direita correspondem a $sdist(\acute{S}, D_B)$.

Caso exista apenas uma série temporal em D_A ou D_B, não se tem capacidade de discriminação, mas sim um outlier ou um padrão universal, ambos indesejáveis. A fim de excluir tais candidatos a u-shapelet, verifica-se se a proporção de D_A e D_B está dentro do

intervalo [38]

$$\left(\frac{1}{k}\right) < |D_A|/|D_B| < \left(1 - \frac{1}{k}\right),\ k:\ \text{número de clusters}$$

Algoritmo 6.1 Extrair as u-shapelets

Dado um conjunto de séries temporais D e um conjunto de tamanhos das u-shapelets $TamS$:

1. Seja ts a primeira série temporal de D;
2. Seja $\acute{S}$ o conjunto de u-shapelets e DIS a matriz de distância, inicialmente vazios;
3. Enquanto é possível dividir D:
   (a) Para cada tamanho sl de $TamS$:
      i. Para cada subsequência de ts com tamanho sl:
         A. Calcular a máxima gap, através do algoritmo 6.2, e guardar no vetor GAP;
   (b) Adicionar a $\acute{S}$ a subsequência $\hat{s}$ tal que GAP é máximo;
   (c) Seja dis o vetor de distâncias $sdist$ de $\hat{s}$ a cada série temporal de D;
   (d) Adicionar o vetor coluna dis a DIS;
   (e) ts passa a ser a série temporal tal que dis é máximo;
   (f) Excluir do conjunto D as séries temporais semelhantes a $\hat{s}$, ou seja, que têm dis menor que θ, onde $\theta = \mu_A + \sigma_A$;
4. Retornar o conjunto de u-shapelets $\acute{S}$ e a matriz de distância DIS.

Algoritmo 6.2 Calcular Gap

Dado um candidato a u-shapelet $\hat{s}$ e um conjunto de séries temporais D:

1. Calcular o vetor de distâncias $sdist$ de $\hat{s}$ a cada série temporal de D;

2. Seja dt o ponto que maximiza gap, ou seja, pontos à esquerda de dt representam $sdist(\acute{S}, D_A)$, enquanto que pontos à direita correspondem a $sdist(\acute{S}, D_B)$;

3. Para cada possível localização de dt:

   (a) Se $(1/k) < |D_A|/|D_B| < (1 - 1/k)$, então calcular gap;

4. Retornar a *máxima gap* e o dt associado.

Uma questão que pode surgir é: *porquê usar θ em vez de dt para remover o conjunto D_A de D?* O facto é que o uso de θ é mais seletivo que o uso de dt. Considere-se o seguinte exemplo [38]:

- Suponha-se que se quer agrupar as seguintes palavras: *Earth Day*, *Hoover Dam*, *Memorial Day*, *Fink Nottle*, *Alamo Dam*, *Labor Day*, *Bingo Little*;
- Seja *Day* a u-shapelet encontrada;
- Obtêm-se as "distâncias" observadas na figura 6.4. Usando θ, é-se mais seletivo e apenas se removem as frases que contêm a palavra *Day*, enquanto que usando dt, para além destas frases, também se eliminariam as que contêm a palavra *Dam*, palavra esta muito semelhante a *Day*.

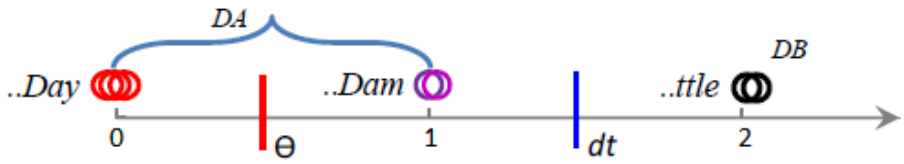

Figura 6.4: Distâncias das frases à palavra *Day* [38]

Note que o algoritmo 6.1 apenas exige ao utilizador um parâmetro, o conjunto dos tamanhos das u-shaplets. Conjuntos menores de tamanhos melhoram a velocidade do algoritmo, contudo podem comprometer a exatidão se evitarem que as subsequências mais informativas sejam consideradas. Para acomodar a execução do algoritmo 6.1, sem nenhum conhecimento especializado dos dados, [15] define um algoritmo simples para estimar o tamanho mínimo e máximo das shapelets (algoritmo 6.3).

Algoritmo 6.3 Estimar o tamanho mínimo e máximo das shapelets

Dado um conjunto de séries temporais D:

1. Seja $D1$ um subconjunto aleatório de séries temporais de D, 40% de D;
2. Aplicar o algoritmo 6.1 a $D1$ e a um grande conjunto de tamanhos $TamS$;
3. Seja $shap$ o vetor ordenado por tamanho das u-shapelets devolvidas;
4. Retornar *min*, igual ao quantil 25% de $shap$, e *max*, igual ao quantil 75% de $shap$.

Por fim, aplicando um algoritmo de clustering à matriz de distância DIS devolvida pelo algoritmo 6.1, obtém-se o agrupamento das séries temporais de D, por forma.

Uma vez que as séries são normalizadas, o método apenas agrupa as instalações por consumos energéticos com formas semelhantes. Assim, após se obterem estes grupos, agrupam-se as instalações dentro de um mesmo grupo por escala. Neste último agrupamento utilizam-se os algoritmos de clustering (secção 6.3) para agrupar as instalações de um mesmo grupo (de forma) usando os consumos energéticos em dias comuns.

Neste trabalho utilizou-se o código disponível no site [35], criado em `Matlab`. Contudo, foi necessário adaptar o código para `R`, alterando-o para receber como input a tabela de séries temporais com três colunas (*Nome da instalação*, *Data* e *Energia*), em vez de uma tabela $[m \times N]$, onde m é o número de observações por série e N é o número de séries temporais. O código também foi alterado para receber séries temporais de tamanhos diferentes e com possíveis falhas.

Este método deve ser utilizado quando as séries temporais estão livres do efeito das variáveis externas no consumo energético. A seguir apresenta-se um método de clustering de séries temporais multivariadas, onde se pode utilizar as variáveis externas em vez de retirar o efeito destas no consumo.

6.2.2 Fatores de similaridade

O clustering de séries temporais multivariadas consiste em agrupar conjuntos de dados que têm características semelhantes. Nesta secção, uma nova metodologia de clustering

de séries temporais multivariadas é apresentada.

A metodologia é baseada no cálculo de um grau de similaridade usando *análise de componentes principais* (PCA) e de um fator de similaridade de distância [26].

Fatores de similaridade podem ser usados em vez da distância Euclidiana para medir a semelhança entre dois conjuntos de dados multivariados. Krzanowski [20] desenvolveu um método para medir a similaridade de dois conjuntos de dados, $\mathbf{X_1}$ e $\mathbf{X_2}$, usando o *fator de similaridade PCA* que é calculado usando as x componentes principais que mais explicam a variância de cada conjunto de dados multivariado. As componentes principais (PCs) são os vetores próprios da matriz de covariância de um conjunto de dados multivariado.

O fator de similaridade PCA, S_{PCA}, é definido como [20]

$$S_{\text{PCA}} = \frac{1}{x} \sum_{i=1}^{x} \sum_{j=1}^{x} \cos^2 \theta_{ij}$$

onde x é o número de PCs selecionadas em ambos conjuntos de dados e θ_{ij} é o ângulo entre a i-ésima PC de $\mathbf{X_1}$ e a j-ésima PC de $\mathbf{X_2}$. O número de PCs, x, pode ser escolhido de forma a que as x PCs descrevam pelo menos 95% da variância de cada conjunto de dados.

Uma vez que S_{PCA} pesa igualmente todas as PCs, este pode não capturar o grau de similaridade entre os conjuntos, quando apenas uma ou duas PCs explicam a maior parte da variância. Assim, é natural definir um fator de similaridade PCA modificado, S^{λ}_{PCA}, que pesa cada PC através da variância explicada.

O S^{λ}_{PCA} é definido como [26]

$$S^{\lambda}_{\text{PCA}} = \frac{\sum_{i=1}^{x} \sum_{j=1}^{x} (\lambda_i^{(1)} \lambda_j^{(2)}) \cos^2 \theta_{ij}}{\sum_{i=1}^{x} \lambda_i^{(1)} \lambda_i^{(2)}}$$

onde $\lambda_i^{(1)}$ e $\lambda_i^{(2)}$ são os i-ésimos valores próprios de $\mathbf{X_1}$ e $\mathbf{X_2}$, respetivamente.

O *fator de similaridade de distância*, S_{dist} [26], compara dois conjuntos de dados que podem ter orientação espacial semelhante, mas estar localmente afastados.

O S_{dist} é definido como [26]

$$S_{\text{dist}} = 2 \times \frac{1}{\sqrt{2\pi}} \int_{\Phi}^{\infty} e^{-z^2/2} dz = 2 \times \left[1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\Phi} e^{-z^2/2} dz \right]$$

onde $\Phi = \sqrt{(\bar{x}_2 - \bar{x}_1) \sum_1^{*-1} (\bar{x}_2 - \bar{x}_1)^T}$, $\bar{x}_1$ e $\bar{x}_2$ são vetores com as médias amostrais de cada variável de $\mathbf{X_1}$ e $\mathbf{X_2}$, respetivamente, $\sum_1$ é a matriz de covariância de $\mathbf{X_1}$ e $\sum_1^{*-1}$ é

a pseudo-inversa de $\sum_1$ calculada usando SVD.

Inclusão de um conjunto de medidas

Para alguns problemas práticos, cada conjunto de dados inclui um conjunto de medidas, por exemplo, a uma instalação está associado um conjunto de dados que contém o consumo energético diário, o comprimento do dia e o ponto de orvalho máximo, ao longo do tempo. Contudo, uma outra característica desta instalação é o horário de trabalho, que não é uma série temporal. Então, esta é referida como uma *medida adicional* da instalação em questão. Estas medidas serão referidas como *medidas adicionais* ou *dados* $\mathbf{Y}$ para distinguir dos conjuntos de séries temporais, referidos como *dados* $\mathbf{X}$. As dimensões de $\mathbf{X}$ são $m \times n_x$, m observações e n_x séries, enquanto que as dimensões de $\mathbf{Y}$ são $1 \times n_y$, onde n_y é o número de variáveis de medidas. O fator de similaridade para os dados $\mathbf{Y}$ é baseado na distância Euclidiana entre dois conjuntos, $\mathbf{Y_1}$ e $\mathbf{Y_2}$: $\Phi_y = \|\mathbf{Y_1} - \mathbf{Y_2}\|$. Assumindo que $\mathbf{Y}$ segue uma distribuição de probabilidade Gaussiana, o fator de similaridade de distância para $\mathbf{Y}$, S^y_{dist}, é definido como [26]

$$S^y_{\text{dist}} = \sqrt{\frac{2}{\pi}} \int_{\Phi_y}^{\infty} e^{-z^2/2} dz$$

Note que ambos, S_{dist} e S^y_{dist} variam entre zero e um.

Combinação dos fatores de similaridade

Quando mais do que um fator de similaridade é usado para calcular a similaridade entre conjuntos de dados, é necessário ponderá-los, combinando-os numa única medida de grau de similaridade

$$SF = \alpha_1 S^{\lambda}_{\text{PCA}} + \alpha_1 S_{\text{dist}} + \alpha_3 S^y_{\text{dist}} (\alpha_1 + \alpha_2 + \alpha_3 = 1) \tag{6.1}$$

Quando Y não é incluído, o último termo $\alpha_3 S^y_{\text{dist}}$ não existe, logo $\alpha_1 + \alpha_2 = 1$.
O fator de similaridade SF pode ser usado como medida de similaridade entre conjuntos de dados e também no clustering. Cabe ao utilizador escolher os pesos $\{\alpha_i\}$, que definem a importância de cada fator, consoante a aplicação. Segundo [26], a experiência tem demonstrado que uma boa correspondência de forma pode ser obtida para um vasto intervalo de $\{\alpha_i\}$.

Algoritmo de clustering k-means usando fatores de similaridade

Como referido, o grau de similaridade pode ser usado como medida de semelhança em algoritmos de clustering. Neste trabalho será seguido o algoritmo de clustering descrito em [26], que usa o algoritmo k-means (ver secção 6.3). O pseudo-código é apresentado a seguir:

Algoritmo 6.4 Clustering k-means usando fatores de similaridade

Dado: Q séries temporais multivariadas (STMs), $\mathbf{X_1}, ..., \mathbf{X_q}, ..., \mathbf{X_Q}$, para serem agrupadas em k clusters:

1. Dividir aleatoriamente o conjunto de Q STMs em k clusters.
2. Seja $\mathbf{X_j^{(i)}}$ a j-ésima STM no i-ésimo cluster. Calcular o conjunto de dados agregados $\mathcal{X}_i$ $(i = 1, ..., k)$, para cada cluster como,

   $$\mathcal{X}_i = \left[(X_1^{(i)})^T ... (X_j^{(i)})^T ... (X_{Q_i}^{(i)})^T\right]^T$$

   onde Q_i é o número de STM em $\mathcal{X}_i$. Note que $\sum_{i=1}^{k} Q_i = Q$.
3. Para $q = 1, 2, ..., Q$ fazer:
   (a) Calcular a dissimilaridade entre $\mathbf{X_q}$ e cada um dos k conjuntos agregados $\mathcal{X}_i$ $(i = 1, ..., k)$ como

   $$d_{i,q} = 1 - SF_{i,q}$$

   onde $SF_{i,q}$ é o fator de similaridade entre a q-ésima STM e o i-ésimo cluster, como descrito na equação 6.1.

   (b) Mover a STM $\mathbf{X_q}$ para o cluster tal que $d_{i,q}$ é mínimo, ou seja, o cluster menos dissimilar de $\mathbf{X_q}$.
4. Calcular a média da dissimilaridade entre cada STM e o cluster a que pertence como

   $$J(k) = \frac{1}{Q} \sum_{i=1}^{k} \sum_{\mathbf{X_q} \in \text{ cluster } i} d_{i,q}$$

5. Se o valor de $J(k)$ é diferente do valor da iteração anterior, então voltar ao passo 2. Caso contrário, pára-se o algoritmo.

Cada STM é constituída por a observações por série e b séries, ou seja, uma STM é uma matriz $[a \times b]$. No passo 2 do algoritmo anterior, as STM são agregadas por linha. Por exemplo, seja $\mathbf{X_1}$ e $\mathbf{X_2}$ duas STM com a_1 e a_2 observações por série, respetivamente, então o conjunto agregado destas STM é de dimensão $[(a_1 + a_2) \times b]$, as primeiras a_1 linhas dizem respeito a $\mathbf{X_1}$ e as restantes a $\mathbf{X_2}$. Note que, para este método, as STM não necessitam ter o mesmo número de observações.

Neste contexto, o número de observações por série foi igual ao número de dias de observações disponíveis para a instalação e as séries foram o *Consumo Energético Diário*, o *Comprimento do Dia*, o *Ponto de Orvalho Máximo* e a *Humidade Mínima*. As variáveis categóricas não puderam ser usadas uma vez que não são séries temporais.

Como referido no capítulo 4, os valores anormais de consumo energético podiam ser uma característica da instalação. Outra característica do consumo foi a sua componente de tendência. Através do SSA (secção 3.3.2) foi possível extrair corretamente a tendência. Se as componentes sazonais fossem extraídas corretamente também poderiam ser usadas como características do consumo. Resta salientar que, como mencionado na secção 3.3.2, a tendência era enviesada pelos valores anormais, por isso foi calculada após substituição dos valores anormais (secção 4.4).

Assim, cada STM foi constituída por seis séries temporais: consumo energético, comprimento do dia, ponto de orvalho máximo, humidade mínima, valores anormais e tendência.

Decidiu-se usar a variável *Horário de Trabalho* nos dados $\mathbf{Y}$ uma vez que é uma variável por dia da semana. Assim, os dados $\mathbf{Y}$ para cada instalação formaram uma matriz $[1 \times 16]$, onde as 16 colunas dizem respeito à hora de inicio e fim de trabalho de cada dia da semana (Segunda - Domingo) e feriados.

Nos sites [21] e [22] estão disponíveis os códigos em `Matlab` para calcular os fatores de similaridade PCA e distância, respetivamente. Foi necessário adaptar o código para `R`. Resta salientar que este método é visto como um método de transformação de dados, pois transforma as STM em componentes principais.

6.3 Métodos de clustering

Existe um largo número de algoritmos de clustering descritos na literatura [14][28]. A escolha do algoritmo mais apropriado para um dado tratamento depende essencialmente do tipo de dados disponíveis e do propósito particular da aplicação. Se a análise de clusters é usada como uma ferramenta de exploração descritiva, é possível testar vários algoritmos com os mesmos dados e, com este tratamento, descobrir qual o que levaria a melhores resultados. De uma forma geral os métodos de clustering podem ser classificados nas seguintes categorias:

6.3.1 Métodos de partição

Dada uma base de dados de n objectos, o método de partição constrói k grupos de dados, onde cada grupo representa um cluster. Isto é, os dados são classificados em k grupos, que satisfazem os seguintes requisitos:

1. Cada grupo deve conter pelo menos um objeto;
2. Cada objeto deve pertencer a apenas um grupo.

Dado k, o número de grupos a construir, o método de partição constrói os grupos iniciais aleatoriamente e de seguida é usada uma técnica de reagrupamento iterativo que tenta melhorar a partição movendo os objetos de um grupo para outro. O critério geral para um bom agrupamento é que os objetos no mesmo cluster sejam semelhantes, enquanto os objetos de diferentes clusters sejam muito diferentes.
Existem vários algoritmos (para detalhes consultar [14][28]):

- Algoritmo *k-means*, onde cada cluster é representado pelo valor médio dos objetos no cluster;
- Algoritmo *k-medoids*, onde cada cluster é representado pelo objeto mais próximo ao centro do cluster.

O algoritmo *k-medoids*, também conhecido por *Partitioning Around Medoids (PAM)*, é mais robusto à presença de outliers uma vez que usa objetos dos dados como centroides em vez de usar a média que pode ser sujeita aos efeitos dos outliers.
Estes algoritmos estão disponíveis na package `cluster` do `R`, nas funções `kmeans()` e `pam()`, respetivamente.

6.3.2 Métodos hierárquicos

Os métodos hierárquicos criam uma decomposição hierárquica de um dado conjunto de objetos. Um método hierárquico pode ser caracterizado pela forma como inicia a decomposição hierárquica, de forma aglomerativa ou divisiva.
O *método aglomerativo*, também chamado *bottom-up*, inicia com cada objeto em um grupo separado. Sucessivamente, junta objetos ou grupos até que todos os grupos estão num só grupo (o nível mais elevado da hierarquia), ou até atingir uma certa condição de paragem.
A diferença entre os dois grupos de candidatos para a junção pode ser medida de várias formas:

- Método *single linkage*: a diferença entre dois grupos é medida pela **menor distância** entre quaisquer duas observações de cada grupo;
- Método *complete linkage*: a diferença entre dois grupos é medida pela **maior distância** entre quaisquer duas observações de cada grupo;
- Método *average linkage*: a diferença entre dois grupos é medida pela **distância média** entre quaisquer duas observações de cada grupo.

O *método divisivo*, também chamado *top-down*, inicia com todos os objetos num mesmo cluster. Em cada iteração sucessiva, um grupo é dividido em subgrupos mais pequenos, até que eventualmente só exista um objecto por grupo ou até que uma dada condição de paragem seja atingida.
Estes algoritmos estão disponíveis na package `cluster` do `R`, nas funções `agnes()` e `diana()`, respetivamente.

6.3.3 Métodos com base na densidade

A maioria dos métodos de partição de clusters baseia-se na distância entre objetos. Tais métodos apenas conseguem encontrar com eficiência clusters de forma esférica, tendo muita dificuldade em fazê-lo para clusters de forma arbitrária. Outros métodos de clustering têm sido desenvolvidos com base na noção de densidade.

A ideia principal é continuar o crescimento de um dado cluster na medida em que a densidade (número de objetos) na sua vizinhança tenha uma proximidade determinada. Isto é, para cada objeto do cluster, se na sua vizinhança com um determinado raio, existir algum objeto não pertencente ao cluster, este deve ser integrado no grupo.

Este método permite criar clusters de forma arbitrária com regiões densas separadas entre si por dados dispersos, que neste método são chamados de dados ruído. O algoritmo *DBSCAN (Density-Based Spatial Clustering of Aplication with Noise)*, é um método típico com base na densidade em que os clusters crescem de acordo com um dado limiar de densidade. Tal método está disponível na package `fpc` do `R`, na função `dbscan()`.

Segundo [28] e [14], existem outros métodos baseados na densidade mais eficientes que DBSCAN, como por exemplo, *OPTICS (Ordering Points to Identify the Clustering Structure)*, *DENCLUE (Clustering Based on Density Distribution Functions)* e *CLIQUE (CLustering In QUEst)*, contudo estes métodos não estão implementados no `R` e pela escassez do tempo não foi possível proceder à sua implementação.

6.4 Avaliação do clustering

Nas secções anteriores descreveram-se vários métodos de clustering. Nesta fase pergunta-se: *Qual o método que terá melhores resultados de clustering?* Em geral, a avaliação do clustering analisa a viabilidade da análise de agrupamento num conjunto de dados e a qualidade dos resultados produzidos por um método. Os principais passos da avaliação do clustering são:

- Determinar o número de grupos nos dados;
- Medir a qualidade do clustering.

6.4.1 Determinar o número de clusters

Determinar o número de grupos está longe de ser um problema fácil, pois o número "certo" é ambíguo. Descobrir o número "certo" de clusters depende muitas vezes da distribuição da forma e escala do conjunto de dados, assim como da resolução de clustering requerida pelo utilizador [14]. Existem vários métodos possíveis para estimar o número de clusters.
Um método simples, mas eficaz e popular [14], é definir o número de clusters como sendo cerca de $\sqrt{n/2}$ num conjunto de n objetos, na expectativa que cada cluster contém $\sqrt{2n}$ objetos [14].
O *método elbow* ("cotovelo") é baseado na observação de que o aumento do número de clusters pode ajudar a reduzir a soma das variâncias "dentro do cluster". Isto porque a existência de mais grupos permite capturar os grupos mais finos de objetos, que são mais similares entre si. No entanto, o efeito marginal de reduzir a soma de variâncias "dentro do cluster" pode baixar se muitos grupos são formados, porque a divisão de um conjunto coeso em apenas dois dá uma pequena redução. Consequentemente, uma heurística para a escolha do número certo de clusters é usar o ponto de viragem na curva da soma de variâncias "dentro do cluster" em relação ao número de grupos [14].
Tecnicamente, dado um número $k > 0$, podem-se formar k grupos do conjunto de dados em questão, utilizando um algoritmo de clustering, por exemplo k-means, e calcular a soma de variações "dentro do cluster" $var(k)$. Pode-se, então, construir a curva de var em relação a k. O primeiro (ou mais significativo) ponto de viragem da curva sugere que é o número "certo" de grupos.
Para o método *Fatores de Similaridade* pode-se usar $J(k)$, dissimilaridade média, em vez de $var(k)$ [26].

6.4.2 Medir a qualidade do clustering

Os métodos que medem a qualidade do clustering podem ser categorizados em dois grupos, caso exista ou não informação dos clusters reais. Os métodos extrínsecos, que comparam o clustering obtido com o real, e os métodos intrínsecos, que não utilizam os clusters reais.
Neste trabalho não existe qualquer informação sobre os grupos das instalações nem a

quantidade de grupos existentes. Assim, só se podem usar os *métodos intrínsecos*, métodos que avaliam a qualidade de um agrupamento, considerando o quão bem os clusters estão separados e compactados.

O *coeficiente de silhueta* (do inglês, *silhouette coefficient*) é tal medida de qualidade do clustering. Para um conjunto de dados D, de n objetos, suponha-se que D é agrupado em k grupos, $C_1, ..., C_k$. Para calcular o coeficiente é necessário [14][28]:

- Para cada objeto i, calcular a distância média de i a todos os objetos do grupo a que pertence, chamando a essa distância a_i;
- Para cada objeto i e qualquer grupo diferente do grupo que contém i, calcular a distância média de i a todos os objetos desse grupo. Obter o valor mínimo dessas distâncias, chamando-lhe b_i;
- O coeficiente de silhueta, s_i é igual a

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

O coeficiente toma valores entre -1 e 1. Idealmente, todos os objetos devem ter valores positivos, e a_i deve ser perto de zero, para $i = 1, ..., n$. Para obter a medida de qualidade do clustering, deve-se calcular a média de $s_i, i = 1, ..., n$, e quanto mais perto esse valor estiver de 1, melhor é o clustering.

O coeficiente de silhueta também pode ser usado para determinar qual o melhor número de clusters. Para vários valores de k, calcular o coeficiente de silhueta - o k que melhor coeficiente tiver é o número mais correto de clusters [14].

Este método está disponível na package `cluster` do `R`, na função `silhouette()`.

6.5 Resultados

Nas secções anteriores foram apresentados dois métodos para agrupar instalações segundo o consumo energético diário. O objetivo principal do estágio foi desenvolver um algoritmo capaz de agrupar um elevado número de instalações. No entanto, inicialmente

utilizou-se uma amostra de 14 instalações para validar o desempenho de cada método. Os consumos energéticos médios diários destas podem ser observados na figura 6.5.

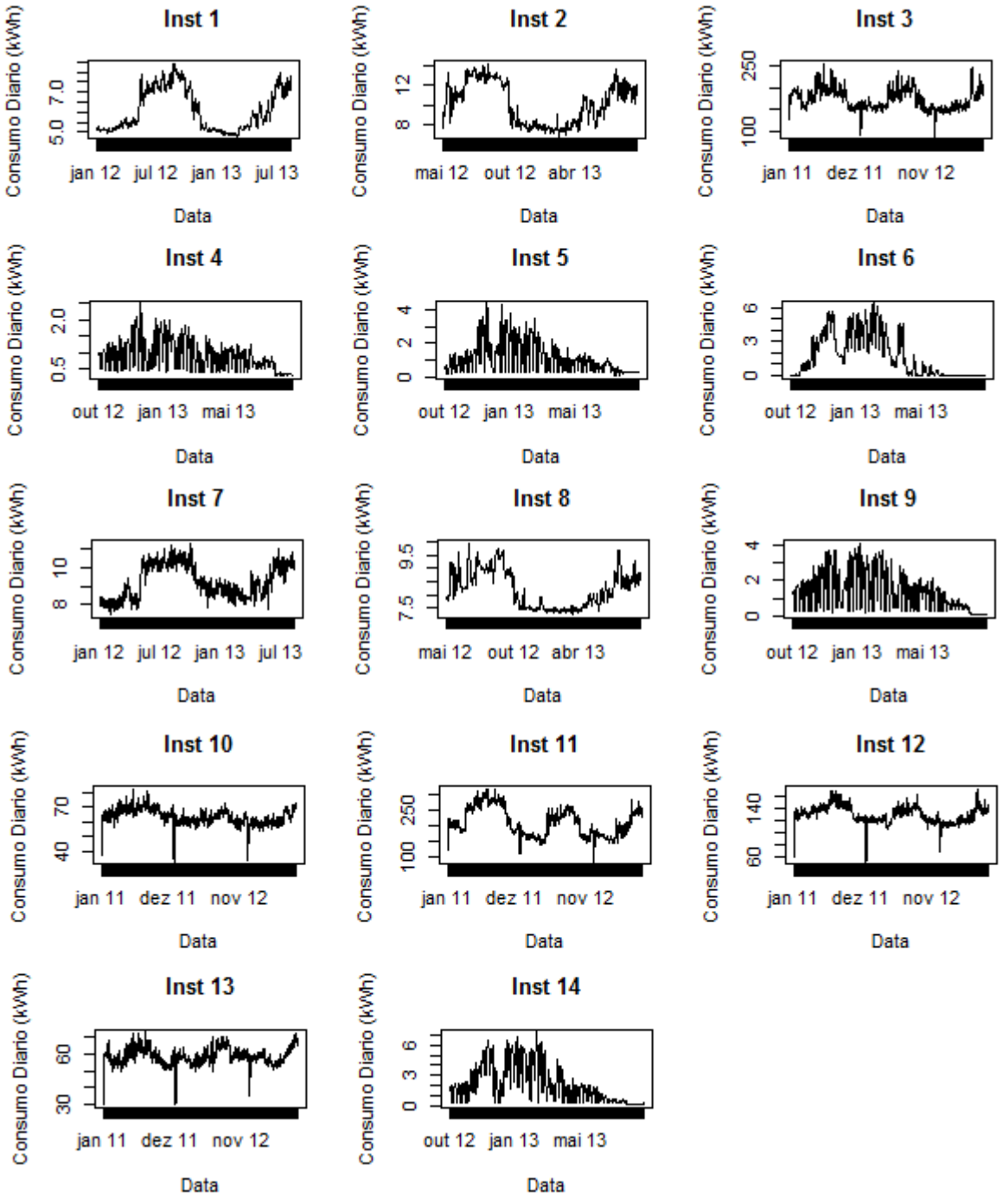

Figura 6.5: Consumos energéticos médios diários de 14 instalações

Shapelets

Como referido na secção 6.2.1, para aplicar o método de u-shapelets foi necessário nor-

malizar os consumos energéticos e, posteriormente, calcular o intervalo de comprimentos e a matriz distância e agrupar a matriz.

No algoritmo 6.3 foi necessário colocar no input um conjunto de tamanhos para se determinar os tamanhos mais adequados ao problema. Em testes iniciais utilizaram-se tamanhos pequenos ($TamS = (30, 40, 50, ..., 90)$, em dias) e concluiu-se que não eram adequados uma vez que o resultado do algoritmo foi $min = max = 90$ dias. Decidiu-se utilizar tamanhos múltiplos de mês, de 3 meses (= 90 dias) a 9 meses (= 273 dias). Apenas foi possível utilizar tamanhos até 9 meses porque para muitas instalações existia menos de 1 ano de observações (cerca de 50 % de instalações no Lote 1 apenas continham 11 meses de observações) e os tamanhos utilizados foram aplicados a todas as instalações do problema. Por este motivo, foi necessário criar um critério para selecionar apenas instalações que continham 9 meses de observações consecutivas.

Um problema que podia surgir com este método era estar a comparar o passado de uma instalação com um passado mais recente de outra instalação. Por exemplo, existiam instalações que tinham período temporal de 1 de Janeiro de 2010 a 31 de Janeiro de 2013 e era possível o consumo energético ter-se modificado bastante ao longo do tempo. Ora, o objetivo foi comparar consumos o mais recentes possível e por isso não faria sentido utilizar consumos muito antigos. Por este motivo, decidiu-se utilizar apenas as observações do ano e meio mais recente disponível para cada instalação. Criou-se um critério que selecionava as observações do ano e meio mais recente e posteriormente aplicou-se o critério anterior. Para a amostra de 14 instalações obteve-se comprimento mínimo igual a 90 dias (3 meses) e comprimento máximo igual a 153 dias (5 meses), logo os tamanho de shapelets utilizados foram 90, 123 e 153 dias (3, 4 e 5 meses).

O próximo passo foi então utilizar o algoritmo 6.1 para calcular a matriz distância. Este requer como input o número de clusters, mas como não se tinha informação do número correto de grupos utilizou-se as abordagens da secção 6.4.1. Segundo [14], o número de grupos para esta amostra foi cerca de $\sqrt{7} \approx 3$, então utilizou-se $k = 2, 3, 4$. Para cada k, calculou-se a matriz distância e aplicaram-se os algoritmos de clustering referidos na secção 6.3.

O clustering da matriz distância forma grupos de forma, ou seja, as instalações pertencentes a um mesmo grupo tinham consumos energéticos com formas semelhantes. Por este motivo, foi necessário agrupar as instalações dentro de um mesmo grupo por escala.

O número de grupos e o algoritmo de clustering mais adequados foram definidos através do coeficiente de silhueta (secção 6.4.2).

Para calcular o coeficiente de silhueta foi necessário calcular uma matriz de distâncias entre séries (consumos energéticos diários). O habitual é utilizar a distância euclidiana, contudo os consumos não tinham os mesmos períodos temporais. Decidiu-se utilizar duas distâncias:

- Distância Euclidiana, utilizando apenas os dias comuns a ambos consumos;
- DWT, admite séries de tamanhos diferentes.

Os coeficientes de silhueta obtidos utilizando o método shapelets apresentam-se na tabela 6.1.

Número de grupos de forma	Número de u-shapelets	Número total de grupos	Coeficiente de Silhueta
2	7 (4 de tamanho 153 e 3 de tamanho 123)	5	Euc: 0.59 DTW: 0.57
3	7 (4 de tamanho 153 e 3 de tamanho 123)	6	Euc: 0.35 DTW: 0.29
4	7 (4 de tamanho 153 e 3 de tamanho 123)	6	Euc: 0.35 DTW: 0.38

Tabela 6.1: Coeficientes de silhueta obtidos utilizando o método Shapelets aplicado a 14 instalações

Observando a tabela 6.1 concluiu-se que o melhor agrupamento (maior coeficiente de silhueta) obtinha-se utilizando 2 grupos de forma, formando no final 5 grupos para as 14 instalações. Para este caso, o melhor algoritmo de clustering foi k-means e as 14 instalações foram agrupadas como apresentadas na tabela 6.2.

Para uma melhor visualização do clustering, na figura 6.6 pode-se observar os consumos energéticos da figura 6.5 agrupados por forma (grupos 1 e 2) e por escala (grupos A, B, C, D e E).

Na figura 6.6 pode-se observar que o clustering obtido foi válido. Observou-se que dentro de cada grupo de forma (grupos 1 e 2) as instalações tinham curvas de consumo com forma bastante semelhante e formaram-se corretamente os grupos de escala.

Instalação	Grupo de forma	Grupo de escala
Inst 1	1	A
Inst 2	1	A
Inst 3	1	B
Inst 4	2	D
Inst 5	2	D
Inst 6	2	E
Inst 7	1	A

Instalação	Grupo de forma	Grupo de escala
Inst 8	1	A
Inst 9	2	D
Inst 10	1	C
Inst 11	1	B
Inst 12	1	B
Inst 13	1	C
Inst 14	2	E

Tabela 6.2: Resultado do agrupamento das 14 instalações através do método Shapelets

Utilizando a amostra de 14 instalações, apesar do coeficiente de silhueta não ser muito próximo de 1, concluiu-se visualmente que o método shapelets agrupou de forma adequada as instalações.

Fatores de similaridade

No método fatores de similaridade o único parâmetro do algoritmo é o vetor $\{\alpha_i\}$, o vetor dos coeficientes que combinam os fatores de similaridade. Diferentes combinações de factores de similaridade foram utilizadas para caracterizar a semelhança entre os conjuntos de dados e os resultados de clustering foram comparados para cada caso.

O procedimento de clustering k-means foi repetido para valores de $k = 2$ até $k = 7$ (uma vez que para o método shapelets obtiveram-se 6 grupos, decidiu-se utilizar este conjunto de ks) e as combinações utilizadas foram $C_1 = (0.45, 0.45, 0.1)$, $C_2 = (0.5, 0.5, 0)$, $C_3 = (0.45, 0.5, 0.05)$ e $C_4 = (0.5, 0.45, 0.05)$. Optou-se pelas combinações anteriores uma vez que tencionava-se agrupar séries com mesma orientação espacial e escala, ou seja, fator PCA e fator de distância com pesos semelhantes, enquanto que os dados Y são menos significativos para o clustering. O número de grupos e a combinação de fatores mais adequados foram definidos através da dissimilaridade média, $J(k)$ (secção 6.2.2 e 6.4).

As dissimilaridades médias obtidas utilizando o método fatores de similaridade apresentam-se na figura 6.7, para os diferentes valores de k e as diferentes combinações.

Como referido na secção 6.4.1, o primeiro (ou mais significativo) ponto de viragem da curva $J(k)$ sugere que é o número de grupos mais adequado. Observando a figura 6.7 concluiu-se que, para todas as combinações, o número mais adequado de grupos foi 3 e

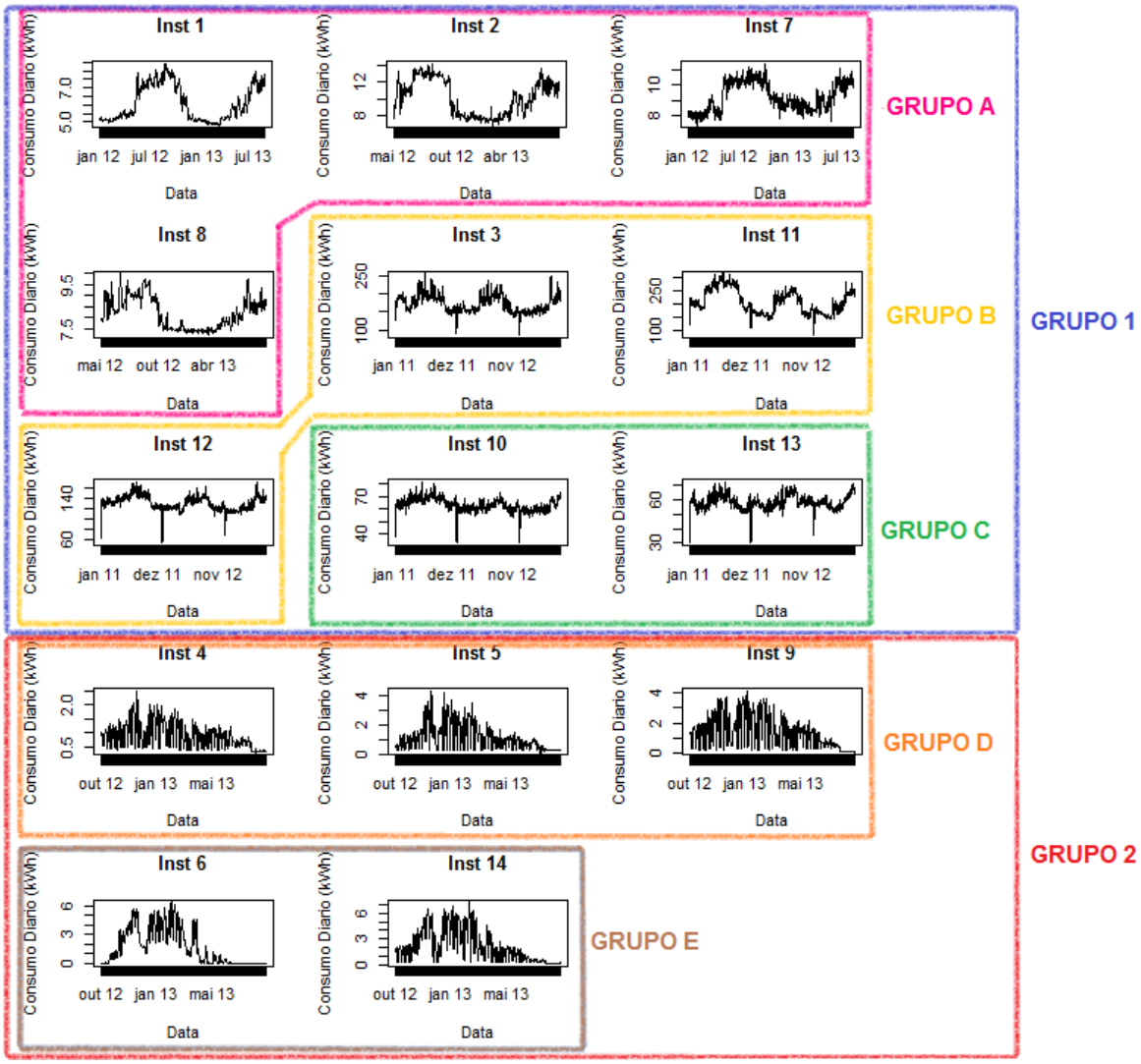

Figura 6.6: Agrupamento das 14 instalações de 6.5 pelo método shapelets. Obtêm-se 2 grupos de forma: Grupo 1 - a azul e Grupo 2 - a vermelho. Dentro do Grupo 1 obtêm-se 3 grupos de escala: Grupo A - a rosa; Grupo B - a amarelo e Grupo C - a verde. Dentro do Grupo 2 obtêm-se 2 grupos de escala: Grupo D - a laranja e Grupo E - a castanho

a combinação mais adequada foi $C_2 = (0.5, 0.5, 0)$, uma vez que a dissimilaridade média é mínima. Para este caso, as 14 instalações foram agrupadas como apresentadas na tabela 6.3.

Para uma melhor visualização do clustering, na figura 6.8 pode-se observar os consumos energéticos da figura 6.5 agrupados pelo método fatores de similaridade.

Analisando a figura 6.8, observou-se que o clustering obtido foi válido. No entanto, no grupo A existiam instalações com formas um pouco diferentes, concluindo-se que através deste método obteve-se um melhor agrupamento por escala do que por forma.

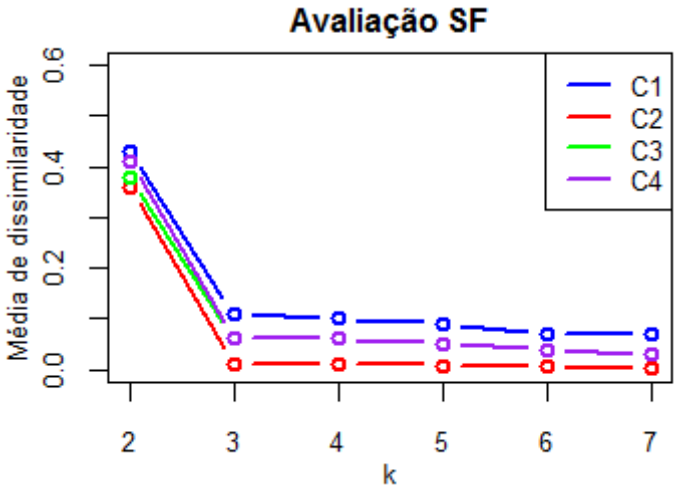

Figura 6.7: Gráfico das médias de dissimilaridade obtidas usando o método fatores de similaridade, para os diferentes valores de k e de combinações, na amostra de 14 instalações

Instalação	Grupo
Inst 1	A
Inst 2	A
Inst 3	B
Inst 4	A
Inst 5	A

Instalação	Grupo
Inst 6	A
Inst 7	A
Inst 8	A
Inst 9	A
Inst 10	C

Instalação	Grupo
Inst 11	B
Inst 12	B
Inst 13	C
Inst 14	A

Tabela 6.3: Resultado do agrupamento das 14 instalações através do método Fatores de Similaridade

Através dos métodos shapelets e fatores de similaridade obtiveram-se dois agrupamentos diferentes das 14 instalações. Para além da amostra utilizada ser pequena, não se pôde concluir qual dos métodos é mais adequado pois utilizam abordagens diferentes.

Utilizando o método shapelets, é possível analisar o agrupamento graficamente uma vez que se utilizam os consumos diários, contudo não se considerou a influência das variáveis externas.

Utilizando o método fatores de similaridade, não se deve avaliar o agrupamento visualizando os consumos diários, uma vez que este considera a influência das variáveis externas e não se tinha acesso à quantidade e forma desta influência em cada instalação.

Concluindo, ambos os métodos devem ser considerados e não existe uma medida comum para os comparar. Poder-se-ia utilizar o coeficiente de silhueta nos resultados do método

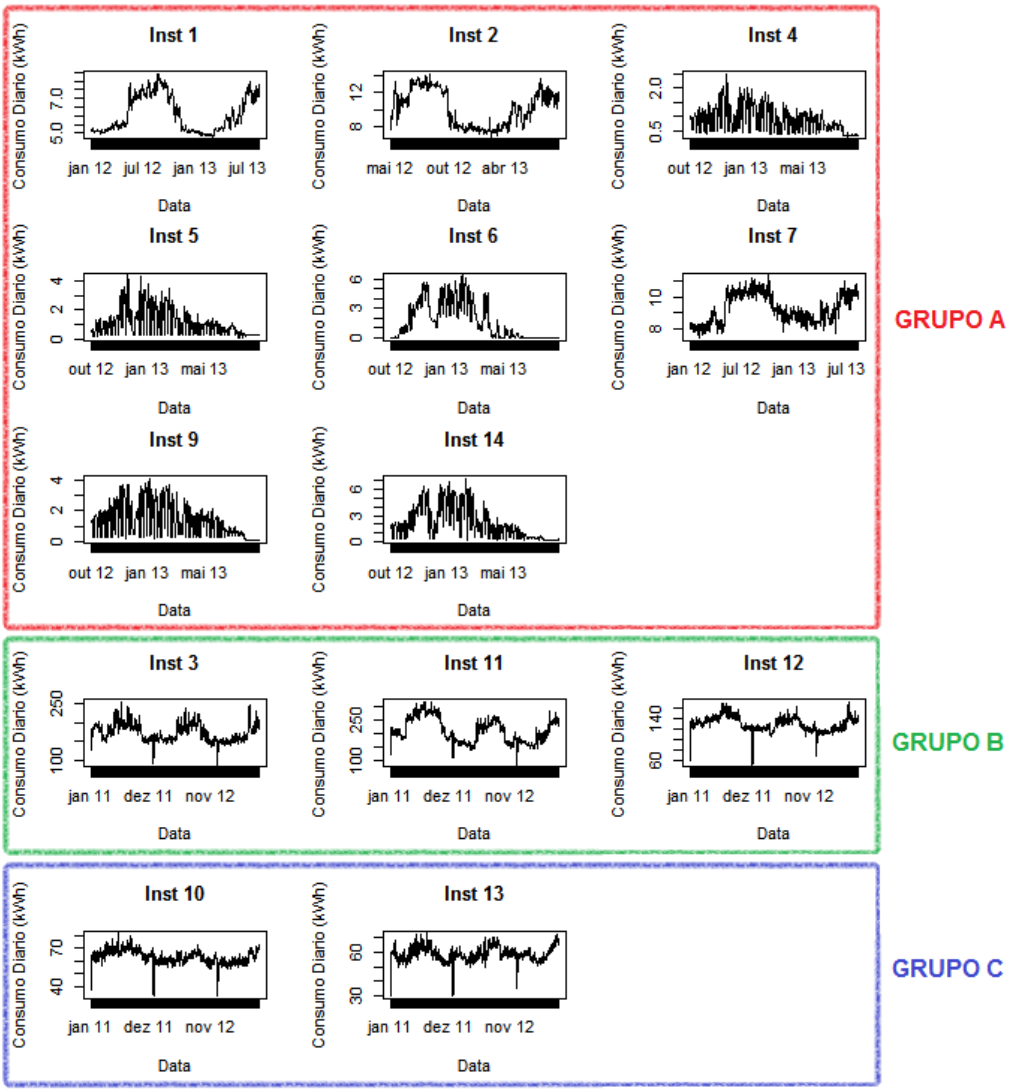

Figura 6.8: Agrupamento das 14 instalações de 6.5 pelo método fatores de similaridade. Obtém-se 3 grupos: Grupo A - a vermelho; Grupo B - a verde e Grupo C - a azul

fatores de similaridade mas, mais uma vez, não se estaria a considerar a influência das variáveis externas, não sendo esta a melhor medida para quantificar a qualidade do clustering.

Resta salientar que os resultados anteriores foram obtidos utilizando os consumos energéticos médios diários das instalações, no entanto, obtiveram-se os mesmos resultados usando os consumos energéticos totais diários.

Perfil característico

Após obter o agrupamento das instalações foi necessário construir uma curva para caracterizar cada um dos grupos. Decidiu-se utilizar a *mediana geométrica*. A mediana

geométrica de um conjunto discreto é o ponto que minimiza a soma de distâncias aos pontos amostrais do conjunto. É também conhecida como *mediana-L1*, *mediana espacial* ou *ponto de Torricelli* [30].

O perfil característico de um grupo é a curva que melhor se aproxima aos consumos energéticos diários das instalações contidas no grupo. Por este motivo, construiu-se o perfil característico aplicando a mediana geométrica aos consumos energéticos existentes no grupo para cada um dos dias, ou seja, para cada dia tinha-se um conjunto de valores, que dizem respeito aos consumos energéticos nesse dia de cada instalação, e calculou-se a mediana geométrica nesse conjunto, obtendo o valor do perfil característico nesse dia.

O período temporal do perfil foi a união dos períodos temporais disponíveis para cada instalação contida no grupo. Por exemplo, considere-se um grupo constituído por 3 instalações, I_1, I_2, I_3. Para as instalações I_1 e I_2 o período temporal disponível é 01-01-2012 a 31-12-2012 e para a instalação I_3 é 01-04-2012 a 31-12-2012. Então, o período temporal do perfil é 01-01-2012 a 31-12-2012 e para os primeiros 3 meses apenas se podem considerar os consumos das instalações I_1 e I_2 para calcular o perfil através da mediana geométrica. No estágio calcularam-se os perfis característicos de cada grupo obtido, para ambos os métodos.

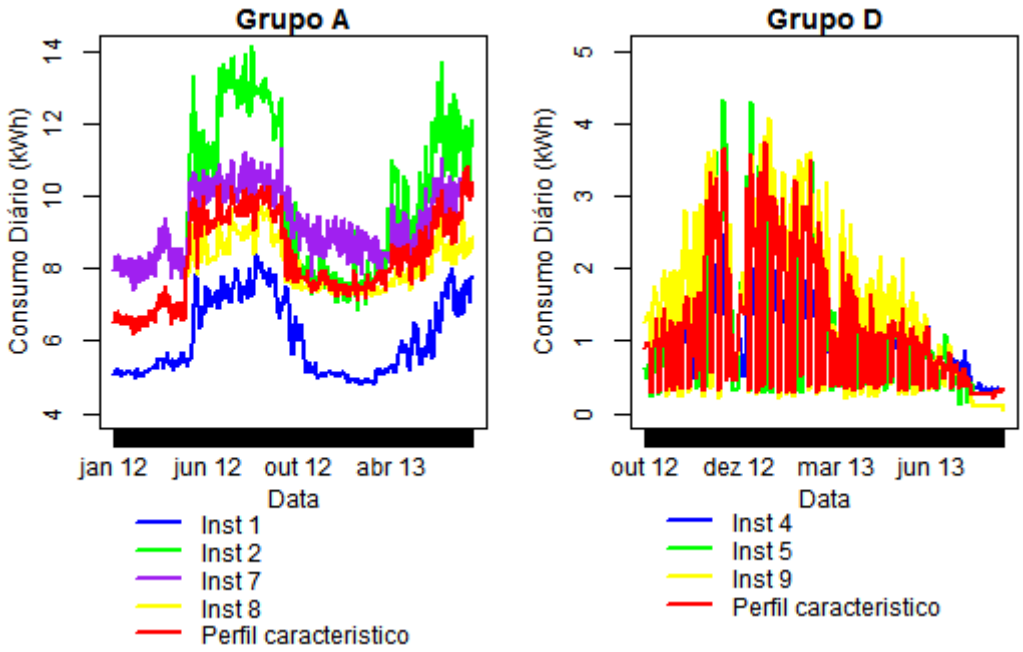

Figura 6.9: À esquerda: Consumos energéticos das instalações contidas no Grupo A e perfil característico do Grupo A (a vermelho). À direita: Consumos energéticos das instalações contidas no Grupo D e perfil característico do Grupo D (a vermelho)

Na figura 6.9 podem-se observar os perfis característicos (curva a vermelho) dos grupos A e D, obtidos pelo método shapelets. Observou-se que os perfis estimados através da mediana geométrica são adequados para caracterizar os grupos.

Agrupamento do Lote 1

Após utilizar a amostra de 14 instalações anteriores, utilizou-se o Lote 1 de instalações. Como mencionado no capítulo 5, o Lote 1 continha 98 instalações, mas com os critérios necessários selecionaram-se 86.

Segundo [14], o número de grupos para o Lote 1 foi cerca de $\sqrt{86/2} \approx 6$, então utilizou-se $k = 2, 3, 4, 5, 6, 7, 8$ para ambos os métodos.

Mais uma vez, a escolha do número de grupos mais adequado foi definido através do coeficiente de silhueta, quando aplicado o método shapelets (escolhendo também o melhor algoritmo de clustering), ou da dissimilaridade média, quando aplicado o método fatores de similaridade.

Shapelets

Aplicando o algoritmo 6.3 ao Lote 1 obteve-se comprimento mínimo das shapelets igual a 123 dias (4 meses) e comprimento máximo igual a 245 dias (8 meses), logo os tamanhos utilizados foram 123, 153, 183, 215 e 245 dias (4, 5, 6, 7 e 8 meses).

Para cada k, calculou-se a matriz distância, utilizando o algoritmo 6.1, e aplicaram-se os algoritmos de clustering referidos na secção 6.3, obtendo-se os grupos de forma. De seguida, agruparam-se as instalações dentro de um mesmo grupo por escala. Na tabela 6.4 apresentam-se os coeficientes de silhueta obtidos.

Observando a tabela 6.4 concluiu-se que o melhor agrupamento (maior coeficiente de silhueta) obtinha-se utilizando 2 grupos de forma, formando no final 4 grupos para as 86 instalações. Para este caso, o melhor algoritmo de clustering foi k-medoides.

Os grupos de forma obtidos tinham características semelhantes aos grupos de forma obtidos para a amostra de 14 instalações, ou seja, um dos grupos continha 37 instalações com consumos diários semelhantes (em forma) às instalações do grupo 1, incluindo as instalações do grupo 1, e o outro continha 49 instalações semelhantes às instalações do grupo 2, incluindo estas instalações (ver os grupos 1 e 2 na figura 6.6).

Número de grupos de forma	Número total de grupos	Coeficiente de Silhueta
2	4	Euc: 0.40; DTW: 0.38
3	5	Euc: 0.28; DTW: 0.25
4	6	Euc: 0.05; DTW: 0.05
5	7	Euc: 0.02; DTW: 0.02
6	8	Euc: -0.03; DTW: -0.5
7	12	Euc: 0.04; DTW: 0.02
8	12	Euc: 0.007; DTW: -0.08

Tabela 6.4: Coeficientes de silhueta obtidos utilizando o método Shapelets aplicado ao Lote 1

Cada um dos grupos de forma dividiram-se em dois grupos de escala, consumos altos e consumos baixos.

Observando os grupos obtidos, visualizando os gráficos dos consumos energéticos diários, concluiu-se que o agrupamento é válido, no entanto, não é possível apresentar os gráficos neste trabalho pela grande quantidade de instalações.

Fatores de similaridade

Para aplicar o método fatores de similaridade utilizou-se, mais uma vez, as combinações $C_1 = (0.45, 0.45, 0.1)$, $C_2 = (0.5, 0.5, 0)$, $C_3 = (0.45, 0.5, 0.05)$ e $C_4 = (0.5, 0.45, 0.05)$.

As dissimilaridades médias obtidas apresentam-se na figura 6.10, para os diferentes valores de k e as diferentes combinações.

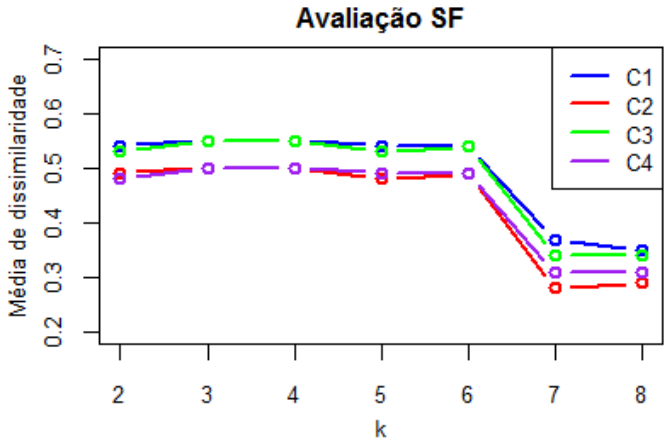

Figura 6.10: Gráfico das médias de dissimilaridade obtidas usando o método fatores de similaridade, para os diferentes valores de k e de combinações, no Lote 1

Como referido anteriormente, o primeiro (ou mais significativo) ponto de viragem da curva $J(k)$ sugere que é o número de grupos mais adequado. Observando a figura 6.10 concluiu-se que, para todas as combinações, o número mais adequado de grupos foi 7 e a combinação mais adequada foi $C_3 = (0.45, 0.5, 0.05)$, uma vez que a dissimilaridade média é mínima. Para este caso, obtiveram-se dois grupos com 6 instalações, um com 49 instalações, um com 11 instalações, um com 9 instalações, um com 3 instalações e um com 2 instalações.
Observando os grupos obtidos, visualizando os gráficos dos consumos energéticos diários, concluiu-se o mesmo que para a amostra de 14 instalações, para este método. Existia num mesmo grupo instalações com consumos energéticos com formas diferentes.

Os métodos utilizados devolvem bons resultados, no entanto demoram bastantes tempo. Para aplicar o método das shapelets ao Lote 1 foram necessárias cerca de 34 horas e para aplicar o método fatores de similaridade foram necessárias cerca de 38 horas.

Capítulo 7

Conclusão e trabalho futuro

Neste trabalho foi apresentado o desenvolvimento do estágio curricular *Clustering de instalações, com (re)definição de segmentos em função do comportamento energético*.
Para alcançar o objetivo final foi necessário ultrapassar alguns problemas:

1. Compreensão dos consumos energéticos;
2. Extrair os fatores externos correlacionados com o consumo;
3. Seleção de instalações a utilizar;

No primeiro ponto foi necessário realizar algumas análises aos consumos energéticos. Foi possível perceber que existiam várias variáveis que podiam explicar o consumo energético de cada instalação e características de séries temporais, como a tendência e componentes sazonais. Concluiu-se que os diagramas de carga teriam que ser agregados e que os consumos têm aspeto complexo, não sendo possível estimar as componentes sazonais através dos métodos utilizados.
No segundo ponto foi necessário dividir Portugal Continental por regiões para extrair as variáveis mais significativas do consumo. Utilizaram-se os métodos *Backward*, *Random Forests* e *Correlação* e concluiu-se que as variáveis significativas são: *Ano*, *Feriado*, *Estação*, *Dia da Semana*, *Comprimento do Dia*, *Ponto de Orvalho Máximo* e *Humidade Mínima*.
No terceiro ponto criou-se uma lista de critérios necessários para selecionar instalações para o problema. Concluiu-se que o processamento dos dados (agregação, junção de

tabelas, seleção) devia ser realizado em `SQL Server`.

Por fim, para agrupar as instalações utilizaram-se dois métodos. Para cada método existiam parâmetros (escolha do número de grupos, do algoritmo de clustering, do conjunto de tamanhos de shapelets - método shapelets e do conjunto de combinações de fatores - método fatores de similaridade). A escolha destes parâmetros foi realizada através de duas medidas: *coeficiente de silhueta* - método shapelets e *dissimilaridade média* - método fatores de similaridade. Ambos os métodos são válidos para o agrupamento das instalações, contudo obtiveram-se grupos diferentes.

Não foi possível definir qual dos métodos é melhor. Para além de não existir uma medida comum, são abordagens diferentes e não se deve descartar nenhuma. O método fatores de similaridade utiliza os consumos energéticos diários e a influência das variáveis externas, enquanto que o método shapelets não considera esta influência.

Os métodos apenas foram aplicados ao Lote 1 de instalações, porque demorou bastante tempo a retornar os resultados.

Existe ainda um conjunto de pontos neste trabalho que podem ser realizados para melhorar os resultados, sendo estes:

- Otimizar os algoritmos dos métodos shapelets e fatores de similaridade e/ou utilizar outro software (por exemplo `Matlab`, verificou-se que o `Matlab` é mais rápido que o `R`, contudo não estava disponível no estágio);
- Retirar a influência das variáveis externas. Como mencionado, duas instalações podem ter consumos relacionados com o trabalho semelhantes, mas devido a estarem situadas em locais diferentes, podem ter consumos totais diferentes;
- Detetar instalações anormais. Instalações que não são semelhantes a nenhuma outra devem ser consideradas como anormais;
- Classificar instalações. Após clustering obtêm-se um conjunto de grupos. Caso mais instalações adiram ao programa é necessário re-agrupar as instalações. Para não fazer clustering novamente, é necessário identificar o grupo (dos que já existem) das novas instalações;
- Prever. Prevendo o perfil característico de cada grupo é possível ter uma estimativa dos consumos futuros das instalações pertencentes ao respetivo grupo.

Bibliografia

[1] *Academia de Astrologia, Localidades de Portugal Continental*. Disponível em: `http://www.academiadeastrologia.com/recursos/coordenadas/portugal.htm` Consultado em Novembro de 2013, Março e Abril de 2014.

[2] Bagnall, A. J.; Janacek, G. J.; *Clustering Time Series from ARMA Models with Clipped Data*, 2004.

[3] Breiman, L; *Manual On Setting Up, Using, And Understanding Random Forests V3.1*, 2002. Disponível em: `http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf` Consultado em: Março de 2014.

[4] Brockwell, Peter J.; Davis,Richard A.; *Introduction to Time Series and Forecasting* Second Edition, 2002, 1996 Springer-Verlag New York, Inc.

[5] *Calendário 365: Épocas / Estações do Ano*. Disponível em: `http://www.calendario-365.com.br/epocas-estacoes-do-ano.html` Consultado em Março de 2014.

[6] Casella, George; Berger, Roger L.; *Statistical Inference* Second Edition, Duxbury Advanced Series, 2002.

[7] Cressie, Noel; Wikle, Christopher K.; *Statistics for Spatio-Temporal Data*, 2011 by John Wiley & Sons, Inc.

[8] *EXPLAINING DEWPOINT AND RELATIVE HUMIDITY TO THE PUBLIC*. Disponível em: `http://www.theweatherprediction.com/habyhints/190/` Consultado em Abril de 2014.

[9] Gan, Guojun; Ma, Chaoqun; Wu, Jianhong; *Data Clustering, Theory, Algorithms and Applications*, 2007.

[10] Genuera, Robin; Poggi, Jean-Michel; Tuleau-Malotc, Christine; *Variable Selection using Random Forests*, 2012.

[11] Golyandina, Nina; Korobeynikov, Anton; *Basic Singular Spectrum Analysis and Forecasting with R*, 2013.

[12] Golyandina, Nina; Nekrutkin, Vladimir; Zhigljavsky, Anatoly A.; *Analysis of Time Series Structure: SSA and Related Techniques*, 2011.

[13] Guilford, J. P.; Fruchter, B.; *Fundamental statistics in psychology and education*, Tokyo: McGraw-Hill Kogakusha, LTD, 1973.

[14] Han, Jiawei; Kamber, Micheline; Pei Jian, *Data Mining: Concepts and Techniques* Elsevier Inc. All rights reserved, 2012.

[15] Hills, Jon; Lines, Jason; Baranauskas, Edgaras; Mapp, James; Bagnall, Anthony; *Classification of time series by shapelet transformation*, Springer, 2013.

[16] Hollander, M.; Wolfe, D.A.; *Nonparametric Statistical Methods* New York: Wiley, 1973.

[17] *Instituto Português do Mar e da Atmosfera, Normais Climatológicas*. Disponível em: `http://www.ipma.pt/pt/oclima/normais.clima/` Consultado em Março de 2014.

[18] Jönsson, Per; Wohlin, Claes; *An Evaluation of k-Nearest Neighbour Imputation Using Likert Data* Proceedings 10th International Symposium pn Software Metrics, pp. 108-118, Chiaco, USA, Setembro 2004.

[19] Keogh, E.; Kasetty, S.; *On the need for time series data mining benchmarks: A survey and empirical demonstration*, 2002. Disponível em `http://www.cs.ucr.edu/~eamonn/selected_publications.htm`.

[20] Krzanowski, W. J.; *Between-Groups Comparison of Principal Components*, American Statistical Association, 1979.

[21] *Matlab Central: PCA similarity factor*. Disponível em: `http://www.mathworks.com/matlabcentral/fileexchange/45478-pca-similarity-factor/content/similarity_pca.m` Consultado em Abril de 2014.

[22] *Matlab Central: similarity factor*. Disponível em: `http://www.mathworks.com/matlabcentral/fileexchange/45479-similarity-factor/content/simil_dist.m` Consultado em Abril de 2014.

[23] Murteira, Bento J. F.; Müller, Daniel A.; Turkman, K.Feridun; *Análise de Sucessões Cronológicas*, 2000.

[24] Seem, John E.; *Using intelligent data analysis to detect abnormal energy consumption in buildings*, Johnson Controls, Inc., 507 East Michigan Street, Milwaukee, WI 53202, USA, 2006.

[25] Shumway, Robert H.; Stoffer, David S.; *Time Series Analysis and Its Applications With R Examples* Third Edition, Springer Science+Business Media, LLC 2011.

[26] Singhal, Ashish; Seborg, Dale E.; *Clustering multivariate time-series data*, John Wiley & Sons, Ltd, 2006. Disponível em `http://www.me.ucsb.edu/~ceweb/faculty/seborg/pdfs/Singhal_JChemometrics.pdf`.

[27] *Sun or Moon Rise/Set Table for One Year*. Disponível em: `http://aa.usno.navy.mil/data/docs/RS_OneYear.php` Consultado em Novembro de 2013, Março e Abril de 2014.

[28] Tan; Ateinbach; Kumar; *Introduction to Data Mining* (1ª edição) Pearson Education Limited, 2014.

[29] Torgo, Luís; *Data Mining with R: Learning with Case Studies* Taylor and Francis Group, LLC, 2011.

[30] Vardi, Y.; Zhang, C.-H.; *The multivariate L1-median and associated data depth*, 1991.

[31] Verzani, John; *Using R for Introdutory Statistics*, 2005 by Chapman & Hall/CRC Press.

[32] Verzani, John; *Using R for Introdutory Statistics*, 2005 by Chapman & Hall/CRC Press, pág. 238-244.

[33] Wang, Xiaozhe; Smith, Kate A.; Hyndman, Rob; Alahakoon, Damminda; *A Scalable Method for Time Series Clustering*.

[34] *Weather Underground*. Disponível em: `http://www.wunderground.com` Consultado em Novembro de 2013, Março e Abril de 2014.

[35] Webpage do artigo [38]: https://sites.google.com/site/icdmclusteringts/, password: ICDM2012.

[36] *WORLD MAPS OF KÖPPEN-GEIGER CLIMATE CLASSIFICATION, Climate shifts*. Disponível em: `http://koeppen-geiger.vu-wien.ac.at/shifts.htm` Consultado em Abril de 2014.

[37] Ye, L.; Keogh, E.; *Time Series Shapelets: a new primitive for Data Mining*, 2009. Disponível em `http://www.cs.ucr.edu/~eamonn/selected_publications.htm`.

[38] Zakaria, Jesin; Mueen, Abdullah; Keogh, Eamonn; *Clustering Time Series using Unsupervised-Shapelets*, 2012. Disponível em `http://www.cs.ucr.edu/~eamonn/selected_publications.htm`.

Apêndice A

Análise Espectral Singular

A técnica de SSA básica divide-se em duas etapas complementares: decomposição e reconstrução da série temporal. Cada etapa é composta por dois passos, constituindo os quatro passos da técnica: *Embedding*, *Decomposition*, *Grouping* e *Diagonal Averaging*. Estes passos serão descritos nas próximas subsecções.

A.1 Decomposição

Na etapa decomposição, a série temporal inicial é decomposta numa soma de poucas subséries, de modo que cada subsérie possa ser identificada e interpretada como uma componente principal (padrão temporal).

Mergulho (Embedding)

Considere uma série temporal unidimensional real e não nula, $X_N = (x_1, ..., x_N)$ de comprimento N.

Inicialmente, a série X_N unidimensional é representada como uma série multidimensional, de dimensão L. L chama-se o *window length* ou comprimento da janela, e é o único parâmetro da etapa decomposição - representa a dimensão do *espaço de mergulho*, $\mathbb{R}^L$, e portanto o número de componentes em que a série original é decomposta. Tal parâmetro deve ser um valor inteiro entre $1 < L < N$ [12][11].

A série temporal multidimensional, é uma sequência de $K = N - L + 1$ vetores (lags)

$X_i = (x_i, ..., x_{i+L-1})^\top$ de comprimento L, constituídos por elementos da série original X_N. Com estes vetores construímos uma matriz $L \times K$, apresentada na expressão A.1, denominada *matriz trajetória* [12][11].

$$\mathbf{X} = [X_1 : ... : X_K] = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_K \\ x_2 & x_3 & x_4 & \cdots & x_{K+1} \\ x_3 & x_4 & x_5 & \cdots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \cdots & x_N \end{pmatrix} \tag{A.1}$$

A matriz trajetória $\mathbf{X}$ é uma *matriz de Hankel* (matriz simétrica com entradas constantes ao longo das diagonais perpendiculares à diagonal principal). Note que, a matriz transposta $\mathbf{X}^\top$ é também uma matriz trajetória da série X_N, contudo o comprimento da janela é igual a K, em vez de L.

Em suma, o passo embedding é considerado como um mapeamento que transforma uma série unidimensional X_N numa série multidimensional $\mathbf{X}$.

Decomposição (Decomposition)

No passo da *decomposição do valor singular* (Singular value decomposition = SVD), é realizada a decomposição da matriz trajetória $\mathbf{X}$ numa soma de matrizes elementares.

Seja $\mathbf{S}$ o produto entre a matriz trajetória e a sua transposta, $\mathbf{S} = \mathbf{X}\mathbf{X}^\top$. Ao realizar a SVD da matriz $\mathbf{S}$, obtém-se os seus valores próprios, que podem ser ordenados de acordo com as suas magnitudes ($\lambda_1 \geq ... \geq \lambda_L \geq 0$), e os correspondentes vetores próprios $U_1, ..., U_L$, ortogonais e normalizados [12].

A transformação dada por $V_i = \dfrac{\mathbf{X^T}U_i}{\sqrt{\lambda_i}}$, $i = 1, ..., L$, permite escrever a matriz trajetória na forma [12]:

$$\mathbf{X} = \mathbf{X_1} + ... + \mathbf{X_L} = \sum_{i=1}^{L} P_i Q_i^\top \tag{A.2}$$

onde $P_i = U_i$ e $Q_i = \sqrt{\lambda_i} V_i = \mathbf{X}^\top P_i$, e $\mathbf{X_i} = \sqrt{\lambda_i} U_i V_i^\top$ representa uma matriz elementar. O triplo $(\sqrt{\lambda_i}, P_i, Q_i)$ é chamado de i-ésimo triplo próprio (*eigentriple*) da matriz $\mathbf{X}$; $\sqrt{\lambda_i}$ é conhecido como valor singular da matriz $\mathbf{X}$ e o conjunto $\{\sqrt{\lambda_i}\}$ representa o espectro de $\mathbf{X}$. Os vetores U_i e V_i são identificados como vetores singulares da matriz $\mathbf{X}$.

A.2 Reconstrução

Nesta etapa escolhe-se os grupos mais semelhantes para a formação das componentes e posteriormente reconstrói-se a série temporal.

Agrupamento (Grouping)

O procedimento *agrupamento* (*grouping*), tem como principal objetivo a distinção das componentes aditivas da série temporal em termos de matrizes separáveis. Por outras palavras, neste passo, identifica-se as componentes mais correlacionadas entre si, com o intuito de organizá-las num mesmo grupo. Seja $d = \max\{j : \lambda_j \neq 0\}$. Uma vez obtida a expansão A.2, o passo *grouping* particiona o conjunto de índices $\{1, ..., d\}$ em subconjuntos disjuntos $I_1, ..., I_m$, o que corresponde à representação [12]

$$\mathbf{X} = \mathbf{X}_{I_1} + ... + \mathbf{X}_{I_m} = \sum_{p=1}^{m} \mathbf{X}_{I_p} \tag{A.3}$$

onde $\mathbf{X}_{I_1}, ..., \mathbf{X}_{I_m}$ são conhecidos como *matrizes resultantes*. Cada matriz resultante é obtida a partir da soma de matrizes elementares num conjunto particular de índices I_p, isto é, $\mathbf{X}_{I_p} = \sum_{i \in I_p} \mathbf{X}_i, p = 1, ..., m$.

Na expressão A.3 tem-se uma nova decomposição de matrizes, que é denominada como *decomposição agrupada*. A organização dos índices $\{1, ..., d\}$ em grupos $I_1, ..., I_m$, deve ser de tal forma que as suas matrizes satisfaçam A.3 e sejam próximas de matrizes de Hankel, para que assim possam ser transformadas em matrizes trajetórias de séries que são separáveis pela expansão A.2.

O procedimento que consiste em escolher os conjuntos $I_1, ..., I_m$ é chamado *eigentriple grouping*. Este fornece o ultimo parâmetro da técnica SSA, cuja escolha utiliza o conceito de separabilidade, que é apresentado na secção A.3.

Mediação diagonal (Diagonal Averaging)

A operação realizada neste último passo transforma cada matriz resultante numa componente aditiva da série original, ou seja, cada matriz da decomposição agrupada A.3 é convertida numa nova série de tamanho N. Isto permite recuperar uma série unidimensional, considerada como uma aproximação da série inicial.

A transformação das matrizes resultantes em séries unidimensionais, consegue-se aplicando o operador linear de Hankelização ($\mathcal{H}$). Tal operador consiste dos passos seguintes

- transforma uma matriz arbitrária Y numa matriz de Hankel, calculando as médias ao longo das diagonais perpendiculares à diagonal principal de Y
- associa uma série unidimensional a esta matriz de Hankel

O operador $\mathcal{H}$ calcula pois as médias dos valores nas diagonais perpendiculares à diagonal principal das matrizes $\mathbf{X}_{I_p}$, para $p = 1, ..., m$. Considere Y uma matriz $L \times K$ cujos elementos são y_{ij}, $1 \leq i \leq L$ e $1 \leq j \leq K$, em que $L \leq K$. O resultado da aplicação do operador $\mathcal{H}$ na matriz Y é a matriz de Hankel $\mathcal{H}Y$, que é a matriz trajetória da série obtida no resultado de *diagonal averaging* [12].
Seja $A_s = \{(l,k) : l + k = s + 1, 1 \leq l \leq L, 1 \leq k \leq K\}$. Os elementos $\tilde{y}_s$ da matriz $\mathcal{H}Y$ são obtidos usando:

$$\tilde{y}_s = \sum_{(l,k)\in A_s} y_{lk}/|A_s|$$

Esta expressão corresponde à média dos elementos y_{ij} das diagonais perpendiculares à diagonal principal de Y. A série resultante $\tilde{\mathbf{X}}^K = \left(\tilde{\mathbf{X}}_1^{(K)}, ..., \tilde{\mathbf{X}}_N^{(K)}\right)$ é produzida ao aplicar o procedimento de Hankelização em cada matriz resultante $\mathbf{X}_{I_p}$, $p = 1, ..., m$. Ou seja, se o operador $\mathcal{H}$ é aplicado a todas as componentes da matriz A.3, obtém-se a última expansão da técnica SSA:

$$\mathbf{X} = \tilde{\mathbf{X}}_{I_1} + ... + \tilde{\mathbf{X}}_{I_m} = \sum_{p=1}^{m} \tilde{\mathbf{X}}_{I_p}$$

em que, $\tilde{\mathbf{X}}_{I_1} = \mathcal{H}\mathbf{X}_{I_1}$. Portanto, a série inicial $X_N = (x_1, ..., x_N)$ é decomposta na soma de m séries reconstruídas $x_n = \sum_{k=1}^{m} \tilde{x}_n^{(K)}$, $n = 1, ..., N$ [12].
Detalhes teóricos sobre o operador $\mathcal{H}$ podem ser encontrados na seção 6.2 de [12].

A.3 Informações Adicionais

A escolha dos parâmetros da técnica SSA depende do objetivo da análise e das informações preliminares sobre a série temporal. Na literatura especializada há algumas informações complementares que auxiliam na escolha do parâmetro L, para atingir uma boa separabilidade das componentes, assim como na forma de agrupamento, que possibilita a

identificação adequada dos triplos próprios mais importantes para a extração das componentes (tendência, sazonalidade e ruído) [12].

A.3.1 Separabilidade

Os procedimentos *Decomposição* e *Agrupamento* apoiam-se na propriedade denominada *separabilidade* (das diferentes componentes da série temporal). Sendo assim, a decomposição e reconstrução da série X_N tem êxito se as suas componentes aditivas são separáveis umas das outras.

Avalia-se a qualidade da separabilidade através de uma medida natural de dependência entre subséries. Por *"correlação ponderada"* ou *"w-correlação"*, entende-se a função que quantifica a dependência linear entre duas subséries $X_N^{(1)}$ e $X_N^{(2)}$, conforme definido por

$$\rho_{1,2}^{(w)} = \frac{\left\langle X_N^{(1)}, X_N^{(2)} \right\rangle_w}{\left\| X_N^{(1)} \right\|_w \left\| X_N^{(2)} \right\|_w}$$

onde a norma da i-ésima subsérie é dada por $\left\| X_N^{(1)} \right\|_w = \sqrt{\left\langle X_N^{(1)}, X_N^{(2)} \right\rangle_w}$ e o produto interno entre um par de subséries é $\left\langle X_N^{(1)}, X_N^{(2)} \right\rangle_w = \sum_{c=1}^{N} w_c x_c^{(i)} x_c^{(j)}, i, j = 1, 2$ sendo que, os ponderadores w_c são representados por $w_c = \min\{c, L, N-c\}$, e assume-se que $L \leq N/2$.

Caso o valor absoluto da correlação ponderada seja pequeno, tem-se que as duas séries são quase ortogonais. Desta forma, uma correlação ponderada entre duas componentes reconstruídas igual a zero ($\rho_{1,2}^{(w)} = 0$) significa que estas componentes são separáveis. Por outro lado, se o valor absoluto da correlação ponderada é alto, então as séries não são (tão) separáveis. Ou seja, valores de $\rho_{1,2}^{(w)} \approx 0$ indicam que as componentes devem ser reunidas num mesmo grupo, correspondendo à mesma componente na decomposição SSA.

A.3.2 Comprimento da janela (L)

O primeiro passo da técnica SSA exige a entrada de um valor para o parâmetro principal L, a sua escolha inadequada dificulta um bom agrupamento e diminui a precisão na aproximação da série.

Segundo [12], valores de $L \leq N/2$ são suficientes e a decomposição da série temporal torna-se mais detalhada à medida que o comprimento da janela aumenta. Todavia, ao trabalhar com séries temporais periódicas é necessário ter atenção à escolha de L. Para alcançar separabilidade suficiente das componentes, sugere-se utilizar um comprimento L proporcional ao período de sazonalidade dos dados.

Para uma série com uma estrutura complexa, um comprimento L muito grande pode produzir uma decomposição indesejável das componentes da série. Esta é uma situação desagradável, uma vez que uma redução significativa de L pode conduzir a uma má qualidade de separação (aproximada) das componentes. Na prática, nestas circunstâncias, utiliza-se um valor de L mais pequeno para a extração da tendência, uma vez que esta é uma curva mais suave, e de seguida aplica-se a técnica SSA com um comprimento L maior para extrair as restantes componentes. Para mais detalhes, ver [12] onde se ilustra a influência de L através de um exemplo.

A.3.3 Escolha dos triplos próprios

O segundo parâmetro da técnica SSA é estrutural, ou seja, refere-se à forma de agrupamento dos triplos próprios. O processo de formação dos grupos também tem uma grande influência na decomposição da série - espera-se que um agrupamento adequado leve a uma boa separação (aproximada) das componentes da série temporal.

O comportamento dos vetores e valores singulares da SVD da matriz trajetória, assim como a análise das correlações ponderadas, auxiliam no agrupamento adequado dos triplos próprios. Algumas indicações para o agrupamento são apresentadas a seguir.

Valores singulares

O comportamento dos valores singulares pode ser observado através de um gráfico em que os seus L valores são apresentados de forma decrescente de magnitude ($\sqrt{\lambda_1} \geq ... \geq \sqrt{\lambda_L}$), exibindo o espectro de $\mathbf{X}$.

O primeiro valor singular, aquele com maior valor absoluto, está sempre associado à componente de tendência. A presença de valores próprios com valores suficientemente próximos, podem ser identificados como um “par” que pode ser associado a uma componente sazonal da série. Teoricamente, uma série puramente residual produz uma sequência lentamente decrescente de valores singulares. Assim, se um ruido é adicionado a um sinal,

composto por triplos próprios com valores singulares altos, então observa-se uma quebra no espectro de $\mathbf{X}$ [12].

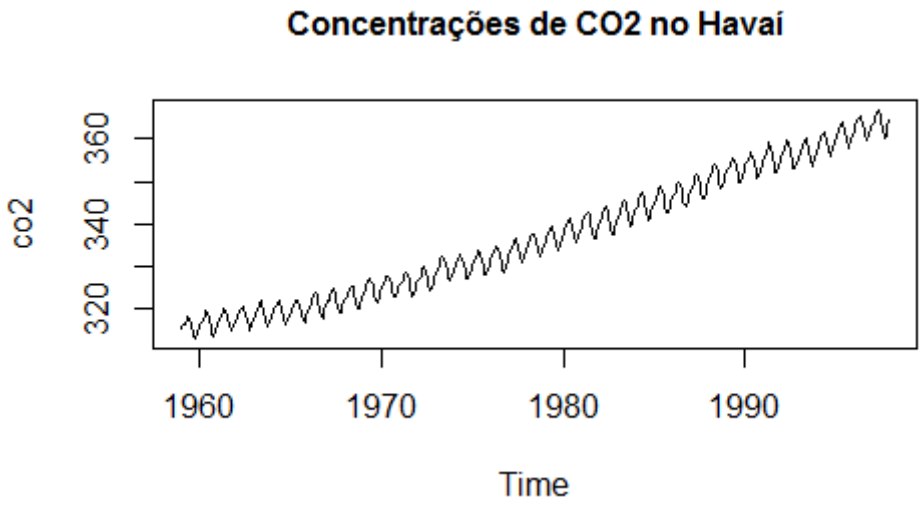

Figura A.1: Nº mensal de concentrações atmosféricas de CO2 no Havaí entre 1959 e 1997

Com o propósito de exemplificação, considere a série temporal disponível na linguagem `R`, "CO2". A série retrata as concentrações atmosféricas de CO2 de Mauna Loa Observatory, no Havaí, e contém 468 observações mensais de 1959 a 1997 (Keeling e Whorf, 1997). Observando a figura A.1, as principais características da série são a tendência linear crescente e uma sazonalidade aditiva.

Na package `Rssa` do `R` existe o comando `ssa(data,L)` que realiza a etapa *decomposição* da técnica SSA sobre os dados `data` e necessita do parâmetro `L` para o embedding. Uma vez que o objetivo deste exemplo é o agrupamento, será utilizado o comprimento de janela segundo [11], $L = 120$.

Na figura A.2 apresenta-se o espectro de $\mathbf{X}$. Observa-se que a partir do valor singular 7 a sequência decresce de forma lenta, ou seja, como dito anteriormente, este valor representa um limite entre o sinal e o ruído. Pode-se identificar naturalmente a componente de tendência, representada no gráfico pelo valor singular de maior magnitude.

Posteriormente, pode-se procurar pares de valores singulares. Como mencionado anteriormente, estes pares formam triplos próprios e correspondem às componentes sazonais da série temporal. De acordo com a figura A.2 existe dois triplos próprios: (2-3) e (5-6).

Apesar da não evidência de mais pares de triplos próprios, o valor singular 4 pode ser considerado também na formação da tendência ou do comportamento sazonal da série.

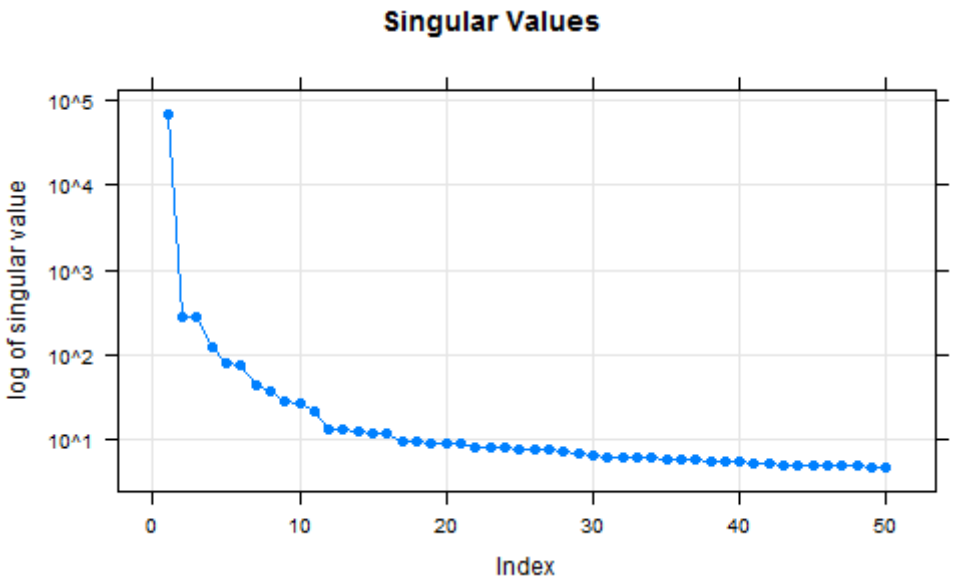

Figura A.2: Valores singulares da decomposição da série em 120 componentes

Vetores singulares

Os vetores singulares associados à componente tendência apresentam uma variação lenta e sem oscilações, como a tendência. Assim, reúne-se todas as matrizes elementares relacionadas com os vetores singulares que apresentam variação lenta para extrair a tendência. O mesmo acontece para as componentes sazonais, ou seja, as componentes e os vetores próprios apresentam formas semelhantes.

Uma outra forma de identificar os vetores singulares associados às componentes sazonais é usar um scatterplot (gráfico de dispersão) de pares de vetores próprios (e vetores de fator). Segundo [12], se o scatterplot apresentar pontos que formem um círculo então os vetores próprios têm a forma de sequências de seno e cosseno (puros) com o mesmo período e a mesma fase. Na prática, as componentes sazonais não são senos e cossenos puros mas podemos identificar os vetores singulares associados caso o scatterplot apresente uma forma idêntica a de algum polígono regular.

Na figura A.3 estão apresentados os gráficos dos 10 primeiros vetores singulares (à esquerda) e os scatterplot's dos 10 primeiros pares de vetores singulares (à direita), para a série exemplificada anteriormente.

Como mencionado anteriormente, os vetores singulares que aparentam ter forma semelhante à da tendência ou à das componentes sazonais da série original estão associados às componentes reconstruídas de tendência ou sazonais. Observando a figura A.3 à

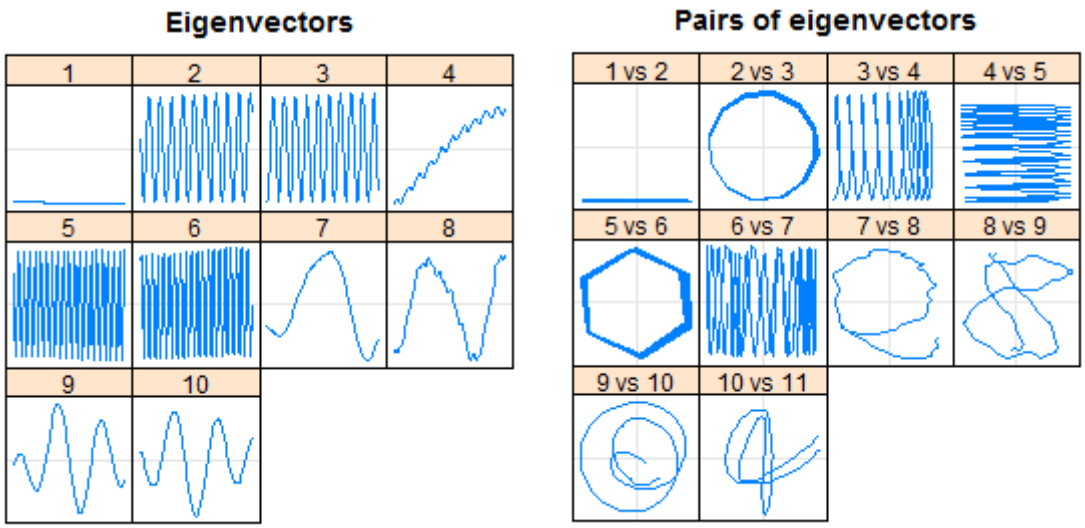

Figura A.3: À esquerda: Gráfico dos 10 primeiros vetores próprios de decomposição da série temporal A.1. À direita: scatterplot's dos 10 primeiros pares de vetores singulares

esquerda, não há dúvidas de que o primeiro vetor próprio está associado à componente de tendência. Os vetores próprios 2, 3, 5 e 6 aparentam ter a forma das componentes sazonais da série, então considera-se que estes estão associados a essas componentes. O vetor próprio 4 tem uma tendência e algumas oscilações, pelo que pode estar associado à tendência ou à sazonalidade. Observando a figura A.3 à direita, podemos concluir que os pares (2-3) e (5-6) estão associados às componentes sazonais, uma vez que são os únicos pares que apresentam um scatterplot muito semelhante a um polígono regular.

Correlação ponderada

Como apresentado na seção A.3.1, as componentes altamente correlacionadas pertencem a um mesmo grupo de triplos próprios. Assim, será usada a *matriz w-correlação*, que indica as correlações entre as componentes da SVD através de uma escala de cores variando do branco ($\rho = 0$) ao preto ($\rho = 1$), como um indicativo de como realizar um agrupamento adequado. Espera-se uma divisão clara das componentes em duas partes: a primeira é constituída por componentes altamente correlacionadas (tonalidade escura), o que caracteriza o grupo dos triplos próprios relacionado ao sinal da série; a segunda parte retrata o ruido, exibindo muitas componentes com correlações baixas (tonalidades claras). A figura A.4 mostra a aplicação da *matriz w-correlação* para a série exemplificada nas subseções anteriores. Através da figura à esquerda, nota-se que a partir da décima nona componente existe mais componentes com correlações maiores, representadas através

das tonalidades mais fortes, ou seja, fazem parte do ruído da série. Como é difícil visualizar mais informações através da matriz, apresenta-se à direita a matriz w-correlação restringida às 20 primeiras componentes do SVD (figura A.4).

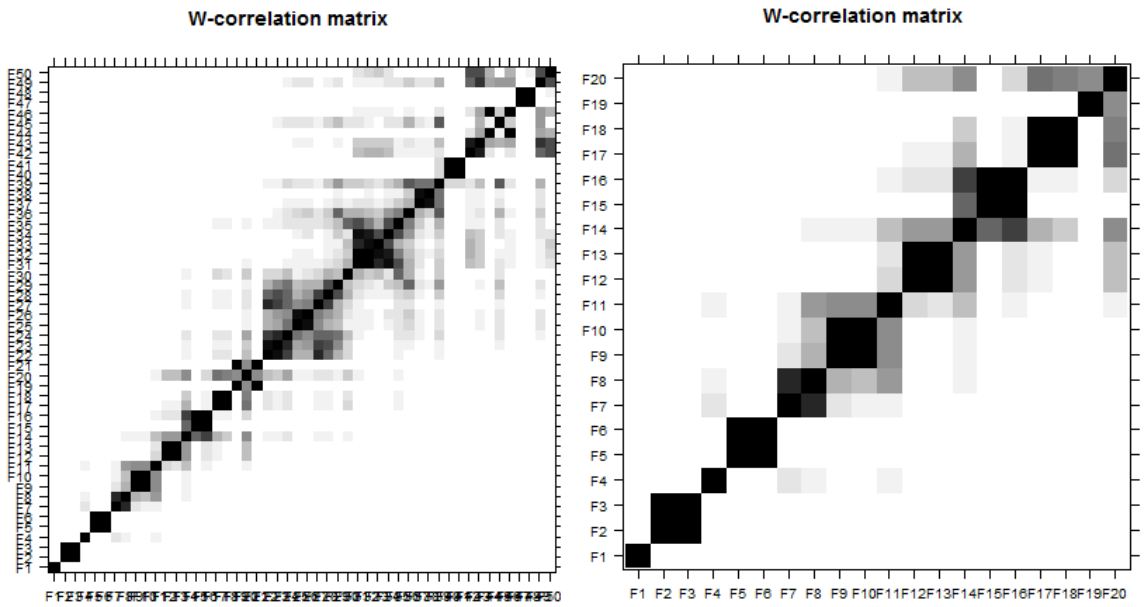

Figura A.4: Matriz w-correlação das componentes SVD resultantes da etapa decomposição da técnica SSA aplicada aos dados de concentrações de CO2 no Havaí. À esquerda: 50 componentes. À direita: 20 componentes

Observando a figura A.4 à direita, conclui-se que o sinal da série pode ser formado pelos seis triplos próprios principais, indicando que as seis primeiras componentes são suficientes para reconstruir a série original, a partir da sétima componente observa-se a existência de mais componentes com correlações maiores, ou seja, fazem parte do ruido da série. Observa-se que os triplos próprios 2 e 3 estão altamente correlacionados, assim como os triplos próprios 5 e 6, uma vez que apresentam uma tonalidade bastante escura.

Conclusão do exemplo

Com a análise apresentada anteriormente pode-se passar para a última fase da técnica SSA, a reconstrução da série, usando o comando `reconstruct` da package `Rssa`. Concluiu-se que o primeiro triplo próprio está associado à componente de tendência, existe dois pares de triplos próprios associados às componentes de sazonalidade ((2,3) e (5,6)), e o quarto triplo próprio está associado à tendência ou à sazonalidade.

Caso o quarto triplo próprio seja agrupado sozinho obtém-se uma componente para a qual não se tem interpretação; caso se junte a um dos grupos associados às componentes sazonais, a respetiva componente reconstruída apresenta sazonalidade mas também

tendência; caso se junte ao grupo associado à tendência obtém-se as componentes como se esperaria, uma componente associada à tendência e duas associadas às sazonalidades como se pode observar na figura A.5.

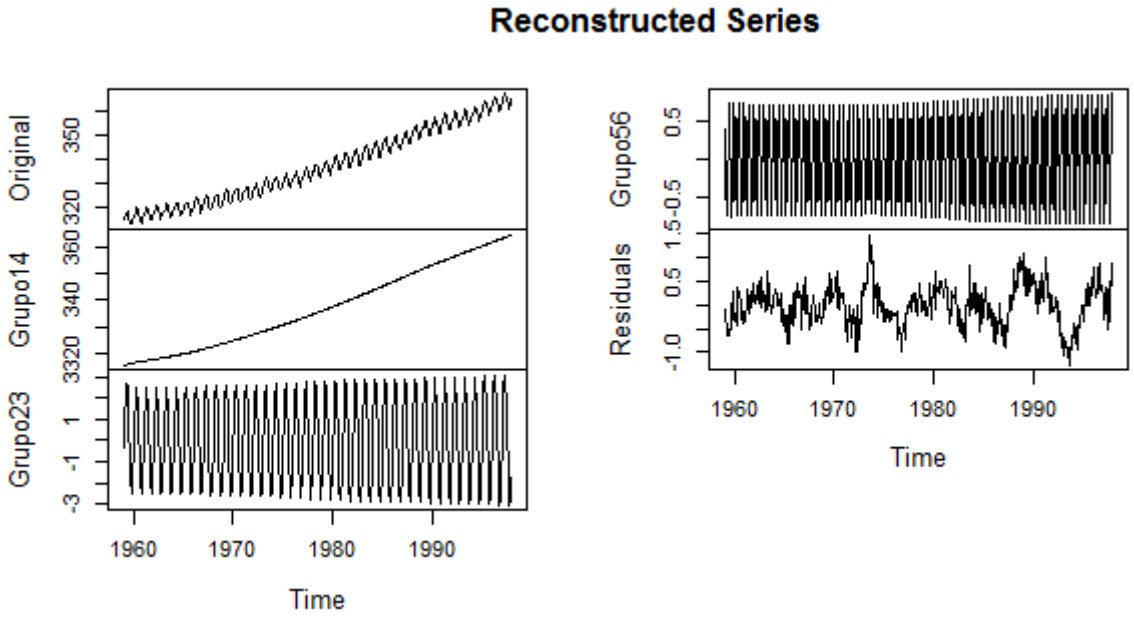

Figura A.5: Série original reconstruída em 4 componentes usando os grupos (1,4), (2,3), (5,6), sendo a última o ruído

Apêndice B

Variáveis Climáticas

B.1 Descrição

Nesta secção é apresentada a análise descritiva do conjunto de variáveis climáticas na região de Lisboa. As variáveis climáticas são: *Comprimento do Dia* (h) - ComprDia, *Temperatura Máxima* (ºC) - TempMax, *Temperatura Média* (ºC) - TempMedia, *Temperatura Mínima* (ºC) - TempMin, *Ponto de Orvalho Máximo* (ºC) - PtOrvMax, *Ponto de Orvalho Médio* (ºC) - PtOrvMedio, *Ponto de Orvalho Mínimo* (ºC) - PtOrvMin, *Humidade Máxima* (%) - HumMax, *Humidade Média* (%) - HumMedia, *Humidade Mínima* (%) - HumMin, *Pressão ao nível do Mar Máxima* (hPa) - PresMax, *Pressão ao nível do Mar Média* (hPa) - PresMedia, *Pressão ao nível do Mar Mínima* (hPa) - PresMin, *Visibilidade Máxima* (km) - VisibMax, *Visibilidade Média* (km) - VisibMedia, *Visibilidade Mínima* (km) - VisibMin, *Velocidade do Vento Máxima* (km/h) - VelVentoMax, *Velocidade do Vento Média* (km/h) - VelVentoMedia, *Velocidade Máxima de Rajada de Vento* (km/h) - VelRajVentoMax, *Precipitação* (%), *Cobertura de Nuvens*, *Eventos* e *Direção do Vento* (graus). O período de tempo das observações aqui apresentadas é de 1 de Janeiro de 2010 a 31 de Agosto de 2013, ou seja, existe 1339 observações para cada variável.

Na figura B.1 pode-se observar o sumário de cada variável climática. Observa-se que todas as variáveis são numéricas à exceção da variável *Eventos*. Esta variável é constituída por 8 eventos: *Chuva*, *Trovoada*, *Nevoeiro*, *Chuva-Trovoada*, *Nevoeiro-Chuva*, *Nevoeiro-Chuva-Trovoada*, *Nevoeiro-Chuva-Granizo-Trovoada* e *Nenhum*.

```
  Compr_Dia        Temp_Max      Temp_Media       Temp_Min     Pt_Orvalho_Max
Min.   : 9.45   Min.   : 8.0   Min.   : 6.00   Min.   : 1.00   Min.   :-8.00
1st Qu.:10.50   1st Qu.:16.0   1st Qu.:13.00   1st Qu.:10.00   1st Qu.:11.00
Median :12.47   Median :21.0   Median :17.00   Median :14.00   Median :14.00
Mean   :12.32   Mean   :21.4   Mean   :17.07   Mean   :13.21   Mean   :13.34
3rd Qu.:14.15   3rd Qu.:26.0   3rd Qu.:21.00   3rd Qu.:17.00   3rd Qu.:16.00
Max.   :14.88   Max.   :40.0   Max.   :32.00   Max.   :27.00   Max.   :21.00
                NA's   :1      NA's   :2       NA's   :1       NA's   :1
Pt_Orvalho_Medio Pt_Orvalho_Min     Humidade_Max    Humidade_Media
Min.   :-11.00   Min.   :-13.000   Min.   : 24.00   Min.   : 16.00
1st Qu.:  8.00   1st Qu.:  5.000   1st Qu.: 83.00   1st Qu.: 64.00
Median : 12.00   Median :  9.000   Median : 88.00   Median : 72.50
Mean   : 10.94   Mean   :  8.292   Mean   : 88.58   Mean   : 71.33
3rd Qu.: 14.00   3rd Qu.: 12.000   3rd Qu.: 94.00   3rd Qu.: 81.00
Max.   : 19.00   Max.   : 18.000   Max.   :100.00   Max.   :100.00
NA's   :1        NA's   :1         NA's   :1        NA's   :1
 Humidade_Min      Press_Max     Press_Media     Press_Min      Visib_Max
Min.   :  8.00   Min.   : 989   Min.   : 987   Min.   : 985   Min.   : 3.000
1st Qu.: 40.00   1st Qu.:1016   1st Qu.:1014   1st Qu.:1012   1st Qu.:10.000
Median : 51.00   Median :1019   Median :1017   Median :1015   Median :10.000
Mean   : 51.27   Mean   :1019   Mean   :1017   Mean   :1015   Mean   : 9.978
3rd Qu.: 63.00   3rd Qu.:1022   3rd Qu.:1020   3rd Qu.:1019   3rd Qu.:10.000
Max.   :100.00   Max.   :1038   Max.   :1036   Max.   :1035   Max.   :10.000
NA's   :1        NA's   :1      NA's   :1      NA's   :1      NA's   :164
 Visib_Media      Visib_Min      Vel_Vento_Max   Vel_Vento_Media Vel_Rajada_Max
Min.   : 1.000   Min.   : 0.000   Min.   : 8.00   Min.   : 3.00   Min.   :23.00
1st Qu.:10.000   1st Qu.: 5.000   1st Qu.:21.00   1st Qu.:10.00   1st Qu.:35.00
Median :10.000   Median :10.000   Median :26.00   Median :13.00   Median :42.00
Mean   : 9.491   Mean   : 7.627   Mean   :26.33   Mean   :14.14   Mean   :43.77
3rd Qu.:10.000   3rd Qu.:10.000   3rd Qu.:32.00   3rd Qu.:18.00   3rd Qu.:52.00
Max.   :10.000   Max.   :10.000   Max.   :60.00   Max.   :39.00   Max.   :90.00
NA's   :164      NA's   :164      NA's   :1       NA's   :1       NA's   :785
 Precipitaçao  CoberNuvens                          Eventos    Dir_Vento_Graus
Min.   :0     Min.   :0.000   Nenhum                   :827   Min.   : -1.0
1st Qu.:0     1st Qu.:1.000   Chuva                    :327   1st Qu.:124.0
Median :0     Median :3.000   Nevoeiro                 : 87   Median :292.0
Mean   :0     Mean   :2.787   Chuva-Trovoada           : 56   Mean   :234.5
3rd Qu.:0     3rd Qu.:4.000   Nevoeiro-Chuva           : 31   3rd Qu.:335.0
Max.   :0     Max.   :7.000   Nevoeiro-Chuva-Trovoada:  7   Max.   :360.0
              NA's   :168     (Other)                  :  4
```

Figura B.1: Sumário das variáveis climáticas

Pode-se observar também que existem vários valores desconhecidos (`NA's`). As variáveis *VisibMax*, *VisibMedia*, *VisibMin* e *Cobertura de Nuvens* contêm muitos `NA's` mas a *VelRajVentoMax* é a que contém um maior número de valores desconhecidos (cerca de 60% das observações estão em falta). As observações da variável *Precipitação* são constantes iguais a zero.

Na figura B.2 pode-se visualizar as curvas de algumas variáveis. Para as variáveis que têm máximo, média e mínimo, apenas está apresentada na figura a curva da média, uma vez que as curvas do máximo e mínimo são semelhantes, excepto na escala.

Observando a figura B.2, conclui-se que quase todas as variáveis são sazonais e variam consoante as estações do ano. Na variável *VelRajVentoMax* observa-se a grande quanti-

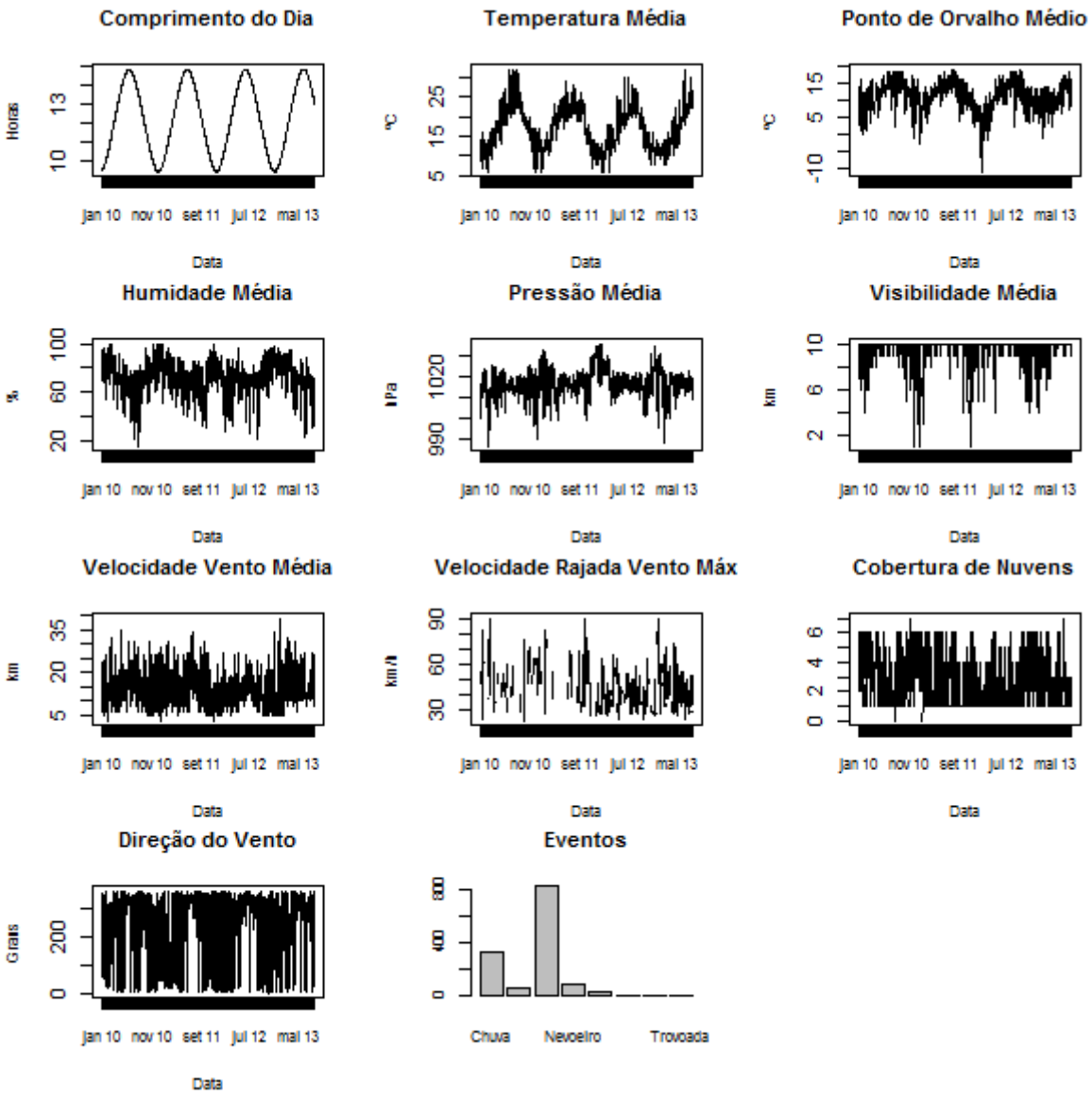

Figura B.2: Gráficos de algumas variáveis climáticas

dade de valores em falta.

Espera-se que algumas variáveis sejam correlacionadas, como por exemplo a temperatura máxima, média e minima estão correlacionadas entre si, assim como as restantes que estão medidas pelo máximo, média e mínimo. Na figura B.3 pode-se observar os diagramas de dispersões das restantes variáveis.

Observa-se que as variáveis *ComprDia*, *TempMedia* e *HumMedia* estão correlacionadas, a *TempMedia* também está correlacionada com o *PtOrvalhoMedio* e a variável *VelVentoMedia* está correlacionada com *VelRajVentoMax*.

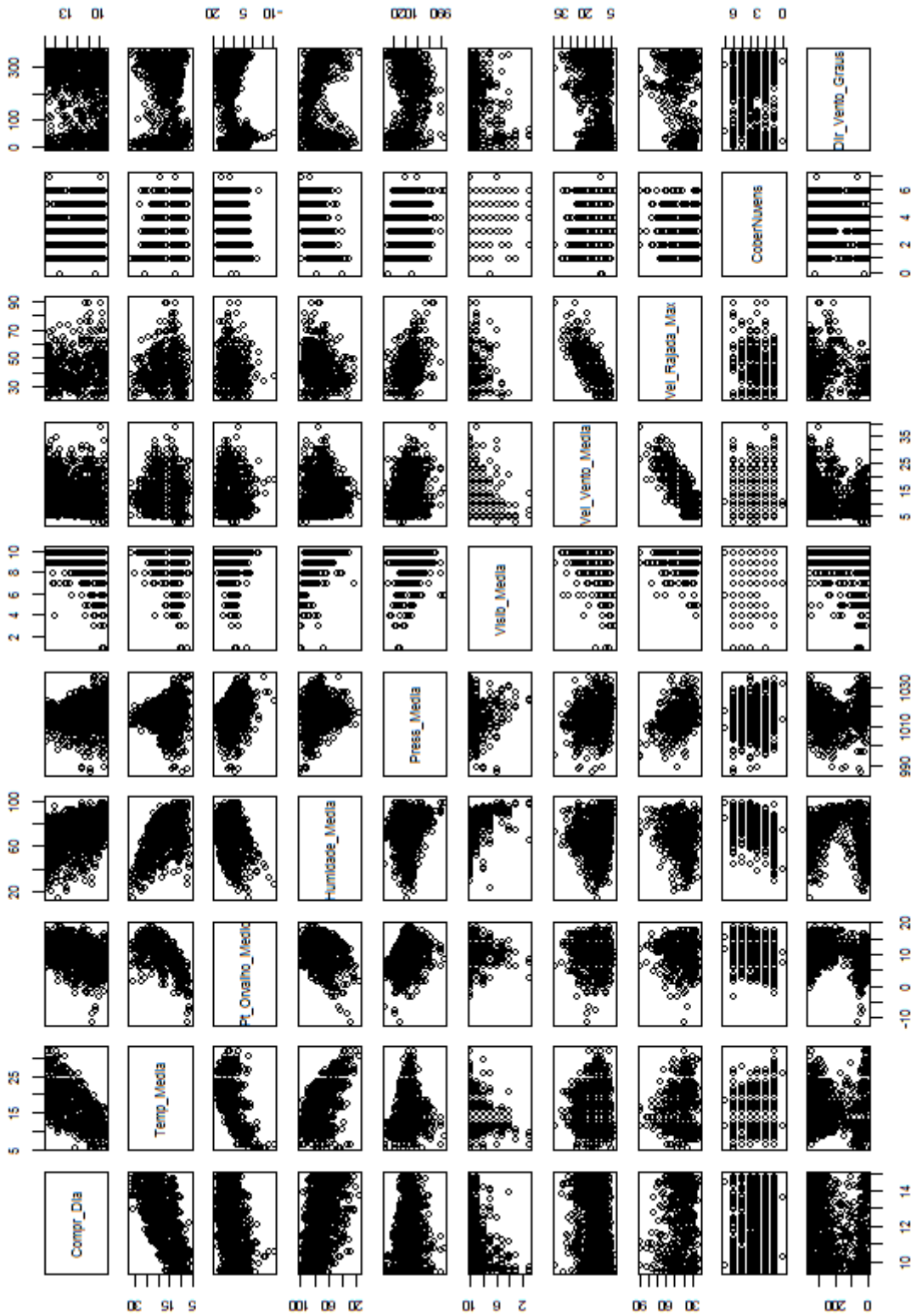

Figura B.3: Diagrama de dispersões de algumas variáveis climáticas

B.2 Tratamento de falhas

Preencher os valores desconhecidos explorando semelhanças entre casos:
Este método assume que se duas observações são similares, e uma delas tem um valor em falta nalguma variável, há uma grande probabilidade que este valor seja semelhante ao valor da outra observação [29]. Para usar este método é necessário definir a noção de similaridade. Esta noção é usualmente definida usando uma métrica sobre o espaço multivariado das variáveis utilizadas para descrever as observações.
Este método está implementado na função `knnImputation()` disponível na package `DMwR` no `R`. A função usa uma variante da distância Euclidiana d para encontrar os k vizinhos mais próximos de qualquer caso [29]:

$$d(x,y) = \sqrt{\sum_{i=1}^{p} \delta_i(x_i, y_i)}$$

onde δ_i determina a distância entre dois valores da variável i da seguinte forma

$$\delta_i(x_i, y_i) = \begin{cases} 1 & \text{se } i \text{ é nominal e } x_i \neq y_i \\ 0 & \text{se } i \text{ é nominal e } x_i = y_i \\ (x_i - y_i)^2 & \text{se } i \text{ é numérica} \end{cases}$$

Estas distâncias são calculadas após normalizar as variáveis numéricas, que é $X_i = \frac{x_i - \bar{x}}{\sigma_x}$.
Após encontrar os vizinhos mais próximos existe duas formas de usar esses valores na função `knnImputation()` (apenas os casos completos podem ser usados para preencher valores desconhecidos [18]). A primeira simplesmente calcula a mediana dos valores dos vizinhos mais próximos para preencher a falha. Caso a variável que contem o valor em falta seja nominal usa-se o valor mais frequente (a moda). A segunda usa uma média ponderada desses valores para preencher o valor desconhecido [29].