Desagregação do consumo energético em conjuntos e caracterização do consumo típico

Elena Selaru
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 Maria Brito, Professora Associada, FCUP
Profª Dra Maria Eduarda Silva, Professora Associada, FEP
Eng Duarte Duarte, DGE, EDP Distribuição – Energia S.A.
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, _____/_____/_____

Resumo

O consumo excessivo de energia, assim como o seu desperdício, são atualmente uma preocupação internacional que requer todos os esforços para poderem ser controlados e diminuídos. Todos podem contribuir nesse sentido, individual e coletivamente, no trabalho ou até mesmo em casa, através de um consumo energético equilibrado que permita diminuir o desgaste da natureza e garantir um futuro melhor para todas as espécies.

Este projeto pode contribuir, ainda que de uma forma pequena, para a persecução desse objetivo, uma vez que a desagregação da energia industrial nas três componentes, energia de base, a útil e a dependente das variáveis externas, pode indiciar as possíveis situações de eficiência energética e contribuir para um planeamento e uma gestão mais eficiente da rede de energia.

Considerando apenas o diagrama de carga e a região geográfica da instalação, a desagregação foi feita, aplicando, essencialmente, a técnica Análise Singular Espectral e os dias de conforto. Estes, por sua vez, foram definidos através das duas variáveis externas Ponto Orvalho Máximo e Humidade Mínima que resultaram ser as mais relacionadas com o consumo energético. As "falhas" entre os dias de conforto, para a determinação da energia útil, foram preenchidas por meio da interpolação linear enquanto que os extremos foram estimados através da regressão linear múltipla.

O algoritmo completo implementado numa função de `R` foi testado em 25 instalações e de seguida aplicado a 471.

Palavras-chave: Energia industrial, desagregação, eficiência energética, conforto.

Abstract

Energy excessive consumption, as well as its waste, are now an international concern that requires every effort in order to be controlled and reduced. Everyone can contribute to this, individually and collectively, at work or even at home, through a balanced energy consumption which allows the reduction of the wear and tear of nature and ensure a better future for all species.

This project can contribute, albeit in a small way, to the pursuit of this goal, since the disaggregation of industrial energy in three components, baseload, useful energy and weather dependent energy may indicate the possible situations of energy efficiency and contribute for the more efficient planning and management of the power grid.

Considering only the load diagram and the facility‘s geographic region, disaggregation was made, by applying essentially the technique Singular Spectrum Analysis and the days of comfort. These, in turn, were defined through the two external variables Maximum Dew Point and Minimum Humidity that turned out to relate the most to energy consumption. The “gaps” between days of comfort, to determine the useful energy, were filled through linear interpolation, while the ends were estimated using multiple linear regression.

The complete algorithm implemented in a function `R` was tested in 25 facilities and then applied to 471.

Kew-Words: Industrial energy, disaggregation, energy efficiency, comfort.

Conteúdo

Resumo i

Abstract ii

Lista de Tabelas vi

Lista de Figuras ix

1 Introdução 1

2 Contextualização do problema 3

2.1 Análise gráfica dos Diagramas de Carga 4

2.2 Modelo físico do consumo energético 8

2.3 Decomposição de uma série temporal 11

3 Critérios de seleção de instalações 18

4 Regiões e variáveis externas 21

4.1 Regiões de Portugal 22

4.2 Variáveis Externas 23

4.2.1 Regressão linear múltipla 26

4.2.2 Floresta aleatória (*random forest*) 27

4.2.3 Correlação cruzada e parcial 29

5 Desagregação do consumo energético **33**

5.1 Índice de aquecimento e Índice de arrefecimento 37

5.2 Determinação da energia de base 40

5.2.1 Mínimo 43

5.2.2 Dias de conforto 46

5.3 Determinação da energia útil 52

5.3.1 Dias de conforto 53

5.3.2 Regressão linear múltipla 56

5.3.3 Associações 58

5.4 Determinação da energia das variáveis externas 65

6 Conclusão **68**

Bibliografia **70**

A Contextualização do problema **75**

A.1 Análise gráfica 75

A.2 Feriados 2011-2013 77

A.3 Decomposição de uma série temporal 78

B Variáveis externas **81**

B.1 Descrição das variáveis climáticas 81

B.2 Floresta aleatória numa instalação 86

B.3 Correlação variáveis climáticas 86

C Desagregação do consumo energético **89**

C.1 Baseload 89

Lista de Tabelas

4.1 Frequência relativa das variáveis selecionadas através da regressão linear múltipla. 26

4.2 Frequência relativa das variáveis selecionadas através da floresta aleatória. . 28

4.3 Frequência relativa das variáveis climáticas com correlação estatisticamente significativa com o consumo total com base em 97 instalações. 30

4.4 As variáveis mais correlacionadas com o consumo total e com correlação inferior a 0.65 entre sí para as 6 regiões. 32

5.1 Número de dias de conforto por ano e por região. 49

A.1 Feriados dos anos 2011 a 2013 77

B.1 Valor p da correlação parcial entre consumo total diário e algumas variáveis climáticas 87

B.2 Correlação cruzada entre algumas variáveis climáticas da região de Beja. . . 87

Lista de Figuras

2.1 Diagrama de carga de Janeiro a Agosto de 2013. 4

2.2 Consumo total diário de Janeiro de 2011 a Agosto de 2013. 5

2.3 Consumo total mensal nos três anos de registo. 6

2.4 Consumo energético de uma instalação. 7

2.5 Consumo total diário (a) / típico (b) e os respetivos consumo estimados através do modelo de regressão. 10

2.6 Algoritmo da Análise Singular Espectral (SSA). 12

2.7 Resultado da primeira fase da SSA sequencial. 14

2.8 Valores singulares da segunda fase da SSA sequencial. 15

2.9 Matriz de correlação W da segunda fase da SSA sequencial. 15

2.10 Pares de vetores próprios da segunda fase da SSA sequencial. 16

2.11 Decomposição da série original com a SSA sequencial. 17

4.1 Divisões de Portugal Continental. 22

5.1 Desagregação do consumo energético residencial. 35

5.2 As três faixas do consumo energético industrial. 36

5.3 Diagrama do cálculo do índice de aquecimento. 39

5.4 Índice de aquecimento no consumo total diário de uma instalação. 39

5.5 Consumo mínimo diário, tendência e a energia de base de duas instalações pertencentes ao conjunto de treino. 41

5.6 Agregação do consumo por dia através da medida mínimo do dia. 43

5.7 Agregação do consumo por dia através do mínimo das médias horárias. . . . 44

5.8 Agregação do consumo por dia a partir do mínimo das médias dos quatro períodos do dia. 44

5.9 Consumos mínimos e as respetivas energias de base. 45

5.10 Consumos mínimos das médias horárias agregados, através da média, por semana, (a) e (b), e por mês, (c) e (d) das duas instalações vistas em cima. . 46

5.11 A aproximação de August-Roche-Magnus. 48

5.12 Consumo total diário de duas instalações e nos dias de conforto. 50

5.13 Consumo total diário de duas instalações, os dias de conforto e a energia de base diária calculada por ano. 52

5.14 Consumo diário de duas instalações, após a componente da energia de base ter sido retirada, e os dias de conforto marcados a vermelho. 53

5.15 Consumo total diário de duas instalações, os dias de conforto, a energia de base diária e a energia útil calculada por dia. 55

5.16 Consumo total diário de duas instalações, os dias de conforto, a energia de base diária e a útil diária calculadas por ano. 58

5.17 Medidas de confiança. $F(\varepsilon) = 4/10$, $\mu(\varepsilon \to \tau) = 4/10$. 61

5.18 Consumo total diário com o agrupamento dos dias. 64

5.19 Desagregação do consumo total diário nas três componentes: energia de base, útil e das variáveis externas. 65

5.20 Consumo total diário o Ponto Orvalho Máximo ao longo do tempo. 66

A.1 Consumo mínimo diário de Janeiro de 2011 a Agosto de 2013 75

A.2 Consumo máximo diário de Janeiro de 2011 a Agosto de 2013 76

A.3 Consumo diário agregado pelo mínimo (preto), média (verde) e máximo (vermelho) 76

A.4 Decomposição da série do consumo usando `decompose()` 78

A.5 Decomposição da série do consumo usando `stl()` 79

A.6 Resíduos da decomposição 80

B.1 Comprimento do Dia 81

B.2 Temperatura Máxima (preto), Média (verde), Mínima (vermelho) 82

B.3 Ponto de Orvalho Máximo (preto), Médio (verde), Mínimo (vermelho) 82

B.4 Humidade Máxima (preto), Média (verde), Mínima (vermelho) 83

B.5 Pressão Máxima (preto), Média (verde), Mínima (vermelho) 83

B.6 Visibilidade Máxima (preto), Média (verde), Mínima (vermelho) 84

B.7 Velocidade Máxima (preto), Média (verde) do vento 84

B.8 Velocidade Máxima da Rajada do Vento 85

B.9 Cobertura com Nuvens 85

B.10 Eventos 85

B.11 Direção do Vento 86

B.12 Importância das variáveis climáticas da região de Beja 86

C.1 Decomposição dos mínimos das médias horárias usando a técnica `decompose` e o resultado após retirar a sazonalidade estimada 90

C.2 Dendrograma de uma instalação 90

Capítulo 1

Introdução

A EDP é o maior produtor, distribuidor e comercializador de eletricidade em Portugal [12]. Com a possibilidade de desenhar um portefólio de serviços que a EDP Distribuição (EDPD) possa utilizar, em complemento aos já prestados aos Clientes a partir da plataforma online, foram propostos dois estágios na Direção de Gestão de Energia cujos temas foram: "*Clustering de instalações*" e "*Desagregação do consumo energético*". O segundo encontra-se exposto neste relatório. Os dois estágios, integrados no mestrado em Engenharia Matemática da FCUP, surgiram com o intuito de contribuir com novos conhecimentos para a optimização do consumo energético.

O objetivo geral deste projeto foi o de decompor o consumo energético de uma instalação em três faixas, correspondentes à energia de base, útil e dependente das variáveis externas, a partir do seu diagrama de carga. Este trabalho tem vantagens tanto para o utilizador de energia como para a empresa fornecedora: ao cliente permite identificar as possibilidades de poupança de energia e melhorar o desempenho do edifício; à EDP Distribuição permite perceber melhor os consumos dos seus clientes sem quaisquer informações adicionais sobre a instalação. Um conhecimento mais aprofundado permite à EDPD prever os consumos com maior precisão o que, por sua vez, contribui para um planeamento e uma gestão mais eficientes da rede.

A descrição detalhada dos métodos testados e algoritmos implementados para fazer a desagregação encontra-se no Capítulo 5. Para poder cumprir o objetivo final, foi necessário estudar e realizar três objetivos específicos, comuns aos dois estágios: conhecer

e perceber o que são os consumos energéticos - Capítulo 2, selecionar as instalações válidas para análise - Capítulo 3, e por fim, escolher as variáveis climáticas que têm uma maior influência sobre o consumo de uma instalação - Capítulo 4.

Price [44] define o diagrama de carga como potência elétrica total usada pelo edifício durante um dado intervalo temporal, que em Portugal é igual a 15 minutos. O diagrama de carga varia ao longo do tempo como resposta à mudança do nível de iluminação, necessidade de aquecimento ou ar condicionado, o uso de computadores, fotocopiadoras e outros equipamentos elétricos presentes na instalação. A curva que representa o diagrama, transmite informações importantes sobre os acontecimentos dentro do edifício. Valores noturnos altos inesperados podem indicar desperdícios (como por exemplo uma luz acesa esquecida quando o edifício está desocupado); uma mudança na curva pode indiciar o mau funcionamento de algum equipamento ou a ocorrência de uma fraude na instalação; sensibilidade inesperadamente elevada à temperatura exterior pode indiciar um mau isolamento do edifício, entre outros. O projeto desenvolvido pode detetar estes comportamentos para a maior parte das instalações em que foi testado.

Durante o estágio recorreu-se a duas ferramentas: `R`, devido à sua forte componente estatística e à familiaridade desenvolvida durante o percurso académico, e `SQL-Server`, devido à sua capacidade de armazenamento de dados de grandes dimensões. Em cada etapa foi usada uma ou outra ferramenta de modo a diminuir o tempo de execução.

O algoritmo desenvolvido foi aplicado apenas a algumas instalações, mais precisamente, 497, de Portugal Continental, que estavam organizadas em lotes e armazenadas em `SQL Server`, uma outra razão para a utilização desta ferramenta. Para fazer testes e tomar decisões utilizou-se apenas o primeiro lote que tinha os diagramas de carga de 98 instalações. Não existia qualquer padrão nas instalações que o constituem, de maneira que não existe enviesamento nos resultados obtidos.

Capítulo 2

Contextualização do problema

Inicialmente foi disponibilizado o diagrama de carga de uma instalação com registos no período de 1 de Janeiro de 2011 a 31 de Agosto de 2013. A série foi estudada sob vários ângulos para se conhecer a área de trabalho e estudar as possíveis abordagens ao tema. A análise inicial foi feita por meio de gráficos, Secção 2.1, seguindo-se a construção de um modelo para o consumo energético, Secção 2.2, e, por fim, a série foi detalhada por meio da decomposição das séries temporais, Secção 2.3.

Algumas conclusões retiradas da avaliação desta única instalação foram generalizadas para as outras e alguns procedimentos que se executaram nesta fase foram repetidos na fase seguinte. Um exemplo disso é a conversão da potência por 15 minutos em energia por 15 minutos, que deve ser feita em todas as instalações.

Esta conversão é requerida porque o objetivo é estudar a quantidade de energia gasta num período específico de tempo e não a sua taxa de utilização (definição de potência) [34]. A transformação foi obtida considerando a fórmula Energia (kWh) = Potência (kW) $\times$ Tempo (h).

Medidas

A medida usada para agregação dos dados foi a soma, já que o objetivo do trabalho era estudar o consumo total típico das instalações. De facto, as outras medidas não servem este propósito uma vez que: as medidas mínimo e máximo permitem avaliar os consumos extremos, enquanto que a média suaviza demasiado a curva.

2.1 Análise gráfica dos Diagramas de Carga

Já com a definição de diagrama de carga enunciada na introdução, pode visualizar-se o gráfico de um diagrama típico na Figura 2.1. Contudo, como a primeira instalação tem registos de 15 em 15 minutos durante 974 dias, isso perfaz um total de 93478 observações. Num gráfico com a largura da página essas observações aparecem em forma de nuvem, de modo que na Figura 2.1 estão representados apenas os 8 meses disponíveis do ano 2013, de Janeiro a Agosto.

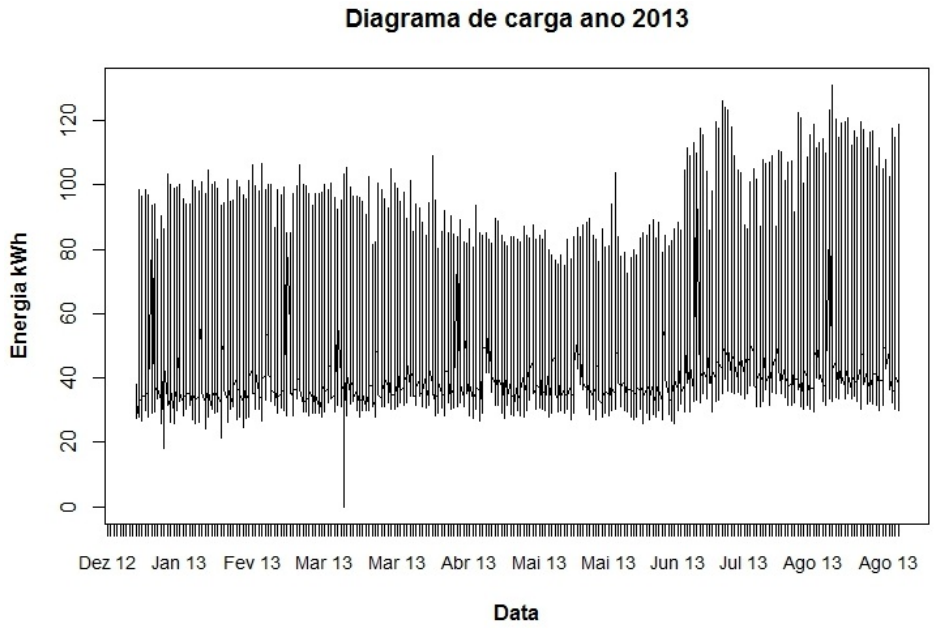


Figura 2.1: Diagrama de carga de Janeiro a Agosto de 2013.

A densidade elevada dos pontos na Figura 2.1 não permite ver o comportamento da curva, de modo que os dados foram agregados por dia através da medida soma, conduzindo à curva de consumo total diário. Outra razão para tomar os dados agregados por dia é diminuir o custo computacional, uma vez que `R` demora 12 minutos a importar os diagramas de carga de 15 em 15 minutos de 98 instalações enquanto que a importação dos dados agregados por dia (agregação feita em `SQL Server`) é quase imediata (menos de 1 segundo).

O consumo diário durante todo o período de registos pode ser visto na Figura 2.2.

A característica da Figura 2.2 que mais se destaca são os valores mínimos extremos, "os poços", sinalizados a verde e que foram identificados como consumos energéticos nos

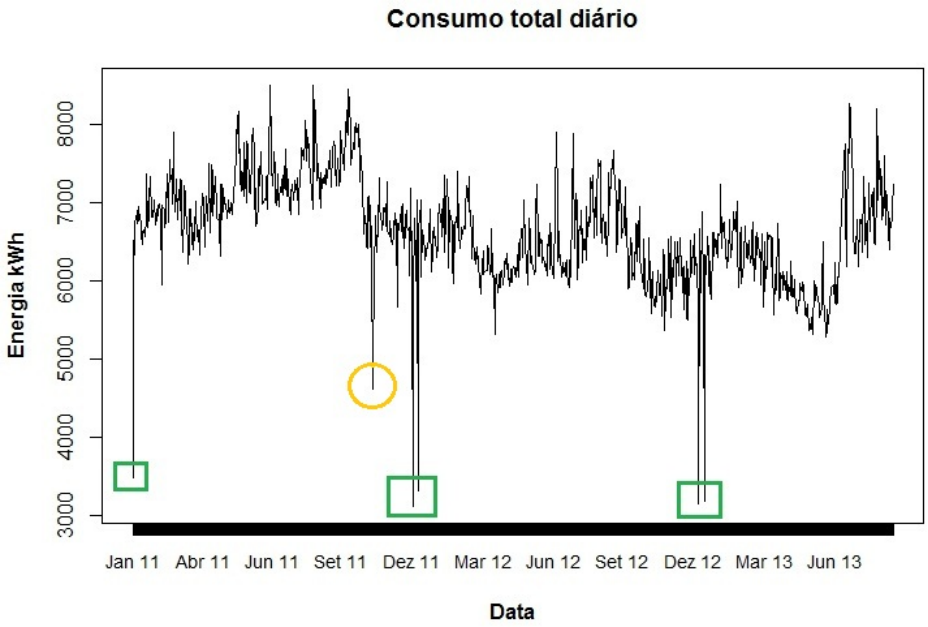


Figura 2.2: Consumo total diário de Janeiro de 2011 a Agosto de 2013.

dias de Ano Novo, 1 de Janeiro, e dias de Natal, 25 de Dezembro. Este comportamento resultou na criação de uma variável binária que indica se o dia é ou não um feriado nacional, chamada **Feriado**. Para isso recorreu-se à lista de todos os feriados nacionais de 2011 a 2013 (Anexo A.2).

Na Figura 2.2 existe um outro consumo excessivamente baixo que está sinalizado a laranja. Explicação para isso são as falhas na energia durante 5 horas em que não houve registos, provocando uma diminuição do consumo total nesse dia. Assim, para se usar a medida soma para a agregação dos dados, deve ser feito um ajustamento dos dados a essas faltas de modo a ter o consumo típico da instalação sem as situações imprevistas. A correção foi feita ajustando o consumo total real ao número de instantes previsto:

$$\frac{\text{Consumo real} \times (\text{Nr instantes previsto} = 96)}{\text{Nr instantes real}}$$

Outras características do consumo que se evidenciam na Figura 2.2 são a diminuição do mesmo ao longo dos anos e ainda uma variação ao longo do ano. Estas anotações são corroboradas pela contemplação da Figura 2.3 onde é representado o consumo total de cada mês para os três anos.

Nota-se uma diminuição do consumo, bastante significativa, comparando os meses de Maio de 2011 e de 2013, onde houve uma redução do consumo em mais de 20%. No

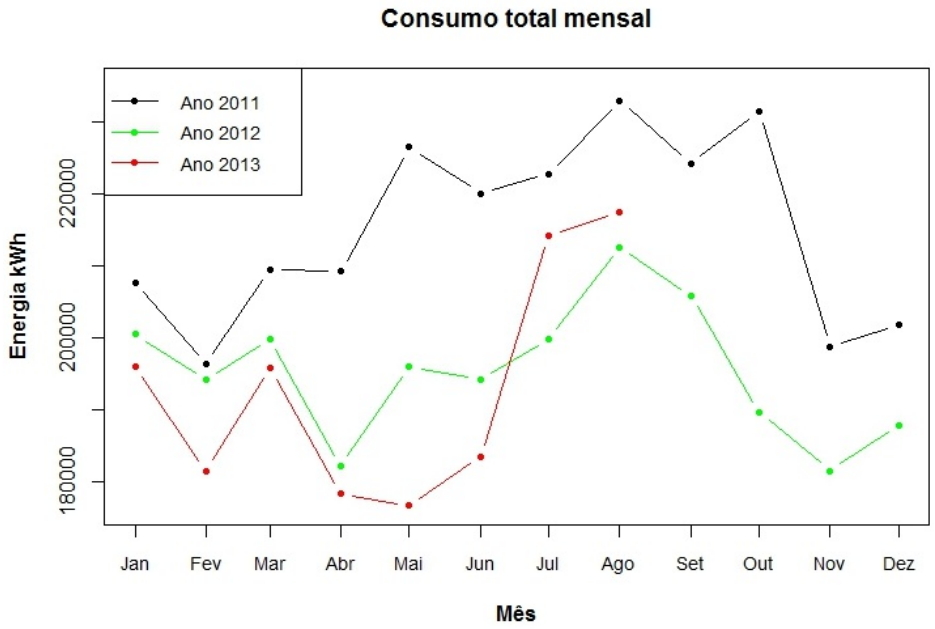
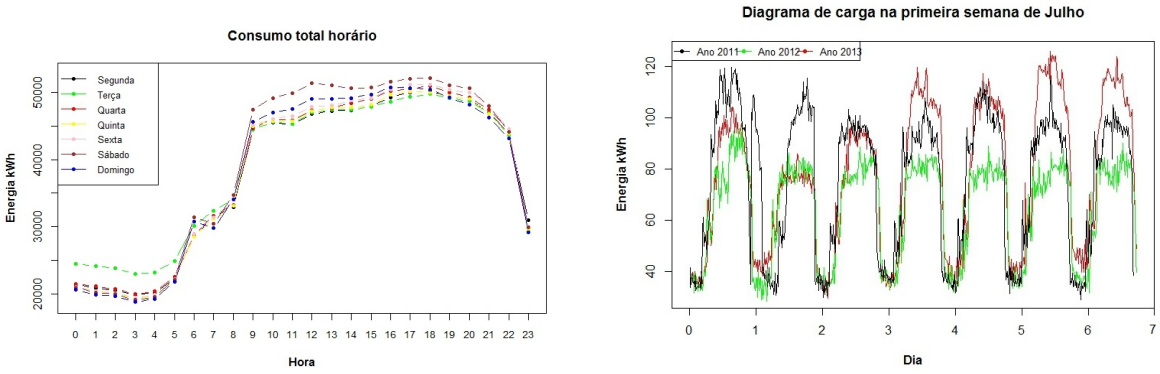


Figura 2.3: Consumo total mensal nos três anos de registo.

ano todo de 2012 o valor da energia consumida é inferior ao ano anterior, que pode ter como justificação as mudanças climatéricas, económicas, sociais, etc. Pensou-se que esta informação de variação ao longo dos anos, poderá ser relevante para caracterizar o consumo, pelo que foi construída uma variável categórica que indica o ano do registo, chamada **Ano**.

As variações no consumo registadas ao longo do ano levaram à construção da variável **Estação**. No entanto, como a sua definição pode ser feita de duas maneiras, a escolha da mais adequada será feita na Secção 2.2, por meio do coeficiente de determinação ajustado, descrito na mesma secção. Os dois tipos de estações são: i). Estações Meteorológicas - para simplificar os cálculos climatológicos e mantê-los uniformes, começam sempre no primeiro dia dos meses Março, Junho, Setembro e Dezembro; ii). Estações Astronómicas - começam por volta do dia 21 dos mesmos meses das estações meteorológicas e delimitam com maior precisão o clima de cada época [54]. Em vez da variável Estação, poder-se-ia criar a variável Mês, que igualmente representaria as variações ao longo do ano, mas optou-se pela variável Estação devido aos valores de consumo próximos dentro de uma estação. Observando, por exemplo, a Figura 2.3, nota-se pequena variação no consumo nos meses de Abril, Maio e Junho, o correspondente à estação astronómica Primavera, registando-se o mesmo comportamento nas restantes épocas.

Ampliando ainda mais a escala do tempo, visualizou-se o consumo agregado (através da média) por hora, nos sete dias da semana. O gráfico da Figura 2.4(a) foi obtido somando os consumos registados durante a respetiva hora em cada um dos sete dias da semana.



(a) Consumo total horário nos sete dias da semana

(b) Diagrama de carga na primeira semana de Julho nos três anos

Figura 2.4: Consumo energético de uma instalação.

Da Figura 2.4(a), pode concluir-se que o horário de funcionamento da instalação é das 9h até às 22h, igual nos sete dias da semana. Existe um pico no consumo entre as 5h e as 6h que também aparece nas instalações exploradas por Price [44] e que ele explica como a ligação automática do ar condicionado. Ao contrário do estudo feito por este, no caso dado não se conhece nem o tipo nem o ramo de atividade da instalação, pelo que esta explicação é apenas uma suposição.

Contemplando a Figura 2.4(b), confirma-se a anotação feita anteriormente de que o consumo varia ao longo dos anos: na primeira semana de Julho, o consumo energético no ano 2012 foi inferior aos outros dois anos. O padrão de consumo ao longo do dia é igual nos sete dias da semana e, juntando a informação revelada pela Figura 2.4(a), pode concluir-se que, para esta instalação, não há grandes diferenças no consumo nos diferentes dias da semana. No entanto, isso pode não se verificar para as outras instalações, de modo que foi criada a variável **Dia da Semana**. Com a mesma linha de pensamento, construiu-se a variável binária **Fim de Semana** que indica se o dia é útil ou é fim de semana. Como estas duas variáveis refletem informação semelhante sobre o consumo, a de menor importância, a averiguar no Capítulo 4, foi retirada.

As cinco variáveis construídas a partir dos gráficos para esta instalação foram também estruturadas para as instalações do primeiro lote, com o intuito de selecionar as que mais influenciam o consumo, com base num maior número de instalações.

Tal como se notou anteriormente, o consumo varia em função da estação e a diferença entre várias estações está nas condições atmosféricas. No inverno, quando as temperaturas estão baixas, existe uma parte do consumo relacionada com o aquecimento, e no verão existe a componente de arrefecimento. Logo, o consumo também varia em função das condições ambientais.

Como tal, foram extraídas 22 variáveis relativas à região de Lisboa, uma vez que a instalação se encontra nesta zona, do site Weather Underground [54] para ver se existia alguma relação entre o consumo e as variáveis externas. Outra variável que poderá ser relevante é o número de horas de luz solar por dia, que foi retirada do Observatório Naval dos Estados Unidos [40], onde era necessário introduzir as coordenadas geográficas da cidade que, por sua vez, foram obtidas da página web da Academia de Estudos Astrológicos de Lisboa [2]. A lista de todas as variáveis e a sua descrição completa encontra-se no Anexo B.1.

Na tabela das variáveis externas existiam valores em falta que foram preenchidos utilizando os casos mais semelhantes. A explicação mais pormenorizada do método pode ser vista na Secção 4.2 ou consultando [28].

Com a tabela que contém o consumo total diário da instalação em causa, as variáveis construídas a partir dos gráficos e as climáticas, fez-se uma análise descrita na Secção 2.2 para determinar quais são as variáveis que mais influenciam o consumo desta instalação e para encontrar uma fórmula que consiga descrever o consumo.

2.2 Modelo físico do consumo energético

Como Price [44] constatou, vários métodos foram aplicados, pelos investigadores ou como parte do Sistema de Informação de Energia, para determinar a relação matemática entre o consumo e as variáveis explicativas, para que, de seguida, seja possível prever o consumo. Um destes métodos é a regressão linear múltipla, que foi vista neste trabalho e que assume a existência de uma relação linear entre a variável resposta e cada uma das variáveis preditoras.

Esta técnica está implementada em `R` na instrução `lm(Y~.,data)`, que constrói o modelo, onde `Y` é a variável resposta, neste caso o consumo total, em função de todas as variáveis explicativas. As variáveis independentes são as cinco construídas anteriormente a partir

dos gráficos e as 23 variáveis climáticas.

A estatística que quantifica o ajustamento do modelo aos dados é o **coeficiente de determinação ajustado** - R^2. Quanto mais próximo de 1 estiver esse valor, tanto melhor é o ajustamento do hiperplano aos dados. Isto é, $R^2 \times 100\%$ da variância estimada de `Y` é explicada pelo hiperplano de regressão [52].

Com isto, o modelo construído com base em 28 variáveis usando, as estações meteorológicas, tem um R^2 ajustado igual a 0.547, as estações astronómicas - 0.576[1]. Através da regressão, pretendia modelar-se o consumo típico, mas como o dia de Natal e de Ano Novo não o são, estes foram retirados e, recorrendo à todas as variáveis explicativas e a estação astronómica, calculou-se novamente o R^2 ajustado, que desta vez foi igual a 67.6%. Assim, pode dizer-se que o modelo de regressão completo explica aproximadamente 68% da variância do consumo total diário típico médio.

Segundo Hastie et al. [21], um modelo de regressão com muitas variáveis explicativas torna-se difícil de interpretar. Uma alternativa é construir um modelo mais simples, com menos variáveis, que evidencia os efeitos mais fortes. A este propósito, existem vários métodos de seleção de variáveis cuja descrição pormenorizada pode ser encontrada em Hastie et al. [21].

Para construir o modelo linear do consumo energético com as variáveis mais significativas, aplicou-se o procedimento *stepwise* que, partindo do modelo completo[2], ou nulo[3], envolve a inclusão e exclusão de variáveis, o procedimento mais frequentemente utilizado, de modo a obter o modelo com o menor conjunto de variáveis explicativas a incluir na regressão [51]. No `R`, a melhoria da qualidade do ajustamento é avaliada através dos critérios de informação AIC ou BIC: quanto menor for o critério, mais preferível é o modelo [21]. Através do comando `step(...,direction="both")` de `R`, do critério AIC e começando com o modelo completo, pois os modelos são estimados sempre com os mesmos dados, as variáveis finais a incluir na regressão são:

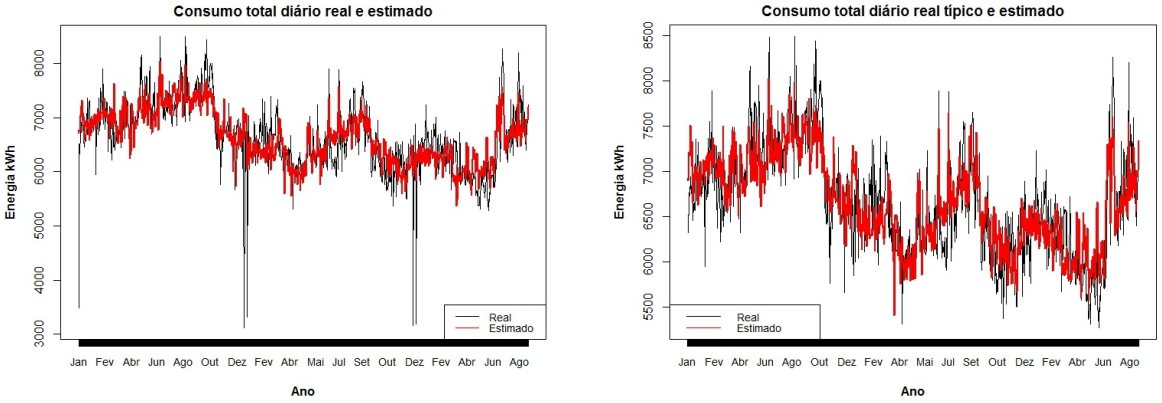
Ano, Feriado, Estação, Dia da Semana, Ponto Orvalho Médio, Humidade Média, Velocidade Máxima do Vento, Velocidade Média do Vento e a Direção do Vento em Graus,

com um R^2 = 0.576, o mesmo que o conjunto maior com todas as variáveis iniciais, ou

[1] Nas análises futuras serão consideradas as estações astronómicas uma vez que têm um maior R^2 ajustado.
[2] Modelo com todas as variáveis explicativas.
[3] Modelo que consta apenas do coeficiente independente.

seja, pode optar-se por um menor número de variáveis que consegue explicar a mesma percentagem de variância do consumo.

Se o objetivo novamente for aproximar o consumo típico, o resultado da seleção automática das variáveis são as anteriores mais o Comprimento do Dia e as três Temperaturas (Máxima, Média e Mínima), que resulta num R^2 ajustado igual a 0.677. Ou seja, tal como anteriormente, o consumo típico médio pode ser modelado usando um menor conjunto de variáveis explicativas.



(a) Consumo total diário

(b) Consumo total diário típico

Figura 2.5: Consumo total diário (a) / típico (b) e os respetivos consumo estimados através do modelo de regressão.

A decisão final sobre a eliminação de algumas variáveis não deve ser feita com base apenas num algoritmo, mas sim, como Kleimbaum [33] refere, as variáveis explicativas finais devem ser especificadas a priori. No entanto, como não existe nenhum conhecimento prévio sobre as variáveis mais significativas para a modelação do consumo energético, e tendo em conta que ainda decorre a fase exploratória dos dados, as variáveis finais para a construção do modelo serão as devolvidas pelo método *stepwise*. Assim, definindo as variáveis indicatrizes para as variáveis categóricas, por exemplo, para a variável Ano, onde o ano 2011 é de referência, tem-se

$$Z_1 = \begin{cases} 1 & \text{se} \quad \text{ano} = 2012 \\ 0 & \text{se} \quad \text{ano} \neq 2012 \end{cases} \qquad Z_2 = \begin{cases} 1 & \text{se} \quad \text{ano} = 2013 \\ 0 & \text{se} \quad \text{ano} \neq 2013 \end{cases}$$

O modelo final que descreve o consumo total diário médio da instalação em questão é:

$Consumo = 7634 - 572Z_1 - 632Z_2 + 66 \times PontoOrvalhoMédio - 21 \times HumidadeMédia + ...$

e o consumo toma um valor exato quando se substituem os valores das variáveis.

Na Figura 2.5 tem-se o consumo total diário 2.5(a), o consumo total diário típico 2.5(b) e as suas estimações (a vermelho), ambas feitas por meio do modelo de regressão linear múltipla, com as respetivas variáveis que foram destacadas como mais significativas.
Agora pode passar-se à exploração dos dados vistos como uma série temporal.

2.3 Decomposição de uma série temporal

Murteira et al. [39] definem a série temporal como um conjunto de observações ordenadas no tempo, não necessariamente igualmente espaçadas. As séries temporais aparecem em vários domínios, sempre que algo ocorre ao longo do tempo. O consumo energético não é exceção uma vez que é registado em períodos de 15 minutos ou, após a agregação, por dia.
A maneira tradicional de explorar uma série temporal é decompô-la nas componentes de tendência, sazonalidade e erro, que Hyndman et al. [26] definem como:

- Tendência (T) - O comportamento da série a longo prazo.
- Sazonalidade (S) - A repetição de um padrão com periodicidade conhecida.
- Erro (E) - Abrange tudo que não foi explicado pelas componentes anteriores da série.

Estas componentes podem ser combinadas de diferentes maneiras, formando o modelo puramente aditivo $y = T + S + E$, multiplicativo $y = T \times S \times E$, etc. A descrição completa e detalhada dos métodos pode ser vista em Hyndman et al. [26].
No gráfico do consumo total diário (Figura 2.2) nota-se a existência de um padrão sazonal anual, de modo que tentou-se decompor a série do consumo por meio dos modelos clássicos de decomposição. Apesar de estarem amplamente desenvolvidos, os modelos clássicos exigem suposições sobre as características da série que, caso não sejam satisfeitas, podem não produzir os melhores resultados. Não obstante a série apresentar uma estrutura complexa, alguns métodos clássicos foram aplicados aos dados usando as funções implementadas em `R`, `decompose` [31] e `stl` [9]. Os resultados podem ser vistos no Anexo A.3. A decomposição aplicando estes métodos não foi satisfatória, de modo que se

recorreu à uma técnica relativamente moderna conhecida como Análise Singular Espectral.

Análise Singular Espectral

Análise Singular Espectral (SSA, do inglês *Singular Spectrum Analysis*) é uma metodologia poderosa de estudo e previsão de séries temporais, que inclui vários métodos diferentes mas interligados entre si. A ideia principal de SSA consiste na aplicação da análise em componentes principais (PCA, do inglês *Principal Component Analysis*) [13] à "matriz trajetória" [19], obtida a partir da série original, com posterior reconstrução da série.
Um dos principais objetivos da SSA é decompor a série numa soma de componentes independentes e interpretáveis que, ao contrário dos métodos tradicionais de decomposição de séries temporais, é feito sem um conhecimento prévio da estrutura da série.
As quatro etapas da técnica estão representadas no diagrama da Figura 2.6; a descrição detalhada e a teoria completa de SSA, tal como numerosos exemplos da sua aplicação, podem ser vistos em Golyandina et al. [19] e Golyandina e Zhigljavsky [20], entre outros.

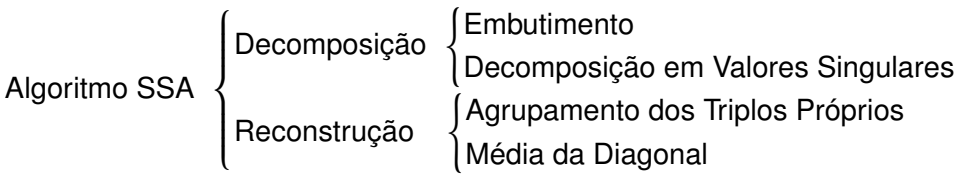


Figura 2.6: Algoritmo da Análise Singular Espectral (SSA).

Neste trabalho, a análise singular espectral completa foi utilizada numa fase exploratória da série do consumo energético total diário, devido à sua generalidade de aplicação e, apesar de não se expor o funcionamento da metodologia, será explicada a escolha dos parâmetros de entrada necessários. Assume-se que a série dada é uma soma de uma tendência, diferentes componentes oscilatórias, e um ruído. Os dois parâmetros a definir são: i) Tamanho da janela, L, ii) Agrupamento dos triplos próprios.

Tamanho da janela, L

O comprimento da janela deve ser suficientemente grande, $L \approx N/2$ onde N é o número de , e se se pretender extrair a componente periódica com período conhecido, produz melhores resultados um comprimento da janela divisível pelo período [20]. Se a série temporal tem uma estrutura complexa, por exemplo, a tendência não é monótona, que é o

caso do consumo energético desta instalação, Golyandina et al. [19] recomendam a SSA sequencial que consiste em duas etapas: na primeira fase, extrai-se a tendência com um comprimento pequeno da janela, e a seguir, detetam-se e extraem-se as componentes periódicas dos resíduos com $L \approx N/2$.

Neste caso, sendo $N = 974$, os 2 anos e 8 meses, existindo componentes periódicas na série (ver figuras da Secção 2.1) e a série tendo uma estrutura complexa, o L a tomar, na primeira fase, será o menor período disponível, uma semana, dado que os registos são diárias. Na segunda etapa, o comprimento da janela será igual a 365 ($< L \approx 487 = 974/2$).

Agrupamento dos triplos próprios

Uma vez executada a etapa de decomposição em valores singulares (SVD, do inglês, *Singular Value Decomposition*), o próximo passo é agrupar os triplos próprios de modo a separar a tendência e as componentes oscilatórias da série, do ruído, ou seja, este não deve apresentar uma estrutura de tendência e/ou sazonalidade. Cada triplo próprio consiste num vetor próprio (vetor singular esquerdo), um vetor fator (vetor singular direito) e um valor singular [19] e representa uma ou parte de uma das componentes da série original.

Quando é feita a SSA sequencial, na primeira etapa, a tendência pode ser extraída, como Golyandina e Zhigljavsky [20] referem, tendo em conta apenas o primeiro vetor próprio, uma vez que corresponde a uma variação suave da série. Por meio do comando `ssa(data,L=7)` do pacote `Rssa` de `R`, extraiu-se a tendência da série e o resultado pode ser visto na Figura 2.7, onde a série original está representada a verde, a tendência a vermelho e os resíduos (dados originais - tendência) a preto.

Examinando a Figura 2.7, pode dizer-se que a tendência cumpre o seu objetivo, ou seja, representa bastante bem a variação suave da série, o que resultou em resíduos com média zero. Se se tivesse usado um comprimento maior da janela, por exemplo 14 ou 21 que são ainda múltiplos de 7 (o período da série), obter-se-ia uma tendência mais suave, com menos pormenor, o que pode ser útil em alguns problemas.

O passo seguinte é extrair as componentes oscilatórias dos resíduos e isso é feito na segunda etapa da SSA sequencial com $L = 365$. Mais uma vez, é necessário agrupar os triplos próprios de modo a separar as componentes da melhor forma possível. Nesta etapa,

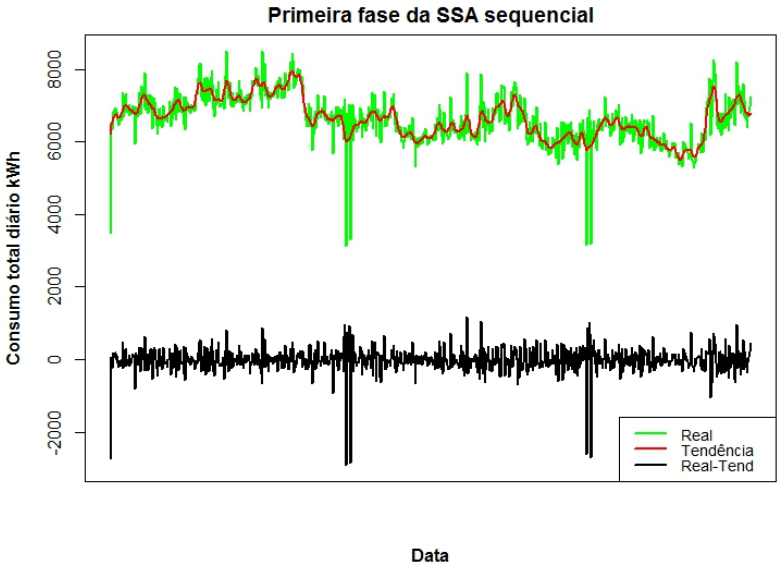


Figura 2.7: Resultado da primeira fase da SSA sequencial.

o agrupamento deve ser feito com base na informação fornecida por vários gráficos:

1. *Valores singulares* - Devem agrupar-se os triplos próprios que têm os valores singulares próximos. Golyandina et al. [19] estabeleceram uma regra: o ruído puro produz uma sequência lentamente decrescente de valores singulares e pode ser detetado através de uma quebra no espectro dos valores próprios. Analisando a Figura 2.8, poder-se-iam formar os seguintes grupos de triplos próprios: (1,2), (3,4), (5,6), (7,8), (9,10), os restantes representando o ruído.
   O agrupamento feito com base no gráfico dos valores singulares é uma ideia inicial que deve ser confirmada com os restantes gráficos.

2. *Matriz de correlação W* - Consiste das correlações ponderadas entre as componentes reconstruídas da série temporal. Examinando a matriz de correlação W, podem encontrar-se grupos de séries reconstruídas elementares correlacionadas e aproveitar esta informação para fazer o agrupamento. Mais uma vez, Golyandina et al. [19] definiram uma regra: não se devem incluir em grupos diferentes componentes correlacionadas. Considerando a Figura 2.9 e agrupando os triplos próprios com correlação elevada, quadrados pretos, formar-se-iam os grupos: (1,2), (3,4), (5,6) pois no restante espaço, estão presentes as tonalidades de cinza, que correspondem ao ruído.

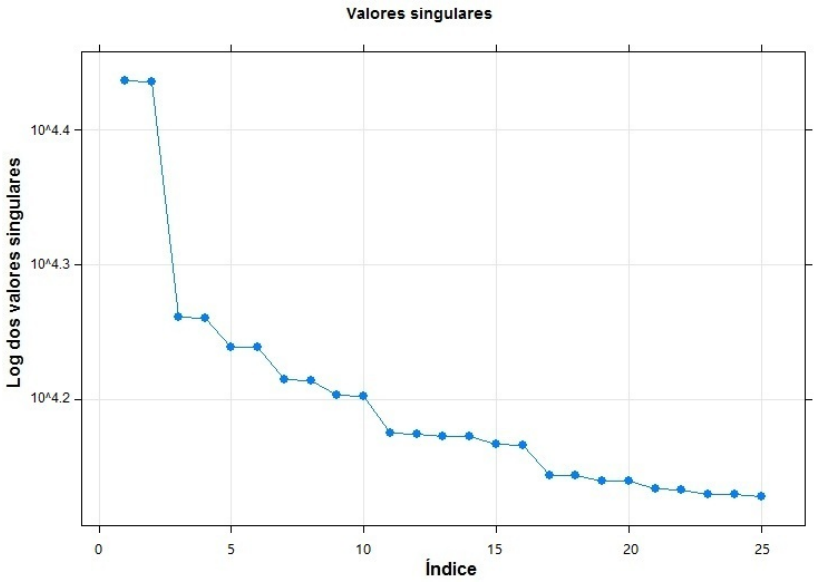


Figura 2.8: Valores singulares da segunda fase da SSA sequencial.

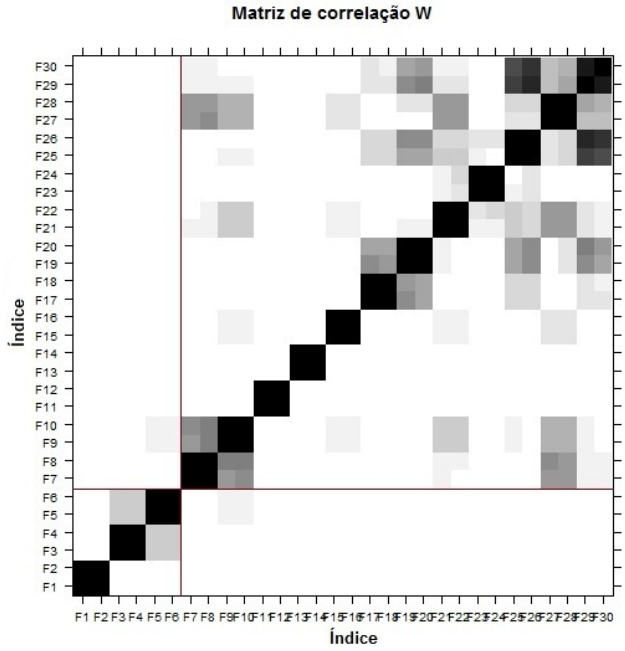


Figura 2.9: Matriz de correlação W da segunda fase da SSA sequencial.

Segundo a matriz de correlação W, como se formaram apenas três grupos, na série sem a tendência existem três componentes oscilatórias a separar. Antes de decidir o agrupamento final dos triplos próprios, deve averiguar-se um último gráfico.

3. *Vetores próprios sucessivos* - Segundo Golyandina e Zhigljavsky [20], no gráfico dos vetores próprios sucessivos, devem-se procurar polígonos regulares, que podem

estar em espiral, e agrupam-se os respetivos triplos próprios. Na Figura 2.10, podem distinguir-se os polígonos formados pelos seguintes pares de triplos próprios: (1,2), (3,4), (5,6), (7,8), (9,10), as mesmas 5 componentes oscilatórias identificadas no gráfico dos valores singulares.

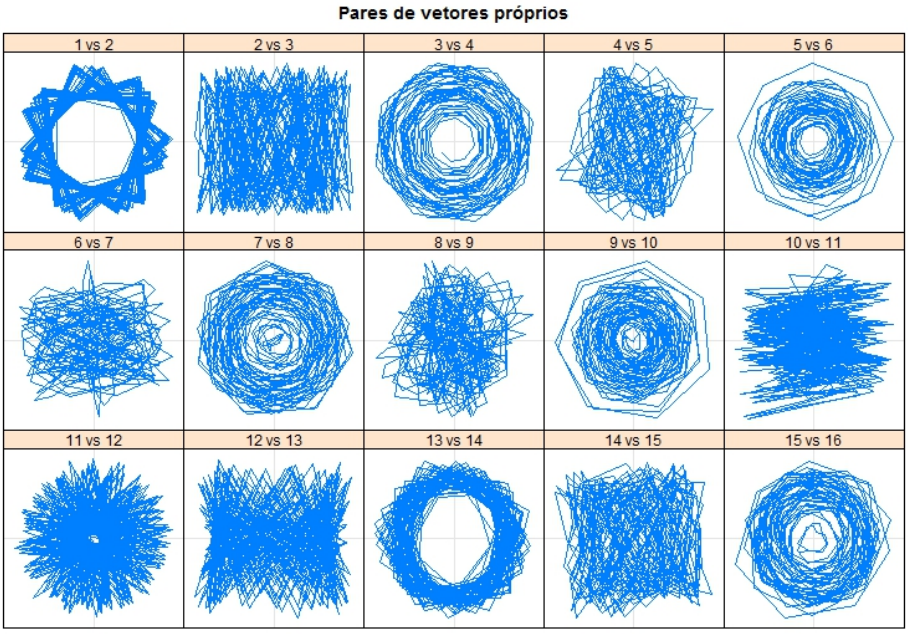


Figura 2.10: Pares de vetores próprios da segunda fase da SSA sequencial.

Uma vez feito o agrupamento, utilizando o resultado do gráfico dos valores e vetores próprios, prosseguiu-se com a reconstrução da série. A decomposição da série original em tendência, as cinco sazonalidades (cinco grupos) e os resíduos podem ser vistos na Figura 2.11.

A interpretação das componentes da série poderá ser: a tendência - variação suave da série, a sazonalidade 1 - representa a variação anual, as sazonalidades 2 e 3 dizem respeito à variação quadrimestral, enquanto que as sazonalidades 4 e 5 referem-se às variações que ocorrem de meio em meio ano. O método SSA não foi capaz de identificar os dias de Natal e de Ano Novo como periódicos devido ao curto período de tempo de que se dispunha. Assim, estas observações foram classificados como anormais e por conseguinte pertencem à componente residual.

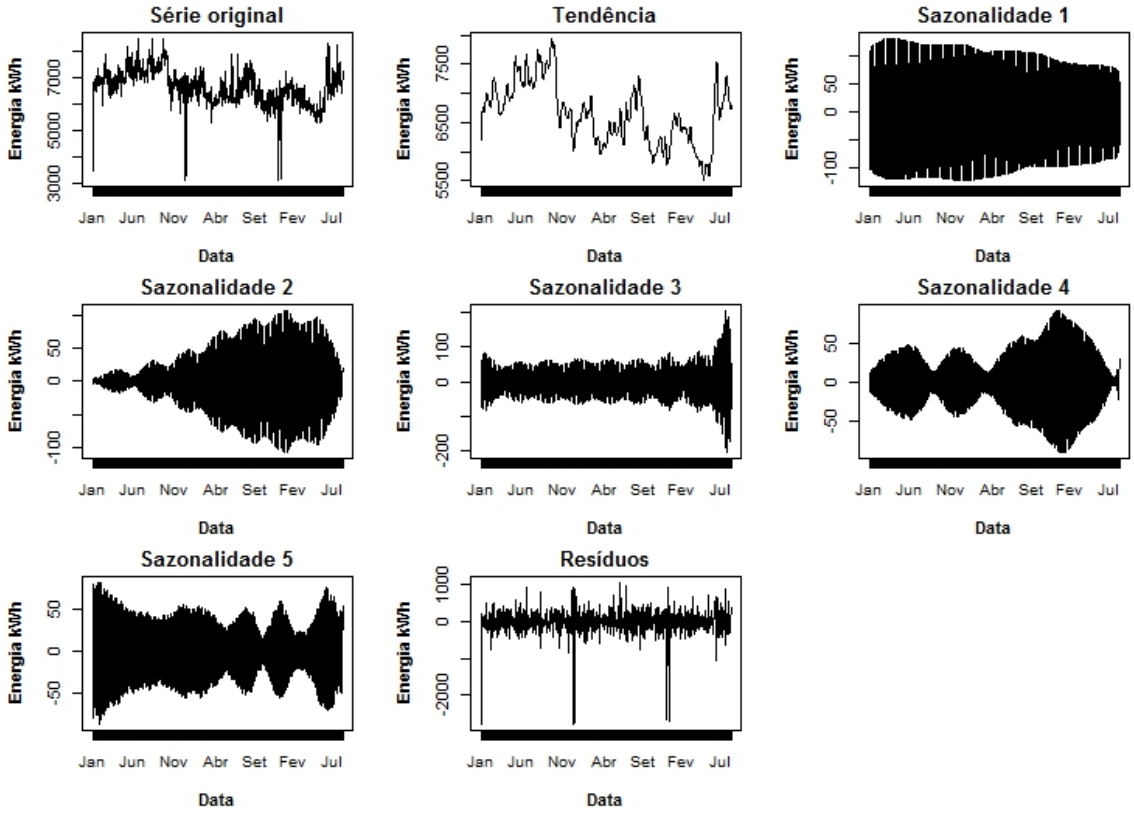


Figura 2.11: Decomposição da série original com a SSA sequencial.

Feito o estudo e a decomposição de uma instalação, pode aumentar-se a dimensão dos dados e abranger um maior número de instalações, mas para tal, é necessário selecionar aquelas que irão produzir resultados válidos. Com esse efeito, no capítulo seguinte estão descritos vários critérios que foram aplicados ao conjunto de 497 instalações.

Capítulo 3

Critérios de seleção de instalações

Os dados relativos às 497 instalações disponíveis em `SQL Server` poderiam inicialmente conter falhas nos dados ou instantes repetidos involuntariamente. Para corrigir isso, foram definidos vários critérios para selecionar o conjunto final das instalações válidas para serem exploradas. Todos eles foram implementados em `SQL Server` uma vez que a operação com datas e horas neste sistema é mais fácil e mais rápida do que em `R`.
De seguida estão apresentados os critérios referidos e, em simultâneo, os objetivos de cada um deles. Os primeiros dois critérios dizem respeito aos casos em que há pelo menos dois registos diferentes ou iguais no mesmo período de 15 minutos. Os restantes quatro servem para selecionar as instalações que podem ser estudadas. Os critérios são aplicados ordenadamente e traduzem-se numa sequência de filtros que, por exemplo, verificam no segundo critério as instalações que passaram no primeiro e assim sucessivamente, até ter os dados no formato desejado. Todas as decisões foram validadas por um perito em contadores e consumos energéticos. No entanto todos os valores de corte e parâmetros definidos, podem ser reajustados conforme o objetivo e o contexto do problema.

1. O consumo energético é registado por contadores que, devido à avarias ou por outras razões, podem ser submetidos à substituições. Quando esta substituição ocorre, pode existir mais de um registo no mesmo período temporal. Uma vez que não é sensato analisar dados com períodos repetidos, o primeiro critério resolve esse problema eliminando o registo do contador antigo nos momentos em que há duplicação de dados.

2. Quando uma instalação possui mais do que um contador de energia, existem consumos diferentes nos mesmos períodos de 15 minutos. Um exemplo de uma instalação deste tipo é uma faculdade composta por vários departamentos, em que cada um pode possuir um contador próprio que regista o respetivo consumo energético. Assim, para obter o consumo total de energia da instalação devem somar-se todos os registos do mesmo instante.

Feita esta prévia seleção de contadores e a eliminação de períodos repetidos, prosseguiu-se com a seleção de instalações válidas para a análise em causa. Os critérios construídos que fazem essa seleção estão enumerados mais abaixo. Os dados dos consumos energéticos de todas as instalações foram processados no Outono de 2013 portanto os últimos registos no máximo datam do dia 31 de Agosto de 2013.

1. Um dos objetivos da desagregação é ajudar o cliente a gerir o seu consumo energético no futuro. Se uma instalação cessou o seu trabalho no período de 1 de Janeiro de 2010 (a data mais antiga disponível) até 31 de Agosto de 2013, não faz sentido estudá-la. Assim, optou-se por avaliar as instalações que estavam ativas pelo menos nos últimos 6 meses, ou seja, em primeiro lugar, irão eliminar-se as instalações cuja última data de registo é anterior a 28 de Fevereiro de 2013, inclusive.

2. As instalações cujo consumo total diário é igual a zero em mais de 80% de dias durante o último ano de registos (de 1 de Setembro de 2012 até 31 de Agosto de 2013), serão excluídas. Mais concretamente, uma instalação com todos os dias presentes e com registos até 31 de Agosto de 2013, é eliminada se tiver mais de 292 dias (corresponde a 80% de 365) com consumo igual a zero. A consequente condição de eliminação foi imposta uma vez que se pretende considerar apenas instalações ativas num passado recente, já que isso aumenta a probabilidade de estarem ativas de momento.

3. Há instalações que têm um maior consumo energético durante o dia ou durante a noite, facto este relacionado com, por exemplo, o horário de funcionamento. Se faltarem vários registos durante alguma parte do dia, o consumo total diário será afetado ou enviesado pela outra parte. Para controlar isso, o número de falhas permitido foi dividido pelos quatro períodos diários definidos como:

- [07h00, 13h00[
- [13h00, 19h00[
- [19h00, 01h00[
- [01h00, 07h00[

Assim, são tomados apenas os dias que, em cada um dos períodos de 6 horas, têm no máximo 1 hora de falhas.

4. Com alguns dias eventualmente eliminados, o próximo passo é verificar se as instalações não têm muitas falhas durante o período todo de registo. O valor de corte estabelecido foi de 10% do número total possível de dias. Se o primeiro dia de registo for 1 de Janeiro de 2011 e o último for 31 de Agosto de 2013, mas a instalação tiver mais de 97 dias em falta (mais de 3 meses), distribuídos por todo o período, ela não será incluída no estudo.

Por fim, após a sequência de critérios descrita, aplicou-se novamente o primeiro critério ou seja, a última data de registo ser posterior a 1 de Março de 2013, uma vez que a eliminação de alguns dias podia colocar a instalação na posição de encerrada.
Um critério específico ao tema da desagregação é existirem pelo menos 2 meses de registos nas instalações em causa. Isto garante o mínimo de confiança nos resultados obtidos. Tal como os restantes critérios, este pode ser ajustado conforme o problema.
Nestas condições, do primeiro Lote, de um conjunto inicial de 98 instalações, restaram 97. `SQL Server` demorou 2 minutos e 20 segundos a aplicar os critérios, a agregar os dados por dia e a juntar com as respetivas tabelas das variáveis externas. Do número total de instalações disponíveis, de 497, restaram 471.

Capítulo 4

Regiões e variáveis externas

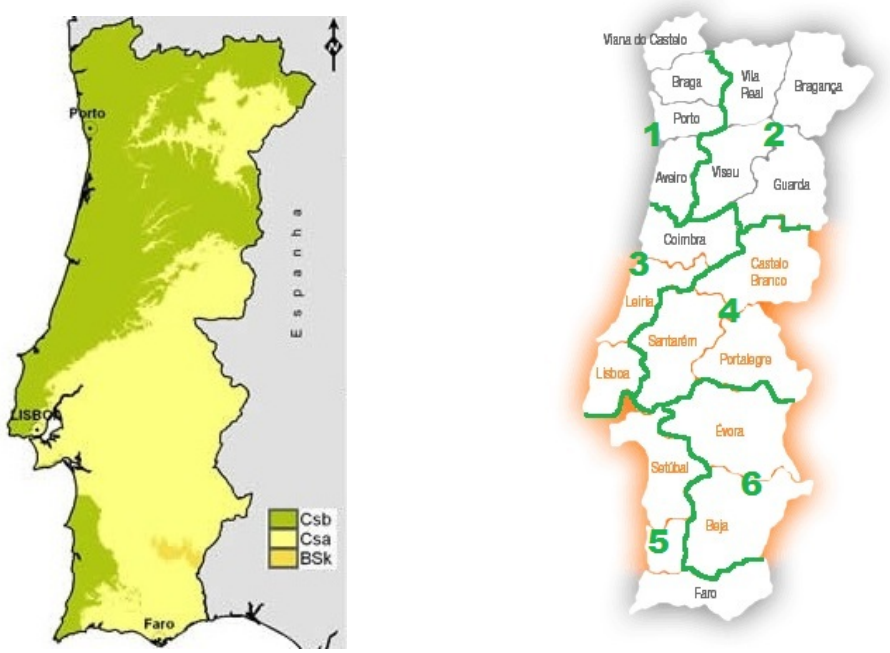
Para determinar a parte do consumo energético relacionada com as variáveis externas, é necessário saber quais são as variáveis que mais o influenciam. No site [54] e [40] dispunha-se inicialmente de 23 variáveis climáticas. A lista completa e a respetiva descrição pode ser vista no Anexo B.1. Como podia haver variáveis altamente correlacionadas ou que fossem simplesmente ruído, foi necessário aplicar vários métodos, descritos na Secção 4.2, de modo a averiguar estes casos. A pergunta que surge de imediato é a cidade ou a região cujas condições climáticas devem ser consideradas.

Cada instalação está situada num dos 18 distritos de Portugal e num dos 278 concelhos. Como é do conhecimento comum, existem algumas diferenças na temperatura, precipitação, humidade, no Norte e no Sul do país, no Litoral e Interior, diferenças estas que podem influenciar o impacto das variáveis climáticas sobre o consumo. Como não era sensato usar um único conjunto de variáveis externas para todos os distritos, pesquisaram-se as possíveis divisões climáticas de Portugal Continental. A descrição dos resultados encontra-se na Secção 4.1.

Em todos os métodos apresentados nas secções seguintes foi usado o consumo total diário. No entanto, tirar-se-iam as mesmas conclusões com o consumo médio diário.

4.1 Regiões de Portugal

Apesar de uma pesquisa exaustiva das possíveis divisões climáticas de Portugal, a única encontrada foi a classificação climática de Köppen-Geiger [27] [43]. De acordo com esta classificação, a maior parte do território tem um Clima Temperado - Tipo C, e apenas no distrito de Beja pode sentir-se um Clima Árido - Tipo B. O mapa de Portugal com as duas regiões climáticas de Köppen pode ser visualizado na Figura 4.1(a). Esta classificação baseia-se no pressuposto de que a vegetação de cada região da Terra é uma representação do respetivo clima [43]. Para tal, foi considerada a sazonalidade e os valores médios anuais da temperatura e da precipitação registados nos anos 1951-2000 [46].



(a) Classificação climática de Köppen

(b) Divisão em 6 regiões

Figura 4.1: Divisões de Portugal Continental.

Como se pretendia ter uma divisão mais refinada, pensou-se em considerar as variáveis climáticas dos 18 distritos de Portugal. No entanto, o site do qual foram extraídas as variáveis climáticas [54], dispõe de informações relativas apenas a 10 distritos de Portugal. A extração (por anos ou por meses) das variáveis para cada um dos distritos, e a organização destes dados em `Excel` de modo a ser possível lê-los em `R` ou `SQL Server`, tinha que ser feita manualmente. Como era um procedimento moroso, após um debate de grupo e uma

análise crítica subjetiva, decidiu-se pela divisão do país nas 6 regiões seguintes:

- Região **Porto** abrange os distritos: Viana do Castelo, Braga, Porto e Aveiro
- **Bragança** – Vila Real, Bragança, Viseu e Guarda
- **Lisboa** – Coimbra, Leiria e Lisboa
- **Castelo Branco** – Castelo Branco, Santarém e Portalegre
- **Faro** – Setúbal, Faro e Litoral de Beja (concelho Odemira)
- **Beja** – Évora e Interior de Beja (restantes concelhos)

O mapa com os distritos e com as 6 divisões aproximadas pode ver-se na Figura 4.1(b).

4.2 Variáveis Externas

Uma vez delimitadas as regiões de Portugal Continental, extraíram-se as variáveis climáticas, para cada uma delas, no período de 1 de Janeiro de 2010 até 31 de Agosto de 2013. No entanto, no Capítulo 2, para uma instalação, viu-se que apenas algumas das 28 variáveis conseguiam explicar a variabilidade do consumo energético. Antes de começar a seleção das variáveis com base num maior número de instalações, foi feita uma análise prévia das mesmas.
Observou-se que, em algumas variáveis, havia muitos valores em falta (>70%) e, como não é viável analisar variáveis deste tipo, elas foram retiradas do conjunto. As variáveis que estavam nestas condições eram:

- A variável Velocidade Máxima da Rajada do Vento, pois na região de Beja apresenta 90% de falhas (de 1339 dias, só há registo de 126), enquanto que na região de Bragança não há nenhum registo desta variável.
- A variável Cobertura com Nuvens, na região de Bragança, tem mais de 85% de valores em falta.

Quando uma variável é constante ao longo do tempo, ela não contem informação relevante, de modo que pode ser omitida do estudo. As variáveis deste tipo são:

- A variável Direção do Vento, que é constante e igual a -1 na região de Bragança e Castelo Branco, em todo o período temporal considerado.

- A variável Visibilidade Máxima - na região de Lisboa, em 3 anos e 8 meses é diferente de 10 em apenas 9 dias, sendo sempre constante no ano 2012 e 2013.

Outra variável retirada previamente foi a Precipitação, cujo significado era duvidoso, uma vez que era igual a zero nos dias quando havia chuva ou neve (existe outra variável, Eventos, que fornece esta informação).
Assim, após esta pequena análise, restaram as variáveis que a seguir entraram no estudo para determinar aquelas que mais influenciam o consumo energético. No entanto, ainda permanecia o problema de haver valores em falta para algumas das variáveis que, caso não fosse tratado, prejudicaria a análise enviesando os resultados [28]. Outra razão para o tratamento de falhas é a incapacidade de algumas instruções de `R`, nomeadamente a correlação, lidarem com valores em falta (as observações que têm NA (abreviação, em inglês, de *Not Available*) são eliminadas e pode correr-se o risco de ficar com poucos dados).

Preenchimento de falhas

Na literatura existem vários métodos para tratamento dessas falhas. Batista e Monard [4] sugerem eliminar dados com falhas, o que pode ser feito (i). nas observações, inaplicável neste caso, uma vez que, em alguns casos, há falhas em mais de um ano de observações, ou (ii). nas variáveis, o que também não é viável, pois podiam estar a eliminar-se variáveis relacionadas com o consumo energético. Outro método apresentado por Torgo [50] é a substituição das falhas pelo valor mais provável da variável em causa - moda se a variável for qualitativa, e média ou mediana se a variável for numérica. A desvantagem deste método é a sensibilidade das estatísticas aos valores extremos ou às distribuições assimétricas das variáveis. Neste caso, por exemplo, se faltar uma temperatura de um dia de verão, o valor imputado será influenciado pelas temperaturas de inverno.
Uma alternativa ao uso de todos os registos é a consideração apenas dos *k* casos mais semelhantes à observação com falhas - *k*-Vizinhos mais Próximos. Neste caso, isto traduz-se em estimar um valor em falta de um dia de verão, usando apenas os k dias mais

parecidos, que eventualmente serão também de verão. Os casos mais semelhantes são os que minimizam a função distância [28]. Apesar de haver várias métricas que medem a proximidade entre observações, a mais frequentemente utilizada, segundo Torgo [50], é a distância Euclideana:

$$d(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{1=1}^{p} \delta_i(\mathrm{x}_i,\mathrm{y}_i)},$$

onde p é o número de variáveis preditivas e $\delta_i(\cdot,\cdot)$ é a distância entre dois valores da variável i, dada por

$$\delta_i(\mathrm{x}_i,\mathrm{y}_i) = \begin{cases} 1 & \text{se } i \text{ é nominal e } \mathrm{x}_i \neq \mathrm{y}_i \\ 0 & \text{se } i \text{ é nominal e } \mathrm{x}_i = \mathrm{y}_i \\ (\mathrm{x}_i - \mathrm{y}_i)^2 & \text{se } i \text{ é numérica.} \end{cases}$$

Como as variáveis podem estar em escalas diferentes e isso pode influenciar a medida de semelhança entre observações, é necessário normalizar previamente os dados:

$$\mathrm{z_i} = \frac{\mathrm{x}_i - \bar{\mathrm{x}}}{\sigma_{\mathrm{x}}}.$$

Neste trabalho, para preencher as falhas, recorreu-se precisamente a este método, uma vez que parecia ser o mais adequado ao problema. Como Jönsson e Wohlin [28] afirmam, vários estudos mostraram que *k*-Vizinhos mais Próximos é tão bom ou melhor do que os outros métodos para preencher falhas. A estatística utilizada foi a mediana, uma vez que é mais robusta a valores extremos: se um dos vizinhos mais próximos tiver um valor anormal na variável em que há falhas, mas for semelhante nas restantes variáveis, o valor imputado não será influenciado pelo outlier. Quando todos os valores são semelhantes, a mediana está próximo da média.

O passo seguinte foi a escolha do número de casos semelhantes a usar. Jönsson e Wohlin [28] provaram que o *k* mais adequado a usar é aproximadamente a raiz quadrada dos casos completos. O menor número de dias sem falhas foi de 869 na região de Bragança, que resultou em 29 vizinhos. No mesmo estudo mostrou-se que uma pequena variação no *k* não alterava a eficácia do método, facto desejável caso *k* fosse par, uma vez que a mediana seria a média dos valores centrais e a maior parte das variáveis devem ser inteiras.

Com a tabela das 18 variáveis climáticas completa, mais as 5 variáveis construídas no Capítulo 2, prosseguiu-se com a escolha das variáveis mais relacionadas com o consumo energético. Para isso foram utilizados três métodos: regressão linear múltipla, floresta aleatória (do inglês *random forest*) e a correlação cruzada e parcial, sendo o resultado final uma combinação de todos. A escolha das variáveis mais significativas foi feita com base nas 97 instalações que passaram nos critérios descritos no Capítulo 3.

4.2.1 Regressão linear múltipla

Na Secção 2.2 do Capítulo 2, já foi feita uma breve descrição da regressão linear múltipla e viu-se também a sua aplicabilidade numa instalação. Neste capítulo, o procedimento de seleção de variáveis mais significativas repetiu-se para as 97 instalações. Tal como anteriormente, aplicou-se a seleção automática (usando o comando `step` de `R`) das variáveis, com o objetivo de obter apenas aquelas que mais influenciam o consumo. A variável resposta continuou a ser o consumo total diário, sendo as variáveis climáticas as preditoras. Os resultados apresentados na Tabela 4.1 indicam a frequência relativa de cada uma das variáveis no conjunto total de instalações.

Variável	Estação	Ano	Feriado	Humidade Mínima	Dia da Semana
Frequência	98	95	77	77	75
Variável	Comprimento do Dia	Temperatura Máxima	Visibilidade Média	Ponto Orvalho Médio	Pressão Máxima
Frequência	71	67	58	56	53
Variável	Pressão Média	Ponto Orvalho Máximo	Pressão Mínima	Humidade Média	Temperatura Média
Frequência	52	51	40	37	37
Variável	Ponto Orvalho Mínimo	Veloc. Média do Vento	Eventos	Temperatura Mínima	Visibilidade Mínima
Frequência	34	34	33	32	29
Variável	Humidade Máxima	Veloc. Máxima do Vento	Fim de Semana		
Frequência	14	10	4		

Tabela 4.1: Frequência relativa das variáveis selecionadas através da regressão linear múltipla.

Da análise da tabela, podem distinguir-se as variáveis que foram reconhecidas como significativas para a estimação do consumo energético em >65% de instalações, nomeadamente, por ordem decrescente de frequência relativa: Estação, Ano, Feriado, Humidade Mínima, Dia da Semana, Comprimento do Dia, Temperatura Máxima. Contudo, os resultados obtidos através da regressão linear múltipla não podem ser seguidos cegamente, pois pode existir correlação elevada entre as variáveis finais. Esta característica é estudada na Secção 4.2.3.
O tempo de construção dos modelos e da seleção das variáveis para as 97 instalações foi de 1 minuto e meio.

4.2.2 Floresta aleatória (*random forest*)

A floresta aleatória é um método frequentemente utilizado na comunidade científica para classificação e regressão, uma vez que consegue lidar com dados com poucas observações e muitas variáveis, com dados que têm interações complexas, e até com os que têm correlações elevadas entre variáveis preditoras [48]. A floresta aleatória consiste num conjunto de modelos baseados em árvores, onde cada árvore é obtida escolhendo, aleatoriamente, durante a sua construção, observações e variáveis dos dados originais. A descrição do algoritmo pode ser vista em Breiman [6], o criador da floresta aleatória.
Uma instrução de `R`, que tem implementado o método da floresta aleatória original, encontra-se no pacote `randomForest` e é `randomForest(Y~.,data,mtry,ntree,importance=T)`, onde `Y` é a variável resposta do conjunto `data`, neste caso o consumo total diário, `mtry` é o número de variáveis amostradas aleatoriamente a tomar em cada nó, `ntree` - número de árvores a construir, `importance=T` para obter a significância das variáveis na previsão do consumo. A floresta aleatória é relativamente robusta à escolha dos parâmetros [7], uma vez que tem valores por defeito, embora o resultado possa ser melhorado ajustando-os ao problema.
Como Breiman [7] advertiu, para haver estabilidade na importância das variáveis, é necessário construir pelo menos 1000 árvores - ao mesmo tempo garante que o algoritmo é suficientemente rápido com um número elevado de árvores, e não cria sobre-ajustamento aos dados. Quanto ao número de variáveis amostradas aleatoriamente consideradas em cada nó, o valor por defeito no algoritmo, para regressão, é igual ao número de variáveis

independentes/3 [35] (o consumo total é uma variável contínua). Como há 23 variáveis climáticas, tomaram-se 8, uma vez que Breiman em [7] aconselha o uso de um maior número quando há variáveis de ruído presentes nos dados. No entanto, para outros valores próximos de 8, os resultados são os mesmos com diferenças mínimas nas frequências relativas.

Um exemplo da aplicação da floresta aleatória numa instalação e da seleção das 10 variáveis mais significativas, pode ser visto no Anexo B.2.

Na Tabela 4.2 tem-se o resultado da aplicação do método floresta aleatória ao conjunto de 97 instalações, selecionando para cada uma as 10 variáveis com maior incremento no erro quadrático médio. Breiman [7] descreveu a estimação da importância de uma variável do seguinte modo: na passagem de um nó para o outro, reordenam-se aleatoriamente todos os valores da variável e calcula-se o incremento no erro médio quadrático do nó anterior para esse novo nó; quanto maior for esse incremento maior é a importância da variável. O erro médio quadrático é definido por $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$ onde n é o número total de observações, $\hat{Y}_i$ é o valor previsto pelo modelo e Y_i é o valor real.

Variável	Comprimento do Dia	Estação	Temperatura Média	Temperatura Máxima	Temperatura Mínima
Frequência	99	99	96	95	91
Variável	Ano	Ponto Orvalho Máximo	Dia da Semana	Humidade Média	Feriado
Frequência	84	78	64	62	52
Variável	Humidade Mínima	Ponto Orvalho Médio	Ponto Orvalho Mínimo	Visibilidade Média	Fim de Semana
Frequência	49	39	23	14	13
Variável	Pressão Mínima	Pressão Máxima	Veloc. Média do Vento	Humidade Máxima	Pressão Média
Frequência	9	8	8	5	5
Variável	Eventos	Veloc. Máxima do Vento	Visibilidade Mínima		
Frequência	4	1	1		

Tabela 4.2: Frequência relativa das variáveis selecionadas através da floresta aleatória.

Os resultados obtidos através da floresta aleatória são semelhantes aos obtidos usando a

regressão linear múltipla: as variáveis Comprimento do Dia, Estação, Temperatura Máxima, Ano, Dia da Semana e Feriado continuam a ser significativas para o consumo em mais de 50% de instalações, apesar de aparecerem por outra ordem. Outras variáveis também foram destacadas como importantes em mais de 50% de instalações para explicar o consumo e estas são: a Temperatura Média e Mínima, Ponto Orvalho Máximo e a Humidade Média.

O tempo que a ferramenta `R` demorou para devolver este resultado foi de, aproximadamente, 19 minutos.

Nas medições diárias das variáveis externas, como é do conhecimento empírico, existe correlação elevada. Por exemplo, a Temperatura Média está fortemente relacionada com a Temperatura Mínima e Máxima. Analisando a Tabela 4.2, pode notar-se que a floresta sinalizou as três temperaturas como as variáveis mais significativas, isso porque, como Strobl et al. [48] observaram, no processo de construção das árvores existe uma predisposição para a seleção de preditores correlacionados. Com base nestas considerações, eles desenvolveram um algoritmo de permutações condicionais que reflete o impacto verdadeiro de cada preditor sobre a resposta, implementado em `R` por Hothotn et al. [23] e Strobl et al. [47] [48] no pacote `party` na instrução `cforest()`. Por meio de `varimp(modelo,conditional=T)`, onde `modelo` é a resposta da função anterior, obtêm-se as variáveis não correlacionadas que mais influenciam a resposta.

Essa nova abordagem testou-se para um ano de registos numa instalação. A construção das árvores demorou 42 minutos, mas para a determinação da importância das variáveis não houve memória suficiente no computador. Como a maior parte das instalações têm mais de um ano de registos e pretendia-se que os resultados fossem validados num maior número de instalações, a aplicabilidade deste método não é viável. Assim, foi estudada uma outra via de seleção das variáveis mais significativas para o consumo energético, correlação cruzada e parcial, descritas na secção seguinte.

4.2.3 Correlação cruzada e parcial

Dado um conjunto de dados com diversas variáveis, se o objetivo for diminuir o seu número, deve ter-se o cuidado de não reter variáveis que contêm a mesma informação, ou seja, não considerar as que estão altamente correlacionadas entre si. Pretende-se ao mesmo tempo

obter as variáveis que mais influenciam o consumo.
Com esse intuito, o primeiro passo foi selecionar, do conjunto total de variáveis climáticas numéricas, apenas aquelas que têm correlação parcial estatisticamente significativa com o consumo total diário. Neste ponto devem ser introduzidas algumas notas informativas:

- A análise de correlação parcial envolve o estudo da relação entre duas variáveis, após a eliminação do efeito dos outros fatores independentes [45]. Foi a utilizada na primeira etapa, uma vez que se pretendia ver a relação do consumo energético com cada uma das variáveis, sem a influência das outras.
- Tomaram-se apenas as variáveis numéricas porque as instruções implementadas em `R` para o cálculo da correlação não aceitam variáveis categóricas.

A instrução em `R` que calcula a correlação parcial é `correl <- pcor()` do pacote `ppcor`. A escolha das variáveis mais correlacionadas com o consumo foi feita usando 97 instalações do seguinte modo: para cada uma das instalações, consideraram-se as variáveis que tinham correlação estatisticamente significativa (valor p, obtido através de `correl$p.value`, inferior a 0.05) com o consumo total. Na Tabela B.1 do Anexo B.3 encontra-se a significância das correlações entre o consumo diário total de uma instalação e algumas variáveis climáticas.

Variável	Comprimento do Dia	Humidade Mínima	Ponto Orvalho Máximo	Temperatura Máxima	Humidade Média
Frequência	77	52	40	28	26
Variável	Visibilidade Média	Ponto Orvalho Médio	Visibilidade Mínima	Ponto Orvalho Mínimo	Veloc. Média do Vento
Frequência	25	20	16	13	11
Variável	Temperatura Mínima	Pressão Máxima	Pressão Média	Temperatura Média	Humidade Máxima
Frequência	10	9	6	6	5
Variável	Veloc. Máxima do Vento	Pressão Mínima			
Frequência	5	4			

Tabela 4.3: Frequência relativa das variáveis climáticas com correlação estatisticamente significativa com o consumo total com base em 97 instalações.

Na Tabela 4.3 está apresentada a frequência relativa das variáveis climáticas numéri-

cas, como resultado do procedimento anterior, com base em 97 instalações, ou seja, a proporção de vezes que a correlação entre uma variável e o consumo energético total foi considerada estatisticamente significativa. O tempo que `R` demorou para devolver o resultado foi de 16 segundos.
Com base nestes resultados, dir-se-ia que apenas as variáveis Comprimento do Dia, Humidade Mínima e Ponto Orvalho Máximo têm uma relação forte com o consumo energético.
O passo seguinte foi: para cada uma das regiões descritas na Secção 4.1, através da matriz de correlação cruzada entre variáveis climáticas (desta vez deseja-se avaliar o efeito de umas variáveis sobre as outras), eliminavam-se as variáveis que tivessem correlação maior que 0.65 com a variável mais frequente da Tabela 4.3. Para ficar mais claro, apresenta-se um exemplo:

1. Consideram-se os dados das variáveis climáticas numéricas de uma das 6 regiões.

2. Usando o comando `rcorr` do pacote `Hmisc` constrói-se a matriz de correlação cruzada entre estas mesmas variáveis. Na Tabela B.2 do Anexo B.3 tem-se a matriz de correlação entre algumas variáveis climáticas da região de Beja.

3. Seleciona-se primeiramente a variável Comprimento do Dia (porque é a primeira na Tabela 4.3, ou seja, nas 97 instalações é a que apareceu mais vezes como correlacionada com o consumo energético).

4. Com base na matriz de correlação construída no passo 2, eliminam-se todas as variáveis que têm correlação com a variável Comprimento do Dia maior que 0.65. Supõe-se que estas variáveis são Humidade Mínima e Temperatura Máxima.

5. A próxima variável a considerar será Ponto Orvalho Máximo, uma vez que é a variável mais frequente a seguir ao Comprimento do Dia, não correlacionada com esta. O algoritmo repete-se até esgotar a lista das variáveis.

Inicialmente, em vez de se ter 0.65 como valor de corte, tentou-se aplicar o mesmo princípio que anteriormente, ou seja, correlações estatisticamente significativas, no entanto era um critério muito rígido uma vez que a única variável final era o Comprimento do Dia. Tentaram-se também outros valores de corte tais como 0.6, 0.7 e 0.75, no entanto, o mais adequado pareceu ser 0.65. O resultado para cada uma das regiões pode ver-se na Tabela 4.4. A

Beja	Bragança	Castelo Branco	Faro	Lisboa	Porto
Comprimento do Dia	Comprimento do Dia	Comprimento do Dia	Comprimento do Dia	Comprimento do Dia	Comprimento do Dia
Humidade Mínima	Humidade Mínima	Humidade Mínima	Humidade Mínima	Humidade Mínima	Humidade Mínima
Ponto Orvalho Máximo	Ponto Orvalho Máximo	Ponto Orvalho Máximo	Ponto Orvalho Máximo	Ponto Orvalho Máximo	Ponto Orvalho Máximo
Visibilidade Média		Visibilidade Média		Visibilidade Média	Visibilidade Média
Veloc. Média do Vento	Veloc. Média do Vento	Veloc. Média do Vento	Veloc. Média do Vento	Veloc. Média do Vento	Veloc. Média do Vento
Pressão Máxima	Pressão Máxima	Pressão Máxima	Pressão Máxima	Pressão Máxima	Pressão Máxima
Humidade Máxima		Humidade Máxima			Humidade Máxima
Veloc. Máxima do Vento		Veloc. Máxima do Vento			

Tabela 4.4: As variáveis mais correlacionadas com o consumo total e com correlação inferior a 0.65 entre sí para as 6 regiões.

obtenção dos resultado em R para todas as regiões foi imediata.

Conjugando os resultados das Tabelas 4.3 e 4.4, conclui-se que as variáveis climáticas numéricas que mais influenciam o consumo são: Comprimento do Dia, Humidade Mínima e Ponto Orvalho Máximo. Como as restantes variáveis apresentam frequências relativas iguais ou inferiores a 25%, não foram consideradas nas análises futuras.

Decisão final

Como foi visto anteriormente, a escolha das variáveis que mais influenciam o consumo total diário, através dos métodos das Secções 4.2.1 e 4.2.2, não fornece os melhores resultados, uma vez que não é avaliada a relação entre as variáveis climáticas. A solução para esta contrariedade foi encontrada na matriz de correlação, Secção 4.2.3, que, no entanto, não avalia as variáveis nominais. Assim, através da regressão linear múltipla e floresta aleatória foram destacadas como variáveis significativas: Estação, Ano, Dia de Semana e Feriado, enquanto que a correlação deu ênfase às variáveis Comprimento do Dia, Humidade Minima e Ponto Orvalho Máximo.

Capítulo 5

Desagregação do consumo energético

A desagregação pode ser definida como a separação de um total em várias partes que o compõem. No contexto do consumo energético, a sua desagregação pode ser feita de duas maneiras:

- no eixo do tempo. Dado o consumo total de um período (mês, por exemplo), o objetivo é decompor esse total em consumos por intervalos menores, dias. Isso pode ser feito através da média aritmética - todos os dias têm o mesmo consumo; a média ponderada - dispondo de alguma informação adicional sobre a instalação, podem associar-se pesos aos dias (por exemplo, nos fins de semana haverá um menor consumo do que nos dias úteis), etc. A desagregação de séries temporais em intervalos no eixo temporal pode ser vista na dissertação de Vitullo [53] que, tendo o consumo mensal e anual, os divide em consumos diários e quadrimestrais, respetivamente.

- no eixo da energia. Dado o consumo energético num intervalo de tempo, o objetivo é decompô-lo nos subconsumos que o constituem, mas mantendo a escala temporal. Este tipo de desagregação é a estudada neste projeto.

A desagregação do consumo energético pode ser feita em vários domínios, sendo de salientar dois: i) residencial ii) industrial.

Desagregação da energia residencial

A gestão racional do consumo energético é extremamente importante. Os custos económicos e os problemas ambientais associados à queima de combustíveis fósseis estão a tornar-se cada vez mais prejudiciais para o bem estar de muitas espécies, inclusivé a humana. O uso energético no setor residencial contribui significativamente para a emissão de gases que aumentam o efeito de estufa. Os estudos provaram que, se cada utilizador doméstico do Reino Unido reduzir o consumo elétrico em 10%, então a emissão anual de CO_2 do Reino Unido irá diminuir em 6 milhões de toneladas [30].

Há duas abordagens típicas para o problema da poupança de energia: eficiência e redução [32]. A eficiência envolve ações singulares (por exemplo, a utilização de aparelhos mais eficientes em termos energéticos) que têm um custo maior. A redução requer uma participação contínua (por exemplo, utilizar menos aquecimento/arrefecimento num dia ameno) que tem um custo menor. De acordo com Attari et al., Gardner e Stern Riche et al., citados por Kim [32], há duas questões gerais que inibem os consumidores de aplicar estas técnicas: 1) O uso energético é um conceito bastante abstrato para a maioria dos consumidores; 2) Os consumidores têm frequentemente pouco conhecimento sobre como a energia é distribuída em casa e, sobre as ações que beneficiam mais a poupança da energia.

Como solução para estes inconvenientes e para ajudar os consumidores a gerir o seu consumo energético, foram desenvolvidos vários métodos que, partindo do consumo energético total da casa, permitem identificar o gasto energético de cada um dos aparelhos. O objetivo final está esquematizado na Figura 5.1.

A este propósito, existem numerosos estudos que, abordando diferentes interpretações, conseguem desagregar o consumo total de uma casa. Por intermédio dos medidores inteligentes (do inglês, *smart meters*) e conhecendo a curva de consumo de alguns dispositivos instalados em casa (computador, máquina de lavar roupa/louça, televisão, etc), Lines et al. [36], Kelly [30], Holcomb [22] e Froehlich [16], dispondo de um conjunto de treino, tentaram dividir o consumo total diário e semanal em partes que correspondessem a cada aparelho. Kelly [30] e Zefman [56] tiveram em consideração também a potência ativa e reativa de cada um dos dispositivos.

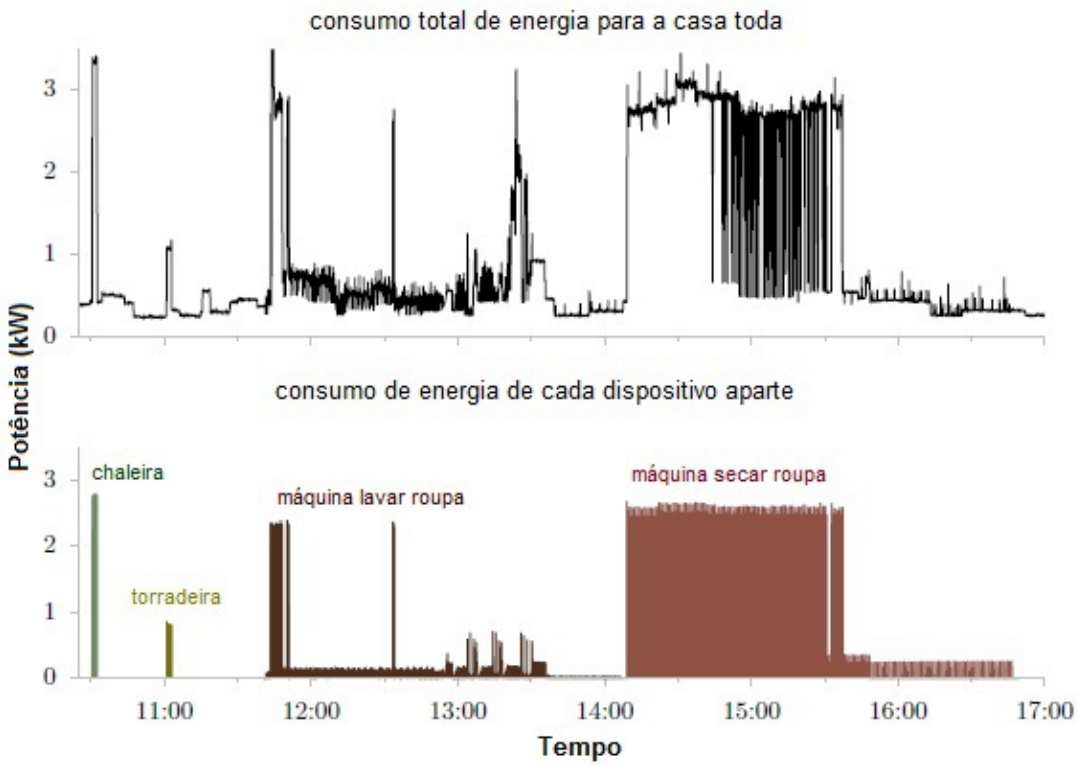


Figura 5.1: Desagregação do consumo energético residencial.

Uma abordagem relativamente diferente pode ser encontrada em Farinaccio e Zmeureanu [14] que usam a potência para definir o evento: eles ligam um aparelho de cada vez e a diferença entre os dois estados da potência irá caraterizar o dispositivo em causa. Uma outra maneira de medir o consumo de cada um dos aparelhos elétricos é instalar medidores separados em cada um deles. Esta técnica está referida em Kim [32] que adverte para o custo elevado e, portanto, para a sua pouca utilidade.

Um problema comum aos estudos que se focaram na desagregação do consumo residencial, foi a presença de ruído gerado pelo funcionamento de vários pequenos aparelhos, ou pela variação da tensão elétrica nas linhas de serviço público, que poderia ocultar o sinal real [14]. Como consequência, todos os resultados têm um erro associado.

Não obstante, um estudo de campo conduzido por Fitch e mencionado por Kim [32], mostrou que uma resposta adequada da desagregação em causa, pode levar a uma redução do consumo energético residencial até 50%, apesar de as poupanças mais comuns se encontrarem na faixa dos 9-20%.

Feita uma breve descrição da desagregação residencial, inaplicável para as instalações disponíveis, uma vez que não se dispunha de qualquer outra informação sobre a instalação, para além dos diagramas de carga e da região geográfica, prossegue-se com a explicação da desagregação industrial, o tema fulcral deste estágio.

Desagregação da energia industrial

Consiste na decomposição do consumo energético em três partes: a componente dependente da produção (será chamada de energia útil), das condições meteorológicas (energia das variáveis externas) e a componente independente (energia de base). Estas componentes podem ser definidas mais detalhadamente como:

Baseload - quantidade mínima de energia elétrica exigida ao longo das 24 horas, todos os dias, por uma instalação ativa. Um exemplo disso seria a energia gasta por um alarme que está ligado constantemente enquanto a instalação estiver funcional.

Energia útil ou **Trabalho** - quantidade de energia elétrica necessária para produzir. Por exemplo, numa fábrica que produz mesas, a energia útil será a quantidade de energia diária necessária para produzir as mesas. Numa faculdade, esta componente irá corresponder à energia utilizada pelos computadores, fotocopiadores, etc.

Energia das variáveis externas - quantidade de energia elétrica gasta devido às condições meteorológicas. Será, por exemplo, a energia gasta por um ar condicionado quando o arrefecimento/aquecimento é necessário.

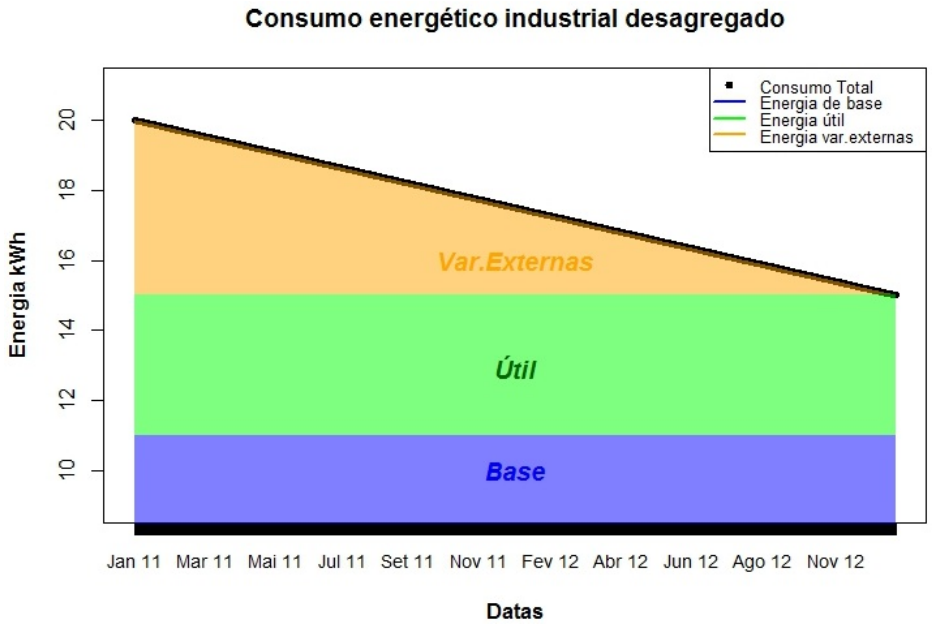


Figura 5.2: As três faixas do consumo energético industrial.

A soma das três componentes, energia de base, útil e externa, resulta no consumo energético total. Na Figura 5.2, tem-se o consumo total diário durante 2 anos e 8 meses e as três faixas descritas anteriormente que o constituem. A energia de base e a útil são constantes ao longo do tempo, embora este comportamento não seja obrigatório nas instalações reais, como se verá neste capítulo.

A desagregação do consumo energético industrial tem diversas aplicações e benefícios tais como: a quantificação da parte energética não utilizada para a produção, a identificação das oportunidades de eficiência energética, um melhor planeamento da rede da energia elétrica, entre outros.

Abels et al. [1] descreveram um método estatístico que desagrega o consumo energético industrial nas três componentes, dispondo da informação sobre o consumo total da instalação, a temperatura exterior média diária e os dados sobre a produção da instalação. No entanto, o método descrito por eles não podia ser seguido pois, mais uma vez, não havia acesso ao ramo de atividade da instalação e, por conseguinte, não se conhecia a produtividade da instalação. Assim, foram testadas várias abordagens para encontrar cada uma das componentes do consumo energético, descritas nas secções seguintes.

De relembrar que o comprimento mínimo da série necessário para poder efetuar a desagregação é de 2 meses, critério mencionado no Capítulo 3. Como o objetivo principal deste trabalho é separar a energia em três partes, energia de base, útil e das variáveis externas, para uma primeira análise, foi aplicado o índice de aquecimento e de arrefecimento que separa a parte do consumo energético dependente das variáveis externas da independente, que neste caso, é a soma da energia de base com a útil.

.

5.1 Índice de aquecimento e Índice de arrefecimento

Ferreira [15] define o Índice de aquecimento/arrefecimento (HDD e CDD, do inglês, *Heating Degree Days* e *Cooling Degree Days*, respetivamente) como a medida que avalia o quanto (em graus) e por quanto tempo (em dias) a temperatura exterior estava abaixo/acima da temperatura de base dada.

O índice de arrefecimento é a medida oposta ao índice de aquecimento: o primeiro é

calculado usando os dias em que a temperatura exterior estava maior do que o valor de referência, enquanto que o índice de aquecimento considera os dias em que a temperatura estava abaixo da de base. Como o procedimento de determinação de cada um dos índices e os problemas são idênticos, neste trabalho está descrito e estão apresentados os resultados relativos apenas ao índice de aquecimento.

Bromley [8] refere que os índices são frequentemente utilizados na determinação do consumo energético necessário para aquecer/arrefecer o edifício. Esta técnica assume que a parte independente das variáveis externas é constante ao longo do ano.

Tendo em conta o facto de que a temperatura interior depende da exterior e que esta, por sua vez, está em constante mudança, os métodos em causa são baseados na ideia de que a quantidade de energia necessária para aquecer o edifício num determinado período é diretamente proporcional ao índice de aquecimento nesse mesmo período: quanto menor for a temperatura exterior, mais energia é requerida para manter o ambiente confortável dentro do edifício.

O maior problema da técnica do índice de aquecimento é a definição da temperatura de base de um edifício, a chave do método, que é a temperatura acima da qual não é necessário o aquecimento. Este parâmetro difere de um edifício para outro e varia ao longo do ano, devido ao sol, vento e padrão de ocupação do edifício, que tipicamente se encontram em constante variação. Atendendo a esta variabilidade, em geral toma-se um valor aproximado para a temperatura de base. Esta instabilidade do método pode facilmente conduzir a resultados imprecisos ou errados.

Para o Reino Unido, a temperatura de base é igual a 15.5°C enquanto que para os Estados Unidos é de 18°C [11]. Se o objetivo for estudar o consumo energético de uma instalação, Bromley [8] avisa que o valor de referência a usar deve ser o mais apropriado possível ao edifício em causa. Não se encontrando uma referência para a temperatura de base para Portugal e não tendo informações sobre as instalações em análise, foram considerados três valores para a temperatura média de conforto, 16, 17 e 18°C. Não estando disponível nenhuma maneira para validar ou rejeitar os resultados, o índice de aquecimento foi utilizado como análise exploratória da desagregação do consumo energético.

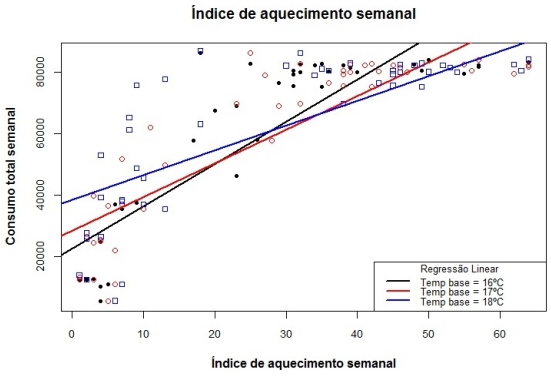
A componente do consumo total semanal, independente das variáveis externas, pode ser calculada através de regressão linear: no eixo dos xx representa-se o índice de aquecimento acumulado semanal (isto é, obtido somando os índices de aquecimento diários

relativos a essa semana) e, no eixo das ordenadas, representa-se o consumo total acumulado durante essa semana, mas somando apenas os valores do consumo nos dias em que o índice de aquecimento é não nulo. Não se consideram as semanas em que o índice de aquecimento acumulado semanal é nulo (por exemplo, semanas nos meses de verão). O ponto de intersecção com o eixo dos yy, da reta de regressão estimada, constitui a parte independente semanal das variáveis externas. O processo para a determinação do índice de aquecimento está representado no diagrama da Figura 5.3 onde a temperatura de base é igual a 16°C. Para o índice de arrefecimento, o procedimento é feito para temperaturas superiores ao valor de referência.

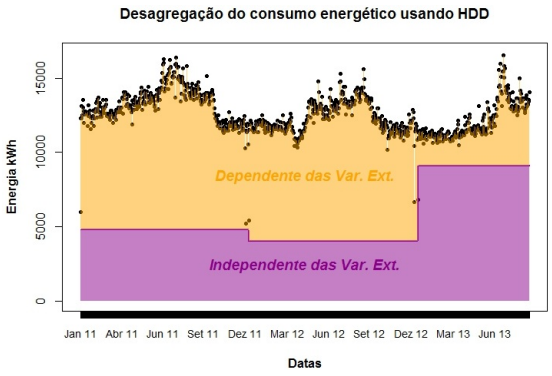
	⋯	07-02-2013	08-02-2013	09-02-2013	10-02-2013	⋯
Temp. Média	⋯	14	16	17	15	⋯
Índice aquecim.	⋯	2	0	0	1	⋯

Figura 5.3: Diagrama do cálculo do índice de aquecimento.

A componente independente das variáveis externas foi calculada apenas para algumas instalações, obtendo-se um valor constante para cada ano de registo. Na Figura 5.4(a) está representada a regressão linear do consumo total semanal de uma instalação, para um ano de registo, contra o índice de aquecimento acumulado semanal, para três temperaturas de base, 16, 17 e 18°C, e na Figura 5.4(b) tem-se o consumo total diário e as duas componentes (dependente e independente das variáveis externas) obtidas através do índice de aquecimento com a temperatura de base igual a 17°C.



(a) Índice de aquecimento para o ano 2012 com três temperaturas de base: 16, 17 e 18°C e as respetivas retas da regressão linear.



(b) Decomposição do consumo energético, com registos de Janeiro de 2011 até Agosto de 2013, em duas partes, através do índice de aquecimento com temperatura de base = 17°C.

Figura 5.4: Índice de aquecimento no consumo total diário de uma instalação.

Obteve-se o valor da componente independente das variáveis externas para cada ano,

através da divisão por sete (o número de dias numa semana), da constante que resulta da intersecção da reta com o eixo vertical. O aumento acentuado da energia independente das variáveis externas do ano 2012 para o ano 2013 pode dever-se ao facto do ano 2013 estar incompleto.

Não se pode especificar a precisão dos resultados obtidos, pois não existem valores de referência. Utilizou-se o índice de aquecimento com o intuito de efetuar uma primeira abordagem à desagregação do consumo energético. Também se testou o índice de arrefecimento, mas os resultados foram semelhantes e, por isso, não foram inseridos neste relatório. Como o objetivo deste trabalho foi determinar as três componentes do consumo energético, prosseguiu-se com a definição de métodos que o possam cumprir.

5.2 Determinação da energia de base

Baseload pode ser visto como a quantidade mínima de energia indispensável para uma instalação estar ativa. O consumo mínimo pode ser detetado fora do horário laboral da instalação, nomeadamente, nos feriados, fins de semana ou durante a noite. No entanto, estes métodos podem não ser caraterísticos para todas as instalações, ou seja, pode haver instalações que só trabalham nos dias de Natal ou Ano Novo, ou apenas durante a noite, etc. Assim, o problema teve que ser visto de uma forma imparcial, de modo a abranger todos os tipos de utilizadores.

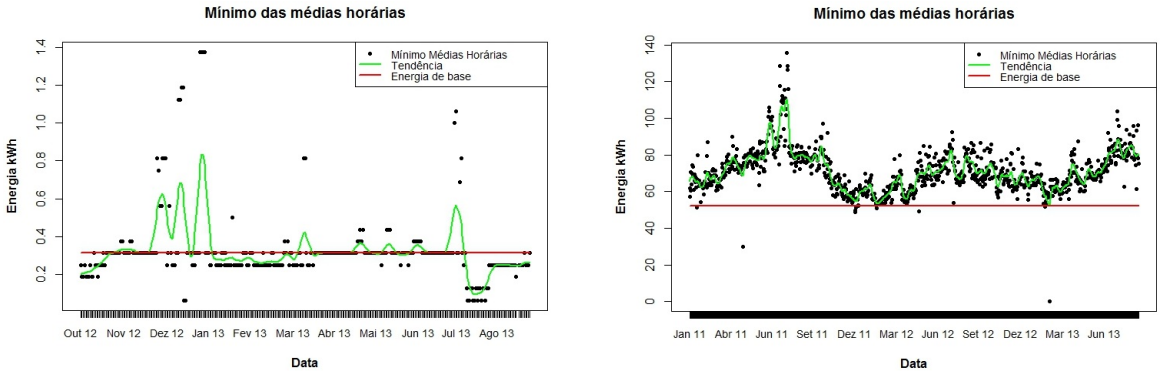
A energia de base foi calculada de várias maneiras e a validação dos resultados obtidos foi feita com base num conjunto de instalações que constituem o conjunto de treino.

Descrição do conjunto de treino

Formado por 25 instalações pertencentes ao primeiro lote. A escolha das mesmas para o treino não foi aleatória. De facto, foram consideradas aquelas cuja energia de base podia ser determinada visualmente a partir da análise do gráfico do consumo diário agregado pelo mínimo, porque não existiam valores de referência. Esta técnica visual para o cálculo da energia de base, abordada por Ferreira [15], não é razoável quando se dispõe de um número elevado de consumos energéticos a desagregar e, por isso, foi necessário construir

um método que o fizesse automaticamente.

Na Figura 5.5 encontram-se, representados pelos pontos pretos, os mínimos das médias horárias (a explicar mais à frente) de duas instalações que pertenciam ao conjunto de treino. A verde tem-se a tendência destes mínimos, determinada através da técnica de Análise Singular Espectral (SSA) descrita na Secção 2.3, e por fim, a vermelho, a energia de base de cada uma das instalações. Para determinar a tendência, tomou-se um tamanho da janela L = 7.



(a) Instalação com registos de Outubro de 2012 até Agosto de 2013

(b) Instalação com registos de Janeiro de 2011 até Agosto de 2013

Figura 5.5: Consumo mínimo diário, tendência e a energia de base de duas instalações pertencentes ao conjunto de treino.

Os mínimos das médias horárias não são a energia de base, pois estes podem incluir algumas observações anormais, como por exemplo, as falhas. Assim, em primeiro lugar, tentou-se amaciar a série dos mínimos utilizando a tendência, de modo a dar um menor peso às anomalias. Se a energia de base for definida como a tendência, ter-se-ão dados diários desta componente do consumo que deverão ser aproximadamente iguais. Uma variação mais acentuada permitirá detetar mudanças que, eventualmente, tenham ocorrido dentro da instalação, como por exemplo, mudanças de equipamentos, que consomem mais ou menos.

O amaciamento dos mínimos através da tendência não foi sempre bem sucedida, como é o caso da instalação da Figura 5.5(a), em que a tendência ainda contém bastante pormenor. Tentaram-se comprimentos maiores da janela, L = 14, 21, que suavizaram bastante a curva da tendência, mas o problema permanecia. Por conseguinte, tal como foi referido anteriormente, a energia de base das instalações de treino foi determinada visualmente.

A energia de base da instalação da Figura 5.5(a) e das outras que têm distribuição do consumo parecida, foi calculada como o valor mais frequente dos mínimos das médias horárias.

O mesmo não se pôde aplicar à segunda instalação devido à elevada variação nos consumos e, portanto, a existência de poucos valores repetidos.

No gráfico da Figura 5.5(b), existe uma variação no consumo ao longo do ano, o mesmo padrão que o consumo total diário. Isto significa que os mínimos das médias horárias também variam em função do clima exterior e portanto, não pode ser esta a curva da energia de base, pois, pela definição, é a quantidade mínima de energia indispensável, independente das condições meteorológicas. Um exemplo de instalação em que a energia de base varia de acordo com as condições climatéricas é um supermercado, em que a energia gasta pelas arcas refrigeradoras durante a noite no verão é superior à do inverno.

Assim, nas instalações de treino em que a sazonalidade está presente nos mínimos das médias horárias, a energia de base foi tomada como o mínimo da tendência calculada anteriormente. Esta ideia não é válida para todas as instalações, pois podia dar-se o caso de a tendência, em algum instante, ser igual a zero e, na realidade, não ser esta a energia de base.

O valor encontrado foi chamado de "energia de base real" e serviu para comparar o desempenho de cada um dos métodos testados e, por fim, escolher o de menor erro. Utilizou-se o erro percentual que é dado por

$$\text{Erro Percentual} = 100\% \times \left| \frac{\text{Valor Real - Valor Estimado}}{\text{Valor Real}} \right|,$$

onde o Valor Real é a energia de base real determinada visualmente e o Valor Estimado é a energia de base calculada através de cada uma das técnicas.

Todos os métodos devolvem o valor da energia de base por um período de 15 minutos. O resultado final é a energia de base ajustada ao dia, ou seja, a energia de base por 15 minutos multiplicou-se por 96 (de relembrar que na Secção 2.1, o consumo total diário foi ajustado ao número de instantes previsto, 96).

Segue-se a descrição, as vantagens e os inconvenientes de cada um dos métodos testados.

5.2.1 Mínimo

No Capítulo 2, foi mencionado que serão analisados os consumos totais diários (= soma das três componentes) das instalações. Contudo, esta medida de agregação não é a melhor para a determinação da energia de base. Assim, para o cálculo desta componente do consumo, na primeira etapa, os diagramas de carga foram agregados através do ***mínimo diário***. O procedimento está esquematizado no diagrama da Figura 5.6.

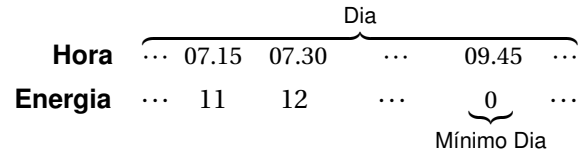


Figura 5.6: Agregação do consumo por dia através da medida mínimo do dia.

Havia instalações em que o diagrama de carga variava da seguinte forma:

Hora	⋯	04.00	04.15	04.30	04.45	05.00	05.15	⋯
Energia	⋯	0	0.25	0	0.25	0	0.25	⋯

Segundo o critério de determinação da energia de base definido em cima, energia de base do dia = mínimo dos 15 minutos do dia, a energia de base do dia desta instalação seria 0, ou seja, a instalação não tem consumo mínimo necessário para funcionar. Contudo, esta afirmação pode não ser verdadeira e na realidade, a instalação pode ter consumos bastante baixos mas diferentes do zero. Isto é, se a instalação consumir menos de 1 kW (potência) durante 15 minutos, o contador irá registar zero (a potência registada em kW toma valores inteiros) mas nos 15 minutos seguintes, registará o valor acumulado. Deste modo, para calcular a energia de base, devem ser considerados vários períodos sucessivos de 15 minutos cada.

Como alternativa, calculou-se o ***mínimo das médias horárias***, Figura 5.7:

O resultado da aplicação desta técnica pode ser visto na Figura 5.5. Tal como foi referido anteriormente, para ter um valor viável da energia de base, é necessário retirar a variação do mesmo ao longo do ano.

Uma primeira tentativa foi calcular a energia de base pelo ***mínimo das médias dos quatro períodos do dia*** (definidos no Capítulo 3) em vez dos mínimos das médias horários. A esquematização do método encontra-se no diagrama da Figura 5.8 e o resultado para

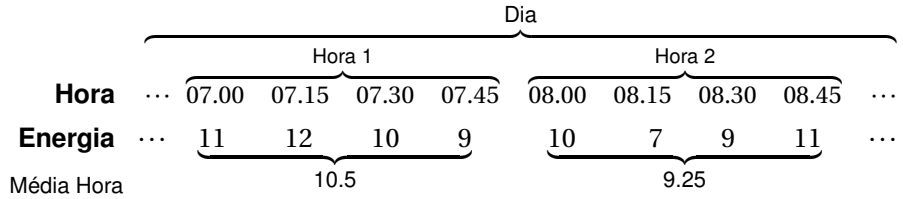


Figura 5.7: Agregação do consumo por dia através do mínimo das médias horárias.

uma das instalações anteriores pode ser visto na Figura 5.9, onde são comparados os consumos obtidos através das duas técnicas: mínimo das médias horárias, (a), e mínimo das médias dos períodos, (b).

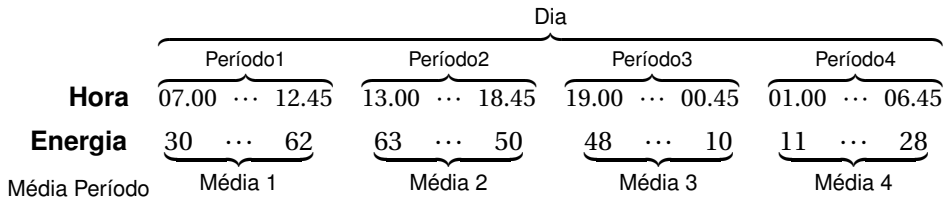
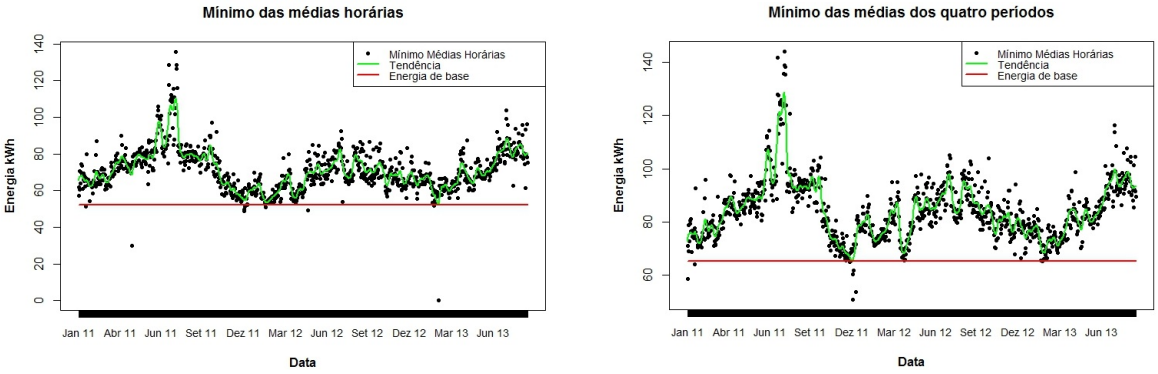


Figura 5.8: Agregação do consumo por dia a partir do mínimo das médias dos quatro períodos do dia.

A energia de base calculada utilizando os mínimos das médias dos períodos (energia de base=65) é 25% maior do que a energia de base determinada aplicando os mínimos das médias horárias (energia de base=52), a estabelecida como valor de referência, mantendo-se a sazonalidade. De seguida, tentou-se diminuir o efeito desta variação amaciando a série através da agregação dos dados diários (mínimos das médias horárias) por semana, e depois por mês, através da média. As tabelas com os dados agregados foram construídas em `SQL Server`. A ideia foi novamente testada no conjunto de treino. O resultado para as duas instalações já vistas em cima pode ser visto na Figura 5.10, onde estão representados os dados semanais/mensais, as respetivas tendências[1], mais uma vez determinadas através de SSA (Secção 2.3), e a energia de base calculada anteriormente.

Analisando os gráficos da Figura 5.10, pode concluir-se o seguinte: i) no caso da primeira instalação, a tendência dos dados agregados por semana e por mês continua a ter excesso

[1] Foi feita uma função que recebendo os consumos de uma instalação devolve a tendência dos mesmos recorrendo a SSA.



(a) Mínimo das médias horárias

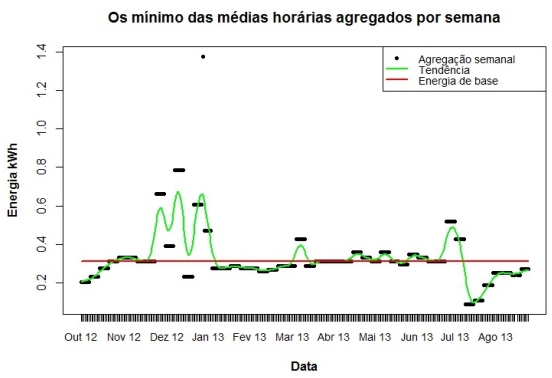
(b) Mínimo dos quatro períodos do dia

Figura 5.9: Consumos mínimos e as respetivas energias de base.

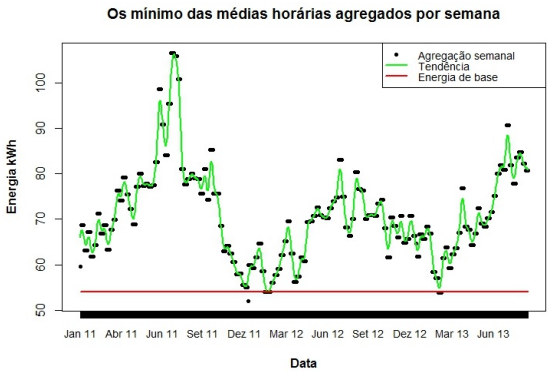
de pormenor e por conseguinte não pode ser a energia de base; ii) quanto à segunda instalação, a sazonalidade permaneceu tanto nos dados semanais como mensais e portanto também não é a energia de base.

Em resumo, para calcular a energia de base das instalações em análise, terão que ser usados os dados diários, nomeadamente, os mínimos das médias horárias. A tendência suaviza bastante estes dados, no entanto, contém sazonalidade indesejada para a energia de base. A agregação dos dados por semana e por mês não solucionou esta questão.

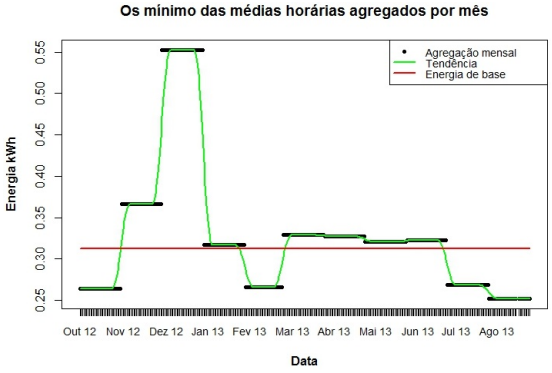
Uma outra tentativa de resolução foi desagregar a série temporal, conceito introduzido na Secção 2.3. A ideia é decompor a série nas suas componentes (tendência, sazonalidade e ruído) e, de seguida, retirar dos dados originais a parte correspondente à variação periódica. Para tal, foram testados os métodos disponíveis em `R`, tais como: i) decomposição clássica, referida na Secção 2.3 e implementada através das funções `decompose` e `stl`; ii) decomposição com o SSA, descrita na mesma secção; iii) decomposição de séries temporais com sazonalidade múltipla, ver De Livera et al. [10], através das funções `msts` e `tbats`. Alguns resultados podem ser vistos no Anexo C. De referir que a sazonalidade não foi bem estimada na sua totalidade pois, após a sua retirada, continuava presente parte desta componente nos dados. Assim, pensou-se em considerar a informação apenas dos dias em que não há gastos de energia relacionados com o ambiente exterior, os chamados *dias de conforto*.



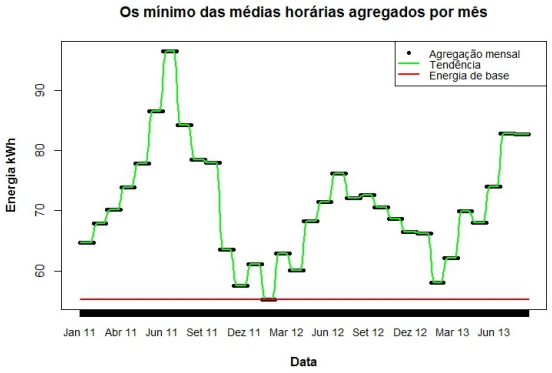
(a) Instalação 1: mínimos agregados por semana



(b) Instalação 2: mínimos agregados por semana



(c) Instalação 1: mínimos agregados por mês



(d) Instalação 2: mínimos agregados por mês

Figura 5.10: Consumos mínimos das médias horárias agregados, através da média, por semana, (a) e (b), e por mês, (c) e (d) das duas instalações vistas em cima.

5.2.2 Dias de conforto

No local de trabalho, a produtividade de uma pessoa ativa depende das condições ambientais que a rodeiam: demasiado frio ou calor contribui de uma forma negativa para a eficiência no trabalho. Assim, várias organizações, tais como American Society of Heating, Refrigeration, and Air-Conditioning Engineers (ASHRAE), European Committee for Standardization (CEN) e International Organization for Standardization (ISO), definiram limites para o conforto humano.

Parsons [42] define o conforto humano como “a condição da mente que expressa satisfação com o ambiente térmico”. O conforto térmico é altamente subjetivo. Kamholz e Storer [29] explicam que, sendo um fenómeno psicológico, é influenciado pelas preferências pessoais, de aclimatização, cultura e outros fatores sociais.

Vários estudos, [15], [5] referem que a temperatura é a variável mais importante para a

definição de conforto num espaço. Durante o metabolismo, o corpo humano gera calor que deve ser dissipado para o exterior. Quando a temperatura do ambiente está elevada, o aparelho termorregulador do corpo recorre à transpiração para resfriar o corpo, com o efeito de refrigeração diretamente relacionado com a taxa de evaporação do suor. Esta taxa depende da humidade do ar e da quantidade de vapor de água que este ainda consegue conter, o que leva à consideração da segunda variável mais importante na definição de conforto, a humidade relativa. Se o ar está saturado, o suor não se irá evaporar o que causará uma situação de desconforto. O ar muito seco também é desconfortável, pois nestas condições a superfície da pele e as mucosas ficam mais secas, levando a queixas sobre a secura do nariz, garganta, olhos e pele [3]. Assim, o conforto pode ser definido como a ausência de qualquer forma de stress térmico.

Outras variáveis tais como, o metabolismo de cada indivíduo, as roupas que veste, as cores envolventes, a superfície da sala [5], entre outras, também influenciam o conforto humano. No entanto, como não se dispunha de qualquer informação sobre o interior da instalação, foram consideradas apenas as variáveis climáticas.

OSHA [41], referindo a temperatura e a humidade como questões do conforto humano, recomenda temperaturas de controle entre 20 e 24.5°C e humidade a 20-60% (valores para o interior do edifício), intervalos mencionados em vários outros estudos, tais como [5]. Contudo, as condições ambientais dentro do edifício estão fortemente relacionadas com as do exterior. A medida em que é afetado o interior depende de fatores tais como, o isolamento do edifício, o material de que é feito, o ramo de atividade da instalação, etc. Como o clima varia constantemente, os valores da temperatura e humidade interior poderão ser diferentes dos de referência. Valores fora dos intervalos de controle provocam mal estar, incómodo e falta de conforto o que diminui a produtividade. Para evitar tal inconveniente, é necessário manter as condições ambientais dentro dos limites estabelecidos, que é feito recorrendo ao ar condicionado, aquecimento ou ventoinhas. Todos estes dispositivos têm um consumo energético extra, em comparação com os dias em que não estão ligados - os dias de conforto.

Como não havia a possibilidade de aceder ao interior da instalação e de medir a temperatura e a humidade interior, nem se dispondo das horas de funcionamento dos aparelhos controladores das condições ambientais, procurou-se uma relação entre o clima do interior do edifício com o do exterior. A única referência encontrada foi um estudo interno feito na

EDP Distribuição, com base num edifício, que encontrou uma temperatura média de 15°C como valor de conforto, isto é, quando a temperatura ambiente é diferente deste valor, dentro do edifício é ligado o ar condicionado/aquecimento. Contudo, na Secção 4.2 viu-se que as variáveis que mais se relacionam com o consumo energético são Ponto Orvalho Máximo e Humidade Mínima.

Goldstein [18] indica o ponto de orvalho como um excelente indicador do conforto, pois reflete a quantidade de vapor de água contida no ar que, por sua vez, influencia o mecanismo natural de arrefecimento, a transpiração, já referido anteriormente. Quando a humidade relativa é igual a 100%, a temperatura é igual ao ponto de orvalho. A relação entre a temperatura, ponto de orvalho e humidade relativa é expressa através da aproximação de August-Roche-Magnus [55], apresentada a seguir, que permite encontrar o valor de uma das variáveis a partir das outras duas.

$$T = \frac{b\left[\frac{aT_d}{b+T_d} - \ln\left(\frac{RH}{100}\right)\right]}{a + \ln\left(\frac{RH}{100}\right) - \frac{aT_d}{b+T_d}}; \qquad T_d = \frac{b\left[\ln\left(\frac{RH}{100}\right) + \frac{aT}{b+T}\right]}{a - \ln\left(\frac{RH}{100}\right) - \frac{aT}{b+T}}; \qquad RH = 100\frac{\exp\left(\frac{aT_d}{b+T_d}\right)}{\exp\left(\frac{aT}{b+T}\right)}$$

onde $a = 17.271, b = 237.7$, T é temperatura em °C, T_D é ponto de orvalho em °C e RH é humidade relativa em %. Esta equação é válida para valores das variáveis nos seguintes limites: 0°C < T < 60°C, 1% RH < 100%, 0°C < T_d < 50°C. Esta relação está representada no gráfico da Figura 5.11.

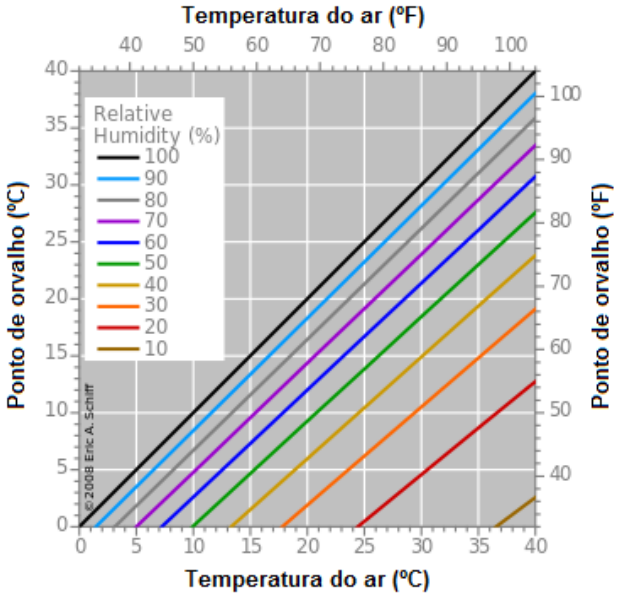


Figura 5.11: A aproximação de August-Roche-Magnus.

Assim, em vez de considerar a temperatura média para definir o conforto, foram utilizadas as variáveis Ponto Orvalho Máximo e Humidade Mínima, as resultantes da seleção feita na Secção 4.2. Deve observar-se que estas variáveis quantificam o momento mais quente do dia, que, em muitos casos, irá pertencer ao horário laboral da instalação e, portanto, são viáveis para definir o conforto humano. É de notar que a temperatura média tem em consideração as variações da mesma, registadas durante a noite, e, não se refere ao mesmo momento que as outras duas variáveis, de modo que, o valor de 15°C da temperatura média corresponde a 20°C da máxima. A transformação foi obtida considerando o valor mais frequente da temperatura máxima quando a média era igual a 15°C.

Tendo em conta o gráfico da Figura 5.11 e a transformação entre as variáveis disponibilizada no site oficial de McNoldy [37], que tem por base a aproximação de August-Roche-Magnus, foram definidos intervalos para o Ponto Orvalho Máximo - de 7 a 10°C e Humidade Mínima - de 45 a 65%, valores que colocam a Temperatura entre 17 e 23°C. Foram considerados intervalos e não apenas um valor para cada uma das variáveis para poder abranger o maior número de dias possível e aumentar a confiança nos resultados. Em vez de intervalos para os dias de conforto testou-se a distância entre os valores observados do Ponto Orvalho Máximo, Humidade Mínima e os correspondentes ao conforto. Contudo, os erros para a energia de base obtidos através deste método eram superiores aos erros utilizando intervalos.

A distribuição dos dias de conforto por ano e por região (Secção 4.1) pode ser vista na Tabela 5.1.

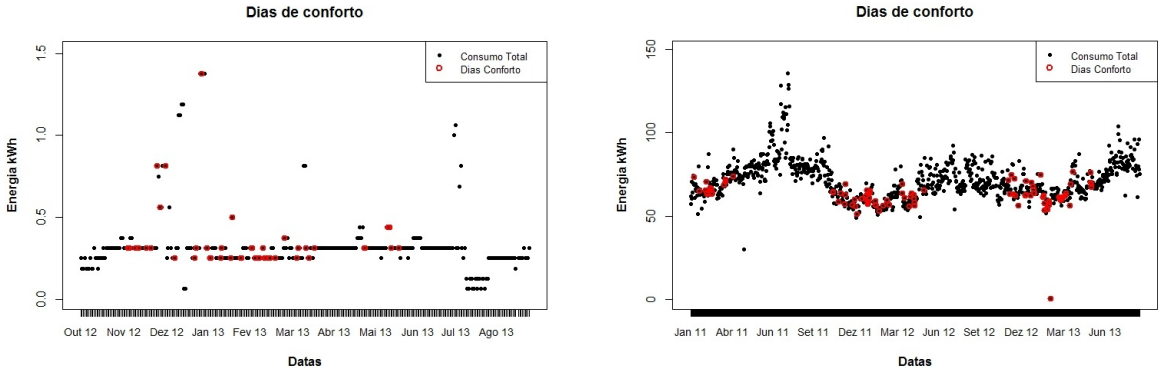
Região	**Ano 2010**	**Ano 2011**	**Ano 2012**	**Ano 2013**
Beja	30	24	40	23
Bragança	22	23	15	14
Castelo Branco	25	38	25	26
Faro	15	14	24	5
Lisboa	25	29	48	30
Porto	21	25	39	31

Tabela 5.1: Número de dias de conforto por ano e por região.

Com as condições de conforto definidas, prosseguiu-se com o cálculo da energia de base.

Para determinar o valor da energia de base por ano[2], foram considerados os consumos nos dias do respetivo ano que satisfizeram as condições ambientais anteriores. Os dados nestes dias foram agregados pelo mínimo das médias horárias, descrito na Secção 5.2.1, pois foi visto ser esta a medida mais adequada para encontrar a energia de base.

Na Figura 5.12 estão representadas os consumos mínimos das médias horárias das instalações já vistas anteriormente e os dias de conforto. É de notar que, nos dias de conforto, foram registados os menores consumos energéticos.



(a) Instalação 1 (b) Instalação 2

Figura 5.12: Consumo total diário de duas instalações e nos dias de conforto.

A energia de base não é a tendência dos mínimos das médias horárias nos dias de conforto, pois podia ocorrer algum consumo anormal (por exemplo, uma lâmpada esquecida durante a noite) que poderia afetar a tendência e já se sabe que a energia de base representa o consumo típico. Assim, os vários valores obtidos tinham que ser traduzidos num só por ano.

No caso da primeira instalação, poderia ser o valor mais frequente de todos os dias de conforto de cada ano de registo. Contudo, esta técnica não produz bons resultados para a segunda instalação devido à alta variabilidade entre os consumos. Assim, pensou-se em juntar valores de consumos próximos. Para tal, foi aplicado um **método de agrupamento (do inglês, *clustering*) hierárquico aglomerativo** implementado em `R` através da instrução `hclust`. Este método é baseado na distância entre observações e pode ser feito por meio de um dos métodos aglomerativos: ward, mínimo, máximo, média, mediana, centroid ou mcquitty. A ideia desta técnica de agrupamento é, partindo de tantos grupos quantas observações existem, em cada passo, juntar as observações mais semelhantes até obter

[2]Quando se dispunha de dados suficientes, o ano foi considerado como o período entre 1 de Janeiro e 31 de Dezembro do mesmo ano; caso contrário, o ano começava/acabava no primeiro/último dia de registo de consumo disponível.

um só grupo que contém todas as observações. Em cada etapa é recalculada a distância entre os grupos já existentes. Este método pode ser representado em forma de árvore, que se denomina por dendrograma. Uma descrição detalhada desta técnica de agrupamento pode ser vista em Everitt e Hothorn [13].

Uma vez finalizado o agrupamento das observações, os grupos podem ser acedidos através da instrução `cutree` de `R`, especificando o número de grupos desejado ou estabelecendo um limite para a distância máxima entre eles. No trabalho desenvolvido recorreu-se à segunda opção, pois não se sabia o número de grupos a formar.

Como a distância entre observações depende da escala das mesmas, construiu-se uma função em `R` que, recebendo os valores a agrupar e a distância, um inteiro maior que 0, executa os seguintes passos:

1). Arredonda as observações em função da sua média, portanto pode haver valores com 2 (média < 1), 1 (1 < média < 10) ou 0 (média > 10) casas decimais;

2). Agrupa as observações através do método hierárquico aglomerativo descrito em cima;

3). Tendo em conta o número de casas decimais da média (Ncd), corta a árvore a uma altura igual a $10^{-Ncd}\times$distância;

4). Retorna a média[3] das observações pertencentes ao maior grupo formado até à altura estabelecida no passo anterior.

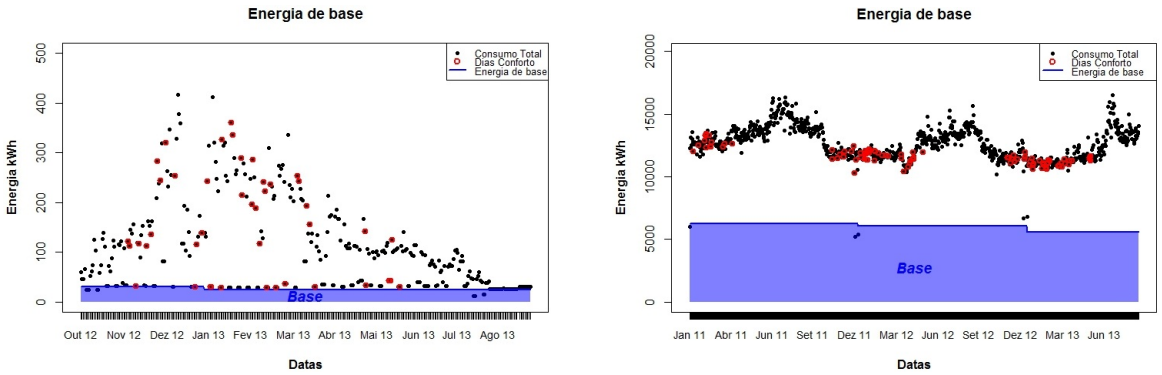
Foram testados alguns valores para a distância: 3, 4, 5 e 6 e foram experimentados todos os métodos aglomerativos enumerados em cima. Os resultados obtidos, para o conjunto de treino (formado por 25 instalações), através de todas as combinações possíveis, foram comparados com a energia de base real. O menor erro relativo foi o da *média*, com uma distância igual a 5. Um exemplo de um dendrograma dos consumos mínimos das médias horárias pode ser visto no Anexo C.

O resultado devolvido no quarto passo é o valor da energia de base num intervalo de 15 minutos. O procedimento é imediato e é repetido para cada ano de observações. Se houver menos de 20 dias de conforto por ano[4], em vez de serem considerados apenas os dias de conforto, são abrangidos todos os dias do respetivo ano. De relembrar que a energia de base final por dia se obtinha através da multiplicação da energia de base, num intervalo de 15 minutos, por 96 (número de instantes previsto). Na Figura 5.13 encontram-

[3] Foram testadas outras medidas como a mediana e o 1º quartil mas o menor erro foi o obtido para a média.

[4] Este valor foi escolhido de modo a poder aproveitar a informação dos dias de conforto sem cometer grandes erros na estimação da energia de base. O parâmetro pode ser ajustado conforme o objetivo do problema.

se representados os consumos totais diários das duas instalações já vistas e as respetivas energias de base obtidos através do procedimento descrito anteriormente.



(a) Instalação 1 (b) Instalação 2

Figura 5.13: Consumo total diário de duas instalações, os dias de conforto e a energia de base diária calculada por ano.

Observando os gráficos da Figura 5.13, podem notar-se alguns dias em que os consumos estão abaixo da linha da energia de base encontrada, ou seja, consumos inferiores ao valor definido como indispensável. Uma possível explicação para este acontecimento é o facto de serem dias de férias dentro da instalação e, por conseguinte, reduzirem o consumo mais do que o habitual. Como a energia de base é o consumo típico, estes dias são vistos como dias especiais. Na Figura 5.13(b) pode ver-se uma diminuição da energia de base ao longo dos anos. Existem numerosos fatores que podem explicar esta variação, como por exemplo: a energia de base refletir as mudanças sócio-económicas externas, ocorrer uma gestão mais eficiente do consumo energético da instalação, entre outras.
Uma vez encontrada a energia de base diária para todos os dias de registos, prosseguiu-se com o cálculo da energia útil.

5.3 Determinação da energia útil

Após a determinação de uma das componentes do consumo energético, o próximo passo foi subtrair estes valores do consumo total diário, para separar a componente definida das ainda desconhecidas. De relembrar que, para executar este trabalho, se supôs que a soma das três partes resultava no consumo total diário.
Como foi visto, pode haver dias em que o consumo total diário é inferior à energia de base

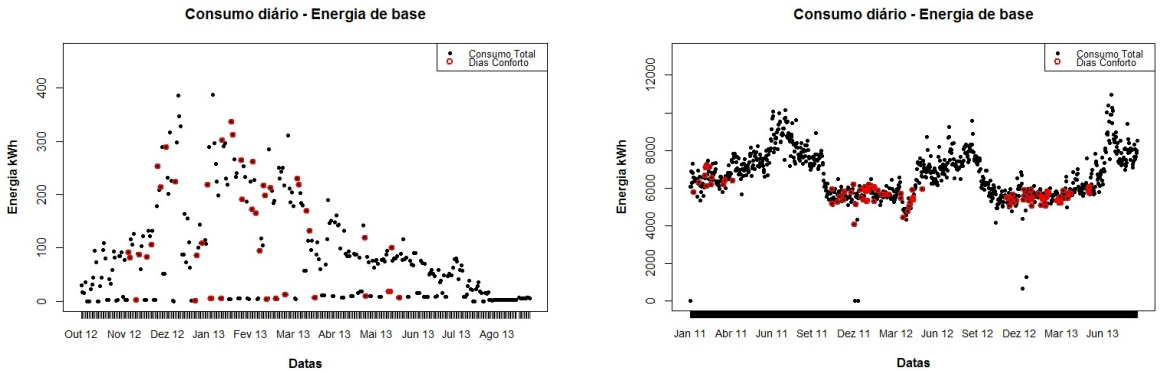
encontrada e, como não pode haver consumos negativos, os dados para calcular o trabalho foram obtidos considerando o máximo entre 0 e a diferença entre o consumo total diário e a energia de base diária. Efetuado este ajustamento, avançou-se para o cálculo da energia útil.

A quantidade de energia gasta com o trabalho depende do tipo da instalação e do seu ramo de atividade. Abels et al. [1] tinham esta e outras informações relacionadas com a produção e aproveitaram-nas para calcular a segunda componente. Como, neste trabalho, só se dispunha do diagrama de carga da instalação, tiveram que ser estudados métodos alternativos aos descritos em [1], mas que conduzissem a resultados aceitáveis para EDP Distribuição.

5.3.1 Dias de conforto

Como já foi referido na secção anterior, nos dias neutros presume-se não haver quantidades de energia relacionadas com as variáveis externas, de modo que, tal como para o cálculo da energia de base, usaram-se os dias de conforto para calcular a energia útil.

As condições de conforto foram as mesmas que anteriormente: Ponto Orvalho Máximo entre 7 e 10°C e Humidade Mínima entre 45 e 65%. Na Figura 5.14 podem ver-se os consumos diários das duas instalações já familiares, após a energia de base ter sido retirada, e os dias de conforto marcados a vermelho.



(a) Instalação 1 (b) Instalação 2

Figura 5.14: Consumo diário de duas instalações, após a componente da energia de base ter sido retirada, e os dias de conforto marcados a vermelho.

Observando os gráficos da Figura 5.14, pode notar-se uma diminuição na escala da energia

(fenómeno já esperado), mais acentuada no caso da segunda instalação. Outro acontecimento lógico é a existência de "espaços" entre os dias de conforto, pois nem todos os dias são amenos.

Para calcular a energia útil estimou-se a tendência dos dias de conforto, calculada através da já conhecida técnica da Análise Singular Espectral, Secção 2.3. Antes de fazer esta aproximação, os espaços existentes entre os dias de conforto tinham que ser preenchidos. Isto foi feito por intermédio da interpolação linear [38] implementada na função `na.approx` de `R`.

Uma outra observação a fazer da análise dos gráficos é o facto de os extremos da série dos dados não corresponderem a dias de conforto. Os intervalos entre o início/fim da série e o primeiro/último dia de conforto não podiam ser preenchidos através da interpolação pois esta técnica exige o conhecimento dos limites do intervalo a completar. Como solução, recorreu-se à metodologia de Hyndman e Khandakar [25] de previsão de séries temporais, programada em `R` na função `forecast`. O comando pode receber como parâmetro de entrada a série ou o modelo da série temporal. A introdução do modelo na função permite prever valores considerando apenas a tendência da série e omitindo a sazonalidade, componente relacionada com as variáveis externas. O modelo foi construído por intermédio da função `tslm(tendencia~trend)`, onde `tendencia` é a tendência dos dias de conforto e `trend` é a caraterística da série a considerar.

Geralmente, a previsão é feita para estimar o futuro. Contudo, neste trabalho, para determinar a energia útil, é necessário estimar o início da série constituída pelos dias de conforto. Hyndman [24] denomina este processo por "backcast", isto é, previsão em tempo inverso e, não estando implementado em `R`, sugere inverter a série dos dados, fazer a previsão para os períodos necessários aplicando a instrução já conhecida, e depois reverter novamente a série já completa.

Os resultados da previsão e da interpolação são consumos diários aproximados que, em alguns casos, podem ser superiores aos consumos observados, isto é, estima-se que o trabalho de uma instalação num certo dia é superior ao consumo total diário real. Um exemplo disto é o dia de Natal na segunda instalação apresentada anteriormente, em que o consumo total diário neste dia é inferior a 10000 kWh, mas a interpolação estimou-o ser superior a 15000 kWh. Ou seja, a energia útil diária final é o mínimo entre a energia útil estimada e o consumo total diário.

O procedimento para a determinação da componente relacionada com o trabalho pode ser resumido nos seguintes passos:

1). Dados os dias de conforto, fazer interpolação linear para completar os dias em falta.

2). Recorrendo à técnica SSA, estimar a tendência da série contínua construída no passo anterior.

3). Construir o modelo linear da série obtida no passo 2 tendo em conta apenas a tendência desta série.

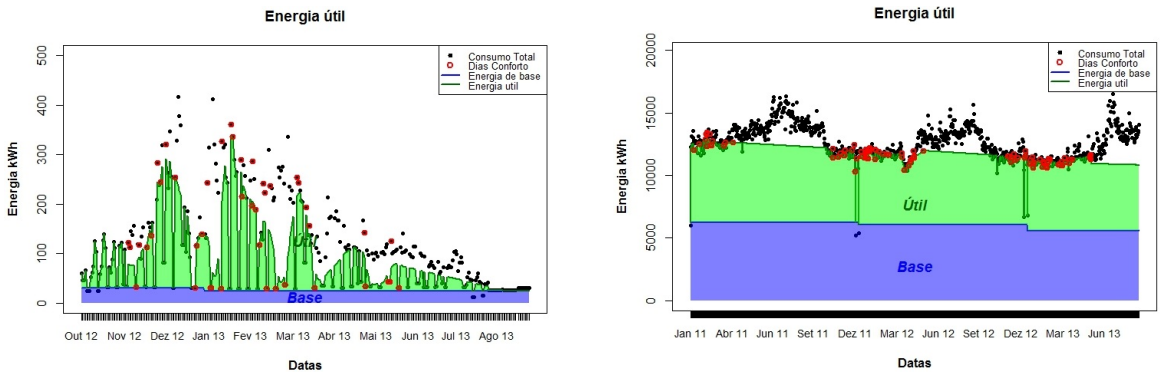
4). Usar o modelo do passo 3 para prever os dias posteriores ao último dia de conforto. Resta estimar os dias anteriores ao primeiro dia de conforto.

5). Inverter no tempo a série do passo 3 e prever novamente os dias em falta.

6). Inverter no tempo a série do passo 5.

7). Ajustar o consumo estimado ao consumo real.

O resultado deste algoritmo é um conjunto de valores com o mesmo comprimento que a série original dos consumos diários e representa a energia útil gasta num determinado dia. Na Figura 5.15 tem-se o consumo total diário das duas instalações anteriores decomposto nas duas componentes já definidas que são a energia de base e útil.



(a) Instalação 1 (b) Instalação 2

Figura 5.15: Consumo total diário de duas instalações, os dias de conforto, a energia de base diária e a energia útil calculada por dia.

Observando a faixa da energia útil da primeira instalação, vê-se uma elevada variação que, contudo, parece ser periódica. Apesar de não se conhecer o tipo de instalação, poderia supor-se que a variação tem um ciclo semanal que corresponde aos dias úteis e aos fins de semana, dias em que o trabalho está reduzido ao mínimo e a única componente presente é a energia de base.

A instalação da Figura 5.15(b), por sua vez, tem a curva do trabalho bastante mais su-

ave. Exceção são os poços ocorridos nos dias de Natal e de Ano Novo, já identificados na Secção 2.1, em que o consumo total diário era bastante inferior ao previsto para a componente da energia útil. Pode notar-se um estreitamento da faixa do trabalho do ano 2011 a 2013 que pode ser explicado pela diminuição da produção devido à crise económica mundial, mudanças dentro da instalação, etc. Contudo, o declive acentuado deve-se também à diminuição da energia de base ao longo dos anos que cria uma ilusão visual de um decréscimo proeminente.

Visualmente, o resultado parece ser bastante bom. Como não havia maneira de validar ou verificar os resultados, o método foi aplicado às 25 instalações pertencentes ao conjunto de treino e os resultados desta fase de decomposição foram avaliados através dos gráficos, à semelhança dos apresentados na Figura 5.15.

Um problema surge quando a instalação não tem dias de conforto ou esse número é bastante reduzido. Para não as excluir da análise, pensou-se num método alternativo que é aplicado quando se tem menos de 20 dias de conforto no período total de registos (diferente da condição da energia de base que exigia ter pelo menos 20 dias de conforto por ano). Do primeiro lote que serviu para elaborar o trabalho, apenas 5 de 97 (equivale a 5%) tinham menos de 20 dias de conforto: 4 delas, todas da região de Faro, com registos de 1 de Maio de 2012 até 31 de Agosto de 2013 tinham 8 dias de conforto e 1, de Lisboa, com registos a partir de 28 de Maio de 2013 até Agosto de 2013, não tinha nenhum dia de conforto.

A técnica incorporada na função de desagregação que soluciona estes casos, com menos detalhe, é a regressão linear múltipla.

5.3.2 Regressão linear múltipla

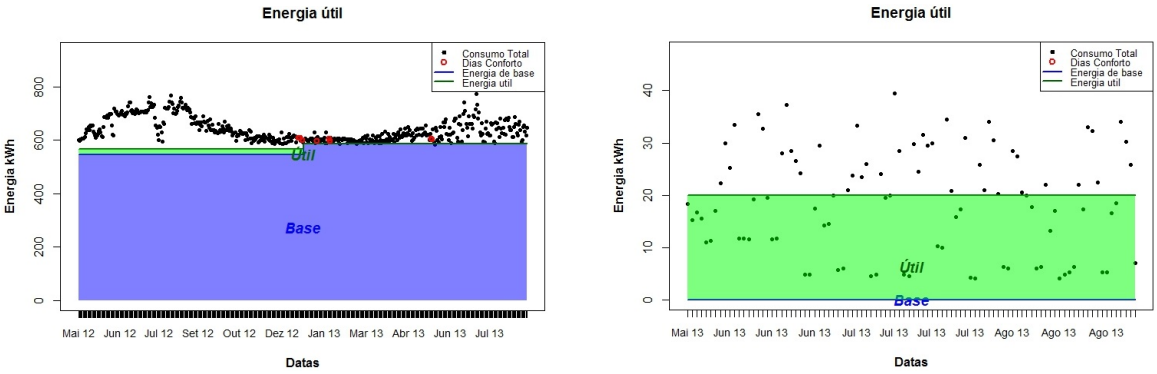
Foi uma das técnicas aplicadas por Price [44] para prever o consumo, tendo a temperatura e a humidade exteriores como variáveis independentes. Contudo, neste trabalho, na regressão linear foram incluídas as variáveis Ponto Orvalho Máximo e Humidade Mínima como preditoras, as variáveis identificadas na Secção 4.2 como as mais relacionadas com o consumo e mais tarde, Secção 5.2.2 as variáveis que definem o conforto humano. Como fundamento para o uso desta técnica relativamente simples (regressão), Price [44] apresenta os seguintes argumentos: i) quando bem construído, o modelo proporciona um bom ajustamento na maior parte dos edifícios, ii) os seus resultados são fáceis de

interpretar, iii) pode ser facilmente modificada, e, não menos importante, iv) o seu peso computacional é reduzido (caraterística que, mesmo com os computadores de hoje, nem todos os métodos de análise possuem).

Tal como no caso dos dias de conforto (secção anterior), a regressão linear múltipla é feita após a subtração da energia de base do consumo total diário. O modelo construído através da regressão linear múltipla contém sazonalidade (ver Secção 2.2). Para evitar a presença desta componente na energia útil foram definidos valores para as variáveis preditoras em que se supõe não haver influência do ambiente exterior, os já conhecidos dias de conforto. Esta abordagem foi possível através da instrução `predict(modelo,conforto)` de `R`, onde `modelo` é o modelo linear do consumo contra as duas variáveis climáticas e `conforto` é o conjunto com os valores de conforto para as duas variáveis explicativas. O resultado é o consumo total diário médio estimado quando as condições ambientais são iguais aos valores considerados. Na Secção 5.2.2, as condições ambientais neutras foram definidas como temperaturas para o Ponto Orvalho Máximo entre 7 e 10°C e Humidade Mínima entre 45 e 65%. Contudo, a instrução `predict` só aceita um valor para cada variável e não intervalos, de modo que, Ponto Orvalho Máximo = 9°C e Humidade Mínima = 50% (valores que podem ser ajustados) definiram os dias de conforto na regressão, condições que resultam numa Temperatura de 20°C (o estado de conforto).

Todo o procedimento é feito por cada ano, ou seja, se uma instalação tiver registos do ano 2012 até 2013, a energia útil terá dois valores, um por cada ano. Como o modelo pode estimar valores negativos para o consumo, a energia útil final é dada pelo máximo entre o valor estimado e zero. Na Figura 5.16 têm-se os consumos totais diários e a energia útil de duas instalações que têm menos de 20 dias de conforto no período total (primeira com 8 dias de conforto e a segunda com 0) .

Da análise do gráfico da Figura 5.16(a), pode dizer-se que a energia de base desta instalação constitui a maior parte do consumo total diário. Isto significa que a instalação funciona 24 horas por dia, o que justifica também a faixa bastante estreita (no segundo ano quase zero) da energia útil, e que é pouco sensível às condições ambientais externas. Na segunda instalação acontece o contrário: a energia de base é sempre zero e a faixa da energia útil é maior. Este comportamento pode ser explicado pelo tamanho reduzido dos dados (há registos de apenas 96 dias) e, por conseguinte, um maior erro na desagregação.



(a) Instalação 3

(b) Instalação 4

Figura 5.16: Consumo total diário de duas instalações, os dias de conforto, a energia de base diária e a útil diária calculadas por ano.

O problema do método baseado nos dias de conforto são os casos quando a instalação produz apenas nos dias que não são de conforto. Para ser mais claro, considera-se uma fábrica de gelados que se supõe funcionar apenas durante o verão. Nesta época ocorrem as temperaturas mais elevadas e, se for aplicado o método dos dias de conforto para a determinação da energia útil, serão utilizados os dias amenos que são os dias em que a fábrica tem um trabalho muito reduzido. A energia útil real, através da interpolação linear, poderá ser atribuída às variáveis externas obtendo-se resultados errados. Procurou-se uma alternativa aos dias de conforto, descrita na secção seguinte.

5.3.3 Associações

A ideia desta técnica é identificar padrões no consumo e nas variáveis externas, relacioná-los e por fim associá-los a consumos dependentes e independentes das variáveis externas. O procedimento encontra-se descrito sumariamente nos seguintes passos:

- **Padrões de consumo energético.** Considere-se a série do consumo energético observado E_t, de tamanho N e uma janela de comprimento L. Considere-se o conjunto de todas as subsucessões de comprimento L

$$\mathscr{S} = \{S_i = (E_i, E_{i+1}, \cdots, E_{i+L}) : i = 1, \cdots, K = N - L + 1\}.$$

Cada $S_i \in \mathbb{R}^L$ munido da distância Euclideana.

Exemplo. Pode considerar-se a série de uma instalação durante 1 ano, agregada por hora, com $L = 24$. Uma hipótese para simplificar é tomar as subsucessões separadas por um salto $v = 24$ horas, isto é, $S_{24i} : i = 1, \cdots, 364$.

- **Agrupamento** (do inglês, *Clustering*). Selecione-se um método apropriado (k-médias, DBSCAN, etc) de agrupamento de $\mathscr{S} \subset \mathbb{R}^L$ em grupos a que chamamos padrões de consumo energético (PCE's).

- **Padrões "climáticos".** Selecione-se uma variável externa "climática", designada genericamente por T, e criem-se padrões típicos (PTT´s) por um processo análogo.

- **Notações.**

  – PCE's: $\varepsilon_\alpha : \alpha \in \{1, 2, \cdots, n\}$

  – PTT's: $\tau_\beta : \beta \in \{1, 2, \cdots, m\}$

- **Discretização.** O processo anterior discretiza a série E_t substituindo-a por uma séria simbólica

$$\mathscr{E} = (\varepsilon_{\alpha(1)}, \varepsilon_{\alpha(2)}, \varepsilon_{\alpha(3)}, \cdots, \varepsilon_{\alpha(364)})$$

onde $\varepsilon_{\alpha(1)}$ é o grupo PTE a que pertence a primeira subsucessão S_1, $\varepsilon_{\alpha(2)}$ é o grupo a que pertence a segunda subsucessão S_2, etc. O mesmo acontece para a segunda série

$$\Gamma = (\tau_{\beta(1)}, \tau_{\beta(2)}, \tau_{\beta(3)}, \cdots, \tau_{\beta(364)}).$$

Por outras palavras, a série original E_t foi substituída por uma série padronizada (ou discretizada), em que a subsucessão $S_1 = (E_1, \cdots, E_{23})$ foi substituída pela subsucessão $\varepsilon_{\mathring{a}(1)}$, que é o representante do grupo a que pertence S_1; $S_2 = (E_{24}, \cdots, E_{47})$ foi substituída pela subsucessão $(\varepsilon_{\mathring{a}(2)}$, que é o representante do grupo a que pertence S_2, e assim sucessivamente. É de notar que os padrões mantiveram os rótulos temporais das horas.

O mesmo acontece para a série climática.

- **Mais notações.**

– Para cada PCE ε

$$\mathscr{D}(\varepsilon) = \{i : \varepsilon_{\alpha(i)} = \varepsilon\} \tag{5.1}$$

é o conjunto dos “dias" cujo PCE é ε. Analogamente, para cada PTT τ

$$\mathscr{D}(\tau) = \{i : \tau_{\beta(i)} = \tau\} \tag{5.2}$$

é o conjunto dos “dias" cujo PTT é τ.

– Para cada PCE ε

$$f(\varepsilon) = \frac{\#\mathscr{D}(\varepsilon)}{365} \tag{5.3}$$

é a respetiva frequência em $\mathscr{E}$. Analogamente, para cada PTT τ,

$$f(\tau) = \frac{\#\mathscr{D}(\tau)}{365} \tag{5.4}$$

é a respetiva frequência em Γ.

- **Associação e confiança.** Determinação e quantificação das associações do tipo

$$\tau \longrightarrow \varepsilon$$

isto é, identificação dos PTT’s que influenciam PCE’s.

– **1.** A cada associação do tipo $\tau \to \varepsilon$, isto é, a cada associação (clima→consumo), atribui-se uma medida de confiança, definida por

$$\mu(\tau \to \varepsilon) = \frac{\#((\mathscr{D}(\varepsilon) \cap \mathscr{D}(\tau))}{\#\mathscr{D}(\tau)} \tag{5.5}$$

onde $\mathscr{D}(\varepsilon) \cap \mathscr{D}(\tau)$ é o conjunto dos “dias"em que os padrões ε e τ ocorrem simultaneamente. Analogamente, a cada associação do tipo $\varepsilon \to \tau$, isto é, a cada associação (consumo→clima), atribui-se uma medida de confiança, definida por

$$\nu(\varepsilon \to \tau) = \frac{\#((\mathscr{D}(\varepsilon) \cap \mathscr{D}(\tau))}{\#\mathscr{D}(\varepsilon)} \tag{5.6}$$

– **2.** Para cada PCE ε,

$$\mathscr{P}(\varepsilon) = \{\tau_{\beta(i)} : i \in \mathscr{D}(\varepsilon)\} \tag{5.7}$$

é o conjunto dos PTT que ocorrem nos “dias" i em que o PCE é ε. Analogamente, para cada PTT τ,

$$\mathscr{P}(\tau) = \{\varepsilon_{\mathring{a}(i)} : i \in \mathscr{D}(\tau)\} \tag{5.8}$$

é o conjunto dos PCE que ocorrem nos “dias" i em que o PTT é τ.

– **3.** Para cada PCE ε, define-se

$$F(\varepsilon) = \frac{\#\mathscr{P}(\varepsilon)}{\#\mathscr{D}(\varepsilon)} \tag{5.9}$$

e para cada PCE ε, define-se

$$F(\tau) = \frac{\#\mathscr{P}(\tau)}{\#\mathscr{D}(\tau)} \tag{5.10}$$

– **4.** Supondo que no dia i o PCE é $\varepsilon_{\alpha(i)}$ a frequência de ocorrência do PTE ε (ou de um PTT τ), nos T dias após i

$$f(\varepsilon|\varepsilon_{\alpha(i)}, T) = \frac{\#\{j \in \{i+1, \cdots, i+1+T\} : \varepsilon_{\alpha(j)} = \varepsilon\}}{f(\varepsilon_{\alpha(i)})}$$

Análogo, trocando os papeis de ε e τ.

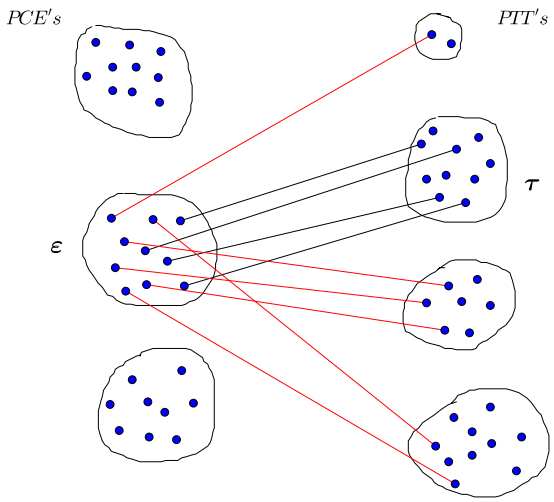


Figura 5.17: Medidas de confiança. $F(\varepsilon) = 4/10$, $\mu(\varepsilon \to \tau) = 4/10$.

- **Notas. Ver Figura 5.17.**

- **1.** Um PCE ε para o qual $F(\varepsilon) \in [0,1]$ seja grande, é um PCE que ocorre associado a vários PTT's distintos e, por isso, deve ser visto como um padrão insensível às mudanças de variável climática. Os E_i's correspondentes, i.e., nos dias i para os quais $\varepsilon_{\alpha(i)} = \varepsilon$ são úteis, por não serem influenciados por variáveis externas.

  Convém aqui definir um valor de corte γ, tal que, para $F(\varepsilon) \geq \gamma$, a conclusão anterior seja fiável.
- **2.** Uma associação (clima→consumo) do tipo $\tau \rightarrow \varepsilon$, com uma medida de confiança $\mu(\tau \rightarrow \varepsilon) \in [0,1]$ alta, é uma associação frequente - ocorre em grande parte dos dias $\mathscr{D}_\tau$.
- **3.** Devem definir-se métricas de avaliação das medidas anteriores.
- **4. Parâmetros para experimentação**
  - Comprimento da janela temporal L. Fixou-se em $L = 24$ horas - parece natural definir ciclos de atividade de 24 horas.
  - Salto υ. Tomou-se também de 24 horas. Isto é, consideraram-se subsucessões de 24 em 24 horas, apenas. Mas é preciso definir o seu início. Todas às 0 horas, por exemplo.
  - A série "climática"que se chamou T. Analisar que variáveis externas devem ser utilizadas.
  - O método de agrupamento - o número k de grupos em k-médias ou outros métodos a considerar. É desejável que os grupos sejam disjuntos o mais possível (alta resolução).
  - Definir outras medidas de associação, sua interpretação e métricas de avaliação.
  - Definir o valor de corte γ.
  - Convém ter os dados organizados em matrizes binárias para acesso rápido do programa.

Uma ideia semelhante pode ser vista em Gautam et al. [17].

O estudo desta técnica começou com o agrupamento dos dias com as observações horárias dos consumos energéticos e com as observações diárias[5] das variáveis externas

[5]Não existem registos horários para as variáveis climáticas.

climáticas (sem as variáveis construídas na Secção 2.1), para cada uma das regiões.

Testaram-se alguns métodos de agrupamento, nomeadamente, hierárquico aglomerativo e divisivo, *Partitioning Around Medoids (PAM)*, *k*-médias e DBSCAN, cuja descrição detalhada pode ser vista em Tan et al. [49]. Todos eles necessitavam de parâmetros de entrada: para k-médias, PAM e os hierárquicos testaram-se diferentes distâncias e valores de k[6] = 1,⋯,10.

Para avaliar o desempenho de cada uma das técnicas e escolher os melhores parâmetros, recorreu-se ao Coeficiente Silhueta, que incorpora as noções de coesão (elementos semelhantes no mesmo grupo) e separação (observações diferentes em grupos distintos), que é dado por

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

onde, para cada observação i, a_i é a distância média a todos os objetos do mesmo grupo, e b_i é a distância média a todos os objetos do grupo mais próximo a que i não pertence. Quanto mais perto de 1 for o valor deste coeficiente, tanto melhor é o agrupamento em causa.

A escolha final do método para agrupar os consumos energéticos horários foi feita com base no Coeficiente Silhueta obtido para as 97 instalações do primeiro lote. Para todas as instalações, o agrupamento com maior coeficiente foi obtido através do método PAM. Para a maioria (86 instalações), o melhor número de grupos a formar foi 2, sendo 3 para as restantes.

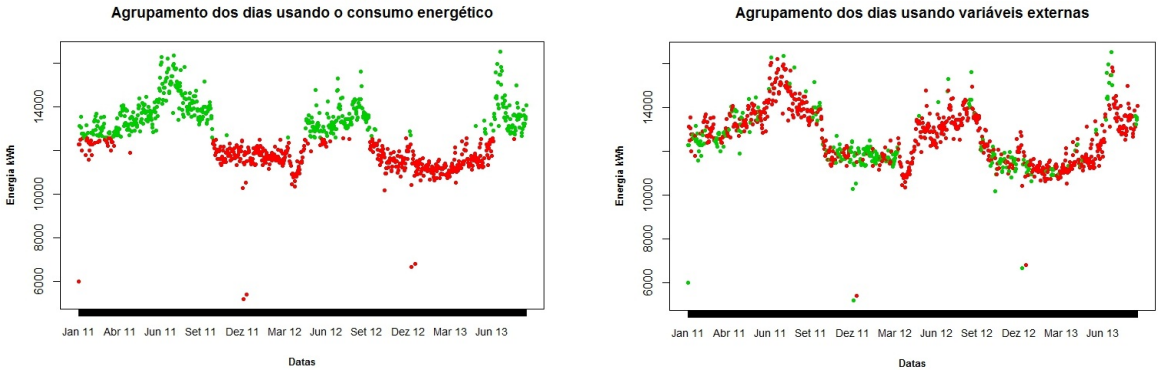
Quanto ao agrupamento das variáveis externas, um dos pontos a experimentar era a seleção daquelas que devem ser consideradas para efetuar o agrupamento. Todos os métodos de agrupamento enunciados anteriormente foram testados para cada uma das 6 regiões (Secção 4.1) com os seguintes conjuntos de variáveis: i). As 17 variáveis climáticas numéricas[7] resultantes da seleção prévia na Secção 4.2; ii) Comprimento do Dia, Ponto Orvalho Máximo e Humidade Mínima (as três finais da Secção 4.2.3); iii) Ponto Orvalho Máximo e Humidade Mínima.

Como as unidades de medida das variáveis climáticas são diferentes, para fazer o agrupamento dos dias foi necessário normalizar previamente as variáveis, procedimento já descrito na Secção 4.2, Preenchimento de falhas.

[6] Número de grupos a formar

[7] Alguns métodos de agrupamento não funcionam com variáveis categóricas, que é o caso da variável climáticas Eventos.

O maior Coeficiente Silhueta total foi obtido com o terceiro conjunto de variáveis (apenas Ponto Orvalho Máximo e Humidade Mínima). Para todas as regiões, o melhor método de agrupamento foi o hierárquico aglomerativo com a distância média e com 2 grupos finais.
Os resultados do agrupamento através dos métodos mais eficientes, com o melhor número de grupos, tanto dos consumos energéticos (PAM, k = 2) como das variáveis externas (Aglomerativo, k = 2), para uma instalação, podem ser vistos na Figura 5.18. Devido à elevada densidade dos pontos dos consumos horários, optou-se por representar o consumo total diário.



(a) Agrupamento dos dias com as observações dos consumos energéticos horários.

(b) Agrupamento dos dias com as observações diárias das variáveis climáticas.

Figura 5.18: Consumo total diário com o agrupamento dos dias.

Observado os gráficos da Figura 5.18, não se podem associar os dois agrupamentos obtidos, pois os consumos baixos (pontos vermelhos da Figura 5.18(a)) ocorrem com diferentes condições ambientais externas, ou seja, as variáveis climáticas não criaram um grupo com condições amenas e outro com extremas. Este comportamento pode ser explicado pelo método de agrupamento utilizado e o número de grupos a formar. Testaram-se outros valores de k, mas os resultados foram semelhantes.
A utilização de métodos de agrupamento mais complexos, como por exemplo CLIQUE e DENCLUE, que são métodos baseados em densidade e que estão descritos em Tan et al. [49], mas que ainda não foram implementados em R, e um estudo mais aprofundado da técnica das Associações poderiam melhorar os resultados.
Devido ao tempo limitado para a conclusão do trabalho, a determinação da energia útil através das Associações não teve continuação e na função final de desagregação do consumo energético foram incluídas as técnicas das secções anteriores (Secção 5.3.1

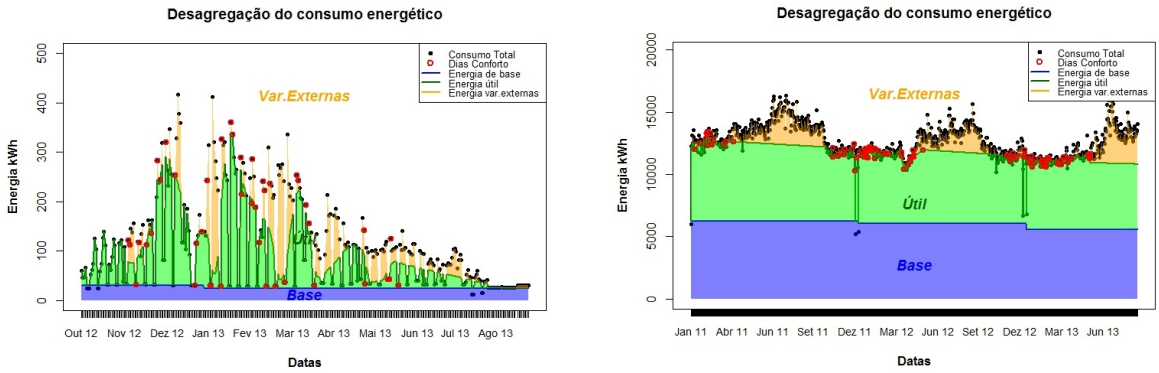
5.3.2). Com a etapa da determinação da energia útil concluída, prosseguiu-se com o delineamento da componente relacionada com as condições ambientais exteriores.

5.4 Determinação da energia das variáveis externas

Por fim, na faixa correspondente à energia relacionada com as variáveis externas, definida no início deste capítulo, foi incluído tudo o que não foi explicado pelas outras duas componentes, ou seja, Externas = Consumo Total - Baseload - Útil. Novamente, pode haver instantes em que a soma da energia de base e da útil é superior ao consumo total real desse dia (acontece quando a energia de base está acima do consumo real), ou seja, a energia das variáveis externas seria negativa. Como não existem consumos energéticos negativos, fez-se um ajustamento e a energia das variáveis externas finais é o máximo entre a diferença acima descrita e zero.

Outra maneira de calcular a componente das variáveis externas foi a descrita por Ferreira [15] que calcula a quantidade de energia desperdiçada por meio do Indicador de Desempenho Normalizado (NPI, do inglês *Normalized Performance Indicator*). Este indicador é dado pelo rácio entre o consumo energético total anual e um fator determinante, geralmente a área bruta interna do edifício (indisponível neste estudo).

A parte correspondente às condições climáticas obtida da diferença pode ser vista na Figura 5.19, em que se tem a desagregação do consumo total diário das duas instalações nas três componentes.

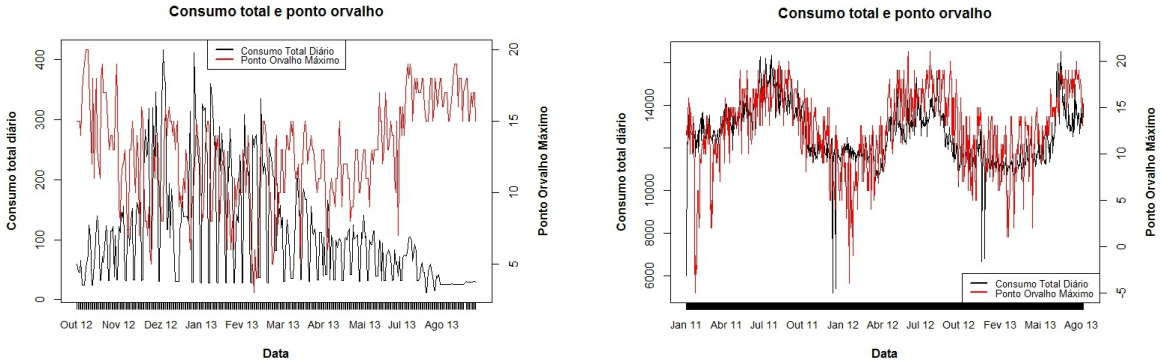


(a) Instalação 1

(b) Instalação 2

Figura 5.19: Desagregação do consumo total diário nas três componentes: energia de base, útil e das variáveis externas.

Para verificar se a energia das variáveis externas foi estimada de forma razoável, pode observar-se a variação do consumo total diário e do Ponto Orvalho Máximo ao longo do tempo, Figura 5.20.



(a) Instalação 1 (b) Instalação 2

Figura 5.20: Consumo total diário o Ponto Orvalho Máximo ao longo do tempo.

No caso da primeira instalação, o consumo total diário parece estar negativamente correlacionado com o Ponto Orvalho Máximo: quando a temperatura aumenta, o consumo diminui e vice-versa. Apesar de o comprimento da série ser inferior a um ano e portanto haver um certo erro ao fazer qualquer tipo de conclusão, esta instalação pode ser um exemplo típico daquelas que são sensíveis às condições ambientais. Quando as temperaturas exteriores estão baixas, menores do que, aproximadamente, 7°C, a instalação liga o aquecimento e, por conseguinte, o seu consumo energético aumenta. Contudo, temperaturas superiores a 10°C não têm qualquer efeito sobre o consumo da instalação. Isto pode ser explicado com, por exemplo, a ausência do ar condicionado dentro da instalação ou com férias na altura mais quente do ano.

Uma relação inversa entre o consumo e o Ponto Orvalho Máximo é registada no gráfico da Figura 5.20(b). A correlação entre as duas séries é positiva: quando a temperatura aumenta, o consumo também aumenta, mas mantém-se constante quando a temperatura desce abaixo dos, já mencionados, 7°C. O comportamento do consumo desta instalação é oposto ao da primeira e pode ser interpretado da seguinte forma: quando as temperaturas estão acima de 10 °C, é acionado o ar condicionado o que aumenta o consumo; quando está frio, não há necessidade de ligar o aquecimento. Um exemplo para este caso seria uma instalação que tem equipamentos sensíveis ao calor, por isso precisa de arrefecimento quando as temperaturas são altas, mas não são sensíveis ao frio, portanto não precisa de

aquecimento. Uma instalação com estas caraterísticas tem salas com servidores, os computadores geram calor e precisam de arrefecimento, mas não é necessário aquecimento quando está frio.

Na ausência de informações adicionais sobre a instalação e devido ao tempo limitado para a realização do trabalho, os métodos descritos ao longo deste capítulo foram os únicos incorporados na função final de desagregação. À exceção de alguns pormenores que foram omitidos neste relatório, tais como eventuais arredondamentos e algumas condições adicionais necessárias para o funcionamento correto do algoritmo, todos os passos essenciais da função que executa a desagregação do consumo energético nas três componentes foram descritos ao longo deste capítulo.
A função final (construída em R) demora 8 segundos e 3 minutos a desagregar os consumos energéticos das 97 (primeiro lote) e das 471 (total) instalações, respetivamente. O resultado são valores diários das três componentes para todos os dias de registo e, caso o utilizador o deseja, o gráfico com as três faixas (ver Figura 5.19).

Capítulo 6

Conclusão

O objetivo principal deste projeto foi delimitar as três componentes do consumo energético industrial, para poder identificar oportunidades de eficiência energética dentro de uma instalação. Ao longo do estudo foi possível identificar as variáveis climatéricas não correlacionadas, com maior impacto sobre o consumo energético, nomeadamente, Comprimento do Dia, Ponto Orvalho Máximo e Humidade Minima. Através das últimas duas definiram-se os dias de conforto que se revelaram primordiais na determinação da energia de base e da energia útil.

Com base em escassas informações sobre a instalação (apenas o diagrama de carga e a região de Portugal onde a instalação se encontrava), conseguiu-se estimar a energia gasta pela instalação para produzir, assim como aferir a sensibilidade da instalação às condições climáticas exteriores.

Propostas futuras

A temperatura exterior de conforto considerada neste projeto foi definida tendo por base apenas uma instalação, o que introduz uma probabilidade de erro bastante elevada nos resultados. Tendo em conta que o sentido de conforto dentro do edifício depende de vários fatores, como o ramo de atividade, o número de ocupantes, o isolamento do edifício, entre outros, pode propor-se o desenvolvimento de um estudo que tenha por base um número superior de instalações, que considere diversos parâmetros, com a finalidade de determinar com maior confiança as condições ambientais exteriores de conforto.

Uma outra forma de determinar a energia de base que se testou neste trabalho foi a decomposição de séries temporais. Esta abordagem permite formular um novo plano de trabalho que consistirá em aprofundar o estudo nesta área e tentar retirar a sazonalidade da tendência estimada. Um êxito neste ponto permitirá ter uma curva da energia de base mais viável e que reflita as mudanças ao longo do tempo, como, por exemplo, a substituição de equipamentos na instalação por outros mais económicos.

Um estudo mais aprofundado das técnicas de agrupamento e a aplicação prática da ideia das Associações pode gerar resultados da energia útil mais precisos do que os obtidos através dos dias de conforto, uma vez que, como já foi referido, esta metodologia não consegue detetar o trabalho que ocorre apenas fora dos dias de conforto.

Bibliografia

[1] Abels, B., Sever, F., Kissock, K. e Ayele, D. (2011). Understanding Industrial Energy Use Through Lean Energy Analysis. *SAE Int. J. Mater. Manuf.*. **4**(1): 495-504.

[2] Academia de Estudos Astrológicos. *Localidades de Portugal Continental*. Acedido em 31-03-2014, em <http://www.academiadeastrologia.com/recursos/coordenadas/portugal.htm>.

[3] American Standards of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). (2005). *ASHRAE Handbook: Fundamentals*. Inch-Pound Edition (Atlanta: ASHRAE, 2005), 8-12.

[4] Batista, G.E.A.P.A. e Monard, M.C. (2003). A Study of K-Nearest Neighbour as an Imputation Method. *In HIS*.

[5] Boduch, M. e Fincher, W. (2009). Standards of Human Comfort: Relative and Absolute. *UTSoA: Meadows Seminar Fall 2009*

[6] Breiman, L. (2001). Random Forest. *Machine Learning*. **45**(1): 5-32.

[7] Breiman, L. (2002). Manual On Setting Up, Using, And Understanding Random Forests V3.1. *Technical Report*. Acedido em 17-10-2013, em <http://oz.berkeley.edu/users/breiman>.

[8] Bromley, M. (2009). *Degree Days: Understanding Heating and Cooling Degree Days*. Acedido em 14-05-2014, em <http://www.degreedays.net/introduction>

[9] Cleveland, R.B., Cleveland, W.S., McRae, J.E. e Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, **6**: 3–73.

[10] De Livera, A.M., Hyndman, R.J. e Snyder, R.D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*. **106**(496): 1513-1527.

[11] *Degree Days - Handle with Care!*. (2008). Acedido em 16-05-2014, em <http://www.energylens.com/articles/degree-days>.

[12] Energias de Portugal (EDP). Acedido em: 07-10-2013, em <http://www.edp.pt/pt/investidores/aedp/Pages/aedp.aspx>.

[13] Everitt, B. e Hothorn, T. (2011). *An introduction to Applied Multivariate Analysis with R*. Springer Science+Business Media.

[14] Farinaccio, L. e Zmeureanu, R. (1999). Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses. *Energy and Buildings*. **30**: 245-259.

[15] Ferreira, V.G. (2009). *The analysis of primary metered half-hourly electricity and gas consumption in municipal buildings*. Ph.D. Thesis. Doctor of Philosophy in Energy and Sustainable Development, Montfort University, Leicester, UK. 194 pp.

[16] Froehlich, J., Larson, E., Gupta, S., Cohn, G., Reynolds, M.S. e Patel, S.N. (2011). Disaggregated End-Use Energy Sensing for the Smart Grid. *Pervasive Computing, IEEE*. **10**: 28-39.

[17] Gautam, D., Lin, K.I., Mannila, H., Renganathan, G., Smyth, P. (1998). Rule discovery from time series. Proceedings of the 4th international *Conference of Knowledge Discovery and Data Mining*. New York. AAAI Press. 16–22

[18] Goldstein, M. (2002). *The Complete Idiot's Guide to Weather*. Alpha Books.

[19] Golyandina, N., Nekrutkin, V. e Zhigljavsky, A. (2001). *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman&Hall/CRC

[20] Golyandina, N. e Zhigljavsky, A. (2013). *Singular Spectrum Analysis for time series*. Springer Briefs in Statistics, Springer.

[21] Hastie, T., Tibshirani, R. e Friedman, J. (2009). *The Elements os Statistical Learning: Data Mining, Inference, and Prediction*. 2ª edição, Springer.

[22] Holcomb, C.L. (2011). *Disaggregation of Residential Electric Loads Using Smart Metered Data*. Thesis. Master of Public Affairs and Master of Arts, The University of Texas at Austin, Texas, USA. 81 pp.

[23] Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A. e Laan, M.V.D. (2006). Survival Ensambles. *Biostatistics*. **7**(3): 355-373.

[24] Hyndman, R.J. (2014). *Backcasting in R*. Acedido em 20-05-2014 em <http://robjhyndman.com/hyndsight/backcasting/>.

[25] Hyndman, R.J. e Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*. **27**(3).

[26] Hyndman, R.J., Koehler, A.B., Ord, J.K. e Snyder, R.D. (2008). *Forecasting with Exponential Smoothing. The State Space Approach*. Springer - Verlag Berlin Heidelberg.

[27] Instituto Português do Mar e da Atmosfera. *Normais Climatológicas*. Acedido em 30-03-2014, em <http://www.ipma.pt/pt/oclima/normais.clima/>.

[28] Jönsson, P. e Wohlin, C. (2004). An Evaluation of k-Nearest Neighbour Imputation Using Likert Data. *10th IEEE International Symposium on Software Metrics (METRICS'04)*. **10**: 108-118.

[29] Kamholz, J. e Storer, L. (2009). Regional and Historic Standards of Comfort. *UTSoA: Meadows Seminar Fall 2009*.

[30] Kelly, D.A. (2011). *Disaggregating Smart Meter Readings using Device Signatures*. Thesis. Master in Computing Science, Imperial College London, London, UK. 75 pp.

[31] Kendall, M. e Stuart, A. (1983). *The Advanced Theory of Statistics*. **3**: 410–414, Griffin.

[32] Kim, H.S. (2012). *Unsupervised disaggregation of low frequency power measurements*. Thesis. Master in Computer Science. Graduate College, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA. 33 pp.

[33] Kleimbaum, D.G. (1994). *Logistic Regression*. Springer Verlag. New York.

[34] *kW and kWh Explained*. (2009). Acedido em: 14-05-2014, em <http://www.energylens.com/articles/kw-and-kwh>.

[35] Liaw, A. e Wiener, M. (2002). Classification and Regression by randomForest. *R News*. **2**(3): 18-22.

[36] Lines, J., Bagnall, A. Caiger-Smith, P. e Anderson, S. (2011). Classification of Household Devices by Electricity Usage Profiles. Em: H. Yin, W. Wang, and V. Rayward-Smith (eds.), *IDEAL*, Springer-Verlag Berlin Heidelberg. 403–412.

[37] McNoldy, B. (2014)*Calculate Temperature, Dewpoint, or Relative Humidity* . Rosenstiel School of Marine & Atmospheric Science. University of Miami. Acedido em 15-04-2014, em <http://andrew.rsmas.miami.edu/bmcnoldy/Humidity.html>.

[38] Meijering, E. (2002). A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*. **90**(3): 319–342.

[39] Murteira, B.J.F., Muller, D.A. e Turkman, K.F. (1993). *Análise de Sucessões Cronológicas*. McGraw-Hill. Lisboa.

[40] Observatório Naval dos Estados Unidos. *Sun or Moon Rise/Set Table for One Year*. Acedido em 30-10-2013, em <http://aa.usno.navy.mil/data/docs/RS_OneYear.php>.

[41] Occupational Safety & Health Administration. *Policy on Indoor Air Quality: Office Temperature/Humidity and Environmental Tobacco Smoke*. Acedido em 18-03-2014, em <https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=INTERPRETATIONS&p_id=24602>.

[42] Parsons, K.C. (1995). “Introduction” from Nicol, Fergus et al (Eds.), *Standards for Thermal Comfort: Indoor Air Temperature Standards for the 21st Century*. pp xiii.

[43] Peel, M.C., Finlayson, B.L. e McMahon, T.A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.*. **11**: 1633–1644

[44] Price, P.N. (2010).*Methods for Analyzing Electric Load Shape and its Variability*. Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-3713E.

[45] Prokhorov, A.V. (2001). Partial correlation coefficiente. Em: M. Hazewinkel (eds.), *Encyclopedia of Mathematics*. Springer.

[46] Rubel, F. e Kottek, M. (2010). Observed and projected climate shifts 1901-2100 depicted by world maps of the Köppen-Geiger climate classification. *Meteorologische Zeitschrift*. **19**(2): 135-141.

[47] Strobl, C., Boulesteix, A.L., Zeileis, A. e Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*. **8**(25)

[48] Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T. e Zeileis, A. (2008). Conditional Variable Importance for Random Forests. *BMC Bioinformatics*. **9**(307).

[49] Tan, P.N., Steinbach, M. e Kumar, V. (2006). *Introduction to Data Mining*. 1ª edição, Pearson Addison Wesley. Boston.

[50] Torgo, L. (2010). *Data Mining with R: Learning with Case Studies*. Chapman & Hall/CRC.

[51] Venables, W.N. e Ripley, B.D. (2002). *Modern Applied Statistics with S*. 4ª edição, Springer. New York.

[52] Verzani, J. (2005). *Using R for Introductory Statistics*. Chapman & Hall/CRC Press.

[53] Vitullo, S. (2011). *Disaggregating Time Series Data for Energy Consumption by Aggregate and Individual Customer*. Ph.D. Dissertations (2009 -). Doctor of Philosophy, Marquette University, Milwaukee, Wisconsin, USA. Paper 169. 141 pp.

[54] Weather Underground. *Historical Weather*. Acedido em 31-03-2014, em <http://www.wunderground.com/history/>.

[55] Wikipedians. *Meteorology*. Pedia Press. Acedido em 19-04-2014, em <http://books.google.pt/books?id=6cRTp4enDxkC&printsec=frontcover&hl=pt-PT#v=onepage&q&f=false.

[56] Zefman, M. (2012). Disaggregation of Home Energy Display Data Using Probabilistic Approach. *Transactions on Consumer Electronics, IEEE*. **58**: 23-31.

Apêndice A

Contextualização do problema

A.1 Análise gráfica

Na figura A.1 está representado o consumo diário agregado pelo mínimo. A forma da curva está próxima da curva do consumo total diário, com excepção do maior número de picos que aparecem e o facto de estarem em outras posições. O mínimo não é uma medida viável para representar a curva do consumo, uma vez que reflete as falhas na instalação (consumo = 0) que não são caraterísticas próprias da instalação.

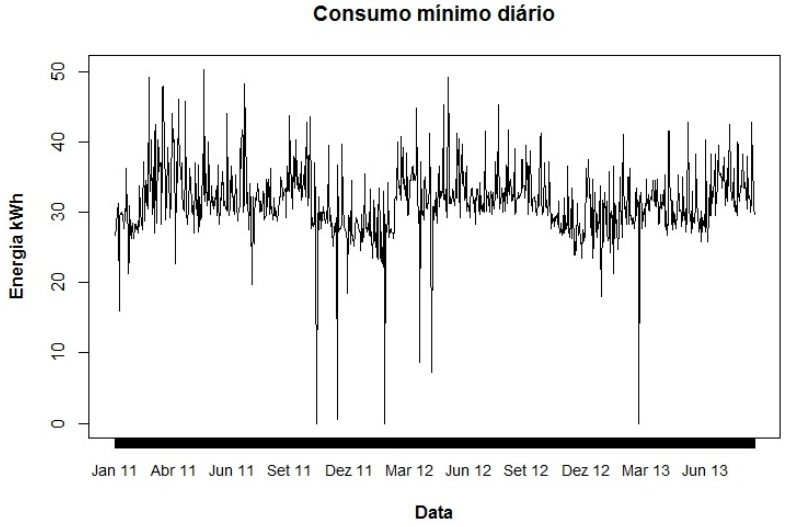


Figura A.1: Consumo mínimo diário de Janeiro de 2011 a Agosto de 2013

Na figura A.2 está representado o consumo diário agregado pelo máximo. A forma da

curva é muito semelhante à do consumo total diário, mais do que é o mínimo. Tanto o máximo como o mínimo são as medidas mais adequadas para identificar as anomalias que são falhas na instalação (consumo = 0 ou consumo é máximo). Como o objetivo deste trabalho é avaliar o consumo típico de uma instalação, estas medidas não serão usadas na agregação de dados.

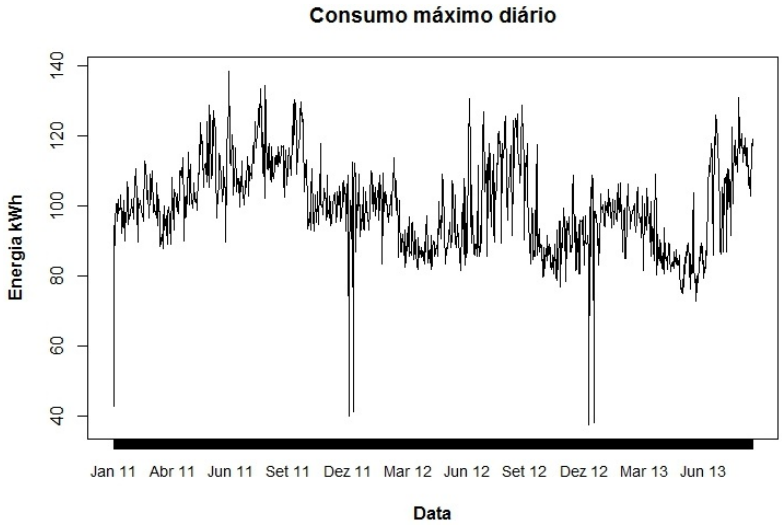


Figura A.2: Consumo máximo diário de Janeiro de 2011 a Agosto de 2013

O consumo diário agregado usando as três medidas (mínimo, média e máximo) pode ser visto na Figura A.3. Apesar de as formas serem semelhantes, existem algumas diferenças (por exemplo, a variação do consumo agregado pelo máximo é muito superior ao agregado pelo mínimo).

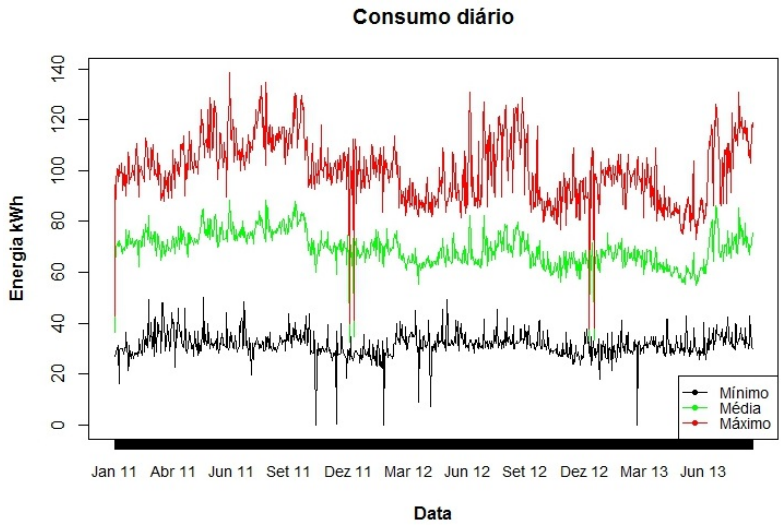


Figura A.3: Consumo diário agregado pelo mínimo (preto), média (verde) e máximo (vermelho)

A.2 Feriados 2011-2013

A seguir tem-se a lista dos feriados nacionais dos anos 2011 até 2013.

Data	Descricao	Data	Descricao
01-01-2010	Ano Novo	08-12-2011	Imaculada Conceição
16-02-2010	Carnaval	25-12-2011	Natal
02-04-2010	Sexta-feira Santa	01-01-2012	Ano Novo
04-04-2010	Páscoa	21-02-2012	Carnaval
25-04-2010	25 de Abril	06-04-2012	Sexta-feira Santa
01-05-2010	Dia do Trabalhador	08-04-2012	Páscoa
03-06-2010	Corpo de Deus	25-04-2012	25 de Abril
10-06-2010	Dia de Portugal	01-05-2012	Dia do Trabalhador
15-08-2010	Assunção de Nossa Senhora	07-06-2012	Corpo de Deus
05-10-2010	Implantação da República	10-06-2012	Dia de Portugal
01-11-2010	Dia de Todos os Santos	15-08-2012	Assunção de Nossa Senhora
01-12-2010	Restauração da Independência	05-10-2012	Implantação da República
08-12-2010	Imaculada Conceição	01-11-2012	Dia de Todos os Santos
25-12-2010	Natal	01-12-2012	Restauração da Independência
01-01-2011	Ano Novo	08-12-2012	Imaculada Conceição
08-03-2011	Carnaval	25-12-2012	Natal
22-04-2011	Sexta-feira Santa	01-01-2013	Ano Novo
24-04-2011	Páscoa	12-02-2013	Carnaval
25-04-2011	25 de Abril	29-03-2013	Sexta-feira Santa
01-05-2011	Dia do Trabalhador	31-03-2013	Páscoa
10-06-2011	Dia de Portugal	25-04-2013	25 de Abril
23-06-2011	Corpo de Deus	01-05-2013	Dia do Trabalhador
15-08-2011	Assunção de Nossa Senhora	10-06-2013	Dia de Portugal
05-10-2011	Implantação da República	15-08-2013	Assunção de Nossa Senhora
01-11-2011	Dia de Todos os Santos	08-12-2013	Imaculada Conceição
01-12-2011	Restauração da Independência	25-12-2013	Natal

Tabela A.1: Feriados dos anos 2011 a 2013

A.3 Decomposição de uma série temporal

Foram usados modelos de decomposição aditivos uma vez que a tendência não apresenta comportamento exponencial, ou seja, a variação ao longo do tempo é suave. Resultado da decomposição da série do consumo usando a instrução `decompose()` de `R` que se baseia no método da média móvel:

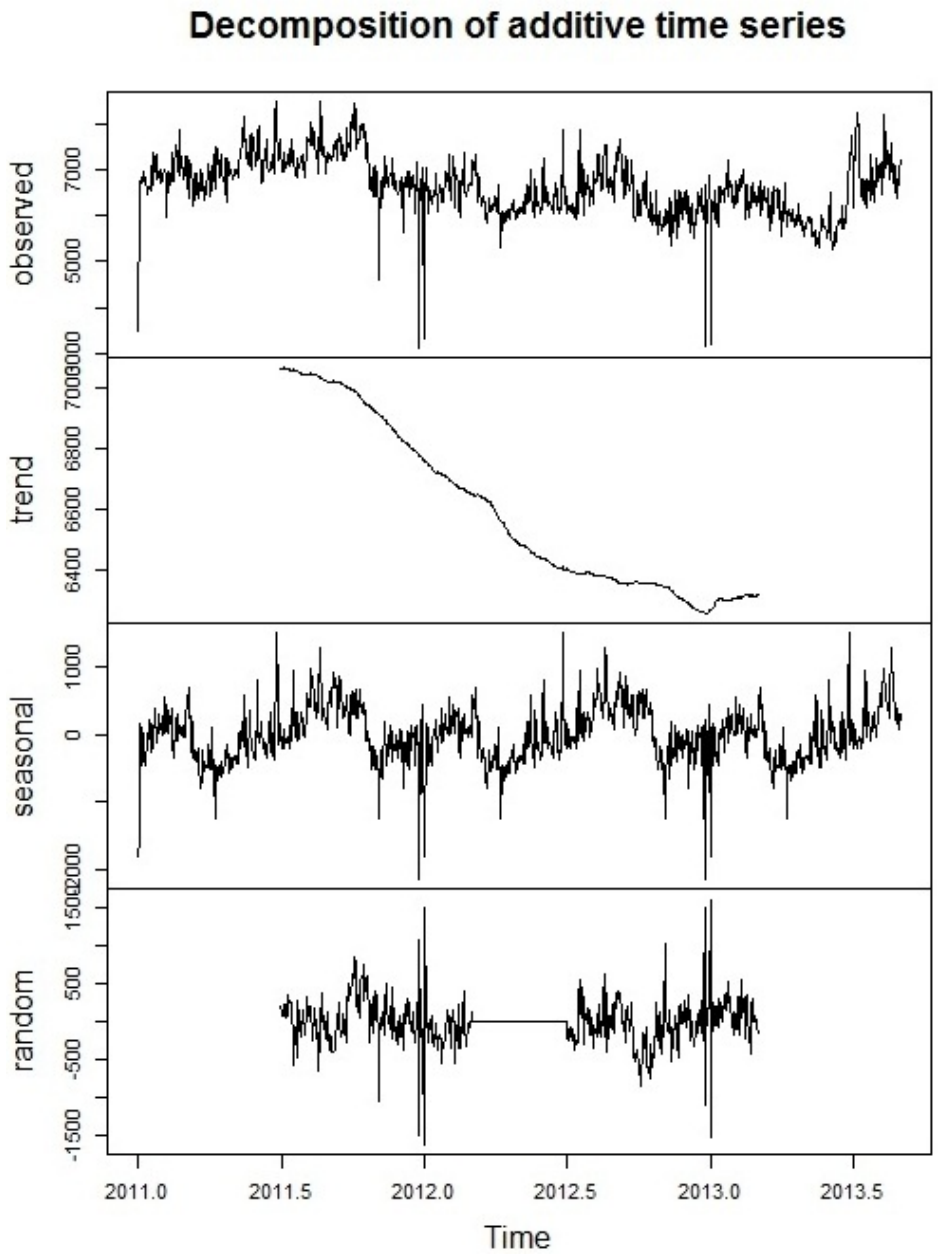


Figura A.4: Decomposição da série do consumo usando `decompose()`

Observando a Figura A.4 pode ver-se que, devido à técnica que está na base desta

decomposição, a tendência foi estimada apenas durante algum tempo, o que não cumpre o propósito.

A decomposição da série usando a instrução `stl()` de `R`, que tem como base a decomposição de Loess, mais detalhe em [9], produz melhores resultados:

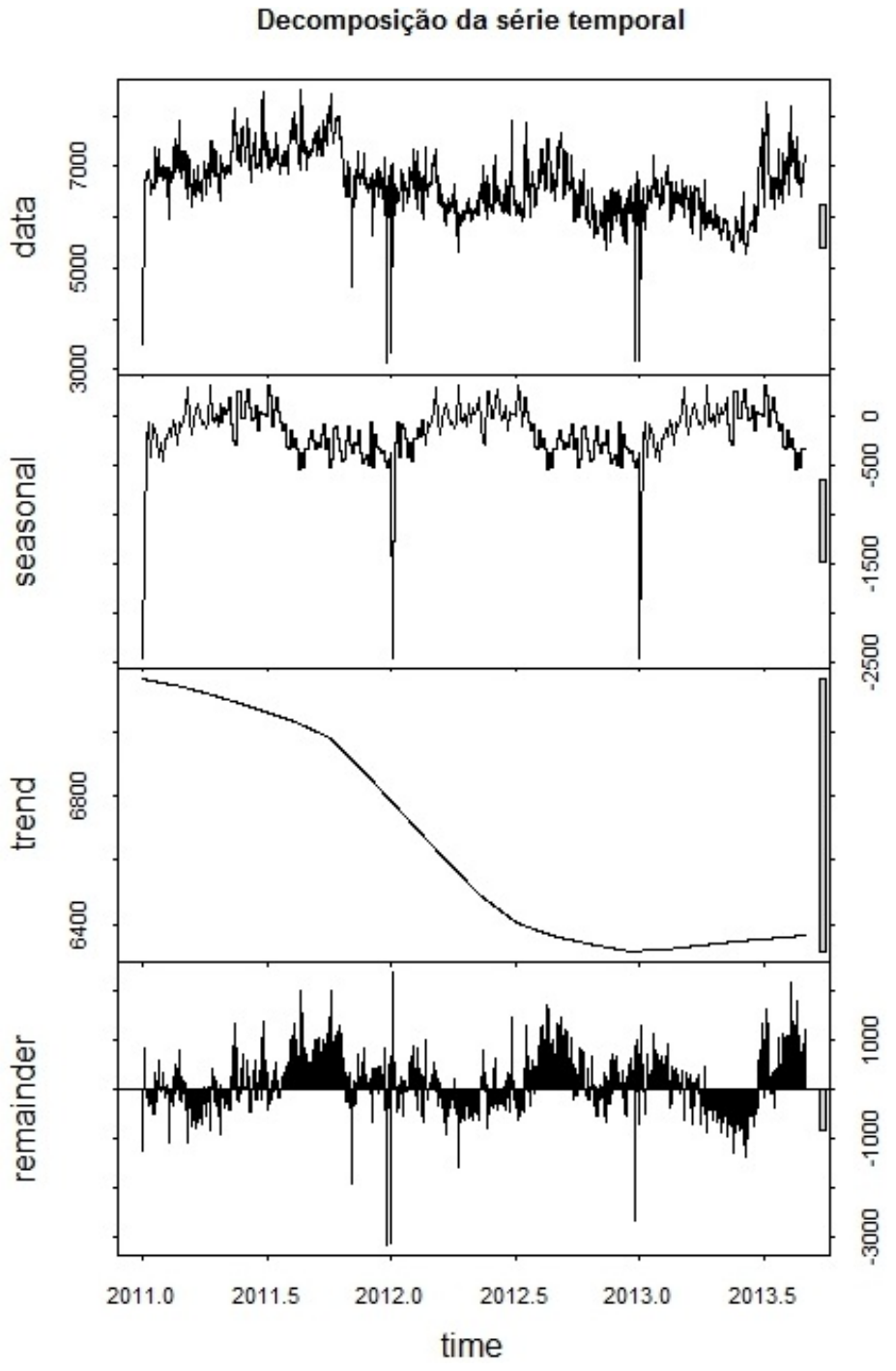


Figura A.5: Decomposição da série do consumo usando `stl()`

Através desta técnica, a decomposição da série foi efetuada com maior precisão, no entanto, deve averiguar-se a distribuição dos resíduos, que deve ser aleatória.

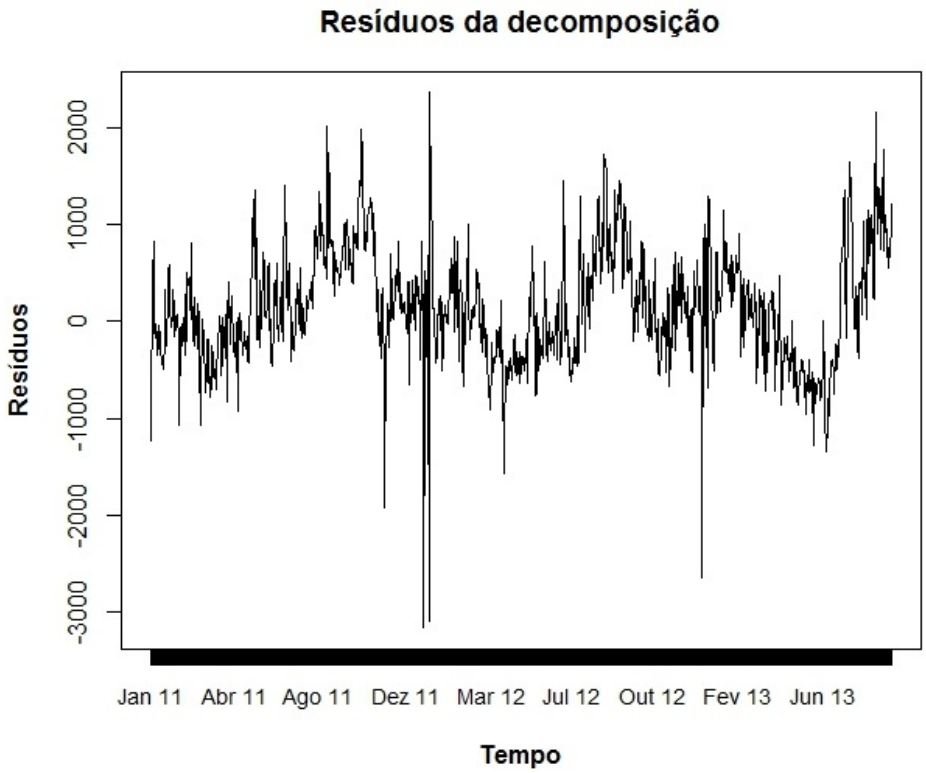


Figura A.6: Resíduos da decomposição

Na Figura A.6 pode ver-se que, após a extração das primeiras duas componentes da série, continua a existir sazonalidade nos resíduos. Portanto esta componente não foi bem estimada usando este método de decomposição.

Apêndice B

Variáveis externas

B.1 Descrição das variáveis climáticas

Segue-se a definição e a descrição estatística das 22 variáveis climáticas extraídas do site [54], mais a variável Comprimento do Dia retirada do Observatório Naval dos Estados Unidos [40]. As definições das mesmas são as dadas pelo Instituto Português do Mar e da Atmosfera [27]. Todas as variáveis têm observações diárias. Os gráficos são as representações das variáveis ao longo do tempo após ter sido feita a substituição ('NA: ' indica o número de falhas antes do preenchimento) dos valores em falta. No caso de se dispor de mais do que uma medida da variável (máximo, médio, mínimo), estas estão representadas no mesmo gráfico.

Comprimento do Dia - Número de horas entre o nascer e o pôr do Sol.

Mínimo: 9.45	1º Quartil: 10.57
Média: 12.37	Mediana: 12.57
Máximo: 14.88	3º Quartil: 14.18
	NA: 0

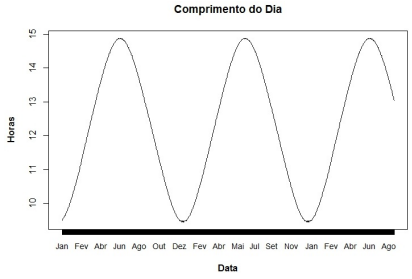


Figura B.1: Comprimento do Dia

Temperatura Máxima/ **Média**/ **Mínima** - Temperatura máxima/ média/ mínima do ar em graus Célcios (ºC) registada no dia.

Máxima	**Média**	**Mínima**
Mínimo: 8	Mínimo: 6	Mínimo: 1
1º Quartil: 16	1º Quartil: 12	1º Quartil: 9
Mediana: 20	Mediana: 17	Mediana: 13
Média: 21.4	Média: 16.9	Média: 12.9
3º Quartil: 26	3º Quartil: 21	3º Quartil: 17
Máximo: 40	Máximo: 32	Máximo: 26
NA: 1	NA: 2	NA: 1

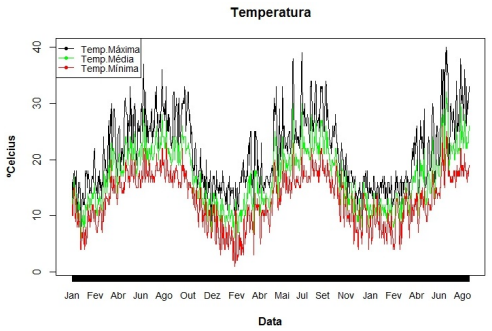


Figura B.2: Temperatura Máxima (preto), Média (verde), Mínima (vermelho)

Ponto de Orvalho Máximo/ **Médio**/ **Mínimo** - Temperatura máxima/ média/ mínima do ar em ºC a que o vapor de água presente no ar passa ao estado líquido na forma de pequenas gotas por via de condensação, o chamado orvalho.

Máximo	**Médio**	**Mínimo**
Mínimo: -8	Mínimo: -11	Mínimo: -13
1º Quartil: 11	1º Quartil: 8	1º Quartil: 5
Mediana: 13	Mediana: 11	Mediana: 9
Média: 13.1	Média: 10.7	Média: 8.0
3º Quartil: 16	3º Quartil: 14	3º Quartil: 11
Máximo: 21	Máximo: 19	Máximo: 18
NA: 1	NA: 1	NA: 1

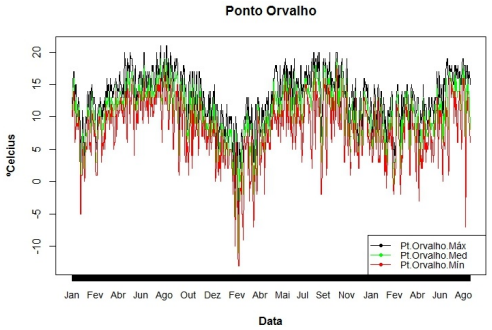


Figura B.3: Ponto de Orvalho Máximo (preto), Médio (verde), Mínimo (vermelho)

Humidade Máxima/ **Média**/ **Mínima** - Medida em %, indica a razão entre a massa real máxima/ média/ mínima, registadas no dia, de vapor de água contida na unidade de volume de ar e a massa de vapor que seria necessária para que este volume de ar ficasse saturado à mesma temperatura. Expressa-se vulgarmente sob a forma de percentagem.

Máxima	**Média**	**Mínima**
Mínimo: 30	Mínimo: 22	Mínimo: 8
1º Quartil: 83	1º Quartil: 64	1º Quartil: 39
Mediana: 88	Mediana: 72	Mediana: 50
Média: 88.6	Média: 70.9	Média: 50.5
3º Quartil: 94	3º Quartil: 81	3º Quartil: 63
Máximo: 100	Máximo: 100	Máximo: 100
NA: 1	NA: 1	NA: 1

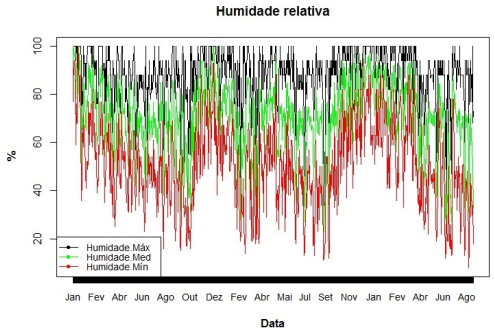


Figura B.4: Humidade Máxima (preto), Média (verde), Mínima (vermelho)

Pressão Máxima/ Média/ Mínima ao Nível do Mar - É a força máxima/ média/ mínima registada no dia a uma altitude de 0m (o nível do mar), exercida sobre uma dada superfície devido ao peso do ar. É medida em hectopascal (hPa).

Máxima	**Média**	**Mínima**
Mínimo: 990	Mínimo: 989	Mínimo: 986
1º Quartil: 1017	1º Quartil: 1014	1º Quartil: 1012
Mediana: 1020	Mediana: 1018	Mediana: 1016
Média: 1020	Média: 1018	Média: 1016
3º Quartil: 1023	3º Quartil: 1021	3º Quartil: 1020
Máximo: 1038	Máximo: 1036	Máximo: 1035
NA: 1	NA: 1	NA: 1

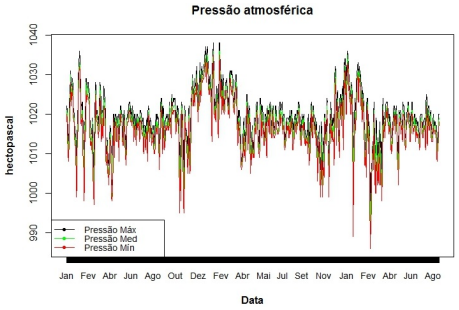


Figura B.5: Pressão Máxima (preto), Média (verde), Mínima (vermelho)

Visibilidade Máxima/ Média/ Mínima - Indica a distância máxima/ média/ mínima no dia numa dada direção, a que uma pessoa de vista normal (sem binóculos,etc.), pode distinguir e identificar contra o céu do horizonte, um objeto de dimensões convenientes, à luz de dia. A unidade de medida desta variável é quilómetros (km).

Máxima	**Média**	**Mínima**
Mínimo: 3	Mínimo: 1	Mínimo: 0
1º Quartil: 10	1º Quartil: 10	1º Quartil: 6
Mediana: 10	Mediana: 10	Mediana: 10
Média: 9.9	Média: 9.5	Média: 7.8
3º Quartil: 10	3º Quartil: 10	3º Quartil: 10
Máximo: 10	Máximo: 10	Máximo: 10
NA: 125	NA: 125	NA: 125

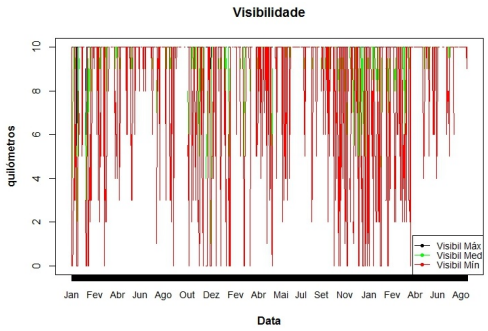


Figura B.6: Visibilidade Máxima (preto), Média (verde), Mínima (vermelho)

Velocidade Máxima/ Média do Vento - O vento é definido como a componente horizontal do movimento do ar à altura de 10 m do solo em terreno aberto. A variável em causa indica a velocidade máxima/ média do vento registada no dia. A unidade de medida é quilómetros por hora (km/h).

Máxima	**Média**
Mínimo: 8	Mínimo: 3
1º Quartil: 21	1º Quartil: 10
Mediana: 26	Mediana: 13
Média: 25.8	Média: 13.9
3º Quartil: 29	3º Quartil: 18
Máximo: 60	Máximo: 39
NA: 1	NA: 1

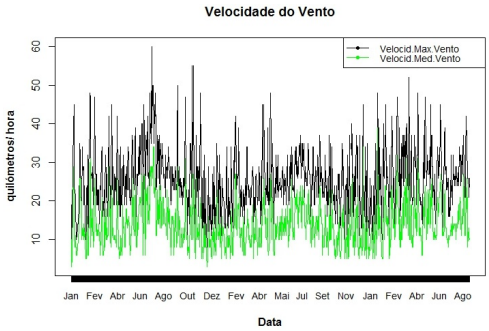


Figura B.7: Velocidade Máxima (preto), Média (verde) do vento

Velocidade Máxima da Rajada do Vento - Rajada define-se como um vento que é dado em forma de pulsos e apresenta-se como um sopro súbito que excede a velocidade média do momento em mais de 5.14 m/s por um período de tempo inferior a 20 segundos. A variável regista a velocidade máxima em km/h da rajada ocorrida no dia.

Mínimo: 24 | 1º Quartil: 34
Média: 42.8 | Mediana: 40
Máximo: 90 | 3º Quartil: 50
NA: 526

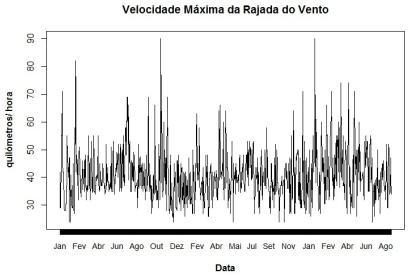


Figura B.8: Velocidade Máxima da Rajada do Vento

Precipitação - Indica a quantidade de água (em estado líquido ou sólido) caída, por unidade de superfície, num dado intervalo de tempo. É medida em milímetros (mm) de altura (1 mm = 1 litro/metro^2). Na região de Lisboa, esta variável é sempre igual a zero, de modo que não se apresenta o seu gráfico.

Cobertura com Nuvens - Indica a quantidade de nuvens no Céu. A unidade de medida é o oitavo (do Céu) e varia entre 0 e 9, onde 0 corresponde ao Céu completamente limpo, 8 - Céu completamente coberto e 9 ao Céu obscurecido. O código numérico 3, por exemplo, irá ser interpretado como 3/8 do Céu estão cobertos com nuvens.

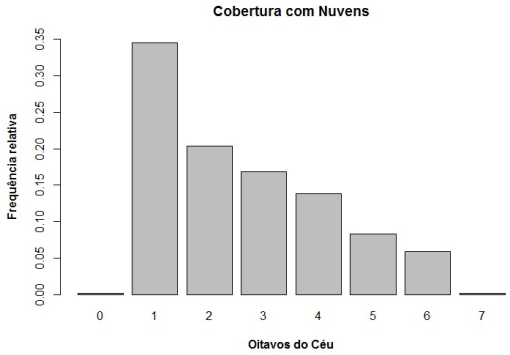


Figura B.9: Cobertura com Nuvens

Eventos - Indica os fenómenos meteorológicos observados no dia, que são explicados pela ciência da meteorologia. Exemplo: Chuva, Neve, Nenhum, etc.

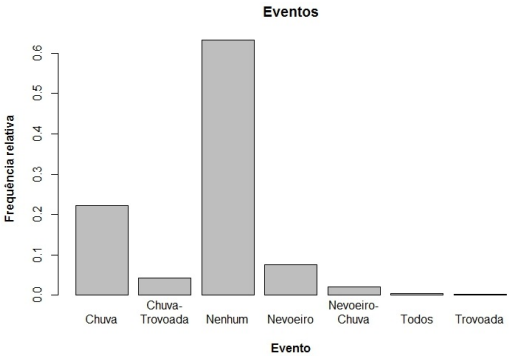


Figura B.10: Eventos

Direção do Vento - É a direção da qual o vento sopra. Exprime-se em graus medidos a partir do Norte geográfico, no sentido dos ponteiros do relógio.

Mínimo: -1 | 1º Quartil: 124
Média: 236.7 | Mediana: 296
Máximo: 360 | 3º Quartil: 337
NA: 0

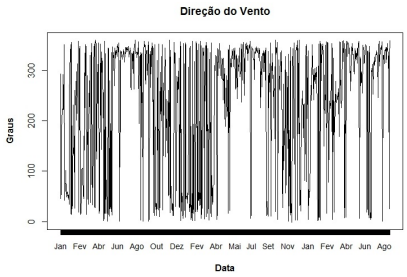


Figura B.11: Direção do Vento

B.2 Floresta aleatória numa instalação

Instruções em R para obter as variáveis mais importantes usando a floresta aleatória:

```
modelo <- randomForest(ConsumoTotal~.,
            data,mtry=8,ntree=2000,
            importance=T)
varImpPlot(modelo)
```

Tal como foi mencionado na Secção 4.2.2, são consideradas as 10 variáveis com maior incremento no erro quadrático médio - as que estão na parte superior da linha vermelha na Figura B.12.

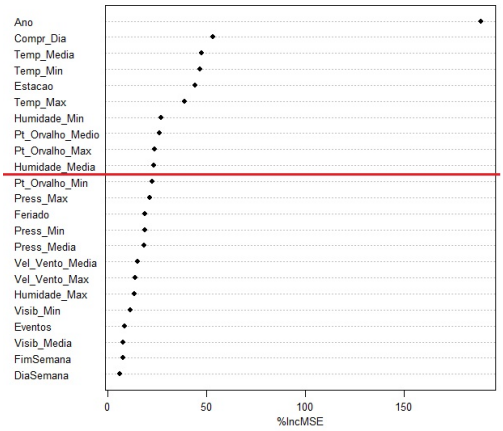


Figura B.12: Importância das variáveis climáticas da região de Beja

B.3 Correlação variáveis climáticas

Correlação entre variáveis climáticas

Na Tabela B.1 tem-se o valor p da correlação parcial entre o consumo total diário e algumas variáveis climáticas, tal como da correlação entre as próprias variáveis externas.

Os valores p da correlação parcial serviram para selecionar as variáveis que têm a correlação com o consumo total diário estatisticamente significativa (valor p inferior a 0.05).

Variáveis	Consumo Total Diário	Comprimento do Dia	Temperatura Máxima	Temperatura Média
Consumo Total	0.000	0.003	0.227	0.508
Comprimento do Dia	0.003	0.000	0.694	0.782
Temperatura Máxima	0.227	0.694	0.000	0.000
Temperatura Média	0.508	0.782	0.000	0.000
Temperatura Mínima	0.429	0.765	0.000	0.000
Ponto Orvalho Máximo	0.004	0.020	0.414	0.699
Ponto Orvalho Médio	0.913	0.019	0.001	0.476
Ponto Orvalho Mínimo	0.328	0.105	0.921	0.872
Humidade Máxima	0.209	0.503	0.799	0.709
Humidade Média	0.973	0.031	0.000	0.438
Humidade Mínima	0.031	0.000	0.019	0.405

Tabela B.1: Valor p da correlação parcial entre consumo total diário e algumas variáveis climáticas

Neste exemplo, as variáveis que satisfazem esta condição são Comprimento do Dia, Ponto Orvalho Máximo e Humidade Mínima, precisamente aquelas que foram declaradas como as mais relacionadas com o consumo.

Tabela da correlação cruzada entre algumas variáveis climáticas da região de Beja.

Variáveis	Comprimento do Dia	Temperatura Máxima	Temperatura Média	Temperatura Mínima
Comprimento do Dia	1.000	0.730	0.718	0.621
Temperatura Máxima	0.730	1.000	0.970	0.827
Temperatura Média	0.718	0.970	1.000	0.934
Temperatura Mínima	0.621	0.827	0.934	1.000
Ponto Orvalho Máximo	0.430	0.577	0.675	0.764
Ponto Orvalho Médio	0.369	0.467	0.590	0.724
Ponto Orvalho Mínimo	0.211	0.234	0.375	0.563
Humidade Máxima	−0.333	−0.516	−0.506	−0.432
Humidade Média	−0.605	−0.775	−0.694	−0.496
Humidade Mínima	−0.645	−0.782	−0.674	−0.434

Tabela B.2: Correlação cruzada entre algumas variáveis climáticas da região de Beja.

Na Tabela B.2 pode observar-se que a correlação entre a Temperatura Máxima e Média (0.970) está muito perto de 1, o valor máximo, o que significa que as variáveis transmitem a mesma informação. Para evitar multicolinearidade, uma destas variáveis terá que ser eliminada. Entretanto foram retiradas ambas uma vez que a correlação destas com o

Comprimento do Dia é maior que 0.65, valor de corte adaptado na Secção 4.2.3 para eliminação de variáveis significativas. Comparou-se com o Comprimento do Dia porque esta é a variável com maior frequência relativa da Tabela 4.3 e portanto é a primeira a ser classificada como variável que influencia o consumo energético.

Apêndice C

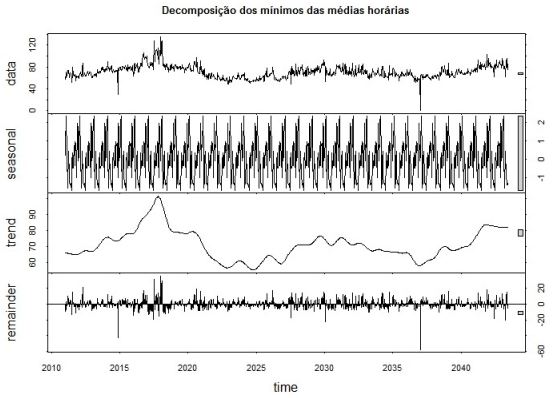
Desagregação do consumo energético

C.1 Baseload

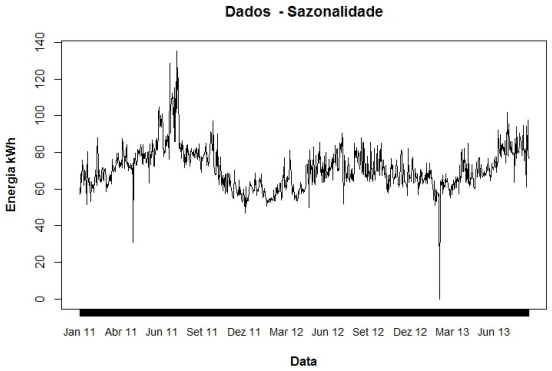
Um dos métodos estudados para determinar o baseload requeria a estimação da sazonalidade e de seguida a eliminação desta componente dos dados, que, neste caso, foram os mínimos das médias horárias (ver Capítulo 5). Isso podia ser feito de duas maneiras: 1. decompor os dados originais, ou 2. decompor a tendência dos dados; sendo o resultado parecido para ambas as abordagens.

Foram testadas algumas das técnicas implementadas em `R`, nomeadamente: i) métodos clássicos de decomposição (`stl` e `decompose`), ii) decomposição usando SSA e iii) decomposição de séries com múltiplas sazonalidades. Os métodos clássicos tiveram respostas parecidas e o resultado da aplicação da função `decompose` pode ser visto na Figura C.1. O mesmo aconteceu usando o SSA e as sazonalidades múltiplas.

A decomposição clássica não consegue estimar suficientemente bem a sazonalidade, de modo que, parte desta componente, ainda permanece nos dados após a sua retirada.



(a) Decomposição dos mínimos das médias horários usando a técnica `decompose`



(b) Os dados menos a sazonalidade estimada

Figura C.1: Decomposição dos mínimos das médias horárias usando a técnica `decompose` e o resultado após retirar a sazonalidade estimada

Dendrograma

Na Figura C.2 tem-se o resultado da aplicação do método de agrupamento hierárquico aglomerativo aos consumos mínimos das médias horárias de uma instalação. A reta vermelha indica o valor de corte pré-definido. Na horizontal têm-se os números dos grupos formados, na vertical - a distância entre os grupos.

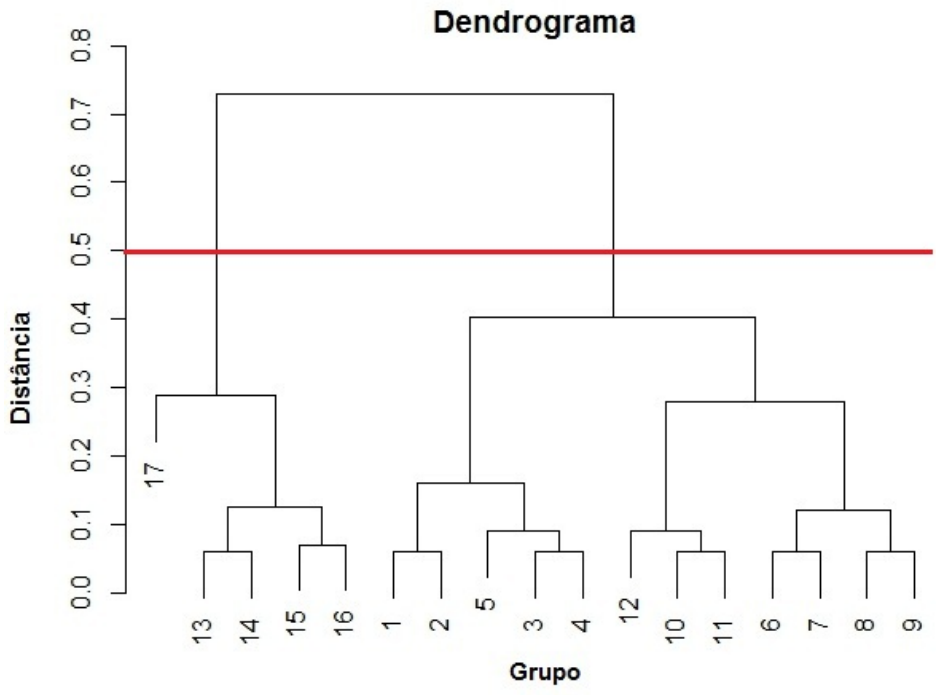


Figura C.2: Dendrograma de uma instalação