Inês Margarida Simões Pessoa

Algoritmos de Análise e Desagregação de Consumos Elétricos em Edifícios

Dissertação apresentada ao Departamento de Física da
Universidade de Coimbra para obter o grau de Mestre em Engenharia Física

Março de 2018

· U C ·

Universidade de Coimbra

• U C •
FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Inês Margarida Simões Pessoa

Algoritmos de Análise e Desagregação de Consumos Elétricos em Edifícios

Dissertação apresentada à Universidade de Coimbra para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Física no ramo de Metrologia e Qualidade

Orientador:
Jorge Afonso Cardoso Landeck (Universidade de Coimbra)

Coimbra, 2018

Este trabalho foi desenvolvido em colaboração com:

www.vps.energy

Para a minha Mãe, Pai e Irmã.

Agradecimentos

À minha família, sem eles nada disto seria possível.

Ao Bernardo, por ser o meu porto seguro.

Aos meus amigos, pelo apoio nestes últimos meses.

Ao Professor Jorge Landeck, pela orientação e acompanhamento ao longo deste projeto.

À Luísa Matos, pela orientação inicial dada neste projeto, e ao resto da equipa da inovação da VPS, pela companhia.

A todos aqueles que fizeram parte deste percurso académico, que termina com este projeto, aos professores pelos ensinamentos e aos colegas, juntos crescemos e criámos memórias.

Resumo

A par do avanço industrial e tecnológico, tem-se verificado um crescente aumento do consumo energético nos países desenvolvidos.

Por consequência da elevada demanda de energia na rede, as fontes limpas não são suficientes para dar resposta às solicitações de energia, tornando-se necessário recorrer a fontes fósseis e a energia nuclear. Existe uma necessidade de privilegiar o consumo de fontes renováveis face às restantes fontes de energia. A necessidade da redução do uso de fontes fósseis e nucleares passa pela sua escassez (fontes fósseis) e poluição libertada para o meio ambiente. Para atingir este fim, deve-se privilegiar o consumo de fontes renováveis face às restantes fontes de energia.

No âmbito do consumo de energia elétrica em edifícios habitacionais e empresariais, a forma óbvia de contornar este problema passa pela redução do consumo energético e a educação do consumidor para usar a energia em períodos de menor saturação da rede.

Este projeto tem como propósito o desenvolvimento de algoritmos, para descoberta de conhecimento a partir do histórico de dados do consumo elétrico de um edifício. Este conhecimento consiste na informação de como a energia elétrica é usada no edifício e dos fatores externos que influenciam os consumos. Estes algoritmos consistem em descobrir padrões nos dados e associações frequentes entre os registos e fatores externos, como por exemplo a temperatura registada ou outros equipamentos a funcionar em paralelo.

O projeto foca-se também no desenvolvimento de algoritmos de desagregação dos consumos totais registados no edifício. Estes algoritmos consistem em previsões de que aparelhos presentes no edifício funcionam, a cada instante de registo, nos consumos agregados.

Assim, com este projeto, pretende-se concluir o máximo de informação da forma como a energia está a ser usada, nas bases de dados disponíveis, para que possa servir de suporte à tomada de decisões para redução de consumos elétricos e deslocamento de carga para períodos de menor demanda da rede, sem que as necessidades e conforto dos

presentes no edifício sejam comprometidas. E também, em desenvolver estratégias para que no futuro a recolha de informação tenha o mínimo custo e seja o menos intrusiva possível, para assim, estar ao alcance de todos e não causar desconforto, convergindo para que a adesão a serviços de monitorização de consumos elétricos seja cada vez maior. Com uma maior adesão a serviços de monitorização, a energia é usada de forma mais eficiente num grande número de edifícios, contribuindo para uma redução considerável na pegada ecológica coletiva.

Palavras-Chave

Assinaturas de carga; Agrupamento não supervisionado; Monitorização não intrusiva de carga; Performance energética de um edifício; Aproximação agregada simbólica.

Abstract

With industrial and technological growth, the consumption of energy in developed countries has also increased.

Because of the high energy demand in the network, there is a massive amount of energy consumption and not enough clean energy to overcome it, from which arrives the need to use fossil sources and nuclear energy. There is a big need to favor renewable energies instead of other kind of energy sources. The need to avoid these kinds of sources comes from its shortage and pollution released to the environment. For this, it is necessary to use renewable sources instead of other sources.

In the energy consumption context in residential buildings and companies, the obvious way to solve this problem is reducing the energetic consumption and educate the consumer to use it when the energy saturation on the network is lower.

The purpose of this project is to develop algorithms to obtain knowledge from the building consumption's energy data. This consists in the information of how the energy is used on a building and the external factors that influence that consumption. These algorithms find patterns on data and frequent association between data records and external factors, such as the registered temperature or other parallel working equipment.

The project also focuses on developing desegregation algorithms of total consumption registered on the building. These algorithms are predictions of which equipment is in use, in the building, at any time, in the aggregated data.

The goal of this project is to get the maximum information of how energy is being used and on the available databases, which can be used to help in the decision making to reduce energy consumption and displace it for smaller demand times of the network without affecting the needs and comfort of others in the network. And also to develop strategies to make the information gathering less expensive and intrusive in the future, to make it available to everyone without causing discomfort, causing a bigger support to monitoring services. This way, the energy is used in a more efficient way in a lot of buildings, contributing to a substantial reduction on the collective ecological footprint.

Keywords

Building energy performance; Clustering; Load signatures; Non-Intrusive Load Monitoring (NILM); Symbolic Aggregate Approximation (SAX); .

Glossário

cloud	neste contexto, representa uma abstração física de um lugar em que estão armazenados os dados que se pretendem aceder
cluster	um grupo de dados, dos grupos resultantes de um algoritmo de clustering*
clustering	métodos matemáticos para executar uma classificação, não supervisionada, dos dados em grupos, chamados de clusters* [8]
clusters	plural de cluster*
data mining	processo de extração de conhecimento útil numa base de dados [1]
datasheet	documento com as informações técnicas de um equipamento
datasheets	plural de datasheet*
features	informação medível [4]
framework	neste contexto, representa uma metodologia genérica para abordar um determinado assunto
hub	dispositivo que recebe ou transmite informação numa rede
input	dados de entrada
machine learning	modelação computacional durante uma fase de aprendizagem proporcionada por uma base de dados [2]
outlier	observação que se afasta de tal maneira da média que é suspeito não ser regida pelas mesmas leis que as restantes observações [19]

outliers	plural de outlier*
smart meter	contador geral de eletricidade inteligente que permite monitorizar os consumos elétricos agregados do edifício em que está instalado
smart meters	plural de smart meters*
smart plug	dispositivo associado a um equipamento elétrico que permite monitorizar os seus consumos elétricos, consumos desagregados
smart plugs	plural de smart plug*
software	conjunto de instruções interpretáveis e executáveis por uma máquina
tag	etiqueta única identificadora
tags	plural de tag*
threshold	valor limiar

Acrónimos

ANN	Artificial Neural Network
FDD	Fault Detection and Diagnosis
FFT	Fast Fourier Transform
PAA	Piecewise Aggregate Approximation
PCA	Principal Components Analysis
RAM	Random Access Memory
SAX	Symbolic Aggregate Approximation
SOM	Self-Organizing Map
VPS	Virtual Power Solutions

Lista de Figuras

Lista de Tabelas

Conteúdo

Capítulo 1

Introdução

1.1 Contextualização

O desenvolvimento da *internet of things* permitiu que equipamentos comuns incluíssem uma tecnologia que lhes permitem receber e transmitir dados. Temos assim, equipamentos do dia-a-dia com capacidade de se conectarem à rede e permitirem a sua monitorização e controlo de forma remota.

A monitorização e controlo remoto dos equipamentos mostra-se de grande utilidade para conduzir a um aumento da eficiência energética dos edifícios e habitações. Com a informação de como é que a energia está a ser gasta, existe um maior suporte à tomada de decisões, no sentido de evitar gastos desnecessários e evitar picos na rede de distribuição de energia. Com o controlo remoto, ainda é permitido que a tomada de decisões seja feita à distância.

Nem todos os equipamentos estão atualizados com estas tecnologias. Os *smart meters** e as *smart plugs** permitem monitorizar os consumos energéticos, sendo que o último equipamento referido também permite ligar e desligar equipamentos de forma remota.

Um *smart meter** é um contador de eletricidade inteligente que regista os consumos elétricos agregados, isto é, regista o somatório do consumo elétrico total, num instante

temporal e em tempo real, não discriminando os consumos individuais dos aparelhos presentes num edifício.

Uma *smart plug** tem uma função semelhante aos *smart meters*. Este aparelho é colocado numa tomada e regista os consumos individuais do aparelho ligado, em tempo real, permitindo também o seu controlo remoto.

Os *smart meters* permitem ter o espetro geral dos consumos de um edifício, de forma não intrusiva e pouco dispendiosa, mas não disponibilizam a informação de que equipamentos estão ativos em cada instante no edifício, nem o seu consumo associado. É necessário submeter os dados do consumo geral a algoritmos de desagregação, para que se consiga ter uma estimativa de que equipamentos estão ativos, em cada instante.

Por outro lado, as *smart plugs* fazem uma monitorização individual do equipamento, mas é um método intrusivo e dispendioso.

A empresa Virtual Power Solutions (VPS)* centra o seu trabalho na eficiência energética, fornecendo poupança energética como um serviço. Isto é conseguido pelo fornecimento de serviços de monitorização remota de consumos elétricos, com recurso aos aparelhos indicados, *smart meters* e *smart plugs*, instalados no edifício que o cliente pretenda monitorizar, acedendo ao histórico de consumo armazenado na *cloud**, bem como a sua análise e sugestões de poupança, tendo em atenção as necessidades dos clientes e não comprometendo o seu conforto.

Este projeto centra-se no desenvolvimento de algoritmos de *data mining**[1] e *machine learning**[2], para descoberta de informações escondidas no histórico de dados dos clientes VPS. Pretende-se compreender como é que a energia nos seus edifícios está a ser gasta e suportar as suas decisões, conduzindo a uma maior eficiência energética.

Este projeto enquadra-se no ramo da qualidade, visto que os resultados dos algoritmos desenvolvidos no âmbito deste projeto, aplicados ao histórico de dados de consumos elétricos de edifícios, terem como objetivo conduzir a ações que privilegiam sempre a poupança energética. Fica o cliente a ganhar, pois a fatura de eletricidade sai reduzida, e o ambiente por se reduzir a eletricidade gasta, que obriga ao recurso de fontes fósseis e nucleares, pois os recursos renováveis não são suficientes para fazer face à demanda.

1.2 Motivação

A par do avanço industrial e tecnológico, tem-se verificado um crescente aumento do consumo energético nos países desenvolvidos. A energia disponível ao consumidor provem de combustíveis fósseis, reações nucleares ou de fontes renováveis. Dada a escassez das fontes fósseis e os problemas ambientais associados ao uso desta energia e da energia nuclear, é essencial o desenvolvimento de estratégias que reduzam a sua utilização, privilegiando o uso de fontes renováveis.

A energia proveniente de fontes renováveis é volátil, no sentido que a sua geração depende de fatores externos ambientais, havendo uma grande produção num certo instante de um dia mas, num outro instante a produção pode ser escassa, face as necessidades de consumo. Adicionado a falta de capacidade de armazenamento e transporte de energia, torna-se imperativo criar estratégias para aliviar os picos de procura energética, para momentos do dia em que a rede está mais disponível, isto é, que a energia disponível seja superior à procura [3].

As estratégias de redução de consumo elétrico passam pela educação do consumidor, informando-o detalhadamente dos seus consumos, para que possa reduzir os consumos desnecessários e deslocar o consumo de carga, nos picos de sobrecarga na rede, para zonas temporais em que a energia na rede seja menos procurada, fazendo-se um maior uso de energia renovável e menos de outras fontes. Um cliente que esteja informado da forma como está a usar a energia e que lhe seja dadas sugestões de poupança e de deslocação de carga, em que a alteração do seu consumo em favor da poupança económica e ambiental, não afete o seu conforto, estará à partida disponível para alterar o seu comportamento.

As falhas dos equipamentos também representam uma fonte significativa de desperdício energético. É necessária a deteção precoce de equipamentos que gerem falhas sistemáticas para se tomar as devidas medidas para evitar falhas futuras, protegendo também o equipamento de avarias e um gasto extra para o cliente.

Pretende-se também desenvolver algoritmos de desagregação de consumos elétricos,

obtidos por *smart meters*, para que a monitorização de consumos elétricos seja o menos intrusiva e dispendiosa possível para o cliente.

1.3 Objetivos

Este projeto tem como objetivo desenvolver algoritmos de *data mining* e *machine learning*, que aplicados ao histórico de dados de consumos elétricos agregados e desagregados efetuados num edifício, permitam obter informação que sustente a tomadas de decisões do cliente no sentido de um aumento da poupança energética.

Pretende-se então desenvolver algoritmos que descubram padrões de consumo dos equipamentos, rotinas de consumo e os fatores externos que estão associados aos padrões e rotinas detetadas. Também se pretende, detetar o uso simultâneo de aparelhos que possam estar associados a atividades recorrentes no edifício.

Com o conhecimento do que é um padrão de consumo do aparelho, e a rotina esperada, pretende-se detetar falhas no equipamento e fazer o seu diagnóstico.

Por último, pretende-se desenvolver algoritmos com a finalidade de identificar, em cada instante, quais os aparelhos que estiveram em atividade, tendo apenas o histórico de dados dos consumos agregados como informação disponível.

1.4 Metodologia

Para se iniciar um projeto, em que se pretende concluir conhecimento a partir de um histórico de dados, é necessário começar sempre por se estabelecer as questões de partida. Para este projeto, as questões de partida vão de encontro aos objetivos já apresentados.

Posto isto, a metodologia seguida para o desenvolvimento deste projeto, começa por analisar, de uma forma geral, a base de dados. Qual o intervalo de tempo de aquisição de dados, as unidades em que o consumo elétrico é registado, se a base de dados contem falhas (valores nulos)... Posteriormente, deve-se fazer o tratamento necessário à base de dados, caso sejam detetadas falhas.

Com a base de dados bruta tratada, deve-se passar para a discretização da base de dados e construção de *features**[4]. A necessidade de discretizar a base de dados passa pela existência de um grande volume de dados, que se traduzem numa larga despesa computacional. A necessidade de criação de *features* passa por criar variáveis novas, associadas à base de dados bruta, que acrescentem informação útil, ou de forma mais clara, para que ao aplicar os algoritmos pretendidos, estes conduzam à descoberta de informação relevante.

Após se aplicar os algoritmos à base de dados construída, os resultados devem ser visualizados na forma mais conveniente às questões de partida. Caso os resultados sejam não conclusivos, o algoritmo deve ser reavaliado bem como as questões de partida.

Para os algoritmos de *machine learning*, a base de dados deve ser dividida numa fase de treino e noutra de teste. A fase de treino serve para que o algoritmo possa aprender com o histórico de dados, e a de teste para que o conhecimento adquirido possa ser aplicado para realizar previsões. Na fase de teste é importante saber o resultado real para que se possa comparar com a previsão e, então, avaliar a exatidão do algoritmo.

Capítulo 2

Estado da Arte

2.1 Padrões, Rotinas e Associações de Consumo Elétrico

As bases de dados, onde se pretende aplicar algoritmos de *data mining* e *machine learning*, são no geral extremamente heterogéneas e de largas dimensões, o que muitas vezes ultrapassa a capacidade da Random Access Memory (RAM)*, e estrangula o disco aquando da execução dos algoritmos referidos. O artigo [5] apresenta uma *framework** que se propõe a contornar estas características das bases de dados, permitindo que um computador comum consiga processar estas grandes quantidades de dados. A abordagem é feita para bases de dados de séries temporais.

Primeiramente, para reduzir a dimensão da base de dados, segmenta-se a série temporal de dados de dimensão n em k segmentos de iguais dimensões. Cada segmento é representado pela sua média, criando-se assim uma nova variável Piecewise Aggregate Approximation (PAA)* de dimensão k.

A figura 2.1 representa a tradução de uma série temporal C, de dimensão n=60, para uma série PAA de dimensão n=6, reduzindo, assim, a dimensão dos dados em uma ordem de grandeza.

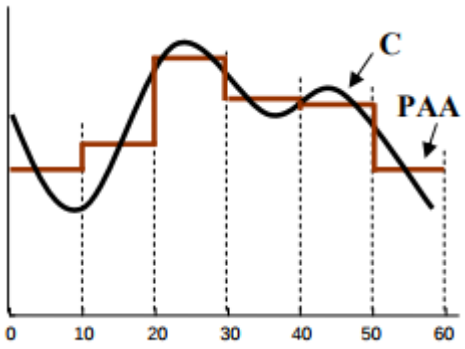


Figura 2.1: Aplicação da metodologia PAA a uma série temporal C [6].

Para reduzir a heterogeneidade dos dados e facilitar a sua manipulação, estes são representados por um baixo número de símbolos distintos, através da aplicação da metodologia Symbolic Aggregate Approximation (SAX)*. Os valores dos consumos são segmentados em patamares de consumo equiprováveis, atribuindo um símbolo único a cada patamar (com a média do patamar associado a si), diminuindo assim, o ruído da base de dados e aumentando a exatidão dos resultados da aplicação dos algoritmos. Esta representação simbólica revela-se de extrema utilidade na aplicação de algoritmos de *machine learning*, como por exemplo árvores de sufixos ou cadeias de *Markov*, e também, na descoberta de padrões de consumo aquando da aplicação de algoritmos de *data mining*.

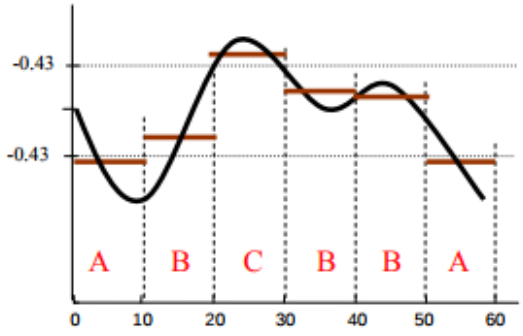


Figura 2.2: Aplicação da metodologia SAX a uma série PAA obtida a partir de C (figura 2.1) [6].

Tem-se assim, a série de símbolos apresentada na figura 2.2, tendo associado a cada símbolo o valor médio do patamar.

Para o caso em estudo, esta metodologia de redução da dimensão e heterogeneidade da base de dados não é ideal. Os dados de consumo elétrico são adquiridos em intervalos de cinco minutos ou de uma hora, e correspondem ao consumo acumulado nesse período de tempo. As análises efetuadas aos dados devem ser feitas no intervalo de tempo de aquisição mais próximo da ordem do período típico de estado ativo de um equipamento. Portanto, não é de estranhar que dois dados sucessivos representem estados do equipamento distintos. Como tal, aos segmentar os dados pela metodologia referida é incluído no mesmo segmento PAA vários estados do equipamento, perdendo-se informação relevante.

O artigo [7] apresenta uma metodologia para deteção de padrões de consumo em séries temporais de consumos elétricos. Tal como no artigo [5], o artigo [7] usa uma linguagem simbólica SAX para representar os dados, mas, aplicada à série temporal original (série C na figura 2.1), isto é, não aplica a metodologia PAA. Assim, apenas é reduzida a heterogeneidade dos dados, não é reduzida a dimensão da base de dados.

Traduzindo uma série de registos de consumos sucessivos no tempo numa série de símbolos, ambas de dimensão n, esta série de símbolos pode ser representada numa árvore de sufixos e daí concluídos os padrões de consumo (na forma de séries simbólicas) frequentes e *outliers**. A metodologia apresentada no artigo referido tem um tempo de computação da ordem da dimensão da série de símbolos, introduzidos para análise.

Por resultado da aplicação da árvore de sufixos, são determinados quais são os padrões de consumo presentes na árvore e a frequência de ocorrência.

Este algoritmo revela informação sobre os dados, de uma forma dispersa. No fim da sua execução é obtida uma lista de séries de símbolos com uma frequência associada. Mas, a título de exemplo, dois padrões com dimensões elevadas, com tamanho semelhante e com poucos símbolos distintos, podem representar o mesmo padrão. Logo, é necessário mais algum tratamento aos resultados da árvore de sufixos. Muito destes padrões são redundantes, por isso o agrupamento de padrões similares em *clusters** irá

permitir concluir informação sobre os consumos rotineiros, de forma muito mais rápida e objetiva.

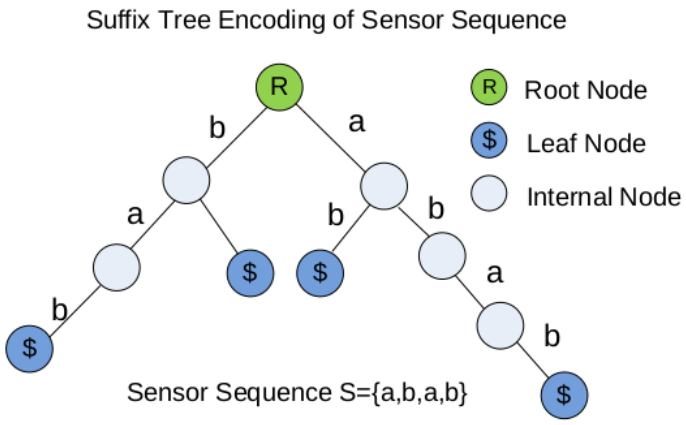


Figura 2.3: Tradução de uma série de simbolos abab numa árvore de sufixos [7].

Os padrões detetados pela árvore de sufixos são classificados de acordo com a sua dimensão (número de símbolos que constituem a série). Posto isto, os algoritmos de *clustering** [8] são aplicados a cada classe, em que dentro da mesma classe todos os padrões têm a mesma dimensão. Os padrões de cada classe são divididos por *clusters*, de acordo com o critério da distância euclidiana que separa cada par de padrões possível de formar. A distância euclidiana contabiliza três parâmetros: o módulo da diferença entre as frequências de cada padrão, o somatório do módulo das diferenças entre o valor associado a dois símbolos nas mesmas posições em padrões distintos, e o módulo da diferença do somatório do módulo das diferenças entre os valores de dois símbolos consecutivos num padrão.

Pode-se concluir que com a metodologia apresentada em [7], obtém-se padrões da base de dados agrupados, para cada grupo de padrões de igual dimensão, de acordo com a frequência de ocorrência, distância entre padrões e mudanças de estado ocorridas (mudança de símbolo ao longo do padrão).

Esta metodologia permite que toda a base de dados seja analisada. No entanto,

acaba por não ser uma metodologia muito viável para o agrupamento de padrões semelhantes. Não são comparados padrões com dimensões distintas, que podem na verdade representar o mesmo padrão, visto que um equipamento pode ter oscilações no tempo de funcionamento. O parâmetro frequência de um padrão é usado como critério de agrupamento, o que é errado, pois dois padrões muito frequentes podem ser completamente distintos, e pode haver um padrão muito frequente, muito semelhante a um pouco frequente. Para além disso, entre pares de padrões, apenas são comparados os pontos equivalentes, não é tido em conta os pontos vizinhos nas séries de símbolos.

O artigo [9] também se propõe a dividir a série de consumos por padrões, com recurso a uma representação simbólica SAX. Mas, este artigo evita o uso de árvores de sufixo para a descoberta de padrões. Começa por distinguir os consumos ON dos consumos OFF, isto é, os valores de consumos que representam um estado ativo do equipamento daqueles que não representam qualquer trabalho realizado. Tal é conseguido por aplicação de um algoritmo de *clustering* do tipo *k-means* com k=2, isto é, os dados são distribuídos por duas *clusters*, de acordo com a distância que estabelecem entre si.

O artigo referido considera que, apenas é importante o tratamento das séries de estado ON do equipamento. Negligenciando o estado OFF, consegue-se assim, ter uma redução drástica da dimensão da base de dados a analisar.

Também se aplica uma representação simbólica aos dados de consumo. Para tal, traduz-se as séries de estados ON pelos seus símbolos correspondentes, temos assim determinados os padrões de consumo de uma forma simples e rápida. Esta metodologia é chamada de modelo de pacote de palavras. Cada série de estados ativos (ON) é representada por uma palavra.

De forma semelhante ao ocorrido para o artigo [7], os padrões de consumo são sujeitos a agrupamentos. O agrupamento consiste num algoritmo de *clustering* hierárquico, que tem por base os histogramas de ocorrências de cada símbolo numa palavra para agrupar os padrões. Isto é, caso seja escolhido como fator de semelhança o símbolo com maior ocorrência numa palavra, as palavras serão agrupadas de acordo com o seu

símbolo dominante.

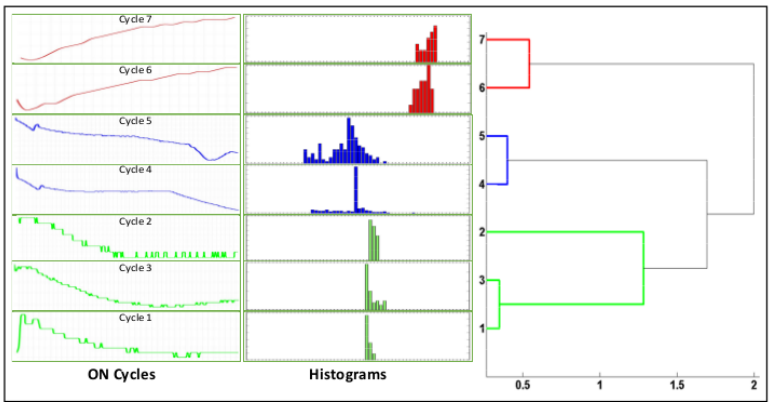


Figura 2.4: Agrupamento hierárquico dos ciclos ON do equipamento representados pelo modelo do pacote de palavras [9].

Este método, relativamente ao anterior, não analisa toda a base de dados, apenas tem em conta os ciclos ON de cada equipamento. Acaba por ser computacionalmente mais rápido e apresentar padrões mais objetivos, mais fáceis de analisar, visto se saber à partida que se trata de um ciclo de atividade do aparelho. A determinação dos valores de consumo correspondentes a um consumo ON ou OFF do equipamento é conseguida por avaliação dos dois grupos de valores que mais se correlacionam entre si. Caso o aparelho tenha um espetro muito elevado de consumos e/ou funcione por patamares, este critério de agrupamento pode levar a conclusões não verdadeiras. O agrupamento de padrões é flexível a nível da dimensão e frequência de cada padrão, pois o agrupamento dos padrões é feito com o critério de frequência de cada símbolo, numa sequência representativa do estado ON do equipamento. Tem a desvantagem de não se ter em conta a ordem com que cada símbolo se dá num padrão e inclusão de patamares não característicos do funcionamento de uma certa classe, por se apresentarem em baixa frequência na série de consumos ON.

Neste projeto é proposta uma metodologia de diminuição da dimensão da base de dados semelhante à exposta em [9], em que a gama de consumos que não representa atividade é identificada de forma direta, não dando espaço para erros na identificação do

estado ON do equipamento. A heterogeneidade da base de dados é diminuida por aplicação da metodologia SAX, com a diferença de que o número de patamares de consumo não é definido à *priori*. O algoritmo de identificação dos patamares, em que se divide a gama de valores de consumo dos equipamentos é não supervisionada. O número de patamares de um equipamento não é igual para todos os equipamentos, e é difícil de saber à *priori*. Esta metodologia contorna essa limitação dos métodos apresentados em [7] e [9].

Quanto ao agrupamento dos padrões de consumo, o projeto apresenta uma metodologia em que todos os padrões são comparados entre si e são agrupados de acordo com a sua semelhança. É tida em conta a sequência de símbolos para a comparação e é contabilizado também a vizinhança.

Para além de determinar os padrões de consumo, importa também traçar rotinas de consumo e perceber quais os fatores que as influenciam.

O artigo [10] propõe uma metodologia para descoberta de rotinas de consumo. Esta metodologia passa por, a cada dia presente na base de dados, associar um vetor de *features*, como por exemplo, a energia média consumida num dia ou os picos de consumo a cada hora. Estes vetores de *features*, para que sejam eliminados efeitos sazonais ao vetor em questão, deve ser subtraída a média dos sete vetores vizinhos, incluindo a si próprio. Seguidamente, os vetores das *features* correspondentes aos dias da base de dados, são distribuídos por sete classes, cada uma correspondente a cada dia da semana. Dentro de cada classe são identificados e removidos os *outliers*. Para terminar, as classes com dados redundantes são combinadas e, assim, é descoberto os dias tipo relevantes.

A metodologia apresentada para descoberta de padrões é computacionalmente rápida, mas acaba por descartar as tendências sazonais da equação. Apenas, identifica rotinas semanais. É um sistema fixo, no sentido de que as relações determinadas entre os dias da semana são válidas, por generalização, para toda a base de dados. Isto é, se tendencialmente se verificar dois dias tipo, numa classe se encontre os dias úteis e na restante os fins-de-semana, vai se generalizar uma rotina fixa que distingui os dias úteis dos fins-de-semana. Para além disso, este algoritmo resume a comparação entre as

classes representativas de cada dia da semana à comparação de um conjunto de *features* e não às séries de consumos no tempo verificadas para cada dia. Esta abordagem é rápida, mas é muito generalista.

O artigo [11], para determinar o período em que se deve segmentar os dados, recorre à Fast Fourier Transform (FFT)*, para calcular o período predominante na série temporal de consumos. Com o período determinado, a base de dados é segmentada, de acordo com esse critério e analisados os segmentos de consumo entre si.

Os segmentos de dados são submetidos à análise das suas componentes principais (Principal Components Analysis (PCA)*), que traduz os segmentos de dados nos seus vetores próprios. O cálculo das componentes principais de cada vetor permite reduzir a dimensão dos mesmos, por determinação das suas componentes ortogonais e independentes, que traduzem a variação dentro de cada vetor, criando um vetor próprio associado.

Num segmento de consumos, um consumo a uma dada hora está relacionado com o consumo da hora anterior e da seguinte. Portanto, é de esperar que exista uma autocorrelação entre os valores registados para cada segmento. O valor próprio associado a cada vetor próprio é um medidor da autocorrelação do segmento. Visto o objetivo ser determinar rotinas, que se mantenham durante todo o ano, independentemente da amplitude dos picos de energia que podem variar de acordo com a altura do ano, são excluídos todos os vetores próprios com valores próprios inferior a um, como sugerido pela *Kaiser rule*.

Aos vetores próprios com maior autocorrelação identificados é aplicada uma rotação, *varimax rotation*, para que sejam evidenciados os instantes de consumo mais relevantes para cada vetor próprio.

Posto isto, os valores próprios que sofreram rotação são sujeitos a um algoritmos de *clustering* que agrupa os dados em dias tipo, não tendo em conta sazonalidades. Visto não serem analisados todos os dados, a tendência de distribuição dos dias da semana pelas *clusters* é generalizada para o resto da base de dados. Portanto, esta metodologia, apesar de, ao contrário da anterior, usar toda a informação relevante de um dia,

também impõe uma rotina fixa e não tem em conta sazonalidades. O interessante seria classificar cada dia da base de dados, de forma individual, com um dia tipo, considerando sazonalidades e usando o máximo de informação.

O artigo [12] recorre a Self-Organizing Map (SOM)*, que é um algoritmo de Artificial Neural Network (ANN)* não supervisionado de *clustering*, para identificação de rotinas numa empresa. Neste algoritmo, os neurónios são organizados numa arquitetura de dois níveis.

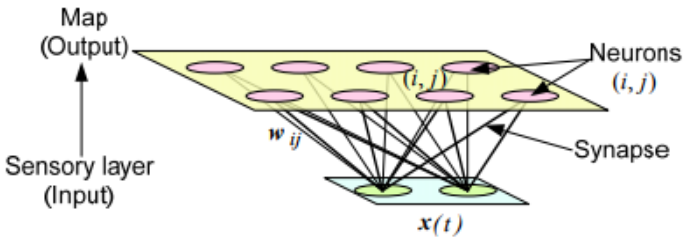


Figura 2.5: Arquitetura SOM [12].

O primeiro nível é constituído por neurónios em igual número ao número de variáveis, dos dados a analisar. O segundo nível tem tantos neurónios quantos os definidos. A arquitetura deste algoritmo é apresentada na figura 2.5. A ligação entre o primeiro e segundo nível é feita por sinapses. Para se treinar a rede neuronal, as sinapses (com pesos associados) têm de ser inicializadas. No caso exposto no artigo, os dados de entrada são uma série de 24 valores de consumo acumulado, registados a cada hora num dia, o dia da semana, o mês e se trata de um dia útil ou não. Tem-se assim, 27 variáveis de entrada para o algoritmo SOM, significando que existem 27 neurónios no primeiro nível.

A cada entrada de dados é verificado o peso da sinapse mais similar com o *input**. Este peso tem a indicação das coordenadas do neurónio pertencente à segunda camada, para onde se deve encaminhar os dados de entrada. Posto isto, o peso correspondente a esta entrada é recalculado para que a sua semelhança com estes dados de entrada seja

superior, e, numa futura entrada de dados equivalentes, a correlação entre o peso e os dados de entrada seja ainda mais forte. Os pesos dos neurónios vizinhos também são atualizados recorrendo a uma função gaussiana.

Assim, o algoritmo SOM ao identificar o peso mais adequado para cada entrada de dados e ao encaminhar os dados para o neurónio correspondente a esse peso, agrupa os dias em *clusters*. Tem-se assim, um algoritmo de redes neuronais que usa grandes quantidades de dados para detetar rotinas de consumo, sem impor uma rotina fixa e incluindo sazonalidade.

Para treinar um algoritmo de redes neuronais, como é o caso do algoritmo SOM, é necessária uma grande quantidade de dados. No presente projeto, apenas são analisados dados de nove meses, o que é insuficiente para detetar rotinas sazonais.

A abordagem aos dados neste projeto passa pela classificação dos dias tipo, não desprezando as sazonalidades, num algoritmo de *clustering* hierárquico que tem como critério de agrupamento a semelhança entre os perfis ON e OFF, associados às séries temporais de cada um dos dias da base de dados.

Para entender o comportamento de consumo é importante também classificar os consumos de acordo com as atividades desempenhadas e perceber o contexto de utilização dos equipamentos. Para o efeito, a abordagem passa por aplicar algoritmos de associação aos consumos de um edifício, que identifiquem valores de variáveis distintas que ocorrem frequentemente em simultâneo, por exemplo, um fogão e um forno, ambos elétricos, às 20 horas, que corresponde provavelmente numa habitação à atividade de fazer o jantar. O algoritmo de associação deve ser sensível à frequência dos valores que as variáveis tomam, isto é, um dispositivo que tenha um estado ativo de forma constante no tempo não irá ter nenhuma relação causal com um outro dispositivo, que funcione em curtos períodos de tempo, mas em simultâneo com o primeiro. Por exemplo, um frigorífico tem um estado ativo muito frequente mas, o fogão não. É provável que quando o fogão funcione, o frigorífico também funcione, mas é uma consequência do constante estado ativo do equipamento, não existe nenhuma relação causal.

O artigo [13], após os dados serem processados e se ter calculado todas as variáveis

relevantes associadas à série temporal de consumos (*features*), aplica um algoritmo que usa a métrica *J-Measure* com um valor *threshold** para identificar as associações relevantes entre valores de variáveis distintas. Uma associação é definida por um par X $\rightarrow$ Y, em que X é o antecessor e Y o sucessor. O antecessor e um sucessor podem ser compostos por um ou mais valores. Este par não tem uma relação causa $\rightarrow$ efeito, mas sim, que o sucessor se verifica quando o antecessor se verifica, e não necessariamente o contrário.

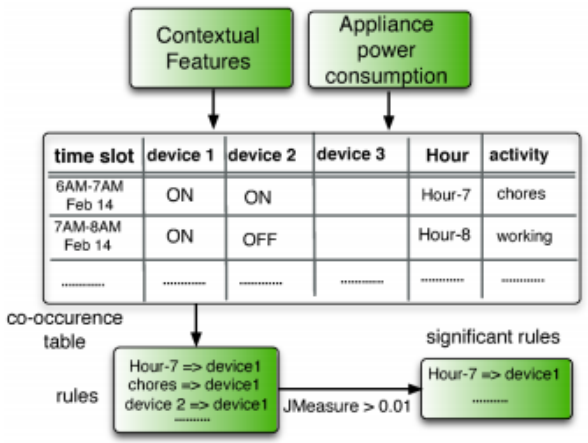


Figura 2.6: Associações de valores de variáveis distintas [13].

A figura 2.6 representa a identificação de um conjunto de associações, e a sua identificação com uma atividade. As associações relevantes são determinadas pelo cálculo da métrica *J-Measure* (2.1) para todos os pares de associações possíveis, e identificadas aquelas que têm um valor para esta métrica superior ao valor *threshold* definido.

$$(X \rightarrow Y) = P(X \cap Y) * log(\frac{P(Y|X)}{P(Y)}) + P(X \cap \bar{Y}) * log(\frac{P(\bar{Y}|X)}{P(\bar{Y})}) \qquad (2.1)$$

A métrica *J-Measure* tem em conta a direção do par antecessor $\rightarrow$ sucessor, e avalia se realmente existe uma relação de associação entre o par, a sua dependência, e não de que a associação é devida à elevada frequência de ocorrência de um dos elementos do par.

A metodologia apresentada neste projeto permite usar métricas de cálculo de associação, que avaliam a frequência de ocorrência do par, a frequência de ocorrência individual, e o nível de dependência de forma individual. Em (2.1) todos estes parâmetros são avaliados em conjunto.

2.2 Deteção e Diagnóstico de Falhas

O objetivo dos algoritmos Fault Detection and Diagnosis (FDD)* é de detetar e diagnosticar falhas dos equipamentos para evitar avarias e consumos desnecessários.

O artigo [14] descreve de uma forma geral as metodologias para deteção e diagnóstico de falhas, partindo da análise de dados.

Quando se tem uma série temporal de dados e uma associação entre estes dados e conhecimento, por confronto torna-se possível a averiguação se a série se encontra dentro das normas. Caso o sinal seja consistente com as normas, o processo que os dados representam ocorreu sem falhas.

O processo analítico FDD consiste no processamento de séries temporais para encontrar correlações entre estes dados e o conhecimento que se possui acerca das falhas. Este conhecimento pode ser obtido de duas formas, por um modelo que padronize um sinal sem falhas ou conhecimento implícito.

As metodologias FDD são então baseados em séries de dados, e conduzidas por modelos ou conhecimento implícito, dividindo-se em três categorias:

1. Análise de dados *online* com base em modelos;
2. Métodos baseados em sinais;
3. Métodos baseados no histórico de dados;

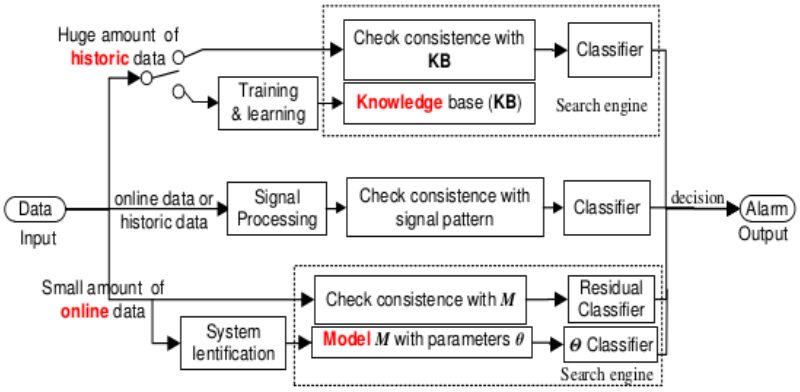


Figura 2.7: Esquema discritivo das metodologias FDD [14].

A primeira metodologia (1.) é apresentada no último nível da figura 2.7. Recorre-se a ela quando se tem disponível uma quantidade de dados muito baixa, apenas os dados obtidos no instante em análise e poucos dados antecessores. Existe um modelo M, pré-definido com base no que é o comportamento padrão do equipamento. Os dados de entrada são comparados com a previsão dos mesmos, efetuada pelo modelo M, e gerados resíduos θ. Os resíduos são classificados para se averiguar se existe falha ou não, caso sim é identificada o seu tipo. Para que o modelo seja bom, este deve saber distinguir falhas de desvios do comportamento padrão não correspondentes a falhas.

Os métodos baseados em sinais para FDD (2.), estão representados no nível intermédio da figura 2.7. Este método baseia-se, apenas, nos dados disponíveis para perceber quando é que ocorrem falhas. Os padrões nos dados são detetados e os desvios são classificados de acordo com a sua causa. Estes desvios correspondem a falhas do sistema. Assim sendo, quando novos dados entram no sistema são confrontados com os padrões e desvios já classificados. Se for caso disso, é detetada falha (caso se correlacione mais com uma classe de desvios) e a sua origem.

Quando existe uma grande quantidade de dados (3.), e demasiado heterogéneos e complexos para serem modelados analiticamente sem ambiguidades, o diagnóstico e deteção de falhas deve ser feito com recurso a algoritmos de inteligência artificial. Este caso está representado na figura 2.7 pelo nível mais alto. Pela aplicação de algoritmos

de *machine learning* à base de dados, constroem-se modelos de padrões e falhas que constituem a base de conhecimento para posterior classificação de dados de entrada.

O artigo [15] apresenta uma metodologia para deteção, diagnóstico e previsão de falhas que se baseia no histórico de dados. Utiliza um algoritmo de *machine learning* para identificar e prever falhas em dados *online*. O algoritmo a que o artigo recorre são as cadeias de *Markov* escondidas. Este algoritmo usa o histórico de dados, em que se possui uma série de dados no tempo (estados observáveis) e a informação associada, nomeadamente, se a cada instante existe falha ou não (estados escondidos), e caso exista falha, que tipo se trata. Com esta informação, a cadeia de *Markov* com estados escondidos associados (há falha ou não) é treinada. Concluída a fase de treino, obtém-se a caracterização desta cadeia pelas probabilidades de transição, emissão e de estados iniciais. As cadeias de *Markov* escondidas são descritas em detalhe na secção 3.8.1. Assim sendo, com a cadeia caracterizada pelo histórico de dados, para novos dados é determinado o estado escondido mais provável, associado.

São treinadas tantas cadeias de *Markov* quanto as falhas que podem ocorrer mais o estado normal. Assim, com as cadeias treinadas, novos dados são submetidos às cadeias treinadas e calculados os estados escondidos mais prováveis para cada cadeia, isto é, cada cadeia (associada a uma falha ou estado normal) verifica se ocorre o estado escondido que esta pretende detetar, ou não, com uma probabilidade associada.

Cada falha tem associado a si uma série de estados de degradação que conduz a ela. Então, as cadeias de *Markov* escondidas associadas a cada falha, para além de identificarem estas falhas como estados escondidos, também devem identificam se o sistema se encontra em alguma das fases de degradação que leva à ocorrência de falhas.

Os dados *online* são submetidos às previsões de todas as cadeias. Todas elas apontam um estado mais provável associado, com uma probabilidade associada à sua ocorrência. É escolhido o estado, entre todos os estados previstos pelas diferentes cadeias, que tem uma probabilidade mais alta.

A previsão de falhas acaba por ter uma importância superior à deteção. É preferível ter a informação de que a falha está para ocorrer e esta ser prevenida do que, apenas a

detetar quando esta já se deu.

Este algoritmo mistura a abordagem (1.) com a (3.), usa uma grande quantidade de dados para treinar um algoritmo de *machine learning*, mas também faz uso de conhecimento externo, neste caso, não para modelar os dados, mas sim para identificar os instantes em que ocorreram as falhas, e de que tipo. Esta acaba por ser a abordagem ideal, visto que, para além de detetar e classificar falhas com bastante precisão, ainda atua na sua previsão.

No entanto, não é possível aplicar esta última abordagem para diagnóstico e deteção de falhas neste projeto, visto não existir informação associada aos dados, isto é, informação que padronize o comportamento normal dos equipamentos, possíveis falhas que podem ocorrer ou identificação dos momentos em que as falhas ocorrem. A juntar a estes fatores, está a frequência de aquisição de dados, que é baixa (de 5 em 5 minutos ou de hora em hora). A frequência baixa acaba por poder camuflar falhas.

Neste projeto é utilizada a metodologia (2.). Visto não se ter informação adicional dos dados e se trabalhar com frequências tão baixas, acaba-se por detetar consumos anormais e não falhas propriamente ditas.

2.3 Desagregação de Consumos Elétricos Agregados

A informação de todos os consumos de um edifício é muito importante para se entender os hábitos de consumo, para traçar planos de poupança minuciosos, sem que se comprometa o conforto e necessidades dos seus ocupantes.

Existem duas formas de se conseguir monitorizar todos os equipamentos. A primeira passa por monitorizar todos os equipamentos de forma individual, com recurso a *smart plugs*, obtendo assim os consumos desagregados de cada equipamento. A segunda hipótese passa por monitorizar todos os consumos de forma agregada, com um *smart meter*, e posterior desagregação, com recurso a algoritmos desenvolvidos para esse fim. A primeira hipótese apresenta uma informação exata sobre o consumo de cada equipamento mas é extremamente intrusiva e dispendiosa, visto exigir a presença

de um equipamento de monitorização em todos os aparelhos. Já os *smart meters* apenas exigem o equipamento indicado, abrangendo todos os consumos da habitação. No entanto, é necessário definir estratégias e algoritmos para que se consiga identificar os instantes de uso de cada um dos equipamentos presentes numa habitação, e o respetivo consumo associado a cada registo.

Existem algumas lacunas que os algoritmos de desagregação de consumos elétricos devem ultrapassar. A começar pela correta identificação do(s) equipamento(s) em cada instante, uso simultâneo de dois equipamentos, equipamentos distintos com uma assinatura de consumo bastante similar, ruído de fundo causado por flutuações no consumo de equipamentos de uso continuado, e uso de aparelhos de baixo consumo que se confundem com o ruído de fundo.

O artigo [16] apresenta uma metodologia para desagregação dos consumos elétricos agregados, com uso de assinaturas dos equipamentos presentes numa habitação. As assinaturas são registadas numa monitorização individual dos equipamentos, num período de teste de uma semana. São também registadas as mudanças de estado associadas aos registos de consumo elétrico, que corresponde à diferença entre dois consumos consecutivos. As mudanças de estado associadas aos consumos agregados são posteriormente comparados com as características dos equipamentos identificadas, na fase de registo (treino) dos consumos individuais de cada aparelho.

A desagregação é iniciada pela identificação do início das séries de estados ON ou OFF, nos consumos agregados, possíveis de representarem o estado do equipamento a desagregar. A deteção do início das séries referidas, consiste em identificar as mudanças abruptas de estado nos consumos agregados e comparar com o registado na fase de treino.

Com o início das séries de estados ON ou OFF identificadas para o equipamento, é calculada a norma deste vetor e comparada com a referência, devidamente normalizadas. Também é comparada a duração das séries identificas com as referências.

Posto isto, a cada comparação dos parâmetros referidos com as referências, são avaliadas e determinadas quais são as séries de estados nos consumos agregados que

correspondem a uma série de estados correspondentes (ON/OFF) no equipamento.

Depois da determinação das séries de estados ON ou OFF, com maior aproximação com a referência, ainda é avaliada e comparado com a referência a correlação entre os trinta minutos antecessores e sucessores da série, e o intervalo mínimo entre o início da série de estados ON/OFF com a de estados OFF/ON sucessora. Por fim é avaliado, durante a série estimada como ON para o equipamento, nos consumos agregados, o consumo mínimo dessa série, que tem de ser superior ao consumo de fundo do equipamento na fase de testes.

Esta metodologia assume da fase de treino que não há intersecção entre o uso deste equipamento e outros presentes no edifício. O que não é praticável em casos reais. O algoritmo deve ser tolerante a mudanças de consumo elétrico nos consumos agregados, não provocadas pelo aparelho em análise, durante todo o período de funcionamento ativo do mesmo. Também é usado como referência o tempo comum de uso de um aparelho e as horas de frequente utilização. Por um lado ajuda a distinguir equipamentos com assinaturas semelhantes, mas por outro, o aparelho pode ter um consumo aleatório ou um funcionamento em intervalos pouco padronizados, portanto, numa amostra de uma semana é difícil estabelecer padrões do que será mais comum para aparelhos de funcionamento mais aleatório. Esta abordagem seria interessante desde que a fase de treino fosse mais longa no tempo.

O artigo [17] apresenta uma metodologia que recorre a cadeias escondidas de *Markov*, e a informações sobre os equipamentos presentes num edifício para desagregar os consumos registados por *smart meters*.

Para a cadeia de *Markov* escondida, os estados observáveis correspondem às mudanças de estado registadas pelo *smart meter*, isto é, à diferença entre dois registos de consumos agregados consecutivos. Os estados escondidos associados correspondem ao estado do equipamento a cada instante que se pretende desagregar. Estes estados podem tomar três valores, ON, que corresponde ao estado ativo do equipamento, OFF, ao estado inativo e pico que corresponde aos registos de consumo máximo nas séries de atividade do equipamento.

Uma cadeia de *Markov* escondida precisa de se submeter a uma fase de treino para se determinar os parâmetros que a caracterizam. Segundo a metodologia proposta, durante a fase de treino, apenas são registados consumos agregados em que o único equipamento que liga e desliga é o equipamento para que se pretende desagregar os consumos elétricos. Os parâmetros que caracterizam a cadeia de *Markov* são a probabilidade de cada um dos valores de estados escondidos ocorrerem no início da cadeia de *Markov*, a probabilidade de transição que indica qual a probabilidade de transição entre todas as combinações possíveis de dois estados escondidos, incluindo a manutenção de estado, esta probabilidade é proporcional ao tempo despendido em cada estado e ao intervalo entre estados distintos, e a probabilidade de emissão que caracteriza a probabilidade de um estado observável ter associado a si um determinado estado escondido.

Os ciclos de atividade de cada equipamento são padronizados e traduzidos por leis normais, com uma média (pico) e um desvio padrão associados. A probabilidade de emissão é função da soma de um com a diferença do consumo agregado num dado instante e o pico de consumo, normalizado ao desvio padrão, para cada modo de funcionamento do equipamento. Tendo este treino efetuado, é possível determinar os estados escondidos para este equipamento, associados aos estados observáveis, e posteriormente subtrair a série de consumos estimados para este equipamento aos consumos agregados, para uma mais fácil desagregação dos restantes consumos.

Esta metodologia acaba por ser mais simples de programar que a anterior, pois apenas depende dos consumos em estado ativo típicos, traduzidos numa lei normal, associados a estados observáveis nos consumos agregados. É avaliada a duração dos consumos ativos e o período decorrido entre estados ativos consecutivos. Quanto à sobreposição de consumos, o algoritmos não atua nesse aspeto.

Neste projeto, para desagregação de consumos elétricos registados por *smart meters*, recorre-se a duas abordagens baseadas nos artigos indicados. Uma primeira abordagem faz uso da assinatura do equipamento, de uma única amostra de mudanças de estado registadas nos consumos agregados quando o aparelho que se pretende identificar estava a realizar trabalho. Esta abordagem serve apenas para ver qual a exatidão das previ-

sões obtidas, com o uso da menor informação possível. Não há garantia que, aquando da tirada da assinatura, não estavam a trabalhar outros equipamentos em paralelo, e, aquando da sua identificação, o algoritmo é bastante sensível a interferências e variações na duração do consumo.

A segunda abordagem apresentada neste projeto utiliza cadeias de *Markov* escondidas para a desagregação de carga. Mas, ao contrário do artigo [17], a metodologia apresentada não padroniza os consumos do aparelho que se pretende identificar numa lei normal. A cadeia é treinada, isto é, calculam-se os parâmetros que a caracterizam, numa fase de teste em que se conhece as mudanças de estado dos consumos agregados e se sabe todos os instantes em que o aparelho a identificar esteve ativo. Esta abordagem não categoriza os consumos típicos do aparelho e é flexível a funcionamentos em paralelo de outros equipamentos, desde que esse funcionamento em paralelo seja recorrente.

Capítulo 3

Algoritmos de Análise e Desagregação de Consumos Elétricos

Para este trabalho, os dados disponíveis para análise são dados de consumos elétricos desagregados do *open-space* da empresa VPS, empresa piloto, obtidos por *smart plugs* ligadas a vários equipamentos. E dados agregados e desagregados de uma habitação, casa piloto, obtidos por um *smart meter* e *smart plugs* ligadas a alguns eletrodomésticos, respetivamente.

Para se poder atingir os objetivos propostos em 1.3, é necessário aplicar uma série de algoritmos de *data mining* e *machine learning* a dados de consumos elétricos agregados e desagregados.

Este capítulo pretende descrever o processo de aquisição e preparação dos dados de consumo elétrico e os algoritmos a aplicar-lhes para que os objetivos apresentados em 1.3 sejam atingidos.

3.1 Aquisição e Registo de Dados

Os dados das *smart plugs* e *smart meters*, pertencentes aos clientes da VPS, estão armazenados na *cloud*. O acesso à *cloud* é feito através de uma API, *ienergy*, sendo que, cada cliente tem associado a si um nome de utilizador e uma palavra-passe para que a

sua privacidade seja assegurada. Por sua vez, o consumo associado a cada *smart plug* e/ou *smart meter* instalado corresponde a uma *tag**. As *tags** associadas a cada *smart plug* ou *smart meter* permitem ter acesso ao registo do consumo elétrico armazenado na *cloud*, com o *software** adequado, indicando o nome de utilizador, palavra-passe, *tag* da/do *smart plug/smart meter* e janela temporal em que se pretende receber os dados.

Cada evento $e_{t,d}$, consumo registado pela/pelo *smart plug/smart meter* para o instante t={hora,minuto}, no dia d={dia,mês,ano}, corresponde ao somatório de toda a energia (kWh), consumida pelo equipamento associado nos cinco minutos passados ou hora passada ao instante indicado, $energia_{(t,t-1),d}$, seja o modo de aquisição indicado no pedido à API base ou horário, respetivamente.

$$e_{t,d} = energia_{(t,t-1),d} \tag{3.1}$$

Para ter um espetro diversificado de dados, o pedido à API é feito numa janela temporal de um ano, para cada *smart plug* ou *smart meter* que se pretende analisar. Para *smart plugs* e *smart meters* presentes no mesmo edifício, deve-se ter em atenção que os instantes de aquisição devem ser iguais para todos os dispositivos. Estes dados são guardados numa base de dados relacional para posterior análise.

As bases de dados construídas para cada *smart plug/smart meter* devem ter identificadas os dispositivo que monitorizam, bem como, para cada evento registado, a data, hora, o local de registo, a temperatura máxima registada nesse dia, humidade e precipitação.

A informação diária do estado do tempo é obtida com recurso a uma segunda API, *Weather Underground*, disponível gratuitamente [18].

Tendo a base de dados dos consumos elétricos registada para todos os dispositivos de monitorização, por um período de um ano, deve-se proceder ao tratamento dos dados brutos para futura análise.

3.2 Tratamento de Dados

Antes de iniciar a análise dos dados propriamente dita, é necessário eliminar possíveis incongruências.

Deste modo, deve começar-se por verificar se existem registos de dados em duplicado, caso existam, eliminar-se a informação redundante.

Não é feito um tratamento de dados de consumo *outliers*. Para o caso em estudo não interessa a sua eliminação, mas sim a sua identificação e análise.

De seguida, é necessário apurar se algum dos eventos pedidos à API não se encontra registado na base de dados. São apurados dados nulos em instantes que, por algum motivo o registo não foi efetuado, por falha de comunicação entre a/o *smart plug/smart meter* e o *hub** ou falha na medição do consumo elétrico por parte da/do *smart plug/smart meter*.

O processo de procura e tratamento de nulos é de extrema importância. A existência de nulos não identificados, numa posterior análise de dados, gera incongruências nos padrões detetados e erros de código, na tentativa de acesso a esses eventos que era pressuposto estarem na base de dados.

Para que os dados nulos sejam identificados, em toda a janela temporal, sabendo o intervalo de tempo de aquisição, base ou horário, tentado o acesso a todos os dados, os instantes temporais em que não são registados quaisquer eventos são adicionados a uma tabela de nulos, com a informação de data, hora e minuto associados ao registo em falta.

Para que não existam problemas com a existência de dados nulos, são adicionados à base de dados previsões para os dados em falta. Tais previsões são conseguidas com recurso ao histórico de dados registados, para o mesmo mês em que se registou a falha. São analisados apenas dados do mesmo mês para garantir que os consumos são feitos em condições semelhantes ao da ocorrência do nulo. Por exemplo, não faz sentido prever dados para um aquecedor elétrico, para o mês de Dezembro, usando dados do consumo do mês de Agosto.

Existem dois tipos de nulos possíveis, associados aos dois tipos de falhas indicadas. No primeiro tipo de nulos, associado a falhas de comunicação entre o dispositivo de monitorização e o *hub*, o dado seguinte à ocorrência de um nulo ou série de nulos, corresponde ao somatório dos consumos desde o instante em que foi registado o último evento. Isto é, caso dois eventos sucessivos, presentes na base de dados, distem de três horas, o segundo evento corresponde ao somatório da energia consumida nesse intervalo de tempo.

Menos frequente é o segundo tipo de erros, falhas de medição, em que o evento seguinte a uma falha corresponde apenas ao somatório dos consumos entre o instante em causa e o último suposto de existir (cinco minutos ou uma hora, de acordo com o modo de aquisição) mas em falta. Este tipo de falhas ocorre quando há falhas na medição.

O tipo de falhas nos dados de uma/um *smart plug*/*smart meter* é identificado por análise dos registos às série de nulos. Caso este registo de consumo seja um *outlier** [19] (em comparação com a vizinhança) significa que estamos perante o primeiro tipo de falhas indicadas, caso não seja, estamos perante o segundo.

Para simplificação da exposição dos algoritmos da previsão de nulos é considerado que cada nulo, ou série de nulos, começa e acaba no mesmo dia d, ou seja, uma série de nulos não inclui dias distintos.

3.2.1 Previsão de Nulos: Falha de Comunicação

Para o caso dos nulos em que o evento registado imediatamente a seguir a um nulo, ou série de nulos, corresponde ao somatório da energia consumida nesse intervalo, esse consumo deve ser registado, bem como o intervalo de tempo em que ocorreu a falha, incluindo o instante do evento seguinte. Este caso verifica-se quando existem falhas de comunicação. O dia em que ocorre o(s) nulo(s) em análise é representado por d_n.

Para se prever os nulos é necessário construir as seguintes variáveis:

- Série de instantes em que se registaram o(s) nulo(s) em análise $[t_0, ..., t_{j-1}]$ mais o instante sucessor (t_j):

$$time = [t_0, ..., t_{j-1}, t_j] \tag{3.2}$$

- Consumo total efetuado para os instantes da série (3.2) num dia d:

$$area_d = \sum_{i=0}^{j} energia_{(t_i, t_i - 1), d} \tag{3.3}$$

- Série de dias, com o mesmo dia da semana que o dia em que ocorreu o nulo ou série de nulos (feriados são vistos como domingos):

$$days = [d_0, ..., d_i] \tag{3.4}$$

- Séries de eventos (3.1), registadas para cada dia da série (3.4). Cada posição corresponde a uma série, para os instantes (3.2) no dia em questão, dos eventos (3.1) correspondentes:

$$events = [e_{d_0}, ..., e_{d_i}] \tag{3.5}$$

Para que a previsão seja possível, no mês em que ocorreu a falha, usam-se os mesmos dias da semana (3.4) para o qual se pretende prever os dados em falta.

Para todas as séries de eventos adquiridas para comparação (3.5), é analisado o somatório da energia e registado, para cada instante, a percentagem, desse somatório, consumida a cada instante (3.2). Ou seja, para cada dia de comparação (3.4), é calculado e registado o somatório de toda a energia consumida (3.3), e a cada instante (3.2), a percentagem sobre esse consumo total, com a hora e o minuto associado. Caso exista mais do que uma série para comparação (todas as séries sem dados nulos presentes são elegíveis), é escolhida a mediana das séries.

A mediana é calculada em casos em que existam três ou mais séries elegíveis para previsão. Para cada série é calculada a diferença, em cada ponto no tempo, estabelecida com as restantes séries, no ponto equivalente. Para cada série é calculada uma

nova série associada, a cada ponto desta série associada corresponde o somatório das diferenças, para os mesmos índices, que a série em análise estabelece com as restantes. A mediana corresponde à série que tiver uma série de diferenças associada com menor somatório.

No caso de apenas existirem duas séries, é escolhida aquela que tem uma melhor distribuição dos consumos, ou seja, aquela que tiver menor máximo. Com isto se reduz a influência de *outliers* na previsão. É preferível não prever um pico de consumo, prevendo uma série de eventos mais uniformes, a prever um pico de consumo que não existe, conduzindo a resultados não tão fiáveis.

No caso da existência de apenas uma série elegível para comparação, é essa que é usada para previsão.

Menos provável de acontecer são os casos em que não existe nenhuma série elegível. Caso ocorra, os dados nulos nas séries para comparação devem ser substituídos por consumos de 0 kWh. O restante procedimento para este caso é o referido anteriormente.

Por fim, as séries de percentagens obtidas são multiplicadas pelo valor de consumos agregados, ocorridos a seguir à falha registada (3.3) a dividir por cem. Assim, se obtém uma previsão para a série de valores em falta.

Algoritmo 3.1 Previsão de Nulos: Falha de Comunicação

events ← *(3.5)*
days ← *(3.4)*
nullday ← *dia* d_n *em que se registou o(s) nulo(s) para previsão;*
nullarea ← *(3.3) para o nullday;*
function PREDICT(events, days, nullarea)
    $listpercents = list()$ ← *lista vazia;*
    for d in days **do**:
        $events_d = events[days.index(d)]$ ← *série de eventos, nos instantes (3.2), para o dia d;*
        $area_d = sum(events_d)$ ← *(3.3), soma da série de eventos, nos instantes (3.2), para o dia d;*
        $percent = \frac{events_d}{area_d} * 100$ ← *série de percentagens elegível para previsão;*
        $listpercents.append(percent)$
    if $len(listpercents) > 2$ **then**
        (cálculo da mediana da listpercents);
        $diferences = list()$
        for p in range(0, len(listpercents)) **do**:
            $d = 0$
            for k in range(0, p) **do**:
                $d = d + sum(abs(listpercents[p] - listpercents[k]))$
            for k in range(p+1, len(listpercents)) **do**:
                $d = d + sum(abs(listpercents[p] - listpercents[k]))$
            $diferences.append(d)$
        $finalpercent = listpercents[diferences.index(min(diferences))]$
        $data = \frac{nullarea * finalpercent}{100}$
    else if $len(listpercents) == 2$ **then**
        (cálculo da série da listpercents com menor máximo);
        $maxitem = [max(listpercents[0]), max(listpercents[1]]$
        $finalpercent = listpercents[maxitem.index(max(maxitem))]$
        $data = \frac{nullarea * finalpercent}{100}$
    else
        $finalpercent = listpercents[0]$
        $data = \frac{nullarea * finalpercent}{100}$
return $data = [e_{t_0,d_n}, ..., e_{t_j,d_n}]$ ← *resultado da previsão;*

O algoritmo 3.1 esboça de forma sucinta o funcionamento da previsão de nulos, feita para as falhas de comunicação. A série de dados que é devolvida representa a previsão de dados para os instantes em falta (3.2), no dia em que se registou a falha de comunicação, d_n.

3.2.2 Previsão de Nulos: Falha de Medição

Em casos, em que o evento seguinte ao nulo e_{t_j,d_n} (3.1), ou série de nulos registados, não corresponde ao somatório total dos consumos dos instantes em falta, o procedimento para previsão difere do anterior exposto (3.3.1). Este caso verifica-se quando existem falhas de medição.

São também selecionados os dias da semana, do restante mês, iguais (3.4). Contudo, as variáveis (3.2) e (3.3) são reformuladas e introduzidas novas variáveis:

- Série de instantes em que se registaram os nulos em análise $[t_0, ..., t_j]$, com exclusão do instante posterior t_j:

$$time = [t_0, ..., t_{j-1}] \tag{3.6}$$

- Consumo total efetuado para os instantes da série (3.6) num dia d:

$$area_d = \sum_{i=0}^{j} energia_{(t_i,t_i-1),d} \tag{3.7}$$

- Série de instantes de um dia, com exclusão dos instantes da série (3.6):

$$day_{time} = [t_0, ..., t_k] \tag{3.8}$$

- Consumo total efetuado para os instantes da série (3.8) num dia d :

$$day_{area,d} = \sum_{i=0}^{k} energia_{(t_i,t_i-1),d} \tag{3.9}$$

- Séries de eventos (3.1), registados para cada dia da série (3.4). Cada posição corresponde a uma série, para os instantes (3.6) no dia em questão, dos eventos (3.1) correspondentes:

$$events = [e_{d_0}, ..., e_{d_i}] \tag{3.10}$$

Para os dias usados como referência para prever os nulos (3.4), em cada dia em análise, para cada instante da série (3.6), é calculado o quociente (percentagem) entre

o consumo para esse instante e o consumo total do dia, fora os instantes para que os nulos ocorrem (3.9).

Algoritmo 3.2 Previsão de Nulos: Falha de Medição

nullday $\leftarrow$ *dia,* d_n*, em que se registou o(s) nulo(s) para previsão;*
events $\leftarrow$ *(3.10);*
days $\leftarrow$ *(3.4);*
$day_{area} = [day_{area,d_0}, ..., day_{area,d_i}] \leftarrow$ *(3.9) para a série (3.4);*
nullarea $\leftarrow$ *(3.9) para o nullday;*
function PREDICT(events, days, area, nullarea)
    $listpercents = list() \leftarrow$ *lista vazia;*
    for d in days **do**:
        $events_d = events[days.index(d)] \leftarrow$ *série de eventos, nos instantes (3.6), para o dia d;*
        $day_{area,d} = day_{area}[days.index(d)] \leftarrow$ *(3.9) para o dia d;*
        $percent = \frac{events_d}{day_{area,d}} * 100.$
        $listpercents.append(percent).$
    $listmean = zeros(len(listpercents[0])) \leftarrow$ *lista de zeros com dimensão de (3.6);*
    for l in range(0, len(listpercents)) **do**:
        for p in range(0, l) **do**:
            $listmean[p] = listmean[p] + l[p]$
    $percent = \frac{listmean}{len(listpercents)} \leftarrow$ *média das séries de percentagens;*
    $data = \frac{nullarea * percent}{100}$
return $data = [e_{t_0,d_n}, ..., e_{t_j,d_n}] \leftarrow$ *resultado da previsão a aplicar no nullday;*

Todos os dias da semana iguais, ao que se pretende fazer a previsão, são elegíveis. É feita uma previsão de nulos prévia para os instantes em falta, nos dias usados para comparação. Portanto, é de esperar que existam várias séries obtidas para fazer a previsão.

Contrariamente ao algoritmo 3.1, para o algoritmo 3.2 é calculada a média das percentagens, para cada instante, de todas as séries calculadas para previsão. Para fazer a previsão, é primeiro feita uma previsão prévia para o dia em que ocorre os nulos em análise, de outros nulos que possam existir mas que não estão em análise na mesma iteração. Posto isto, é calculada a soma total de consumos, sem considerar os instantes dos nulos em análise (3.9). Tem-se então todas as condições reunidas para a previsão

dos consumos em falta. Por meio da aplicação do algoritmo 3.2, torna-se possível prever os nulos para o caso em análise.

Sabendo (3.9) para a série (3.4), que representa o consumo total, fora instantes nulos em análise, para os dias elegíveis para previsão (3.4), sabendo também (3.10), séries de eventos para os instantes de ocorrência de nulos, para os dias referidos, é calculada a média das percentagens de consumo total, fora instantes nulos (3.9), para cada instante (3.6) sobre os valores de (3.10), nos dias (3.4). Posto isto, é calculada a variável (3.9) para o dia em que se deu a falha considerada e aplicada a percentagem calculada, sobre este valor, para cada instante (3.6). Assim, se obtém a previsão de nulos para este caso particular.

3.3 Discretização da Base de Dados

O tratamento anterior efetuado aos dados não é suficiente para se submeter os dados à descoberta de padrões.

Visto que, se está a trabalhar com uma grande quantidade de dados, é necessário desenvolver estratégias para garantir a eficácia e eficiência dos algoritmos. Uma dessas estratégias passa por discretizar os dados.

As bases de dados em causa são extremamente heterogéneas, o consumo elétrico é uma variável contínua que apresenta valores numa janela de grande amplitude, para a maioria dos casos. Por consequência desta heterogeneidade torna-se mais difícil a tarefa de descoberta de padrões. Para contornar esta heterogeneidade, pode-se agrupar dados distintos redundantes nas mesmas categorias. O conjunto de categorias por onde se vai distribuir os dados, corresponde aos novos valores discretos que a variável de consumo pode tomar. Conseguimos, assim, construir uma variável de consumo discreta a partir de dados contínuos.

O processo de discretização segue dois passos. Começa por definir pontos de corte que dividam as séries de dados num baixo número de partições. Estas partições devem ser coerentes, ou seja, os dados inseridos em cada partição devem assemelhar-se. Tal

semelhança, deve ser avaliada por uma função pré-definida que responda às necessidades da discretização. O segundo passo, passa por avaliar a coerência dentro de cada partição, com recurso à função definida [20] [21].

3.3.1 Limiar de Funcionamento

Pela análise direta do consumo de alguns equipamentos dos clientes VPS, monitorizados por *smart plugs/smart meters*, verificou-se que muitos deles tinham um consumo não nulo em períodos de não uso, isto é, um consumo residual. Portanto, o consumo nulo e o consumo não nulo não podem ser generalizados para estados ativos ou inativos, respetivamente, do(s) equipamento(s) sob monitorização.

De forma iterativa concluiu-se que, de grosso modo, não considerando os *outliers*, para as *smart plugs*, o patamar que divide o período de atividade do equipamento do período de inatividade, corresponde a um décimo do pico máximo. Ainda assim, verificou-se que para alguns equipamentos, o período de inatividade correspondia realmente a zero. Para os *smart meters* a divisão entre os patamares é aproximadamente de 15^{-1} do máximo, não considerando *outliers*. Este deve ser um processo iterativo, pelo que, a cada nova série de dados obtida por *smart plugs/smart meters* deve ser visto se são necessários ajustes ao patamar.

A variável (3.1) passa a ter uma nova variável associada que apenas toma dois valores, ON, quando o consumo bruto é associado à atividade do(s) aparelho(s) (consumos desagregados/agregados) e OFF, quando não é detetada atividade:

$$estado_{t,d} = \Big\{ON, OFF\Big\} \tag{3.11}$$

O método descrito é assim supervisionado, para o processo de discretização. É indicado o número de pontos de corte, um, bem como, onde este se dá.

Não se optou por um algoritmo não supervisionado, pois é complicado fazer-se esse processo sem qualquer referência. Para tal ser possível, seria necessário um *input** do cliente, durante um período de teste, de quando está a fazer uso do(s) equipamento(s) que está a monitorizar. Outra hipótese, seria usar um *input* do cliente com a informação

(*datasheet**) do(s) aparelho(s) que está a monitorizar. Ainda assim, a primeira opção é sempre mais viável, uma vez que, o consumo típico de um aparelho pode ter flutuações. Assim sendo, é mais fiável selecionar um período amostral, ver os consumos e o patamar de referência e fazer ajustes se necessário [22].

3.3.2 Patamares de Consumo e Modelo do Pacote de Palavras

A discretização dos valores de (3.1) para os instantes em que a variável (3.11) associada, toma valores ON, tem como objetivo aumentar a eficiência e eficácia dos algoritmos de descoberta de padrões de consumo, deteção e diagnóstico de falhas dos equipamentos e de desagregação de cargas.

Não considerar os valores de (3.1) com valor OFF (3.11) associado para análise, diminui bastante a quantidade de dados a tratar, por consequência, aumenta a eficiência dos algoritmos a aplicar posteriormente.

Os valores (3.1) associados ao valor ON (3.11) são, então, discretizados em patamares de consumo. Cada patamar é representado por uma letra do alfabeto única. Faz sentido, que estes valores sejam divididos em patamares de consumo. Lembrando que (3.1) representa consumos elétricos acumulados entre janelas temporais de 5 ou 60 minutos, o estado ON pode refletir apenas uma fração deste tempo, entre o intervalo de medida, ou então, toda a fração de tempo. A adicionar como argumento para a premissa, de que faz sentido dividir a série de consumos ON em patamares de consumo, está o facto de alguns aparelhos terem um largo espetro de potências de operação.

Para agrupar os dados de forma lógica, é necessário recorrer a algoritmos de aprendizagem não supervisionada. Ou seja, algoritmos em que o programador não tem de introduzir qualquer informação para que o processo se possa dar. O tipo de algoritmos, adequados para o caso, são algoritmos de *clustering*. Com este tipo de algoritmos obtém-se uma base de dados dividida em grupos lógicos, de acordo com a semelhança que os dados estabelecem entre si, podendo caracterizar cada *cluster** pelo seu ponto médio ou pela sua mediana geométrica.

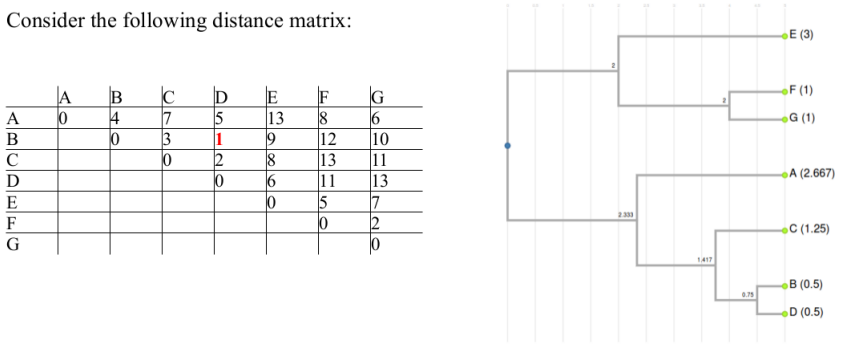
Pode-se dividir os algoritmos de *clustering* em duas categorias. Aqueles em que é necessário definir, previamente, o número de grupos em que queremos dividir os dados, e os algoritmos em que tal não é necessário. Para estes últimos é necessário um registo de todos os agrupamentos possíveis e também a existência de uma função que avalie o número ótimo de *clusters*.

Visto que, interessa um algoritmo genérico, para este projeto, que funcione para diferentes tipos de bases de dados, equipamentos, o tipo de algoritmo ideal, de *clustering*, é aquele em que o número de *clusters* não é definido à partida.

O método recorrido consiste em agrupar a base de dados, de forma iterativa, construindo um hierarquia de *clusters* com sucessivos agrupamentos, denominado algoritmo de *clustering* hierárquico.

Um algoritmo iterativo de agrupamento é iniciado, considerando que cada dado é por si só uma *cluster*, e num processo iterativo, agrupa-se as diversas *clusters* até se chegar a uma única.

Deve-se iniciar o algoritmo pela construção da matriz de distâncias, que contém a distância que todos os pares de *clusters* possíveis de formar, distam entre si.



	A	B	C	D	E	F	G
A	0	4	7	5	13	8	6
B		0	3	1	9	12	10
C			0	2	8	13	11
D				0	6	11	13
E					0	5	7
F						0	2
G							0

Figura 3.1: Representação de uma matriz de distâncias à esquerda e um dendograma à direira, para um caso em que se pretende agrupar letras com um valor associado [23].

A figura 3.1 apresenta uma matriz de distâncias iniciais. O valor a vermelho representa as posições das *clusters* a agrupar na primeira iteração. De notar que só interessa

analisar os pontos acima ou abaixo da diagonal da matriz. Usando a informação acima da diagonal, a que fica abaixo é redundante e vice-versa. A diagonal representa a distância que uma *cluster* estabelece consigo própria.

A cada iteração devem ser agrupadas as duas *clusters* que estabelecem uma distância entre si, mais baixa, de acordo com a métrica definida. A matriz distância deve ser atualizada, a cada iteração, de acordo com o critério de agrupamento definido.

É também representado um dendograma na figura 3.1. Um dendograma é um gráfico que representa os agrupamentos sucessivos entre *clusters* até restar uma única *cluster*. No gráfico apresentado, quanto mais se avança para a esquerda, maior é a distância estabelecida entre as *clusters* agrupadas a cada iteração.

As iterações podem decorrer até a paragem ser forçada, ou, até apenas restar uma única *cluster*, dependendo se é estipulado o número ideal de *clusters* ou não. No caso em estudo, o algoritmo segue até apenas restar uma única *cluster*.

Como já referido, a matriz das distâncias deve conter as distâncias estabelecidas entre todos os pares de *clusters* possíveis de agrupar. A cada iteração, apenas são agrupadas duas *clusters*.

A	**B**	**C**	**D**
{1,2,3,4,7}	{10,13,15,16}	{18,19,21}	{30}

A→B	**B→C**	**C→D**
$10 - 7 = 3$	$18 - 16 = 2$	$30 - 21 = 9$

Tabela 3.1: Representação esquemática dos elementos que cada *cluster* contem e das distâncias estabelecidas entre duas *clusters* consecutivas.

Para o caso em análise, e tendo como exemplo a tabela 3.1, pretende-se agrupar as duas *clusters* que distem de uma menor diferença entre os seus pontos extremos consecutivos, isto é, tendo todas as *clusters* ordenadas por ordem crescente aos consumos que contêem, fazendo a diferença entre o mínimo de uma *cluster* e o máximo da sua antecessora, as duas *cluster* que apresentarem uma menor diferença, nessa iteração, são agrupadas. Na iteração seguinte, representaram uma única *cluster* que contem todos

os dados das *clusters* que lhe deram origem. Assim sucessivamente, até apenas restar um único grupo de dados [24] [25].

Os grupos de dados obtidos, a partir de um valor *threshold*, até ao nível que existe apenas uma única *cluster*, são registados em memória para que posteriormente sejam submetidos a análise, por meio de uma função que devolve o número ótimo de *clusters*.

A função de avaliação analisa, para cada grupo em memória, a semelhança entre os dados presentes em cada *cluster*. A forma mais rigorosa que se encontrou para calcular esta semelhança foi a de fazer a soma do módulo das diferenças, entre todos os pares de pontos possíveis de estabelecer, dentro de cada *cluster*, normalizado ao número de pares. A este valor obtido vamos chamar de variação da *cluster*. Cada agrupamento passa a ter associado o número de *clusters* presentes e a soma total das variações de *cluster* calculadas para o conjunto, sendo respetivamente a abcissa e a ordenada de um ponto, de um conjunto de pontos que servem de entrada para a função de avaliação.

O método de análise dos pontos representantes do agrupamento, a partir do valor *threshold*, é baseado no *L-Method*. Este método, parte de uma representação gráfica destes pontos para achar o joelho da curva que estes formam.

O joelho da curva corresponde ao número ótimo de *clusters*. O joelho de uma curva é identificado pelo ponto de intersecção das duas retas que conferem o melhor ajuste linear ao gráfico.

Para cada ajuste obtém-se duas equação de reta do tipo (3.12).

$$y = m * x + b \leftarrow m \text{ é o declive da reta e } b \text{ a ordenada na origem} \tag{3.12}$$

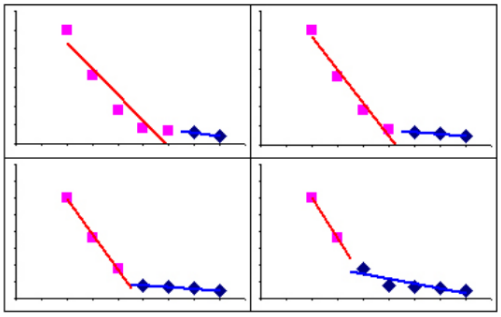


Figura 3.2: Conjunto de todos os ajustes lineares a duas retas possíveis de formar para o conjunto de pontos representados [26].

O melhor ajuste linear a duas retas é identificado pelo cálculo do conjunto de ajustes que gera um menor erro (3.13).

$$erro = \sqrt{\frac{\sum_{i=0}^{j} y_i^{ajuste} - y_i^{real})^2}{j}} \leftarrow n\acute{u}mero\ de\ pontos\ de\ dimens\tilde{a}o\ j \tag{3.13}$$

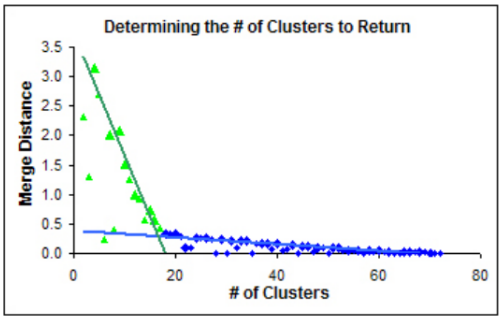


Figura 3.3: Representação do melhor ajuste linear a duas retas efetuado ao conjunto de pontos apresentados [26].

Com isto, se conclui o número ótimo de *clusters*, e por consequência, o conjunto de *clusters* em que se divide os dados, ou seja, os patamares de consumo. Deve ser registado, na base de dados, numa tabela, o mínimo, o máximo, a média de cada patamar e uma letra do alfabeto aleatória que o represente, distinta dos patamares restantes [26].

De notar que em nada acrescenta representar um patamar por uma letra, poderia apenas ser representado pela sua média. No entanto, ao analisar um período de instantes sucessivos em que o dispositivo(s) esteve ativo(s) (3.14) (consumos agregados/desagregados), esta tradução revela-se útil.

Uma série de eventos (3.1) sucessivos, entre os instantes t_i e t_f, para um dia d, com valor ON da variável (3.11) associado para todos os instantes, é traduzida por (3.14).

$$on_{\Delta t,d} = [e_{t_i,d}, ..., e_{t_f,d}] \tag{3.14}$$

A série (3.14) pode ser representada por uma palavra, associando a cada evento (3.1) da série, o caracter do patamar correspondeste, formando uma palavra em que os caracteres em cada índice correspondem aos eventos na série (3.14) na mesma posição.

Os períodos ON, de funcionamento sucessivo (3.14), podem assim, ser representados por palavras. Na vez de se analisarem e compararem séries de dados brutos, em que o equipamento foi identificado como estando em operação, são comparadas palavras, o que facilita imenso a comparação. Esta metodologia é descrita como o modelo do pacote de palavras [9].

3.4 Agrupamento dos Padrões de Consumo

O modelo das palavras descrito em (3.3.2), dá-nos uma descrição dos padrões de consumo no período ON (3.14). Não interessa apenas analisar um consumo isolado. Interessa também analisar todo o período sucessivo de atividade do aparelho. Estas séries (3.14), traduzidas em palavras, dão o potencial da determinação de padrões em consumos agregados, ou desagregados, e de diagnóstico e deteção de falhas de equipamentos.

A análise das palavras resultantes da aplicação do modelo de palavras aplicados a todas as séries (3.14), presentes numa séries de dados, obtidas com recurso a uma/um *smart plug/smart meter*, deve ser iniciada por identificação de todas as palavras associadas, inseridas numa lista e eliminadas palavras em duplicado.

Tendo todas as palavras únicas reunidas, pretende-se agrupá-las de acordo com a sua semelhança. Para isso, submete-se as palavras a um algoritmo de agrupamento hierárquico. O algoritmo segue a mesma estrutura do descrito na secção 3.3.2, muda apenas o critério de agrupamento usado para a construção da matriz das distâncias. O critério escolhido deve traduzir a semelhança entre palavras de forma numérica.

Uma abordagem possível para determinar a semelhança entre duas palavras é calcular a diferença entre um par de palavras, por meio da transformação de uma das palavras do par, na restante. Esta distância é obtida por imposição de uma série de transformações de inserção, remoção e substituição de caracteres, aplicadas à primeira palavra para a transformar na restante. Cada operação tem um custo associado. A distância entre as duas palavras é obtida pelo custo total associado, ao menor número de operações necessárias, para transformar uma palavra na restante. Esta distância obtida é chamada de distância de *Levenshtein* entre duas palavras.

O custo associado a cada operação de inserção, w_i, remoção, w_r e substituição, w_s, por padrão têm um custo de um. Caso os caracteres em comparação sejam iguais então $w_r = 0$.

Seja um par de palavras constituído por $\{S_1,S_2\}$, com um número total de caracteres de i e j, respetivamente, o custo total é obtido pela construção de uma matriz que representa todas as operações para transformar uma palavra na restante.

Algoritmo 3.3 Distância de *Levenshtein*

$S_1 \leftarrow$ *Primeira palavra do par de palavras com i caracteres;*
$S_2 \leftarrow$ *Segunda palavra do par de palavras com j caracteres;*
$D \leftarrow$ *Matriz de zeros com i linhas e j colunas;*
function LEVENSHTEIN(S_1,S_2, D)
    for k in range(1,i) **do**:
        $D[k][0] = D[k-1][0] + w_r$.
    for l in range(1,j) **do**:
        $D[0][l] = D[0][l-1] + w_i$.
    for k in range (1,i) **do**:
        for l in range (1,j) **do**:
            $D[k][l] = min([D[i-1][j] + w_r, [D[i][j-1] + w_i, [D[i][j] + w_s)$
return *distance* $= D[i][j]$

O algoritmo 3.3 descreve o procedimento para determinar a matriz D, representante das operações para transformar s_1 em s_2 ou vice-versa. A distância de *Levenshtein* corresponde ao elemento D[i][j] da matriz D obtida. Este elemento da matriz corresponde ao somatório do custo de todas as operações necessárias para tranformar uma palavra na restante, com menor caminho [27].

Visto que, cada caracter simboliza um patamar de consumo, substituir caracteres de patamares sucessivos não é igual, a nível de custo, a substituir patamares extremos. Para solucionar este problema, em vez de se fazer $w_r = 1$, $w_i = 1$ e $w_s = \{1, 0\}$, atribui-se a cada caracter um valor, que corresponde à média do patamar que representa. O valor de w_r e w_i passa a corresponder ao valor associado ao caracter que se pretende remover ou inserir. O valor de w_s corresponde à diferença dos valores médios associados aos carateres que se pretende substituir. De notar que os custos são referentes aos caracteres correspondentes a cada linha e coluna a que pertence o elemento, em análise, da matiz D. Com esta alteração no algoritmo padrão, consegue-se assim contornar o problema de os patamares não serem equivalentes. Temos, assim, definida uma forma de calcular a distância entre pares de palavras.

Uma vez que, se vai aplicar o algoritmo hierárquico, já abordado em 3.3.2, para agrupar as palavras em *clusters*, com base na semelhança entre as palavras, é necessário definir a cada iteração o centro da *cluster*. O centro de cada *cluster* corresponde à palavra que estabelece uma menor distância de *Levenshtein* com as restantes.

Para determinar a palavra correspondente ao centro de uma *cluster*, é calculada uma matriz de distâncias representativa das distâncias de *Levenshtein* entre todos os pares de palavras possíveis de formar. Para calcular a soma total de distâncias que cada palavra estabelece com as restantes, basta criar um dicionário que a cada linha da matriz, que tem uma palavra associada, faz corresponder a soma de todos os elementos dessa linha, visto que cada coluna representa uma palavras diferente. É equivalente fazer o mesmo procedimento, mas com a soma das colunas. De notar que neste caso em particular, todos os índices da matriz das distâncias devem ser considerados. A palavra que tiver uma menor soma da sua linha corresponde ao centro da matriz. A *cluster*

deve ser representada por essa palavra.

A cada iteração do algoritmo hierárquico de *clustering*, devem ser agrupadas as *clusters* em que os seus centros estabeleçam uma menor distância de *Levenshtein*. Não esquecer que, o centro das *clusters* devem ser atualizados, a cada iteração do algoritmo hierárquico, para as *clusters* que sofrerem alterações.

Para o cálculo do número ótimo de *clusters*, a variação de cada *cluster* é obtida pela soma das distâncias de *Levenshtein* entre todos os pares de palavras possíveis de formar dentro de cada *cluster*, normalizado ao número de pares. Posto isto, o algoritmo é em tudo igual ao *L-Method* descrito na secção 3.3.2.

Temos assim as palavras representativas dos períodos ON (3.14), divididas por *clusters*, baseadas no grau de semelhança que as palavras estabelecem entre si, tendo cada caracter associado a si um custo igual à média do patamar que representa.

3.5 Identificação de Rotinas

Tendo já os padrões de consumo agrupados em *clusters*, interessa agora importar esta lógica de padrões de consumo para todo um dia de consumos. Isto é, agrupar todos os dias em análise, em *clusters*, de acordo com os períodos em que ocorrem as séries (3.14). Ou seja, na vez de analisarmos as palavras associadas às séries (3.14), analisamos as séries de instantes em que elas ocorrem. Com isto, torna-se possível a identificação de rotinas de consumo, isto é, a título de exemplo, para uma empresa que tipicamente fecha ao fim-de-semana, é expectável que os seus dias se dividam em duas categorias, dias úteis (segunda a sexta-feira) e fins-de-semana.

Para se iniciar a identificação de rotinas, cada dia passa a ser representado por uma lista de zeros e uns. Todos os instantes associados às séries (3.14), presentes na base de dados, para o dia em análise, associamos um, aos restantes instantes, zero. O zero corresponde aos eventos (3.1) com valores OFF (3.11) associados, e o um aos com ON (3.11) associados. A lista deve estar ordenada por forma crescente no tempo, ou seja, cada índice representa um evento (3.1), posterior no tempo ao evento (3.1) do índice

que o antecede.

Novamente, as listas representantes dos dias em análise são submetidas a um agrupamento hierárquico igual ao descrito nas secções 3.3.2 e 3.4.

Desta vez, o critério de agrupamento é a correlação entre cada par de listas. É criada uma matriz de distâncias para as listas representantes dos dias. O parâmetro de comparação é a correlação entre cada par de listas.

A correlação entre cada par de listas é calculado por análise dos elementos que correspondem a eventos (3.1) ocorridos nos mesmos instantes. Se os elementos correspondentes de um par de listas forem diferentes, é adicionado um peso ao parâmetro de avaliação da correlação, que inicialmente se igualava a zero. Este peso é tanto maior quanto menor a probabilidade de funcionamento do aparelho ao longo do dia. Analisados todos os pares de elementos, para todos os instantes de um dia, quanto menor o parâmetro de avaliação da correlação, maior a correlação do par de listas.

Uma vez que, existem equipamentos que funcionam muito aleatoriamente no tempo, e/ou em intervalos muito curtos, é importante trabalhar com intervalos de tempo ao invés de instantes. Com a experiência, verificou-se que os instantes em que este tipo de aparelho estava em funcionamento acabavam por passar despercebidos.

O programador tem portanto de introduzir o peso e o intervalo de tempo em que pretende calcular a correlação, sendo que, pode padronizar estes dados para o tipo de equipamento a monitorizar, ou caso se trate de um *smart meter*, parâmetros que correspondam às necessidades dos consumos agregados.

É, então, criada uma matriz de distância representante da correlação entre cada par de listas possível de formar, para iniciar o algoritmo hierárquico de *clustering*.

A cada iteração do algoritmo hierárquico, são agrupados numa *cluster* os pares de listas correspondentes à posição do elemento mínimo da matriz. Este elemento da matriz deve ser substituído por um número superior ao máximo da matriz para que não volte a ser identificado como mínimo, numa iteração seguinte. As *clusters* a que pertencem as listas identificadas como as mais semelhantes, a cada instante, devem ser agrupadas. Caso as listas já pertençam à mesma *cluster*, a junção deve ser ignorada e

o elemento da matriz substituído na mesma. As iterações terminam quando todas as listas estão agrupadas numa *cluster*.

Posto isto, sabendo o número *threshold* de *clusters*, a determinação do número ideal de *clusters* segue os mesmos princípios do exposto na secção 3.3.2.

A variação em cada *cluster*, parâmetro necessário para a descoberta do número ideal de *clusters*, é a soma das correlações de todos os pares de listas possíveis de formar, dentro de cada *cluster*, normalizada ao número de pares.

Temos assim o numero ótimo de dias tipo existentes, divididos por *clusters*. Importa representar cada *cluster* pelo seu centro, que corresponde à lista que apresenta uma menor soma das correlações estabelecidas com as restantes listas presentes na mesma *cluster*.

3.6 Associações entre Valores de Variáveis Distintas

As bases de dados de um edifício, em que se está a monitorizar vários equipamentos, são grandes e diversificadas ao nível de variáveis e valores que estas podem tomar. Com isto, há padrões de associações de valores de variáveis distintas que acabam por passar despercebidos, aquando da visualização de dados, entre a entropia existente na base de dados do edifício.

Não basta, então, fazer uma visualização dos dados para as variáveis brutas, e as construídas durante o processo de descoberta de padrões, por processos de *data mining*. É necessário desenvolver um algoritmo de associação que revele padrões escondidos.

Entenda-se que, ocorre associação, no caso em estudo, de dois ou mais valores de variáveis distintas, quando estes surgem em simultâneo com uma frequência relevante.

Como já referido, para o caso, interessa analisar toda a base de dados do edifício, portanto, as variáveis em causa, para cada instante, são os equipamentos que foram detetados em modo ON (3.14), nesse instante, a hora do registo, o tipo de dia do registo, o mês, a temperatura, humidade e precipitação. A associação entre os valores que estas variáveis podem tomar, em simultâneo, é o assunto desta secção.

Antes de mais nada, é necessário definir alguns conceitos:

- Conjunto de todas as variáveis, de dimensão i, presentes na base de dados de um edifício (inclui vários equipamentos monitorizados):

$$C = \left\{c_1, ..., c_i\right\} \tag{3.15}$$

- Conjunto de todos os valores, de dimensão j, que uma variável $c_n \in C$ (3.15) pode tomar:

$$c_n = \left\{c_n^1, ..., c_n^j\right\} \tag{3.16}$$

- Sub-conjunto de C (3.15):

$$A, A \subset C \tag{3.17}$$

São analisadas e registadas todas as combinações de valores de variáveis distintas, ou seja, todos os sub-conjuntos A (3.17) possíveis de formar. Para cada sub-conjunto A (3.17), são analisados todos os casos possíveis, isto é, todas as combinações possíveis de valores registadas, para este conjunto de variáveis.

Cada combinação de valores de A (3.17), tem um ou mais valores antecessores e um ou mais valores sucessores. Os valor sucessores ocorrem quando os antecessores ocorrem, sem que o contrário seja necessário. De realçar que não existe relação causa-efeito entre os valores das variáveis antecessoras e sucessoras. Para explicar este conceito tomemos i=4 e analisemos o subconjunto de C , $A = \left\{c_1, c_2, c_3\right\}$ (3.17) com $c_1 = \left\{c_1^1, c_1^2, c_1^3\right\}$, $c_2 = \left\{c_2^1, c_2^2\right\}$ e $c_3 = \left\{c_3^1, c_3^2, c_3^3, c_3^4\right\}$ (3.16). Consideramos o caso em que A toma os valores $A = \left\{c_1^2, c_2^1, c_3^4\right\}$ e que c_1^2 antecede c_2^1 e c_3^4. Considera-se que para esta associação de valores c_2^1 e c_3^4 ocorrem quando c_1^2 ocorre. Este acontecimento pode ter uma frequência considerável, mas o acontecimento contrário, de c_2^1 e c_3^4 anteceder c_1^2, não ter uma frequência relevante e não representar uma associação de eventos. Por exemplo, um aluno universitário é necessariamente um estudante, mas, um estudante não é necessariamente um aluno universitário, aliás, no conjunto de todos os estudantes Portugueses, a fração de estudantes que são alunos universitários é mínima. Com isto, pretende-se

mostrar que um par de antecessores e sucessores pode ter uma frequência relevante, mas o par contrário pode ser insignificante em percentagem para ser considerada associação de valores.

Depois de se ter formado todos os sub-conjuntos do conjunto C, (3.17), é necessário então, para cada sub-conjunto, construir os pares de valores ascensores e sucessores possíveis. Para todos esses pares, para se apurar se existe uma associação relevante, é necessário calcular algumas frequências. Entra assim o conceito de *suporte* (3.18), *confiança* (3.19) e *lift* (3.20). Consideraremos o subconjunto $A = \left\{c_1, c_2\right\}$, $c_1 = c_1^3$ e $c_2 = c_2^1$ em que $c_1^3 \rightarrow c_2^1$, ou seja, c_1^3 antecede c_2^1:

- Probabilidade de ocorrência de um valor do conjunto A (3.17):

$$suporte = P(x), x \in A \tag{3.18}$$

- Probabilidade de ocorrência do sucessor, sabendo que o antecessor se verifica:

$$confiança = P(c_2^1|c_1^3) \tag{3.19}$$

- Medida da dependência entre os antecessores e os sucessores:

$$lift = \frac{P(c_2^1 \cap c_1^3)}{P(c_2^1) * P(c_1^3)} \tag{3.20}$$

O processo de deteção de associações relevantes, na base de dados, começa por calcular o *suporte* (3.18) de todos os valores do conjunto A (3.17), a *confiança* (3.19) e o *lift* (3.20) de todos os pares de antecessores e sucessores possíveis de formar a partir do conjunto referido, para todos os eventos registados.

Para se determinar as associações que são relevantes, é necessário que o *lift* (3.20) seja superior a um. O *lift* (3.20) é uma medida da dependência entre os antecessores e os sucessores, quanto maior, maior a dependência. O $lift > 1$ (3.20) garante que a associação não se deu por mero acaso. Também é necessário definir um mínimo de *suporte* (3.18), e o mínimo para a *confiança* do par (3.19). Tendo todos os valores de variáveis associadas a todos os pares possíveis de formar para todos os A (3.17),

impondo estas condições, é feita uma filtragem às associações, resultando apenas as associações relevantes.

Obtém-se, assim, todos os pares de associações relevantes dos registos das séries de dados em análise [28] [24].

3.7 Consumos Anormais

Neste processo, a base de dados é analisada equipamento a equipamento, para um edifício em que se monitoriza vários equipamentos de forma desagregada, ou então, são analisados os consumos agregados obtidos por um *smart meter*. Portanto, não são analisadas associações entre dispositivos distintos. Nesta secção, interessa analisar os padrões individuais monitorizados por um único dispositivo *smart plug/smart meter*, com o objetivo de detetar consumos anormais. Isto é, se a hora para que se deu o padrão de consumo é anormal, se o consumo para a hora de ocorrência não é frequente, se o dia apresenta flutuações da rotina esperada para o dia da semana em questão...

Para a identificação de consumos anormais, a base de dados obtida por cada *smart plug/smart meter* é analisada mês a mês. É normal que um aquecedor funcione nos meses de outono e inverno, mas não é normal que esteja em funcionamento num mês de verão. Neste caso, um consumo de um equipamento como este podia ser confundido como normal, na base de dados geral, caso a etiqueta estação do ano não fosse considerada. Como a etiqueta estação do ano é demasiado geral, por exemplo, no ano que passou, 2017, existiram meses de outono em que se registaram 25ºC de temperatura máxima e dias em o termómetro de temperatura máxima desceu abaixo dos 15ºC. Por estas razões, para cada instante em análise, são considerados trinta dias anteriores à data em análise.

As variáveis a considerar, para os períodos em que foi detetado um estado ON (3.14), para deteção de consumos anormais, é a palavra associada a (3.14) em análise (secção 3.3.2), a *cluster* a que a palavra pertence (secção 3.4), a duração de (3.14) e a hora a que se deu o inicio de (3.14).

Estes dados são comparados com uma tabela de referência que contem, para os trinta dias antecedentes, a cada hora, o centro do conjunto de palavras dos consumos (3.14) que deram início nessa hora, o desvio padrão das distâncias de *Levensthein* (algoritmo 3.3) que as restantes palavras ocorridas para a mesma hora estabelecem com o centro identificado para essa mesma hora, a probabilidade de detetar cada *cluster* existente, a duração média de cada série (3.14), o desvio padrão associado e a percentagem com que se dá consumos na hora em análise, em comparação com as restantes. Com estes dados de referência é possível determinar se algum dos valores detetados é um consumo anormal. Se algum dos valores registados for categorizado como consumo anormal, é reportado.

Para cada dia em análise é também analisado o dia em que foi categorizado o dia da semana em questão (secção 3.5), por consequência da identificação de rotinas, comparado com as restantes frequências que ocorrem todos os dias tipo, para o dia da semana em questão, e visto que se trata de um dia fora da rotina esperada para esse dia da semana ou mês.

Com isto, são detetados consumos anormais que podem ser reportados ao cliente, e este, identificar a sua causa. Caso um consumo anormal se prolongue no tempo, ele é identificado de raiz, uma vez que são analisados os trinta dias para trás do instante em análise. Por exemplo, caso um aparelho entre em sub ou sobre funcionamento, é identificado o instante de início. Por outro lado, aparelhos com consumos sazonais são identificados como consumos atípicos nas primeiras vezes que entram em funcionamento, ou aparelhos com um uso muito pontual. Padrões muito irregulares de consumo podem refletir mau uso ou falha do equipamento. A informação do uso de aparelhos fora de horas típicas e dias de consumo fora da rotina pode ser relevante para o cliente, por exemplo, de uma ocupação do edifício não desejada.

O reporte ao cliente de consumos atípicos revela-se assim de grande valor. Pode conduzir a um aumento da eficiência energética do edifício, possível identificação precoce de falhas dos equipamentos e estados de atividade atípicos do edifício.

3.8 Desagregação de Consumos Elétricos Agregados

A monitorização de um edifício de forma não intrusiva, é conseguida com recurso a *smart meters*. Esta monitorização é não intrusiva porque não é necessária a presença de qualquer equipamento adicional, no interior do edifício que o cliente pretende monitorizar.

Estes dados, visto representarem o somatório dos consumos de todos os equipamentos presentes no edifício sob monitorização, têm o potencial, por meio de aplicação de uma série de algoritmos de desagregação de cargas (consumos elétricos), construir-se uma base de dados, para cada equipamento elétrico individual ativo no edifício sob monitorização.

Para que se consiga fazer uma desagregação eficiente dos consumos elétricos, é necessária alguma informação adicional, como por exemplo, os equipamentos que o *smart meter* monitoriza, durante uma fase de testes, os períodos em que o aparelho esteve em modo ON (3.14)... Tal informação adicional, para as bases de dados analisadas no capítulo 4, existe apenas para um edifício (casa piloto) com um *smart meter*. Para esse edifício existe o registo dos consumos desagregados, para os mesmos instantes dos registos dos consumos agregados, de uma máquina de lavar e de um frigorífico.

A possibilidade de se desenvolver um algoritmo que desagregue completamente os consumos totais e identifique todos os equipamentos em operação é posta de parte, visto que, não existe informação disponível, nas bases de dados que iremos trabalhar, que o permitam. Contudo, é possível desagregar a base de dados para os equipamentos que têm essa informação adicional. Por visualização dos dados, verifica-se, como era de esperar, que o frigorífico, para a base de dados disponível, apresenta um funcionamento contínuo e com um padrão constante no tempo. Com isto, apenas a máquina de lavar é considerada para fins de desagregação de carga. Os algoritmos que são apresentados nesta secção, para identificação de equipamentos na base de dados, a serem usados para identificar a máquina de lavar nos consumos agregados, podem ser generalizados para todos os equipamentos, desde que, exista informação adicional. Caso existisse informa-

ção adicional para todos os equipamentos presentes no edifício em análise, teríamos a série de dados completamente desagregada.

É preciso ter em conta que a variável representante dos consumos brutos agregados, na base de dados dos registos do *smart meter*, é extremamente heterogénea, e que, no mesmo instante, podem estar a funcionar, em paralelo, vários aparelhos elétricos.

Os consumos dos equipamentos (consumos desagregados), durante um período ON (3.14), podem se classificar de uma forma genérica em quatro categorias:

- Equipamentos com apenas dois estados de operação, o estado ON e o estado OFF (3.11) (os valores de (3.1) associados aos valores de (3.11) são aproximadamente constantes), por exemplo uma lâmpada ou uma torradeira.

- Equipamentos que durante o estado ON (3.14) de funcionamento, variam o seu consumo entre patamares bem definidos e em baixo número. Exemplo de um equipamento deste tipo é a maquina de lavar roupa.

- Equipamentos que apresentam um consumo variável no tempo, durante o período ON (3.14), sem patamares definidos. Estes equipamentos podem tomar qualquer valor dentro da gama de funcionamento do aparelho. Um exemplo é um candeeiro que permita ajustar a luminosidade.

- Equipamentos que se mantém ativos durante um longo período de tempo com um consumo constante, um telefone é um exemplo.

O algoritmo de desagregação de carga desenvolvido, começa por criar uma nova variável, obtida a partir dos consumos brutos. Esta variável pretende representar a variação de estado entre dois eventos (3.1) consecutivos. Chamou-se a essa variável de *step* e corresponde à diferença, para um dado instante, do seu evento (3.1) posterior correspondente, e o evento registado nesse instante.

$$step_{t,d} = e_{t+1,d} - e_{t,d} \tag{3.21}$$

Com a variável $step_{t,d}$ (3.21), temos uma representação da mudanças de estado associadas aos consumos brutos (3.1). Esta variável (3.21) é de extrema importância visto que um aparelho numa casa é muito mais facilmente identificável pelas suas sucessivas mudanças de estado características, durante um período de uso (3.14), do que pela sua assinatura de consumos brutos. Descarta-se, assim, a hipótese do uso dos consumos brutos (3.1) para os algoritmos de desagregação de carga, visto que o tempo de uso (Δt em (3.14)), para o mesmo equipamento, pode variar imenso e o facto do *smart meter* detetar consumos agregados, e detetar todo o consumo do edifício para cada instante, cada instante trás associado a si a soma dos consumos de todos os equipamentos, portanto não é possível a identificação com base no consumo bruto.

Para que a variável $step_{t,d}$ (3.21) devolva conhecimento sobre os dados agregados, ela ainda necessita de ser sujeita a algumas transformações. Para isso, começa-se por criar uma lista s (3.23) que contenha todos os valores que a variável $step_{t,d}$ (3.21) toma, os valores negativos devem ser substituídos pelo seu módulo. Devem ser eliminados duplicado. Por fim, resta ordenar a lista de forma ascendente.

- Conjunto de valores, de dimensão k, que a variável (3.21) toma, numa janela temporal, associados aos valores brutos de (3.1) obtidos com recurso a uma/um *smart plug/smart meter*:

$$step = \Big\{ s_1, ..., s_k \Big\} \tag{3.22}$$

- Lista dos módulos dos valores pertencentes a (3.22), sem duplicados, ordenados de forma crescente, de dimensão m:

$$s = \Big[|s|_1, ..., |s|_m \Big] \tag{3.23}$$

De notar que a lista original (3.22) toma valores positivos e negativos, tal é expectável visto que é obtida partindo de valores representativos de diferenças. A lista s (3.23), obtida a partir de (3.22), é submetida ao mesmo algoritmo descrito em 3.3.2, para divisão das diferenças de consumo em patamares e associado um caracter a cada um. Estando a divisão feita, são criados patamares simétricos em relação à diferença

nula e adicionados aos patamares detetados primeiramente. A importância de existir um patamar associado aos valores que a variável $step_{t,d}$ (3.21) toma, prende-se no facto de estarmos a detetar mudanças de estado, portanto, a variável (3.21) necessita de ser discretizada por patamares, visto que, é de esperar que dois valores de (3.21) caracterizantes da mesma mudança de estado apresentem pequenas diferenças de valor, visto que os consumos brutos (3.1), para a mesma ação do aparelho, também oscilam. Os símbolos correspondentes aos patamares com limite superior zero e com limite inferior também de zero, ou seja, os dois patamares que fazem vizinhança com o zero, um do lado dos números negativos, e o outro dos positivos, devem ser substituídos por um símbolo vazio, neste projeto representado por $*$. Quando um aparelho está em modo ON (3.14), nem todos os instantes correspondem a uma mudança real de estado, momentos em que o aparelho apresenta um consumo constante mas com pequenas flutuações. Com a atribuição do símbolo vazio aos patamares vizinhos do nulo, ao aplicar o modelo do pacote de palavras descrito em 3.3.2, os eventos de regime de consumo constante não são representados nas palavras. Assim, ao analisar um período ON (3.14), apenas são representadas as mudanças reais de estado dos equipamentos, registadas de forma ordenada e sucessiva. Obtemos, assim, as sucessões de mudanças de estado significativas e sucessivas para cada estado ON (3.14) dos equipamentos.

Visto que, apenas se tem informação adicional relevante sobre um equipamento (máquina de lavar roupa), para a base de dados, com os dados que se pretende desagregar, criou-se uma alternativa para ter mais algum conhecimento sobre o tipo de equipamentos em operação, a cada período de atividade do edifício Δt (3.14) para os consumos agregados. Tome-se como estado ON (3.11) do edifício, eventos (3.1) em que o consumo se distingue do consumo elétrico de fundo, isto é, eventos que se distinguem do consumo constante do edifício, representando atividade.

Essa alternativa passa por dividir as palavras obtidas pelo modelo do pacote de palavras, representantes das mudanças de estado registadas por um *smart meter*, em *clusters*, com base no princípio descrito na secção 3.4. Assim, tem-se os consumos de período ON (3.14) do edifício agrupados, com base nas sucessivas mudanças de estado,

significantes, verificadas nesse período. Com isto não se tem uma desagregação da base de dados, mas tem-se pistas do tipo, ou tipos de aparelhos que estão em operação para cada *cluster*, por meio da análise dos seus centros [29] [30].

3.8.1 Identificação de Equipamentos em Consumos Elétricos Agregados

O processo de desagregação passa por identificar os instantes, na base de dados, em que cada aparelho presente no edifício esteve ativo. Este processo passa pela análise do patamar a que pertence a variável $step_{t,d}$ (3.21) para cada instante, ao longo da gama de dados disponíveis.

Os patamares correspondentes a cada valor de *step* (3.22), de dimensão igual a *step*, k, estão incluídos no conjunto $step_p$ (3.24).

$$step_p = \left\{ s_p^1, ..., s_p^k \right\} \tag{3.24}$$

Visto que se dispõe de uma série de dados isolada dos consumos de um aparelho que se pretende identificar (máquina de lavar roupa), pode-se tentar dois tipos de abordagem para identificação do aparelho nos consumos agregados.

Numa primeira abordagem usa-se o menor tipo de informação possível sobre o equipamento. Tenta-se uma abordagem o menos intrusiva possível, usando apenas uma assinatura do equipamento.

A segunda abordagem é bem mais intrusiva. É usada uma base de dados de teste em que se sabe quais os instantes em que o equipamento esteve em funcionamento, associados aos eventos registados pelo *smart meter* para aprendizagem do algoritmo. Para o caso, a base de dados indicativa dos instantes de funcionamento do equipamento em análise, é obtida por determinação dos períodos de atividade determinados com recurso a uma *smart plug* associado ao equipamento. Mas, também poderia ser por indicação do utilizador, dos intervalos de tempo em que o equipamento esteve a funcionar, durante o período de teste.

Assinatura dos Equipamentos

A assinatura do equipamento, com o tipo de dados disponível, será os dados de um período aleatório, nos consumos agregados, em que se sabe que o equipamento que se pretende desagregar esteve ativo (3.14). Estes dados devem ser convertidos para as diferenças entre consumos sucessivos (3.21) e associados os patamares correspondentes (3.24), calculados para os consumos agregados. É, assim, construída a palavra que representa as mudanças de estado para uma assinatura de consumos do equipamento em análise.

O algoritmo de identificação de carga, passa por confrontar os dados agregados de mudanças de estado, e seus patamares correspondentes, com a assinatura do equipamento.

O algoritmo 3.4 avalia, para os consumos agregados, a cada instante, as médias dos patamares (3.24) associados às mudanças de estado (3.21) para o próprio instante considerado, e os instantes que o sucedem, perfazendo uma janela temporal igual à da assinatura das mudanças de estado do equipamento a identificar. Estas médias formam uma lista, que é subtraída à lista representante das médias dos patamares da assinatura do equipamento. É aplicado o módulo a cada ponto da lista resultante desta diferença, e feita a soma de todos os elementos. Este resultado é inserido numa lista de erros (*error*), com igual dimensão da lista dos consumos agregados, no mesmo índice que tem o patamar de partida para o cálculo das diferenças (índices iguais entre a lista *error* e *step_sm* representam instantes iguais). Posto isto, é estabelecido o valor máximo (*max*) que este erro pode tomar. Os índices na lista *error* que não ultrapassam *max* são adicionados a uma lista *index*. Estes índices representam os instantes em que se estima que o equipamento iniciou o seu funcionamento. Partindo desta lista, é possível estimar todos os instantes em que o equipamento funcionou, atribuindo uma etiqueta de ON a esses instantes, e aos restantes OFF. O algoritmo devolve uma lista de valores ON e OFF, em que cada índice representa um instante, e o seu elemento a previsão para o estado do equipamento em análise.

Algoritmo 3.4 Identificação: Assinatura dos Equipamentos

pattern ← lista das médias associadas aos patamares representados pelos caracteres da assinatura das mudanças de estado do equipamento, inclui ∗, que se atribui média associada zero;
step_sm ← lista das médias associadas aos patamares representados pelos caracteres (3.24) associadas às diferenças de consumo (3.22) entre instantes consecutivos (3.21), associadas a todos os consumos brutos agregados registados (3.1), por ordem de ocorrência, inclui elementos da lista com o caracter vazio (∗), que se atribui média associada zero;
function ASSINATURA(pattern,step_sm)
$error = zeros.size(len(step\_sm))$ ← lista de zeros de dimensão igual à dimensão da lista *string_sm*
for i in range(0, len(error)-len(pattern) **do**:
$error[i] = sum(abs(step\_sm[i : i + len(pattern)] - pattern))$
$count = 0$
for i in range(len(error)-len(pattern), len(error)) **do**:
$error[i] = sum(abs(step\_sm[i :] - new\_pattern[count :])$
$count = count + 1$
$max = min(error) + 0.25 * max(error)$ ← erro máximo para num dado instante ser considerado que o equipamento deu início ao seu estado de ativação;
$on\_off = [['OFF'] * len(error)]$
$boolean = error <= max$
$index = where(boolean == True)$
$total\_index = list()$
for i in index **do**:
$total\_index.append(i)$
for j in range(1,len(pattern) **do**:
if i+j **not in** index **then**
$total\_index.append(i + j)$ ← como *index* apenas representa os instantes iniciais de funcionamento é necessário incluir numa lista estes instantes mais os sucessores em número igual a len(pattern)-1;
$on\_of[total\_index]$ ='ON'
return on_off

O algoritmo 3.4 falha em alguns aspetos. A assinatura do equipamento impõe uma duração fixa para o funcionamento do equipamento, e, a comparação entre a assinatura e a série de patamares do *smart meter* não admite o funcionamento de outros aparelhos em paralelo. Para adicionar a estes factos, é de notar que um aparelho não tem

necessariamente de se comportar da mesma forma, ou seja, apesar de trabalhar sempre na mesma gama de valores, a sucessão de estados do aparelho não tem necessariamente de ser constante. Portanto, a utilização de apenas uma assinatura do equipamento acaba por ser uma informação muito pobre para permitir a desagregação [30].

Hidden Markov Model

Uma alternativa identificada para detetar os instantes em que o equipamento esteve em funcionamento, passa por usar cadeias de *Markov*. Para explicar o conceito consideremos a variável (3.11), associada a (3.1), que toma valores ON e OFF. Esta variável caracteriza o estado de um equipamento, se está ativo ou não, a cada registo obtido por uma *smart plug*.

As cadeias de *Markov* caracterizam as transições entre estados, os estados iniciais e os finais, de uma série de estados. Isto é, para o caso em análise, considerando uma amostra de vários dias em que se monitorizou o equipamento com uma *smart plug*, as cadeias de *Markov* caracterizam o estado inicial e final, o estado (3.11) associado ao instante inicial e final de um dia, com uma probabilidade de ocorrência referente a cada valor que a variável pode tomar para estes instantes. Também caracteriza a probabilidade de um valor, num dado instante, no instante seguinte se alterar para outro, ou se manter. Isto é, sabendo que num instante o equipamento esteve ON, qual a probabilidade de no instante seguinte se manter no mesmo estado ou alterar o estado para OFF.

Com isto se concluí que para caracterizar uma cadeia de *Markov* é necessário conhecer a probabilidade de cada um dos estados possíveis se dar no início e no fim da sequência, e a probabilidade de todas as mudanças sucessivas de estado.

Para caracterizar uma cadeia de *Markov* é necessário:

- Uma variável V que pode tomar n valores:

$$V = \{v_1, v_2, ..., v_n\} \tag{3.25}$$

- Uma série de i estados da variável V:

$$S = \Big[s_1, s_2, ..., s_i\Big] \tag{3.26}$$

- Matriz de transição, p.e. p_{12} representa a probabilidade de transição do estado v_1 para o estado v_2:

$$\Big[pt\Big] = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \tag{3.27}$$

Para construir uma cadeia de *Markov* é também necessário caracterizar o estado inicial da sequência, que pode ser caracterizado de duas formas:

- O estado inicial para a cadeia de *Markov* e o estado final, respetivamente:

$$\{s_i, s_f\} \tag{3.28}$$

- Série de probabilidades de cada um dos n valores de V (3.25) surgirem como estado inicial:

$$p_i = \Big[p_1, p_2, ..., p_n\Big] \tag{3.29}$$

Uma cadeia de *Markov* permite assim caracterizar as probabilidades de ocorrência de uma sequência de eventos. Mas, muitas vezes, as sequências de eventos que tem interesse caracterizar não são diretamente observáveis. Para eliminar esta lacuna, introduz-se as cadeias de *Markov* escondidas, *Hidden Markov Model*.

Uma cadeia de *Markov* permite associar aos estados observáveis, estados escondidos, numa relação causa-efeito. Para explicar este conceito, imaginemos que durante um período de tempo não existia registo do estado do equipamento (consumo desagregado do aparelho que se pretende identificar nos consumos agregados), mas, existia informação do consumo agregado para todos os instantes, quer dos instantes em falta, quer dos restantes.

O objetivo das cadeias de *Markov* escondidas, para o caso em estudo, é de estimar a variável estado do equipamento (3.11), para cada instante, sabendo o patamar de mudança de estado (3.24), associada a cada registo (3.1), a cada instante correspondente nos consumos agregados obtidos pelo *smart meter*. Generalizando, dada uma sequência de observações (patamares de mudanças de estado associado ao consumo geral (3.24)), pretende-se descobrir a sequência de estados escondidos correspondente (estado do equipamento) (3.11). O modelo, *Hidden Markov Model*, é caracterizada por (3.25) e (3.27-29), representantes dos estados escondidos, e por:

- Uma variável L que pode tomar m valores:

$$L = \{l_1, l_2, ..., l_m\} \tag{3.30}$$

- Uma série de i observações da variável L:

$$O = \Big[o_1, o_2, ..., o_i\Big] \tag{3.31}$$

- Para cada valor de L, estado observável, a probabilidade (probabilidade de emissão), para um evento, de se obter um estado escondido com um valor pertencente a V:

$$\left[pe\right] = \begin{bmatrix} p_{o_1,v_1} & p_{o_1,v_2} & \cdots & p_{o_1,v_n} \\ p_{o_2,v_1} & p_{o_2,v_2} & \cdots & p_{o_2,v_n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{o_m,v_1} & p_{o_m,v_2} & \cdots & p_{o_m,v_n} \end{bmatrix} \tag{3.32}$$

Com recurso às equações (3.25) e (3.27-32) consegue-se caracterizar a cadeia de *Markov* escondida, para uma relação causa-efeito de duas variáveis, uma observável (L) (3.30) e uma escondida (V) (3.25).

Para um modelo com variáveis escondidas, o objetivo final da sua aplicação é dada uma sequência de observações O (3.31), determinar a sequência mais provável de estados escondidos S (3.26). Este processo é chamado de *decoding*. Para o caso em estudo, pretende-se detetar o estado de funcionamento do equipamento por observação dos patamares correspondentes à mudança de estados nos consumos agregados.

O algoritmo de *decoding* a associar ao nosso modelo é o algoritmo de *Viterbi*. Este algoritmo tem como entradas os parâmetros que caracterizam uma cadeia de *Markov* escondida, e uma sequência de eventos observáveis para a qual se pretende descobrir a sequência de estados escondidos mais provável.

A matriz *viterbi* caracteriza, para toda a série de valores observáveis O (3.31), as probabilidades de ocorrência dos valores escondidos possíveis de associar L (3.30). Estas probabilidades dependem do estado observável a considerar no mesmo instante e do estado escondido associado ao instante anterior. Assim sendo, para o cálculo da matriz *viterbi*, para cada valor L (3.30) possível de associar ao valor observável, registado no instante em análise, é considerado que no instante anterior se deu o valor L (3.30) que associa uma maior probabilidade de transição para o valor considerado. O índice do valor anterior em (3.30), para cada valor possível de associar (3.30) ao estado observável, que confere menor probabilidade, é guardado na matriz *breakpointer*. Tendo a matriz

determinada, é analisada segundo ordem inversa das ocorrências dos valores observáveis. Determinando o valor escondido associado ao último valor observável, analisando *breakpointer* para a respetiva posição, sabe-se o valor escondido que se antecede, e por aí em diante, até se chegar ao início da série de observações, com uma série de estados escondidos mais provável de ocorrer associada.

O algoritmo 3.5 devolve a sequência mais provável de estados escondidos associados aos estados observáveis.

Algoritmo 3.5 Identificação: *Hidden Markov Model*

$p_i \leftarrow$(3.29) de dimensão n;
$p_t \leftarrow$ (3.27) matriz de dimensão nxn;
$p_e \leftarrow$ (3.32) matriz de dimensão mxn;
$O \leftarrow$ (3.31) de dimensão i;
$V \leftarrow$ (3.25) de dimensão n;
$L \leftarrow$ (3.30) de dimensão m;
function VITERBI(p_i, p_t, p_e, O, V, L)
    $viterbi = zeros.size(i, n) \leftarrow$ matriz de zeros com i linhas e n colunas;
    $breakpointer = zeros.size(i, n) \leftarrow$ matriz de zeros com i linhas e n colunas;
    $path = zeros.size(i) \leftarrow$ lista de zeros de tamanho i;
    $hiddenstates = list() \leftarrow$ lista vazia;
    $viberbi[0, :] = pi * pe[L.index(O[0]), :] \leftarrow$ probabilidade de emissão, para todos os valores de V escondidos, para o primeiro valor observável na série O, multiplicadas pela probabilidade desses mesmos valores se registarem no estado inicial;
    $breakpointer[0, :] = zeros.size(n) \leftarrow$ lista de zeros de tamanho n;
    for v_O in range(1,i) **do**:
        for v_H in range(0,n) **do**:
            $iteration = [viberbi[v_O - 1, k] * p_{k,v_H} * p_{O_{V_O},V_{V_H}}, for\ k\ in\ range(0,n)]$
            $viberbi[v_O, v_H] = max(iteration) \leftarrow$ máximo da lista *iteration*;
            $breakpointer[v_O, v_H] = argmax(iteration) \leftarrow$ índice do máximo da lista *iteration*;
    $path[i] = argmax(viberbi[i, :]) \leftarrow$ Índice em (3.30) do valor escondido associado ao ultimo valor observável na série O (3.31);
    for t in range(i-1,-1,-1) **do**:
        $path[t] = breakpointer[t + 1, path[t + 1]$
    hiddenstates[:]=[V(l) for l in path]
return *hiddenstates*

Para o caso em estudo, a base de dados para o período em que se conhece os estados

escondidos associados aos observáveis, deve servir de aprendizagem, para construção dos parâmetros caracterizantes da cadeia de *Markov* escondida. Posto isto, para os dados em falta, quando existe apenas a série observável é possível estimar o estado do equipamento [31] [32].

Confusion Matrix

Para testar a exatidão das duas abordagens apresentadas, para a identificação do estado do equipamento, a partir dos dados do *smart meter*, para os mesmos instantes, e sabendo o resultado real do equipamento, pode-se comparar a eficiência dos algoritmos com recurso à *confusion matrix*.

Para descrever a *confusion matrix*, consideramos o exemplo das cadeias de *Markov* escondidas, a série real dos estados escondidos S (3.26), e uma previsão para estes estados, *hiddenstates*, obtidos pelo algoritmo 3.5. Os mesmos índices, de ambas as séries, correspondem aos mesmos instantes. Cada uma das séries toma valores pertencentes a V (3.25).

$$\Big[\textit{confusion matrix}\Big] = \begin{bmatrix} \#_{v_1|v_1} & \#_{v_1|v_2} & \cdots & \#_{v_1|v_n} \\ \#_{v_2|v_1} & \#_{v_2|v_2} & \cdots & \#_{v_2|v_n} \\ \vdots & \vdots & \ddots & \vdots \\ \#_{v_n|v_1} & \#_{v_n|v_2} & \cdots & \#_{v_n|v_n} \end{bmatrix} \tag{3.33}$$

A *confusion matrix* (3.33) compara os valores previstos com os valores reais. Cada linha corresponde a um valor previsto e cada coluna a um valor real. Cada índice de (3.33) representa o número de ocorrências do valor previsto, correspondente à sua linha, sabendo que ocorreu o valor correspondente à sua coluna.

A *confusion matrix* (3.33) mostra-se assim um ótimo recurso para avaliar os resultados de uma previsão [33].

A matriz (3.33), aplicada ao caso da identificação do estado do equipamento, reduz-se a uma matriz 2x2. Seja o valor ON substituído por positivo e o OFF por negativo, nos valores da variável do estado do equipamento.

Quando (3.33) toma a dimensão 2x2, introduz-se os conceitos de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos para a previsão do valor de uma variável, bem como a exatidão dos resultados que é geral para todas as dimensões.

- **Verdadeiro positivo**: Valor previsto e real da variável são ambos positivos;
- **Falso positivo**: Valor previsto para a variável é positivo, mas o real é negativo;
- **Falso negativo**: Valor previsto para a variável é negativo, mas o real é positivo;
- **Verdadeiro negativo**: Valor previsto e real da variável são ambos negativos;

Real / Previsão	**Positivo**	**Negativo**
Positivo	# Verdadeiros Positivos(VP)	# Falsos Positivos(FP)
Negativo	# Falsos Negativos(FN)	# Verdadeiros Negativos(VN)

Tabela 3.2: *Confusion Matrix* de 2x2 quando *V={positivo,negativo}* (3.25)

A tabela (3.2) permite-nos então avaliar os resultados do algoritmo por meio do cálculo dos seguintes rácios:

- Frequência de acertos na previsão para o valor positivo:

$$f_{VP} = \frac{VP}{VP + FN} \tag{3.34}$$

- Frequência de acertos na previsão para o valor negativo:

$$f_{VN} = \frac{VN}{FP + VN} \tag{3.35}$$

- Exatidão dos resultados:

$$accuracy = \frac{VP + VN}{VP + FP + FN + VN} \tag{3.36}$$

Com isto obtem-se três equações (3.34), (3.35) e (3.36) que avaliam de forma completa os resultados dos algoritmos de identificação de equipamentos nos consumos agregados. Podemos, assim, averiguar se os algoritmos expostos são eficiêntes na desagregação de carga, e também qual deles devolve melhores resultados [34].

Capítulo 4

Estudo de Caso

Este capítulo pretende aplicar os algoritmos descritos no capítulo 3, por forma a cumprir os objetivos propostos na secção 1.3, a bases de dados de consumos elétricos agregados e desagregados.

A forma como os dados de consumo agregados e desagregados são adquiridos é descrita em 3.1. Também no capítulo 3, logo no primeiro parágrafo, são expostas as bases de dados para análise. São elas um conjunto de consumos desagregados, registados por diversas *smart plugs*, instaladas numa empresa piloto, VPS, e consumos agregados (monitorizados por um *smart meter*) e desagregados (*smart plugs*) de uma casa piloto.

Bases de Dados	
Empresa Piloto	**Casa Piloto**
Consumos Desagregados	
Impressora Ilha de Computadores Máquina de café Refrigerador de água	Frigorífico Máquina de Lavar Roupa
Consumos Agregados	
Não	Sim

Tabela 4.1: Tipos de consumos analisados.

Na tabela 4.1, tem-se então descrito o que é monitorizado em cada edifício em análise. A janela temporal em análise é igual para todos os equipamentos sob monitori-

zação. É estipulada entre março e novembro de 2017, inclusive. Os dados são registados em ambos os modos de aquisição, modo base ($\Delta t = 5$ minutos) e modo horário ($\Delta t = 60$ minutos). Os consumos elétricos, a cada registo, correspondem ao somatório da energia consumida, no intervalo de tempo de aquisição antecedente ao registo.

4.1 Tratamento de Dados

Os dados de consumo bruto, antes de qualquer análise, como referido na secção 3.2, devem ser tratados. São eliminados eventos (3.1) registados em duplicado, e previstos aqueles que estão em falta, para a janela temporal considerada.

Nesta secção, pretende-se avaliar as previsões feitas para os dados em falta.

Primeiramente, para cada monitorização, para o edifício em análise (tabela 4.1), é preciso perceber qual é o tipo de falhas registadas, se são falhas de comunicação (secção 3.2.1) ou falhas de medição (secção 3.2.2). Para avaliar a exatidão dos algoritmos 3.1 e 3.2, de previsão de falhas, criam-se falsas falhas nos dados, para se comparar a previsão para estes eventos com os eventos reais registados.

Pelo que se conseguiu apurar, por análise direta dos instantes posteriores às falhas registadas, para todas as monitorizações, verifica-se apenas falhas de comunicação, com exceção de um equipamento monitorizado na empresa piloto, que verifica apenas falhas de medição. De notar que os registos após falhas, em falhas de baixa duração ou em períodos em que não há consumos significativos, podem conduzir a um apuramento errado do tipo de falhas, no entanto, e não tendo informação adicional que sustente o tipo de falhas, assume-se a generalização.

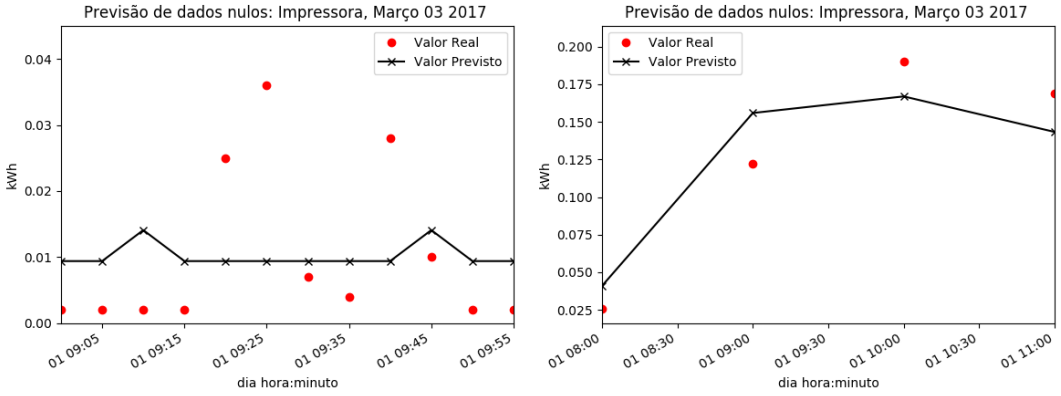
4.1.1 Empresa Piloto

Empresa Piloto-Algoritmos de Previsão de Nulos	
Consumos Desagregados	
Impressora	Algoritmo 3.1
Ilha de computadores	Algoritmo 3.2
Máquina de café	Algoritmo 3.1
Refrigerador de água	Algoritmo 3.1
Consumos Agregados	
Não	**X**

Tabela 4.2: Algoritmos de previsão de nulos a associar a cada monitorização, para a empresa piloto.

Estando identificado o algoritmo a aplicar, para previsão de nulos, aos registos de consumos de cada dispositivo sob monitorização, pode seguir-se para a sua aplicação. Para testar os resultados dos algoritmos, criam-se falhas nos dados, para todas as monitorizações no edifício em estudo, no dia 1/3/2017, entre as 9h:00m e as 9h:55m, para o modo de monitorização base, e entre as 8h:00m e as 11h:00m, para o modo horário. Sendo um edifício do tipo empresarial, espera-se que a hora típica de início de atividade, para um dia laboral, seja por volta das 9h:00m, portanto, para equipamentos que dependam do uso direto, por parte de um utilizador, é expectável uma tendência de subida nos consumos elétricos, para estas janelas temporais consideradas. Tome-se como dia laboral um dia útil. O dia escolhido para análise é um dia laboral, em que se estima que exista atividade no edifício entre as 9 e as 19 horas.

Impressora

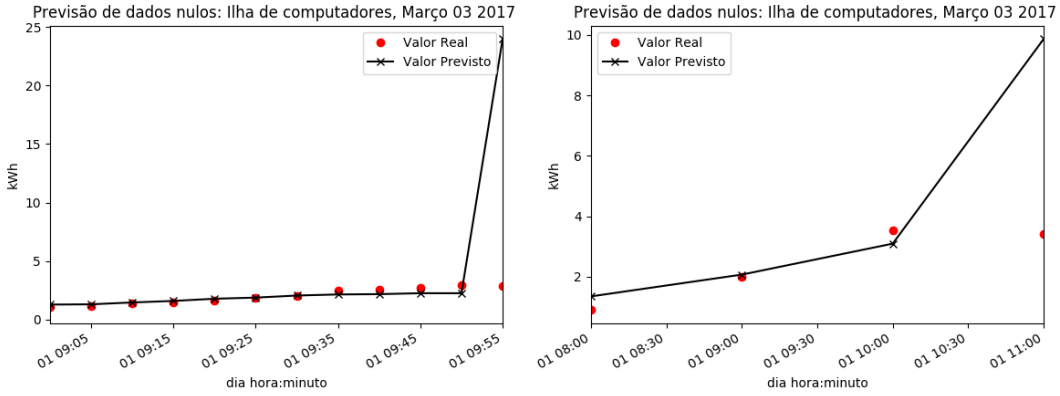


a) Modo de aquisição base; b) Modo de aquisição horário;

Figura 4.1: Previsão de nulos para uma impressora.

Pelo tipo de aparelho em análise, espera-se que ele seja usado de forma aleatória no tempo, durante o período laboral, mas com consumos totais diários não muito variantes. Dado o tipo de uso que o aparelho tem, é de esperar que as previsões para o modo base não tenham uma grande exatidão, contudo, no modo horário, visto o intervalo de tempo entre registos ser maior, é de esperar uma maior exatidão na previsão. Para o modo horário, a figura 4.1 b) mostra uma tendência de subida ao longo do tempo, do consumo do equipamento, que é prevista com sucesso. Esta tendência resulta do facto da janela temporal em análise, para o modo horário, compreender a transição de inatividade para atividade do edifício.

Ilha de Computadores



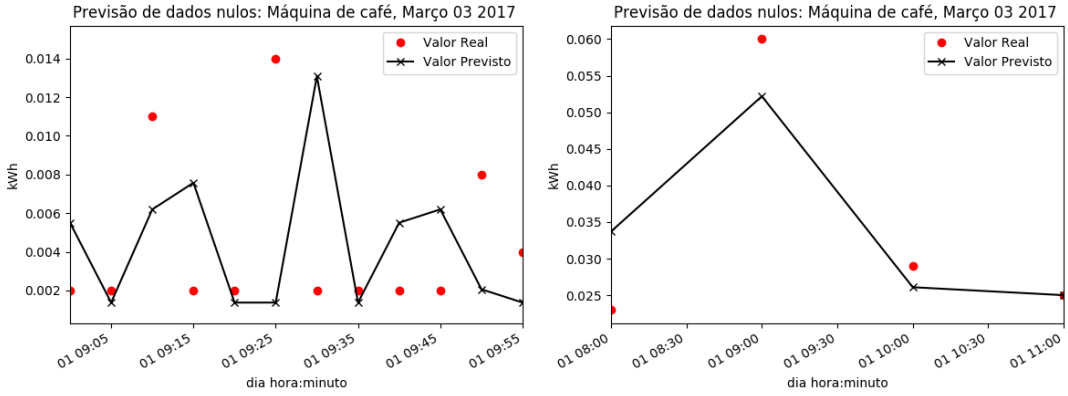
a) Modo de aquisição base; b) Modo de aquisição horário;

Figura 4.2: Previsão de nulos para uma ilha de computadores.

A ilha de computadores, que é um conjunto de computadores em uso pelos trabalhadores da empresa piloto, apresenta um consumo muito constante no tempo. Portanto, quer para o modo base, quer para o modo horário, são obtidos bons resultados para a previsão de nulos. No entanto, é previsto, erradamente, um pico de consumo no final de ambas as janelas temporais.

Também para este equipamento, é prevista a tendência de subida, em ambos os modos de aquisição.

Máquina de Café

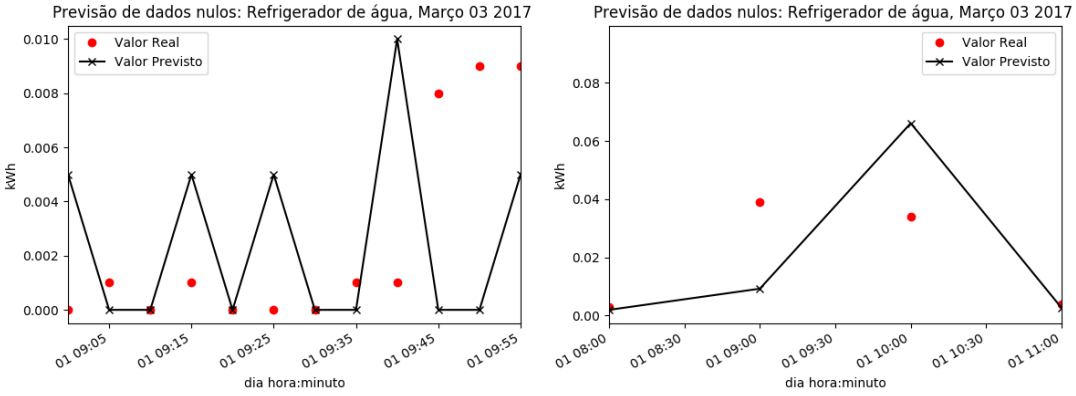


a) Modo de aquisição base; b) Modo de aquisição horário;

Figura 4.3: Previsão de nulos para uma máquina de café.

Por generalização, os hábitos de consumo de café, em termos horários, são tipicamente ao início do dia e no final de uma refeição. Portanto, não é de estranhar que seja expectável um pico de consumo elétrico, da máquina de café, perto das 9h:00m, hora de início de atividade do edifício. Esta tendência é prevista para os dados de consumo horário. Já para os dados de consumo base, é mais difícil prever os consumos com exatidão, visto que, a máquina de café trabalha em instantes muito curtos no tempo e é usada em instantes aleatórios, embora dentro das mesmas gamas temporais.

Refrigerador de Água



a) Modo de aquisição base; b) Modo de aquisição horário;

Figura 4.4: Previsão de nulos para um refrigerador de água.

Por análise da figura 4.4, no modo base, é notória a previsão de um funcionamento cíclico. Para os dados reais registados, novamente no modo base, verifica-se também uma tendência cíclica, de muito menor amplitude que a prevista, no início da janela temporal, e um grande pico, de elevada largura, no fim da janela temporal.

Este equipamento é de difícil previsão exata. A sua atividade não depende da atividade do edifício. O consumo vai variando ao longo do tempo, embora sempre dentro das mesmas gamas de consumo. Um aparelho de refrigeração não depende consideravelmente da sua utilização, trabalha tipicamente em ciclos de arrefecimento, em que a temperatura registada é ajustada à temperatura de referência [35]. Portanto, neste tipo de equipamentos, a previsão por comparação de instantes equivalentes não é eficiente para a previsão de nulos.

Justifica-se o uso de um algoritmo de previsão de nulos, para este equipamento, que traduza o seu funcionamento numa função periódica. Ajustando os dados brutos a esta função, torna-se possível saber a fase do período em que o consumo se encontra imediatamente antes da falha e prever os consumos seguintes.

4.1.2 Casa Piloto

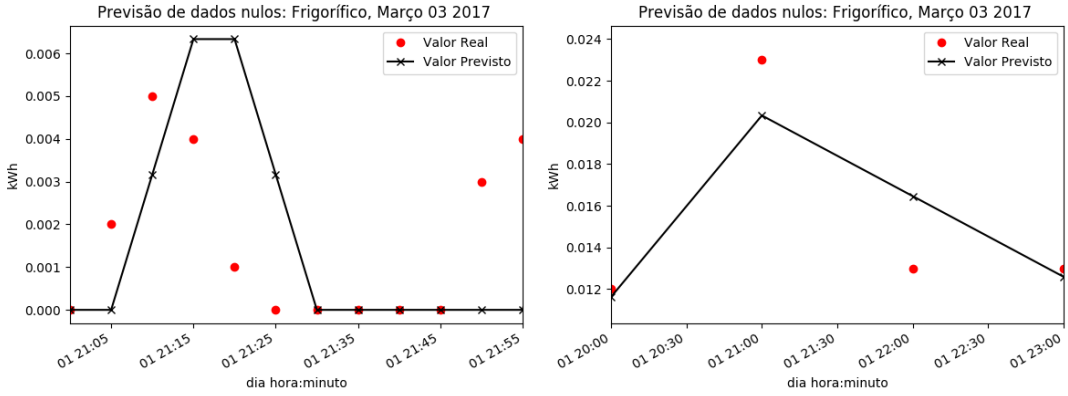
Para a casa piloto, são identificados os tipos de algoritmos a aplicar, a cada monitorização, para previsão dos eventos (3.1) em que não há registos, expostos na tabela 4.3.

Casa Piloto-Algoritmos de Previsão de Nulos	
Consumos Desagregados	
Frigorífico	Algoritmo 3.1
Máquina de Lavar Roupa	Algoritmo 3.1
Consumos Agregados	
Sim	Algoritmo 3.1

Tabela 4.3: Algoritmos de previsão de nulos a associar a cada monitorização, para a casa piloto.

Após a identificação do tipo de nulos que surgem a cada monitorização, as previsões efetuadas devem ser avaliadas. Para o efeito, criam-se falsos nulos nas bases de dados e comparam-se com os valores reais registados. Para dados base, registados de cinco em cinco minutos, cria-se uma falha entre as 21h:00m e as 21h:55m, inclusive, para o dia 1/3/2017. Para os registos horários, a falha é criada para o mesmo dia, mas entre as 20h:00m e as 23h:00m, inclusive. Estas janelas temporais são assim definidas porque correspondem a períodos em que normalmente existe atividade nas residências.

Frigorífico



a) Modo de aquisição base; b) Modo de aquisição horário;

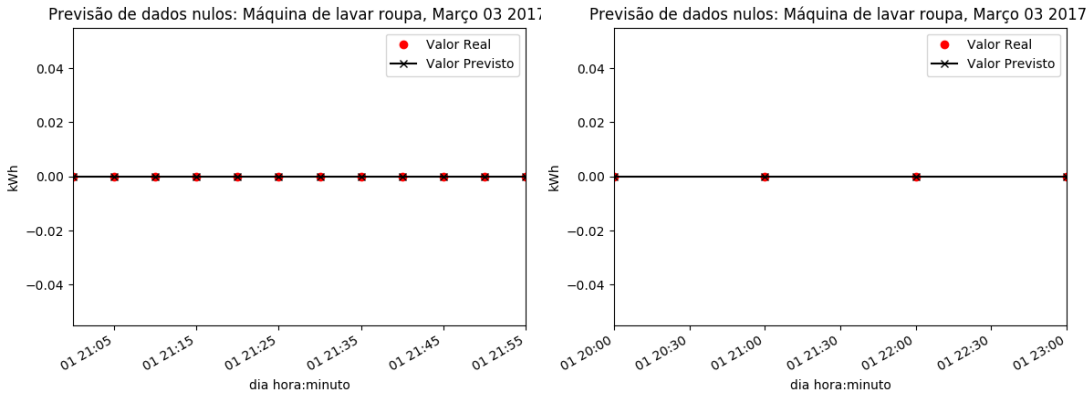
Figura 4.5: Previsão de nulos para um frigorífico.

Visto se tratar de um equipamento de refrigeração, um frigorífico trabalha em ciclos de arrefecimento, tal como o refrigerador de água referido na secção 4.1.1. Portanto, as conclusões obtidas nessa secção, para o equipamento de refrigeração de água, são igualmente válidas para o frigorífico em análise. No modo base, é previsto um possível período de um funcionamento cíclico, mas, nem a fase, nem a amplitude do sinal, correspondem com o ciclo de funcionamento real registado. No modo horário, é previsto, com sucesso, um pico de consumo às 21 horas. Talvez resultado da atividade de fazer o jantar, que pode obrigar a abrir algumas vezes o frigorífico e assim aumentar a sua temperatura interior. Por resultado, a amplitude dos ciclos de arrefecimento será superior nesta hora, para que se possa repor a temperatura de referência.

Apesar de um frigorífico poder consumir mais energia quando a porta esta aberta muito tempo, é muitas vezes aberta, é inserido algo quente no seu interior ou está um dia quente, este consumo extra é inserido nos ciclos de funcionamento periódicos e não no exato momento que se dá o consumo. O que acontece é que a amplitude do ciclo seguinte, ao evento que gera maior consumo, é também maior. Portanto, a adicionar à sugestão de algoritmo de previsão indicado na secção 4.1.1, para o refrigerador de água,

deve ser adicionada também uma modelação da amplitude, de acordo com a fase do dia, altura do ano e outros fatores externos, possíveis de medir, que possam influênciar o consumo do equipamento [35].

Máquina de Lavar Roupa

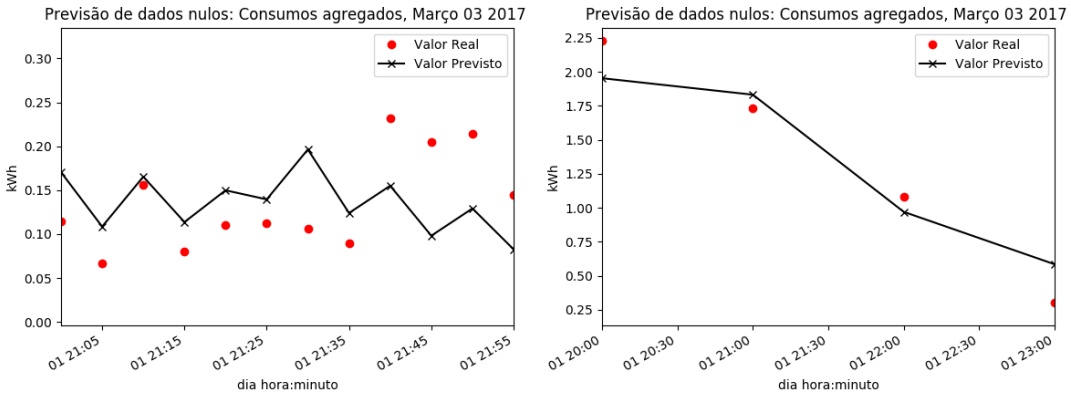


a) Modo de aquisição base; b) Modo de aquisição horário;

Figura 4.6: Previsão de nulos para uma máquina de lavar roupa.

É muito comum as máquinas de lavar roupa serem usadas em horário diurno, uma vez que, são aproveitadas as horas de sol para a secagem ao ar livre. A adicionar a este facto, é um aparelho de uso pontual ao longo de uma semana. Por consequência, não é de estranhar que não seja previsto nem detetado nenhum evento para as horas em análise na figura 4.6. O mais provável de acontecer, caso este equipamento não seja usado sempre dentro dos mesmos horários, é ser quase sempre detetada inatividade do aparelho, quando este está ativo.

Consumos Agregados



a) Modo de aquisição base; b) Modo de aquisição horário;

Figura 4.7: Previsão de nulos para os consumos agregados.

Neste ponto, são monitorizados todos os consumos efetuados na habitação, para cada instante, de forma agregada. Como já referido, é de esperar que haja atividade dento da janela de tempo considerada para análise. Visto se estar a monitorizar uma série de equipamentos, não é de esperar que existam grandes flutuações de consumo, entre instantes correspondentes. Variações de consumo de um aparelho entre iguais instantes, para o mesmo dia da semana, mas em semanas distintas, são mascaradas pelos consumos dos restantes equipamentos. Com isto, não é de estranhar que as previsões apresentadas na figura 4.7, para os consumos agregados, tenham bons resultados ao nível da sua exatidão, para ambos os modos de aquisição.

4.2 Padrões e Rotinas de Consumos Elétricos

Nesta secção, são analisados os dados do consumo elétrico, na janela temporal referida. Os dados são analisados, tanto no modo base como no modo horário, visto que, cada um dos modos pode revelar informação que no outro não está tão visível. No entanto, quando se analisa apenas um modo de aquisição, por o restante ser redundante,

recorre-se sempre ao modo de aquisição com uma janela temporal da ordem do tempo de funcionamento ativo do aparelho sob monitorização.

Para representação e análise dos dados, nesta secção, apenas são expostos (figuras) resultados para três meses, representantes de cada estação do ano, março, agosto e novembro.

Os algoritmos aplicados para a descoberta de padrões, são maioritariamente autónomos, à exceção de dois processos. O processo de escolha do patamar que divide os consumos não ativos dos ativos, e a descoberta de rotinas de consumos, para cada monitorização.

A descoberta do patamar de consumo é obtida pela observação direta dos consumos sujeitos a análise. É gerado um patamar, e ajustado, até se obter uma divisão ideal daquilo que é o consumo de fundo e o consumo que representa atividade do equipamento.

Para a descoberta de rotinas, é atribuído um peso e uma dimensão ao algoritmo (ver secção 3.5).

	Dimensão (b/h)	**Peso (b/h)**
	Empresa Piloto	
Impressora	30/4	150/150
Ilha de Computadores	12/2	10/20
Máquina de café	30/4	150/150
Refrigerador de água	12/2	10/20
	Casa Piloto	
Frigorífico	12/2	10/20
Máquina de lavar roupa	*	*
Consumos agregados	12/2	10/20

Tabela 4.4: Peso e dimensão atribuídos ao algoritmo de descoberta de rotinas (secção 3.5), ao longo de todas as monitorizações, no modo de aquisição base/horário (b/h).

Na tabela 4.4, pode-se concluir um padrão das dimensões e pesos atribuídos. A impressora e a máquina de café apresentam parâmetros iguais, de elevada dimensão e peso. Isto acontece porque estes aparelhos têm um tempo de funcionamento bastante curto, da ordem do intervalo de tempo de aquisição base (5 minutos), e o seu uso, no curso do tempo, é aleatório dentro de uma gama temporal, isto é, por exemplo, a máquina de café é usada normalmente entre as 8 e as 9 da manhã, mas nesse intervalo de tempo, é aleatório o instante em que funciona.

Para a ilha de computadores, refrigerador de água, frigorífico e consumos agregados, a dimensão e o peso atribuídos, para a descoberta de rotinas, são bastante mais baixos. Tal facto, deve-se a todas estas monitorizações apresentarem uma frequência de estados ativos bastante elevada, ao longo do tempo.

A máquina de lavar roupa não tem parâmetros associados, está assinalada com um $*$, pois, para a identificação de rotinas deste equipamento, apenas é considerado se num certo dia a máquina teve pelo menos um estado ativo, ou não, sendo este o único critério de classificação dos dias, em dias tipo de consumo. A classificação é feita deste modo por a máquina de lavar roupa apresentar uma frequência muito baixa de estados ativos, ao longo do tempo, e sem uma hora fixa de arranque.

O conhecimento apresentado, para sustentar a escolha de dimensões e pesos atribuídos, irá ser sustentado nas próximas secções.

Interessa adicionar à base de dados uma variável temperatura, para compreender a forma como esta influência os consumos elétricos. Para isso, é registado para cada dia presente na base de dados, a temperatura máxima verificada nesse dia. A distribuição de todas as temperaturas registadas, na janela temporal considerada, pode ser consultada na figura 4.8.

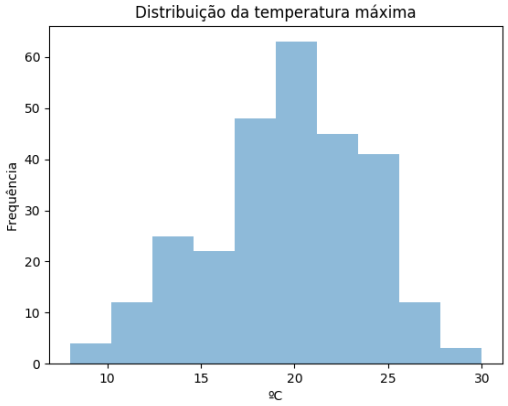


Figura 4.8: Distribuição da temperatura máxima, ao longo da janela temporal em análise.

4.2.1 Empresa Piloto

Estima-se que a empresa piloto, tipicamente, tenha atividade nos dias úteis (dias laborais), segunda a sexta-feira, aproximadamente entre as 9 e as 19 horas, e a pausa para o almoço entre as 13 e as 14 horas.

Impressora

Como descrito no capítulo 3, após se ter tratado os dados brutos, eliminado duplicados e tendo as previsões para os nulos feitas, segue-se para a discretização da base de dados.

Para as séries de consumos da impressora, para o modo base e o modo horário, obteve-se os resultados apresentados na tabela 4.4, para a divisão dos dados brutos em modo ON e modo OFF (secção 3.3.1), usando como critério de divisão 1/10 do valor máximo registado, não *outlier*.

	Mínimo	Máximo	Média
ON	0.007	0.066	0.027
OFF	0.000	0.007	0.002

a) Modo de aquisição base, valores em kWh;

	Mínimo	Máximo	Média
ON	0.036	0.342	0.144
OFF	0.000	0.035	0.020

b) Modo de aquisição horário, valores em kWh;

Tabela 4.5: Divisão dos valores de consumo bruto (3.1) em patamares de consumo ON e OFF (3.11), para a impressora.

É visível na tabela 4.5 que a gama de valores em modo ON, no modo de aquisição horário b), é bastante superior (uma ordem de grandeza) à gama de valores ON adquirida no modo base a). Tal facto, é resultado dos dados registarem o consumo acumulado, e de, provavelmente, a impressora estar a funcionar num período de tempo considerável (no modo base), durante a hora em que é acusado estado ON.

Posto isto, segue-se para a discretização, por patamares, dos valores (3.1) com o valor ON (3.11) atribuído (secção 3.3.2). É de relembrar que a cada patamar é atribuída uma letra, para posteriormente, se aplicar o modelo do pacote de palavras.

	Mínimo	Máximo	Média
A	0.007	0.043	0.022
B	0.044	0.059	0.051
C	0.064	0.066	0.065

a) Modo de aquisição base, valores em kWh;

	Mínimo	Máximo	Média
A	0.036	0.153	0.098
B	0.154	0.283	0.204
C	0.313	0.342	0.328

b) Modo de aquisição horário, valores em kWh;

Tabela 4.6: Divisão dos valores de ON (3.11), em patamares de consumo, para a impressora.

A tabela 4.6 mostra que o patamar A, em ambos os modos de aquisição, concentra a maioria do intervalo da gama de valores que os consumos podem tomar. Isto pode significar que existe uma forte prevalência deste patamar, relativamente aos restantes. Com os patamares da tabela 4.6 identificados, é possível construir as palavras representativas dos períodos ON (3.14), pelo modelo do pacote de palavras (secção 3.3.2). Posto isto, como descrito em 3.4, as palavras são distribuídas por *clusters*, com cada *cluster* a ser representada pelo seu centro.

	Centro de cada *cluster*	
	Dados base	**Dados horários**
C1	A^{263} (0.032%)	$B^{22}A^3$ (0.261%)
C2	A^{36} (0.129%)	AB (16.710%)
C3	A^{25} (1.383%)	A^2BA^2(47.520%)
C4	A^8 (98.456%)	$A^2B^2A^3BA^2$(35.509%)

Tabela 4.7: Centro de cada *cluster* e respetiva percentagem de ocorrência, para a impressora (# 3109 - número total de palavras detetadas no modo base; # 383 - número total de palavras detetadas no modo horário).

A tabela 4.7 apresenta os centros de cada *cluster*, com a percentagem de ocorrência de cada uma, para os dois modos de aquisição. Verifica-se que para o modo base, as *clusters* C1, C2 e C3 ocorrem em número pouco significativo. Apenas a *cluster* C4 é representativa de um padrão. No modo horário apenas C1 é incomum.

É notória a diferença de duração de uma série de estados ON (3.14), representada por uma palavra, pertencente a uma *cluster* frequente, entre o modo base e o modo horário. De relembra que no modo base, um caracter, no modelo de palavras, representa cinco minutos de operação do aparelho, no modo horário representa uma hora. Esta diferença deve-se ao facto dos dados apresentarem consumos acumulados, ocorridos no intervalo de tempo de aquisição. Caso o aparelho esteja em operação, durante dez minutos numa hora, e outros dez na hora seguinte, enquanto que no modo de aquisição base são detetados dois padrões de dez minutos, no modo horário é identificado um único padrão de duas horas.

Conclui-se então que, é comum a impressora trabalhar 40 minutos consecutivos, no patamar A de consumo, no modo base. Então, é de esperar que exista atividade na impressora entre duas a dez horas consecutivas, não necessariamente em toda a hora considerada, durante uns minutos (≈40 minutos), a cada hora. Por consequência, o número total de palavras detetadas no modo horário (#383) é muito inferior às detetadas no modo base (#3109).

Tendo os padrões de consumo identificados, passa-se para a identificação de rotinas, como descrito na secção 3.5. Como indicado na secção referida, a identificação de

rotinas requer a identificação de um peso e uma dimensão. Para o modo base, a dimensão estabelecida é de 30 e o peso de 150, para o modo horário a dimensão é de 4 e o peso novamente de 150, como se pode verificar na tabela 4.4. Foram identificadas várias rotinas, categorizadas de 'A' a 'F', por onde se distribuíram os dias da semana, ao longo dos meses.

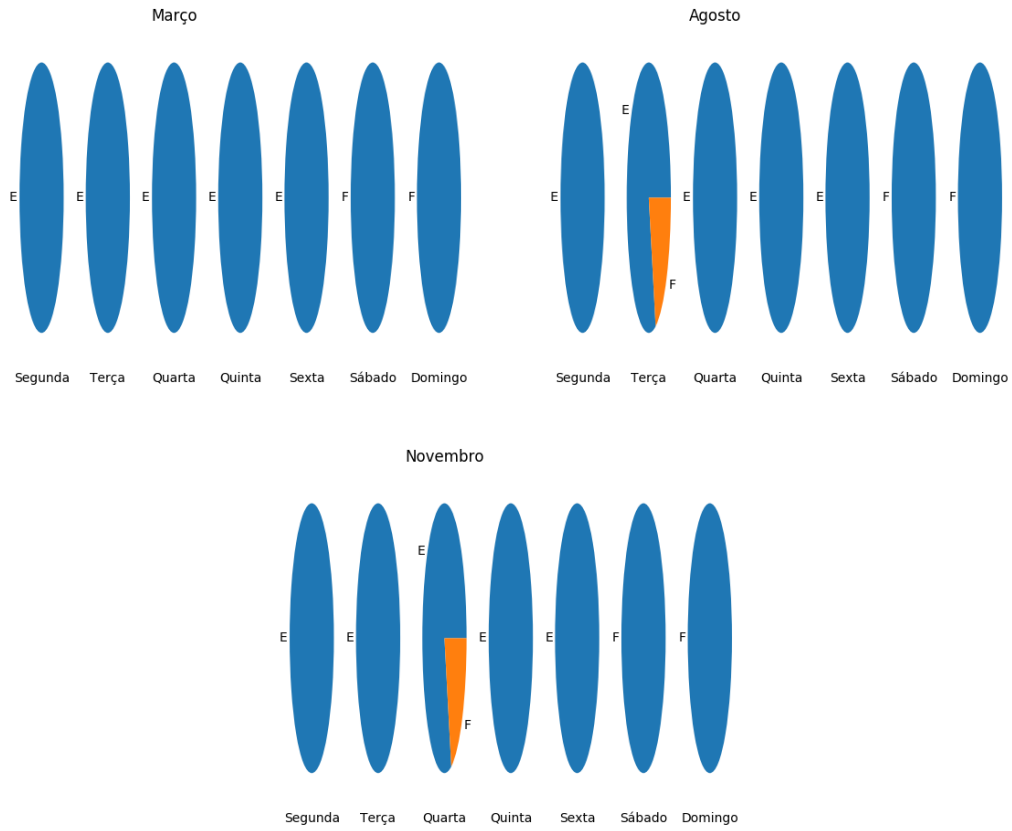


Figura 4.9: Representação da distribuição dos dias tipo, pelos dias da semana, no modo base, para a impressora.

É possível verificar, na figura 4.9, uma predominância do dia tipo 'E' para os dias úteis e do dia 'F' para o fim-de-semana.

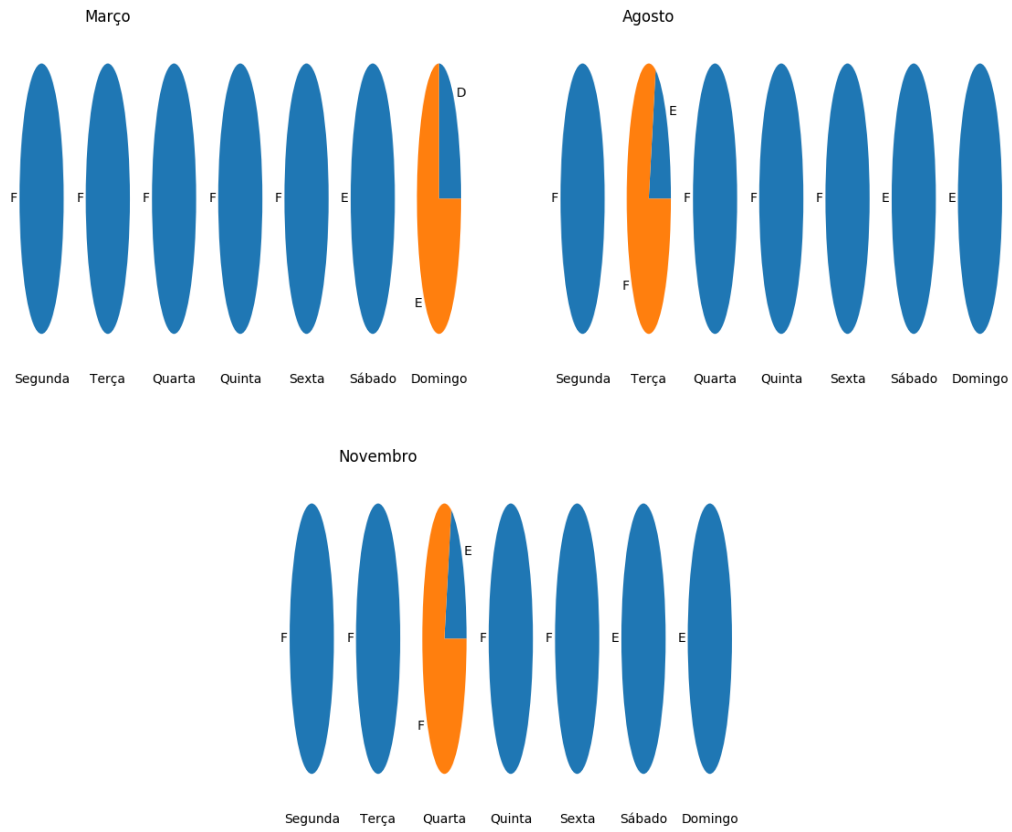


Figura 4.10: Representação da distribuição dos dias tipo, pelos dias da semana, no modo horário, para a impressora.

Na figura 4.10, é possível também notar uma divisão clara entre os dias úteis e os fins-de-semana, com atribuição do dia tipo 'F' para os dias úteis e 'E' para os fins-de-semana. Interessa analisar apenas estes dois dias tipo, que se apresentam em predominância, os restantes dias tipo são considerados *outliers*.

As figuras 4.9 e 4.10, representantes das distribuições das classificações dos dias tipo, apresentam informação redundante, portanto, apenas interessa analisar a distribuição destes dias tipo, aos longo dos meses, num dos modos de aquisição.

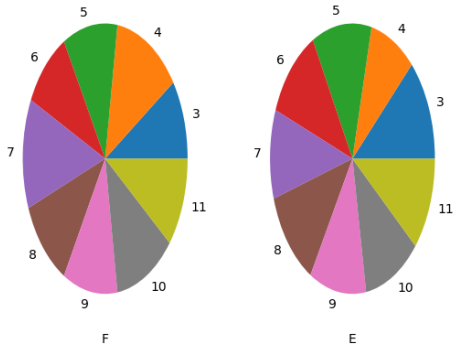


Figura 4.11: Representação da distribuição dos dias tipo relevantes, ao longo dos meses da janela temporal, no modo base, para a impressora.

A figura 4.11, permite ver a distribuição, em número, dos dias tipo ao longo dos meses. A distribuição é aproximadamente uniforme, ou seja, a classificação destes dias tipo não depende do mês em questão.

Sabendo os dias tipo relevantes, pretende-se agora representar o consumo elétrico, ao longo de um dia, para cada dia tipo relevante, para cada mês, em ambos os modos de aquisição.

Nas figuras seguintes 4.12 e 4.13 são representados os dias 'E' e 'F', para cada modo de aquisição. Para cada mês, são selecionados para representação os dias tipo 'E' e 'F' que apresentam uma menor variação relativamente aos mesmos dias tipo, no mesmo mês. A linha horizontal azul representa a divisão entre o consumo ON e OFF, as restantes linhas horizontais correspondem aos limites máximos dos patamares, dentro da gama de valores dos dados representados.

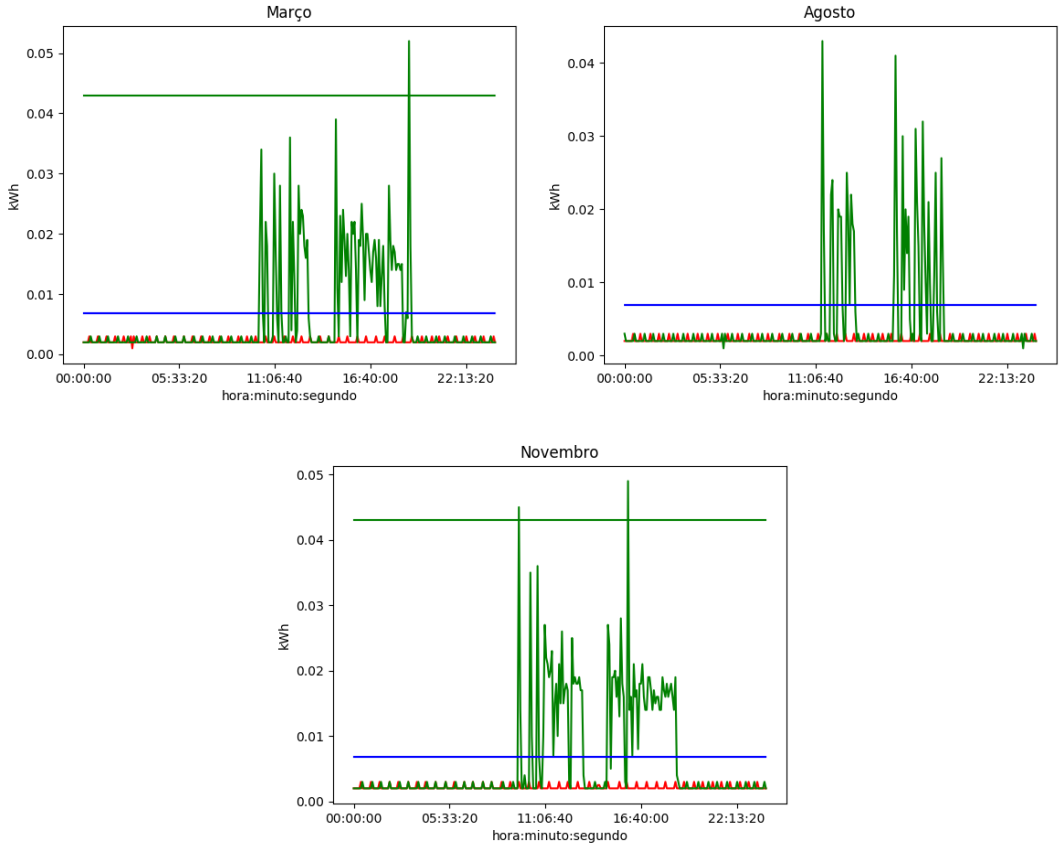


Figura 4.12: Representação, em modo base, dos dias tipo 'E' a verde e 'F' a vermelho para a impressora.

Por análise da figura 4.12, verifica-se que os consumos em modo ON, para o dia tipo 'E' (classificador de dias úteis), que apresentam consumos acima da linha horizontal azul (patamar ON/OFF), encontram-se maioritariamente abaixo do valor máximo do patamar A, linha horizontal verde, o que não é de estranhar, visto que o centro da *cluster*, em modo base, com maior prevalência (tabela 4.7), é constituído por oito A's consecutivos. Também se verifica que, para estes dias tipo, que correspondem maioritariamente a dias úteis, um consumo acentuado dentro do período da manhã e do período da tarde, havendo uma grande densidade de intersecções com a linha horizontal azul nestes períodos, com pausa na hora de almoço. Esta grande densidade de intersecções revela que os consumos da impressora variam muito, num curto espaço de tempo,

aproximadamente da ordem do intervalo de tempo de aquisição. Ou seja, a impressora funciona vastas vezes, mas, em curtos períodos de tempo.

Os consumos de um dia tipo 'F', que corresponde maioritariamente aos fins-de-semana, são residuais, estando a linha que os caracteriza sempre abaixo da linha horizontal azul.

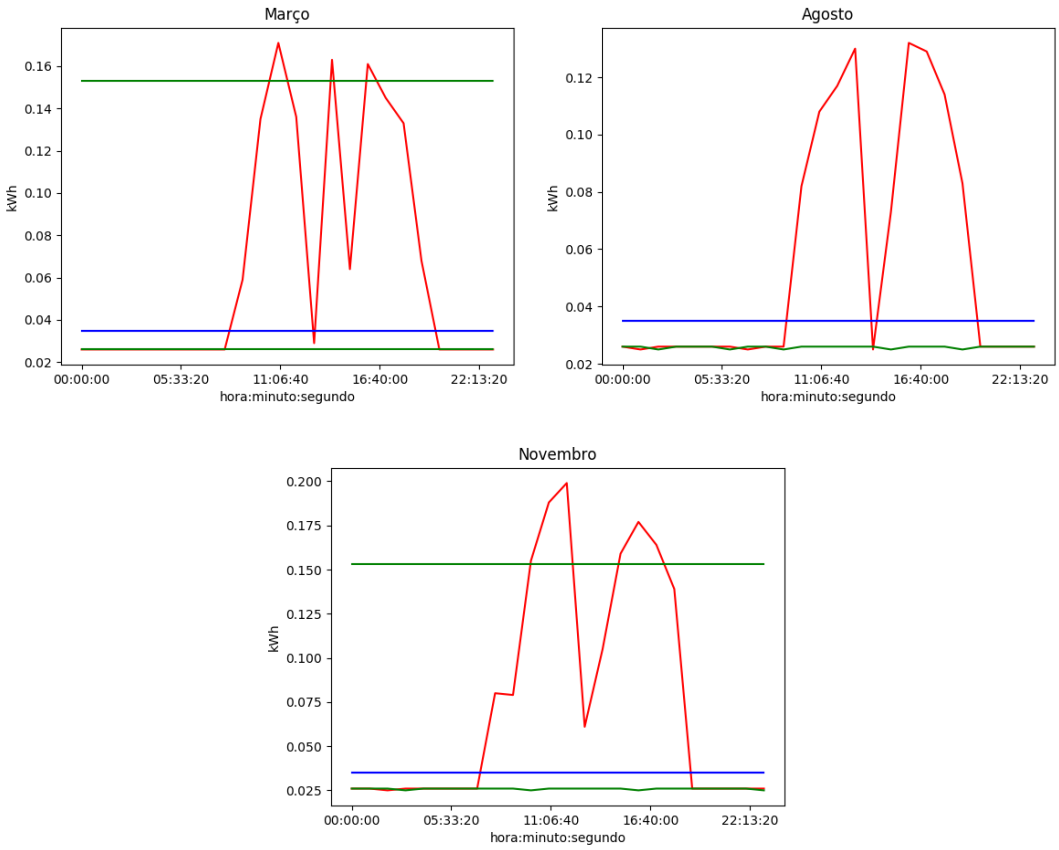


Figura 4.13: Representação, em modo horário, dos dias tipo 'F' a vermelho e 'E' a verde para a impressora.

A informação que se pode obter a partir do gráfico 4.13 é um pouco redundante com as informações obtidas em 4.12. É necessário ter em consideração a troca de atribuição do rótulo dos dias tipo. O 'E' classifica fins-de-semana e o 'F' dias úteis.

Visto a figura 4.13 analisar dados acumulados de uma hora, verifica-se que durante

o período da manhã e do período da tarde, para o dia tipo 'F', a linha de consumo se mantém acima do patamar ON/OFF, isso significa que a cada hora, do período da manhã ou do período da tarde, a impressora trabalhou pelo menos uma vez, havendo um claro decréscimo no período de almoço e no período não laboral, entre as 18 horas e as 8 horas. De notar também que, por consequência de se estar a representar dados horários, a curva dos dias típicos, para os dias úteis, 'F', é muito mais suave.

A informação concluída, a partir das figuras 4.12 e 4.13, sustenta a hipótese de num edifício empresarial, em dias laborais, haver atividade entre as 9 e as 19 horas, com uma quebra na hora de almoço. E em dias não laborais, não haver registo de atividade.

Interessa agora analisar o número de arranques, a cada hora, ao longo dos meses. Esta análise é feita para o modo base, para os dias tipo 'E', ou seja, nos dias tipo em que se espera que haja atividade no edifício. Esta análise não é feita no modo horário, por consequência do tempo típico de funcionamento do aparelho ser inferior a uma hora. No modo horário acaba por ser acusado um funcionamento continuo, quando a impressora arranca duas vezes, em horas consecutivas. Portanto, o modo base representa de uma forma mais real o número de arranques, quer ao longo de um dia, quer ao longo dos meses.

Esta análise serve para verificar as tendências de uso do equipamento, ao longo de um dia e ao longo dos meses.

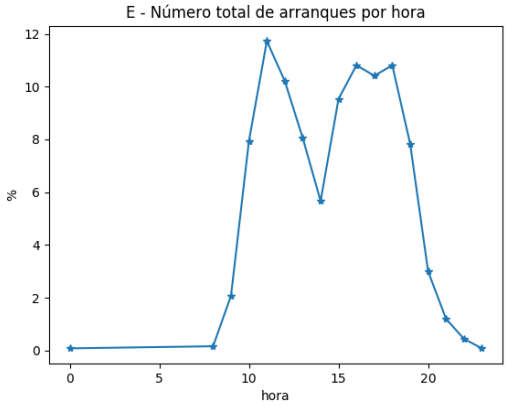


Figura 4.14: Distribuição do número de arranques, para cada hora, no modo base, para a impressora.

Por observação da figura 4.14, conclui-se que os picos dos arranques, para um dia classificado com 'E', no modo base, representante dos dias com atividade da impressora, se dão entre as 11 e as 13 horas e entre as 16 e as 19 horas, aproximadamente.

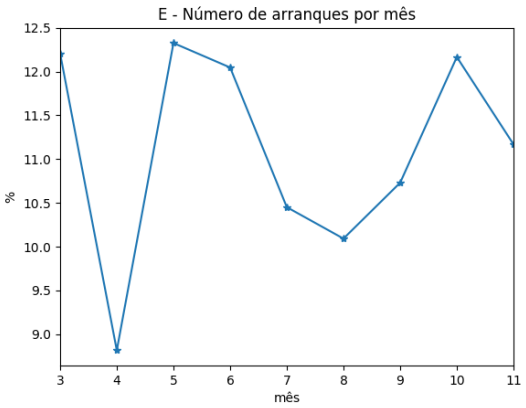


Figura 4.15: Distribuição do número de arranques para cada mês, no modo base, para a impressora.

A figura 4.15 mostra o número de arranques ocorridos, no modo base, ao longo dos meses, para os dias tipo 'E', tipicamente dias laborais (segunda a sexta-feira), dias em que há atividade do aparelho. Verifica-se um menor registo de arranques no mês de abril. Em 2017, neste mês houve a Páscoa, existiu um feriado a uma terça-feira, o que

pode ter levado os trabalhadores a não trabalhar no dia anterior e foi um mês de 30 dias. Todos estes fatores podem ter levado ao menor número de arranques neste mês. Também é notório um menor número de arranques nos meses de Verão, por norma quase todos os trabalhadores tiram férias nestes meses.

O número de arranques, ao longo dos restantes meses, pouco varia.

Conclui-se então que, as variações do número de arranques devem-se a feriados, épocas festivas, férias de verão e ao facto de nem todos os meses terem em igual número os dias laborais.

Para averiguar se existe alguma relação entre os dias tipo e a meteorologia, compara-se a distribuição das temperaturas máximas para cada dia tipo, com a distribuição das temperaturas máximas ocorridas em toda a janela temporal (ver figura 4.8).

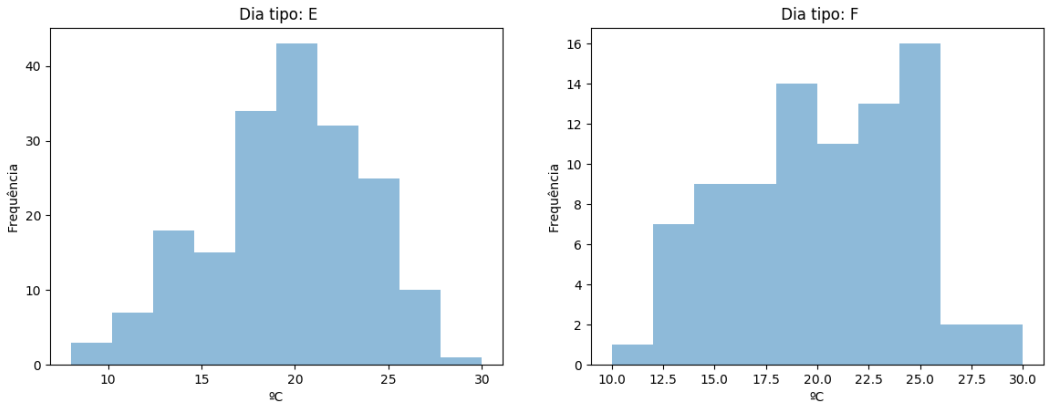


Figura 4.16: Distribuição das temperaturas máximas para cada dia tipo, em modo base, para a impressora.

As distribuições de temperatura, representadas na figura 4.16, seguem uma distribuição semelhante à figura 4.8. Para o dia tipo 'F', no modo base, dias sem consumos, revela-se uma ligeira tendência para dias mais quentes.

Por último, resta analisar um dia típico de consumo, um dia laboral, ao pormenor. Esta análise é feita para os dados adquiridos no modo base, visto ter um intervalo de tempo de aquisição mais próximo com o de um funcionamento singular do equipamento. Para cada hora (índices das tabelas 4.8:10), é analisado o centro dos padrões que tive-

ram início nessa hora, a duração média de cada padrão que se iniciou a essa hora e o número de ocorrências de cada *cluster*.

	Março		
	Duração (min)	# [C3,C4]	Centro
9	21	0,18	AAA
10	17	0,35	AA
11	24	0,39	AAA
12	27	0,27	AAAAA
13	18	0,17	AAA
14	22	1,25	AAA
15	24	0,27	AAA
16	31	2,33	AAAA
17	27	0,31	AAAA
18	23	0,24	AAAA
19	13	0,18	AA
20	10	0,4	AA
21	10	0,3	AA
22	10	0,2	BA

Tabela 4.8: Caracterização do dia tipo 'E', em modo base para o mês de março, para a impressora.

Na tabela 4.8, verifica-se uma predominância da *cluster* C4 (que vai de encontro à *cluster* mais frequente identificada na tabela 4.7) e uma predominância dos arranques entre as 10 horas e as 18 horas, sendo que existe um decréscimo entre as 12h e as 14h. Também se verifica que, um funcionamento do equipamento varia entre os 10 minutos e os 31 minutos. O patamar de consumo mais frequente é claramente o A.

	Agosto		
	Duração (min)	# [C3,C4]	Centro
8	15	0,1	BAA
9	20	0,1	BAAA
10	15	0,18	AA
11	21	0,29	AAA
12	32	2,25	AAAA
13	22	0,24	AAA
14	19	0,15	AAA
15	28	2,23	AAA
16	28	2,31	AAA
17	18	0,26	AA
18	18	0,31	AAA
19	19	0,22	AAA
20	10	0,1	AA

Tabela 4.9: Caracterização do dia tipo 'E', em modo base para o mês de agosto, para a impressora.

Os resultados da tabela 4.9 são mais uniformes que os apresentados na tabela 4.8, no entanto, as mesmas ilações continuam válidas. No geral, nota-se um menor número de arranques, a cada hora, que na tabela 4.7, e que estes arranques se dão mais tarde, isto é, os picos do números de arranques dão-se ligeiramente mais tarde, quer no período da manhã, quer no período da tarde.

	Novembro		
	Duração (min)	# [C2, C3,C4]	Centro
8	13	0,0,3	BAA
9	17	0,0,22	BAA
10	16	0,0,39	AA
11	30	0,1,33	AAAAA
12	33	0,1,23	AAAAA
13	19	0,0,6	AAAA
14	16	0,0,26	AA
15	32	1,1,3	AAAA
16	29	0,1,18	AAA
17	24	0,0,28	AAAA
18	13	0,0,25	AA
19	11	0,0,11	AA
20	10	0,0,3	AA
21	15	0,0,1	AAA
22	8	0,0,2	A
23	15	0,0,1	AAA

Tabela 4.10: Caracterização do dia tipo 'E', em modo base para o mês de novembro, para a impressora.

A tabela 4.10 mostra que para o mês de novembro, existe uma descida muito mais acentuada do número de arranques, à hora de almoço, às 13 horas, mas logo a seguir, às 14 horas, existe uma subida abrupta dos consumos que não era tão visível nos meses anteriores, nas tabelas 4.8 e 4.9. Existe um maior número de arranques, no período da manhã, que no mês de agosto, mas em números semelhantes aos ocorridos em março, no entanto, na parte da tarde, o número de arranques assemelha-se aos ocorridos em agosto.

As tabelas 4.8:10 caracterizam assim o dia tipo 'E', dia laboral, em modo base, para os meses escolhidos como representantes de cada estação do ano.

Ilha de Computadores

Nesta secção, são apresentados os dados obtidos por uma monitorização feita a um conjunto de computadores, usados pelos trabalhadores da empresa, no período laboral. É a única monitorização desagregada em que os dados não são recolhidos com recurso

a uma *smart plug*. Portanto, a aquisição pode apresentar alguma diferenças.

Por iterações aos dados, concluiu-se que a divisão ideal entre os consumos que representam atividade, na ilha de computadores, e os que não representam, corresponde a 1/4 do valor máximo, no modo base, e 1/3 do valor máximo registado, para dados horários, fora *outliers*.

	Mínimo	**Máximo**	**Média**
ON	1.154	5.700	2.231
OFF	0.000	1.154	0.931

a) Modo base, valores em kWh;

	Mínimo	**Máximo**	**Média**
ON	1.310	3.929	2.366
OFF	0.000	1.310	1.046

b) Modo horário, valores em kWh;

Tabela 4.11: Divisão dos valores de consumo bruto (3.1) em patamares de consumo ON e OFF (3.11), para a ilha de computadores.

	Mínimo	**Máximo**	**Média**
A	1.154	2.843	1.877
B	2.845	4.384	3.319
C	4.614	4.929	4.771
D	5.700	5.700	5.700

a) Modo base, valores em kWh;

	Mínimo	**Máximo**	**Média**
A	1.310	1.988	1.569
B	1.990	3.091	2.576
C	3.093	3.930	3.290

b) Modo horário, valores em kWh;

Tabela 4.12: Divisão dos valores de ON (3.11), em patamares de consumo, para a ilha de computadores.

As tabelas 4.11 e 4.12 revelam um pormenor interessante sobre os dados. Os dados horários representam a média dos consumos base registados nessa hora, e não os consumos acumulados, como ocorre para todas as restantes monitorizações.

Contrariamente a todos os equipamentos em análise neste projeto, em que os dados são monitorizados por *smart plugs* ou *smart meters*, como já referido, os dados da ilha de computadores são adquiridos de forma distinta, talvez seja esta a razão pela qual são a única monitorização que utiliza o algoritmo 3.2 para a previsão de nulos, e a aquisição horária representa uma média dos dados adquiridos no modo base, no intervalo de tempo de aquisição horária, e não o consumo total. Este último facto não interfere com os algoritmos aplicados a esta base de dados horária, os resultados obtidos são igualmente válidos.

	Centro de cada *cluster*	
	Dados base	**Dados horários**
C1	$*$ (99.439%)	$*$ (0.379%)
C2	$*$ (0.561%)	$*$ (0.758%)
C3	-	$*$ (2.652%)
C4	-	$*$ (96.212%)

Tabela 4.13: Centro de cada *cluster* e respetiva percentagem de ocorrência, para a ilha de computadores(# 3741-modo base; # 264-modo horário).

As *clusters* assinalados com $*$, na tabela 4.13, visto terem padrões representativos do seu centro, heterogéneos a nível de patamares e ocorrerem entre longos períodos, são representadas nas figuras 4.17 e 4.18 e não na diretamente na tabela 4.13.

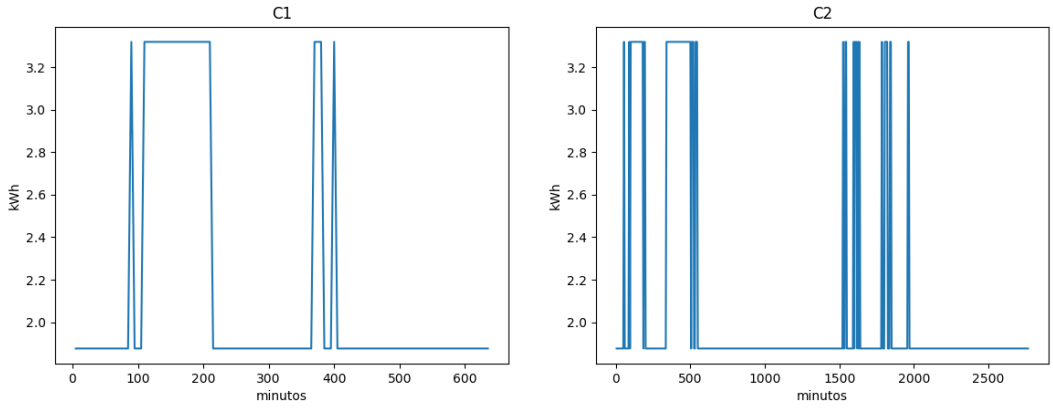


Figura 4.17: Representação do centro das *clusters* C1 e C2, no modo base, para a ilha de computadores.

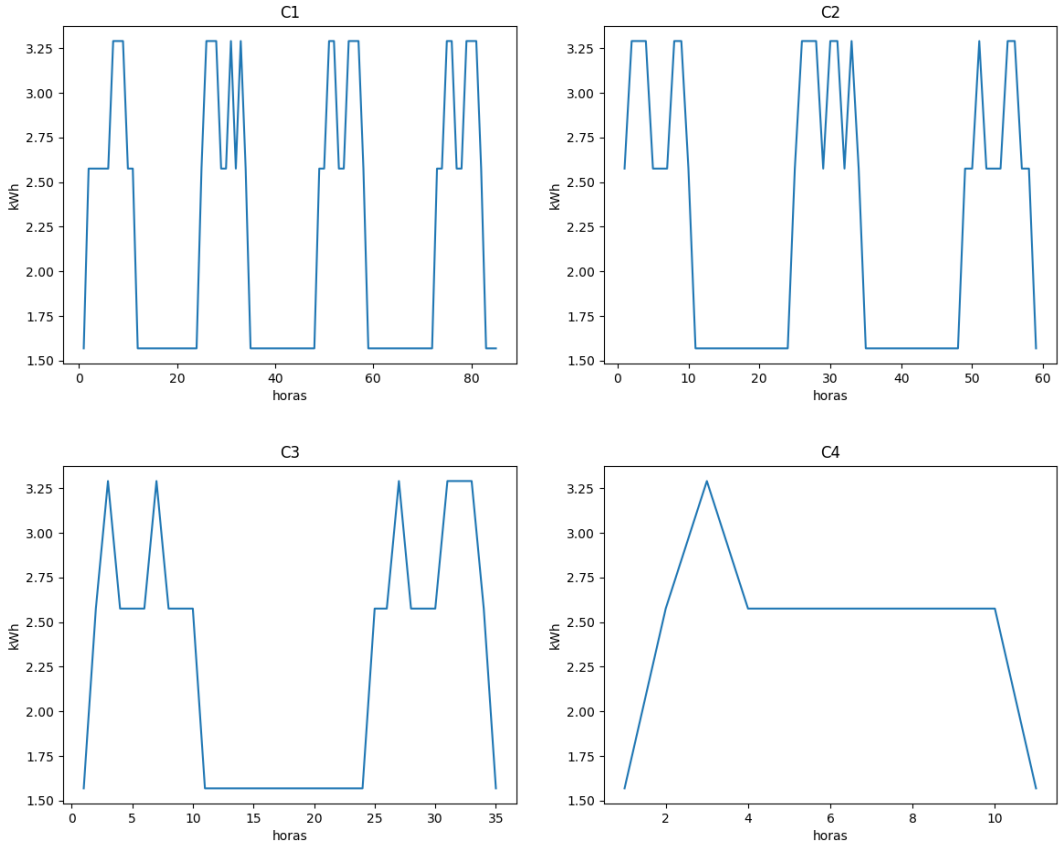


Figura 4.18: Representação do centro das *clusters* C1, C2, C3 e C4 no modo horário, para a ilha de computadores.

Por analise da figura 4.17, conclui-se que o padrão mais provável, para o modo de aquisição base, C1, tem uma duração de 600 minutos, ou seja, 10 horas, que corresponde nada mais do que à janela temporal em que se estima que exista atividade num edifício empresarial, num dia laboral. De notar que se está a monitorizar uma ilha de computadores. Sabendo que os computadores são o principal instrumento de trabalho na empresa em causa, é de esperar que estes estejam ativos durante todo o período laboral. A *cluster* C2, no modo base, tem uma probabilidade de ocorrência demasiado baixa, não pode ser considerada um padrão, representa apenas um período em que a ilha de computadores esteve ativa dias consecutivos.

Para o modo horário, apenas a *cluster* C4, representada na figura 4.18, tem uma

frequência relevante (tabela 4.13). Tal como se concluiu na figura 4.17 para a *cluster* C2, também a *cluster* C4 tem uma duração de 10 horas, mantendo, aproximadamente, o patamar constante. Assim se conclui que, o modo de aquisição mais próximo do período típico de funcionamento da ilha de computadores é o modo horário. Este modo de aquisição, horário, revela-se mais interessante, pois, acaba por não acusar pequenas flutuações, sem significado nos dados. Apresenta os padrões de forma mais clara, e, por a monitorização ter uma ordem de funcionamento horária, não se perde informação relevante.

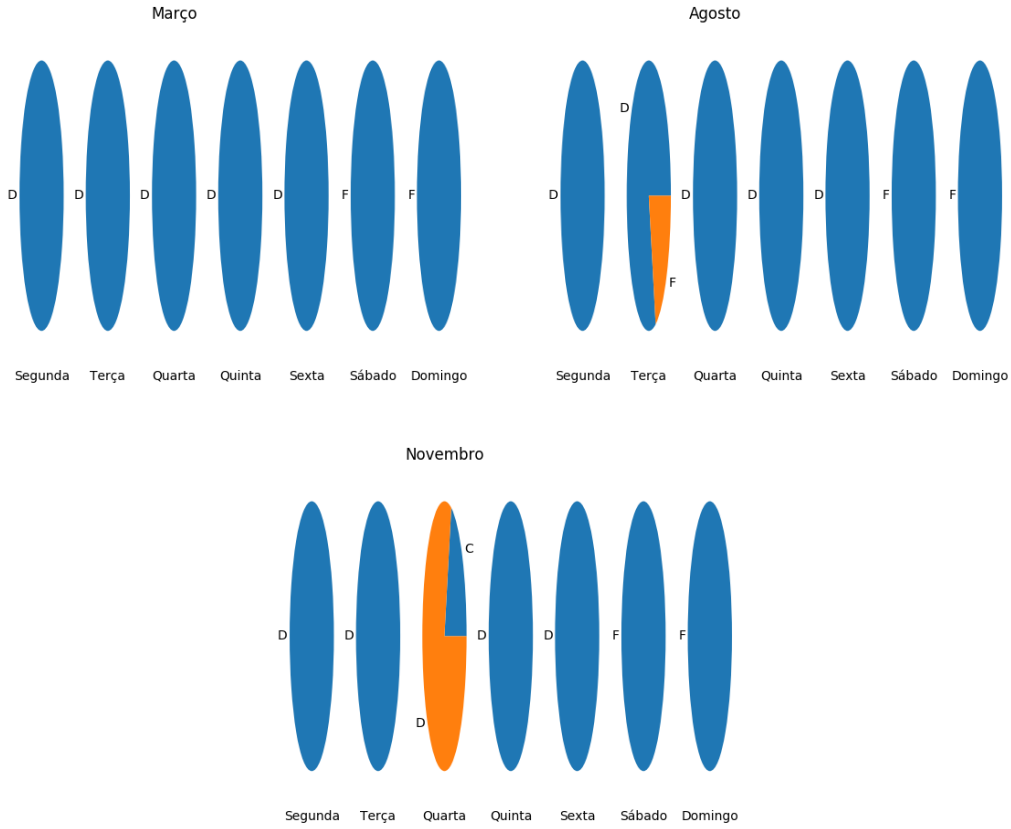


Figura 4.19: Representação da distribuição dos dias tipo, pelos dias da semana, no modo base, para a ilha de computadores.

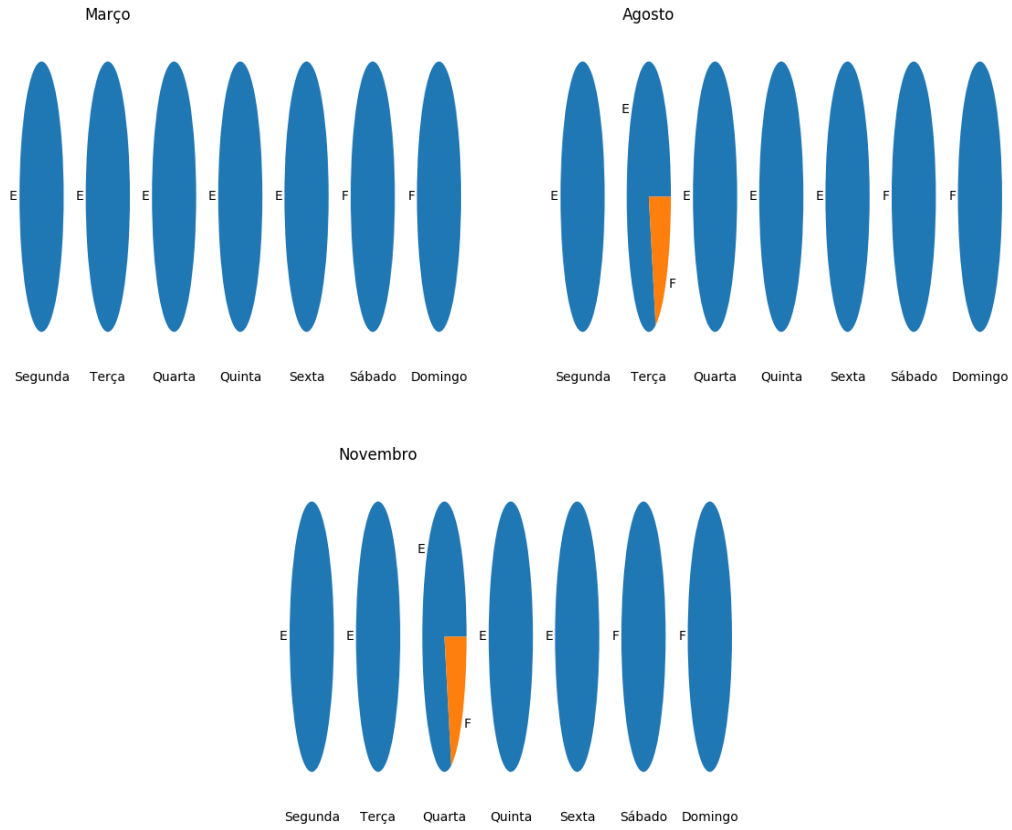


Figura 4.20: Representação da distribuição dos dias tipo, pelos dias da semana, no modo horário, para a ilha de computadores.

Com as figuras 4.19 e 4.20, tal como se verificou para a impressora, conclui-se que para a ilha de computadores, os dias ao longo de uma semana se dividem entre os dias úteis (segunda a sexta-feira) e o fim-de-semana (sábado e domingo). Para o modo de aquisição base, os dias úteis são maioritariamente representados por 'D' e os fins-de-semana por 'F'. Para o modo horário, os dias úteis por 'E' e os fins-de-semana por 'F'.

Ambos os modos de aquisição apresentam uma divisão dos dias tipo equivalente.

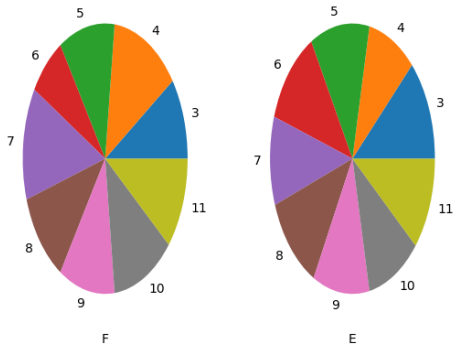


Figura 4.21: Representação da distribuição dos dias tipo relevantes, ao longo dos meses da janela temporal, no modo horário, para a ilha de computadores.

Com a figura 4.21, nota-se uma distribuição uniforme dos dias tipo mais relevantes. Nota-se apenas uma ligeira quebra de dias 'F' (normalmente atribuídos aos fins-de-semana) no mês de junho e uma ligeira prevalência deste dia tipo em julho. Verificou-se que em junho, aos fins-de-semana, apenas 75% dos dias foram classificados com 'F', e , em julho existe uma terça-feira classificada com 'F', resultado de um feriado local (equivalente a um dia de fim-de-semana). Para os dias tipo 'E' não se evidenciam flutuações, visto que, como classificam a maioria dos dias da base de dados, tipicamente dias úteis, não são notórias flutuações.

Estando os dias tipo predominantes identificados, pode-se seguir para a análise dos seus perfis, ao longo dos meses.

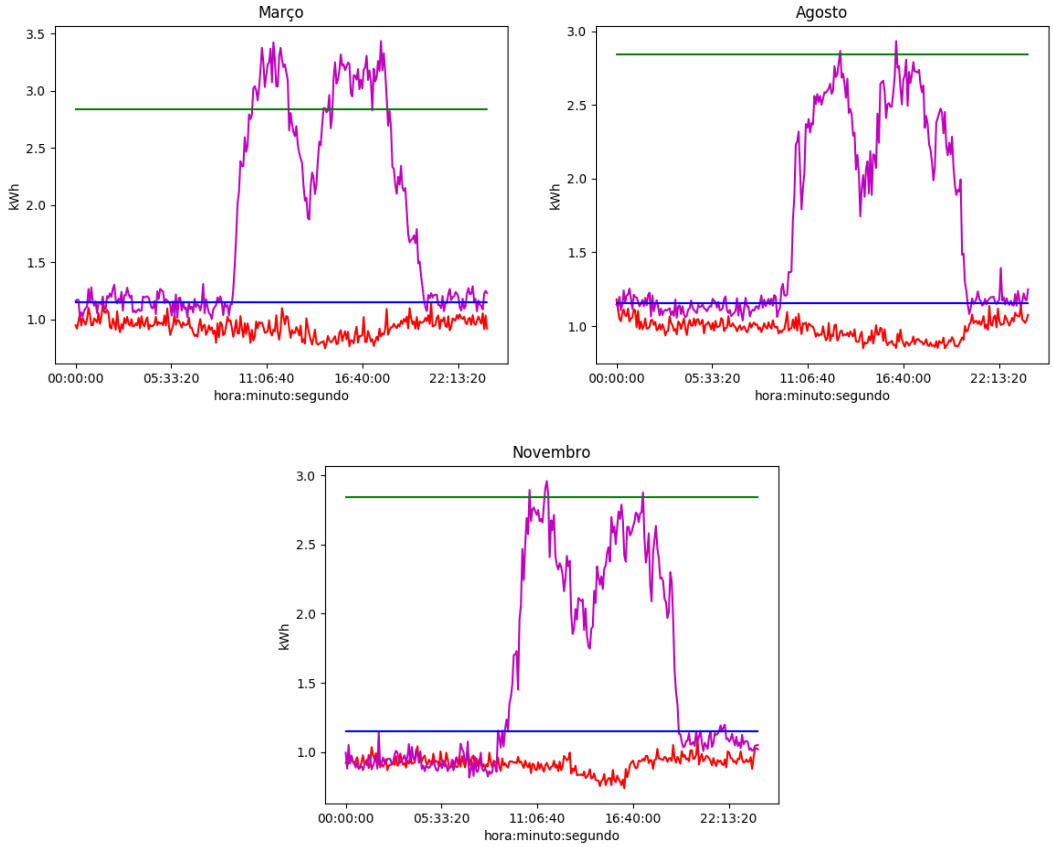


Figura 4.22: Representação, em modo base, para a ilha de computadores, dos dias tipo 'D' a roxo e 'F' a vermelho.

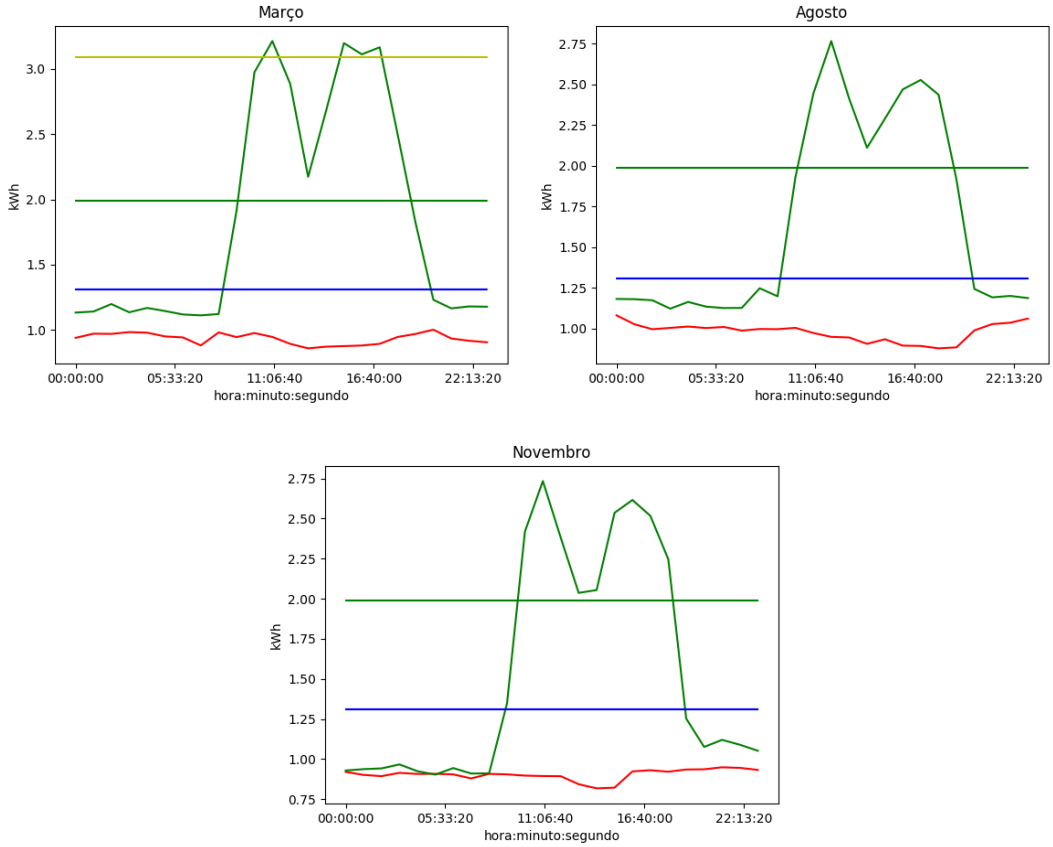


Figura 4.23: Representação, em modo horário, para a ilha de computadores, dos dias tipo 'F' a vermelho e 'E' a verde.

A figuras 4.22 e 4.23 mostram que os consumos típicos durante um dia útil ('D' no modo instantâneo e 'E' no modo horário), durante o período laboral, aproximadamente entre as 9 horas e as 19 horas, se encontra sempre acima da linha azul, que delimita o consumo representativo de atividade, com uma ligeira quebra de consumo na hora de almoço. No entanto, na hora de almoço, continua a ser detetado consumo, o que é normal visto que a hora de almoço é flexível, nem todas as pessoas vão almoçar ao mesmo tempo e está a ser monitorizada uma ilha de computadores, não um computador individual.

Também é de notar que a ilha de computadores apresenta sempre, em períodos de não atividade, um consumo residual.

Os dias tipo 'F', em modo base e horário, representam um consumo residual que se mantém abaixo da linha horizontal azul. São portanto dias que não representam atividade no edifício, o que vai de encontro ao facto de classificarem tipicamente fins-de-semana, dias não laborais.

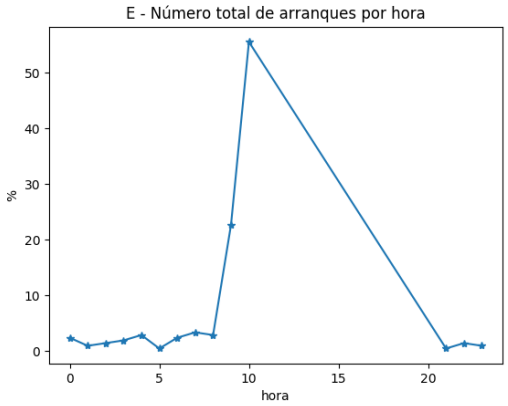


Figura 4.24: Distribuição do número de arranques, para cada hora, no modo de aquisição horário, para a ilha de computadores.

De acordo com a figura 4.24, os arranques verificam-se maioritariamente entre as 9 e as 10 da manhã, para dias em que se espera que haja atividade no edifício, dias tipo 'E', no modo horário. Após esta hora, não há arranques significativos, pois, é expectável que a ilha de computadores apresente um único arranque e se encontre operacional por um período de aproximadamente 10 horas, como já concluído.

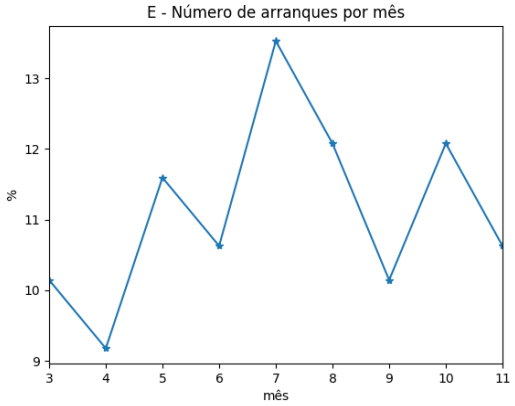


Figura 4.25: Distribuição do número de arranques, para cada mês, no modo de aquisição horário, para a ilha de computadores.

Tal como na figura 4.15 para a impressora, na figura 4.25, verifica-se um decréscimo de arranques em abril e nos meses de verão. As razões suspeitas são as mesmas enunciadas para a figura 4.15. Visto os consumos terem uma duração aproximada de 10 horas e decorrem durante o horário laboral, é de esperar que para cada dia tipo 'E', em modo horário, seja acusado um arranque. Por consequência, a figura 4.25 deveria revelar informação redundante com a figura 4.21, o que não acontece ou pelo menos não é evidente. Visto as probabilidades de ocorrência de um arranque, para um certo mês, não variarem muito, entre 9 e 13 %, conclui-se que as diferenças entre 4.25 e 4.21 devem ser originadas por arranques de fundo, isto é, intersecções da linha horizontal azul com o consumo de fundo (figura 4.23) que não corresponde a uma atividade real. Visto só se verificar um arranque significativo, para cada dia tipo 'E', o gráfico 4.25 torna-se sensível a arranques pontuais.

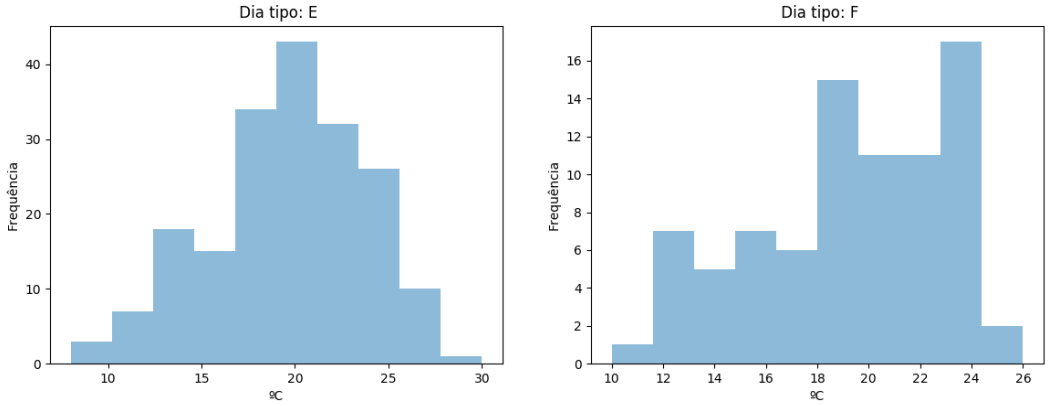


Figura 4.26: Distribuição das temperaturas máximas para cada dia tipo, para o modo de aquisição horário.

Visto os dados horários revelarem a informação de uma forma mais clara, a análise da distribuição da temperatura para os dias úteis ('E') e fins-de-semana ('F'), é feita para este modo de aquisição. Por comparação com a figura 4.8, não se revela nenhuma distribuição fora da referência para os dias tipo 'E'. No entanto, é notória uma certa tendência, dos dias tipo 'F', para dias mais quentes.

Posto isto, e pelas mesmas razões enunciadas, a análise de um dia típico, para um dia em que é expectável existir consumos, é feita para os dados horários (dia tipo 'E').

	Março		
	Duração (horas)	# [C4]	Centro
4	1	1	A
8	12	1	$AB^2C^2B^2C^3BA$
9	12	17	AB^5CB^3A

Tabela 4.14: Caracterização do dia tipo 'E', em modo horário, para o mês de março.

	Agosto		
	Duração (horas)	# [C4]	Centro
2	24	1	$A^9BCB^7A^6$
4	2	2	A
8	7	2	A
9	10	19	AB^8A

Tabela 4.15: Caracterização do dia tipo 'E', em modo horário, para o mês de agosto.

	Novembro		
	Duração (horas)	# [C4]	Centro
9	11	21	AB^3AB^5A
21	1	1	A

Tabela 4.16: Caracterização do dia tipo 'E', em modo horário, para o mês de novembro.

Como era de esperar, as tabelas 4.14:16 apresentam resultados muito similares, uma predominância de arranques para as 9 da manhã, para um dia laboral, com uma duração média do padrão de 10 horas, todos pertencentes à *cluster* C4. No mês de agosto, tabela 4.15, existe um arranque às duas da manhã com duração de 24 horas. Provavelmente representa um dia em que o consumo de fundo foi muito intenso. Todos os outros consumos tabelados, que não têm início às 9 horas da manhã, são consumos pontuais originados por intersecções, de baixa amplitude e curta duração, do consumo de fundo com a linha horizontal azul do gráfico 4.23.

Máquina de Café

Com recurso a uma *smart plug*, é monitorizada a máquina de café, disponível para uso por parte dos empregados, na empresa piloto.

O estado de ativo e inativo deste aparelho é distinguido pelo cálculo de 1/10 do valor máximo dos consumos, sem considerar *outliers*, quer para o modo de aquisição horário, quer para o modo de aquisição base.

	Mínimo	Máximo	Média
ON	0.010	0.131	0.042
OFF	0.000	0.009	0.004

a) Modo base, valores em kWh;

	Mínimo	Máximo	Média
ON	0.014	0.171	0.053
OFF	0.000	0.013	0.006

b) Modo horário, valores em kWh;

Tabela 4.17: Divisão dos valores de consumo bruto (3.1) em patamares de consumo ON e OFF (3.11), para a máquina de café.

Na tabela 4.17, constata-se que a amplitude máxima da gama de valores, no modo ON, para consumos horários, não é muito superior à verificada para o modo base. Sabendo que o modo horário reflete a soma total dos consumos, num período de uma hora e que o modo base representa o mesmo consumo, mas num intervalo de cinco minutos, que é da ordem de tempo de um funcionamento de uma máquina de café típica, suspeita-se que a cada hora haverá poucos arranques da máquina de café e com uma duração na ordem do intervalo de tempo de aquisição do modo base, 5 minutos.

	Mínimo	Máximo	Média
A	0.010	0.043	0.024
B	0.046	0.071	0.058
C	0.073	0.092	0.082
D	0.131	0.131	0.131

a) Modo base, valores em kWh;

	Mínimo	Máximo	Média
A	0.014	0.054	0.035
B	0.055	0.082	0.068
C	0.084	0.097	0.090
D	0.108	0.108	0.108
E	0.126	0.145	0.138
F	0.171	0.171	0.171

b) Modo horário, valores em kWh;

Tabela 4.18: Divisão dos valores de ON (3.11), em patamares de consumo, para a máquina de café.

	Centro de cada *cluster*	
	Dados base	**Dados horários**
C1	A^3 (1.365%)	A^{203} (0.64%)
C2	A^5 (0.195%)	$*$ (1.282%)
C3	AD(0.195%)	$*$ (3.846%)
C4	AB (5.068%)	A^{10} (94.231%)
C5	C (1.754%)	-
C6	A (90.448%)	-
C7	BA (0.975%)	-

Tabela 4.19: Centro de cada *cluster* e respetiva percentagem de ocorrência, para a máquina de café (# 513-modo base; # 156-modo horário).

Verifica-se, com a análise dos padrões, no modo base, identificados na tabela 4.19, uma predominância de consumos de curta duração e de patamar de baixa amplitude. Portanto, pode-se concluir que a duração de um consumo típico da máquina de café, como é de esperar, é da ordem do intervalo de tempo de aquisição do modo base.

Visto os dados horários representarem consumos acumulados, um único consumo, mesmo que de duração da ordem dos 5 minutos, num intervalo de tempo de uma hora, é suficiente para fazer subir o patamar de consumo para um patamar ON, no modo de aquisição horário. Não é de estranhar, visto que de acordo com a tabela 4.17, o limiar de funcionamento do equipamento, para o modo de aquisição base, é muito próximo do horário. Portanto, basta a máquina de café funcionar uma vez, a cada hora, durante 10 horas consecutivas, para se gerar um padrão como o de C4, no modo horário, não estando necessariamente a máquina, efetivamente, em estado ativo todo esse tempo, de forma ininterrupta.

A figura 4.27 representa as *clusters* assinaladas com $*$ na tabela 4.19. Por terem uma elevada duração e serem heterogéneas, não é conveniente a sua representação na tabela referida.

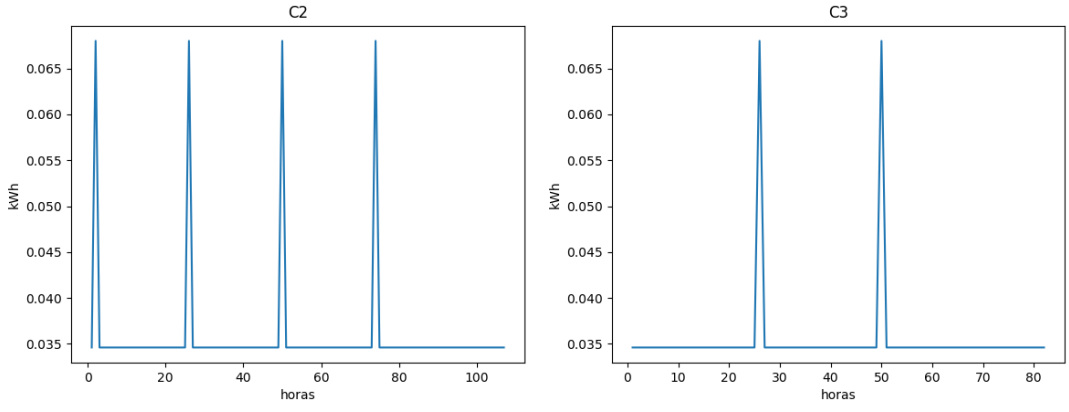


Figura 4.27: Representação da *cluster* C2 e C3, no modo horário.

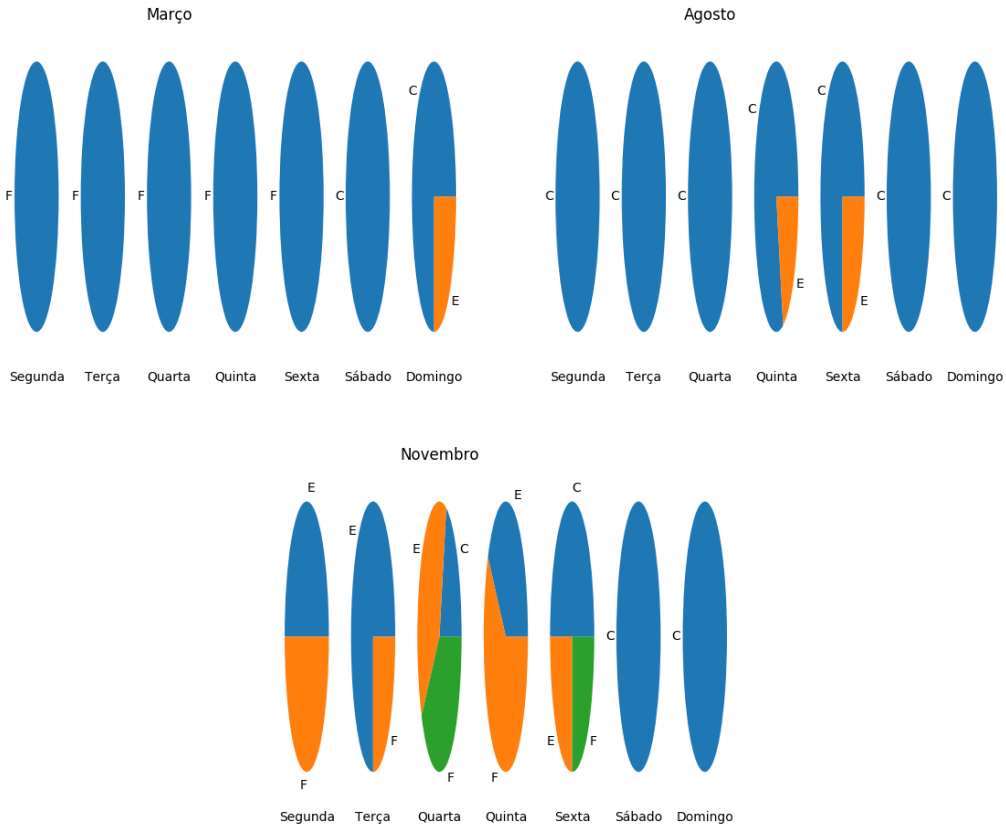


Figura 4.28: Representação da distribuição dos dias tipo, para a máquina de café, pelos dias da semana, no modo base.

Por análise da figura 4.28, no modo base, verifica-se uma predominância do dia 'F' e 'E' para os dias úteis e uma predominância do dia 'C' para os fins-de-semana. No mês de agosto, praticamente todos os dias foram classificados como 'C'.

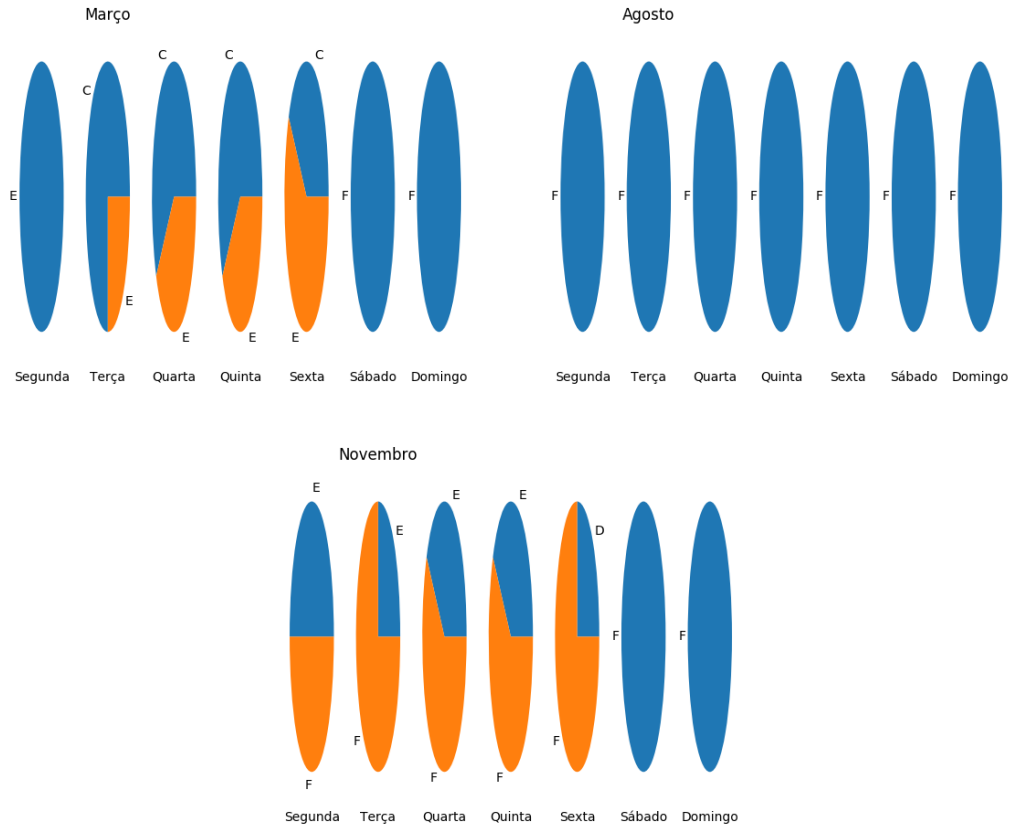


Figura 4.29: Representação da distribuição dos dias tipo, para a máquina de café, pelos dias da semana, no modo horário.

A distribuição dos dias tipo, para os dados horários, figura 4.29, divide-se pala classificação dos dias úteis como 'E' e 'C', e 'F' para o fim-de-semana. Sendo que 'F' também ocorre, em maioria, para os dias úteis no mês novembro e em todos os dias do mês de agosto.

Conclui-se que a distribuição das classificações dos dias tipo não é totalmente equivalente, para os dois modos de aquisição.

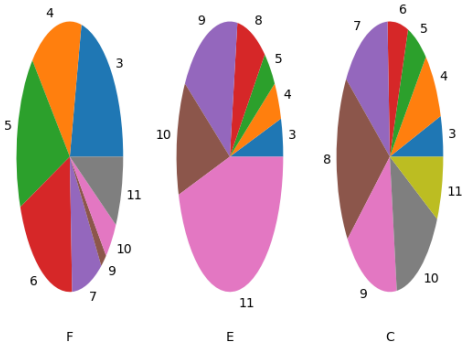


Figura 4.30: Representação da distribuição dos dias tipo relevantes, ao longo dos meses da janela temporal, no modo base.

A figura 4.30 revela, para os dados base, uma clara predominância dos dias 'F' para os meses de março a junho, dos dias 'E' de setembro a novembro e 'C' de julho a outubro.

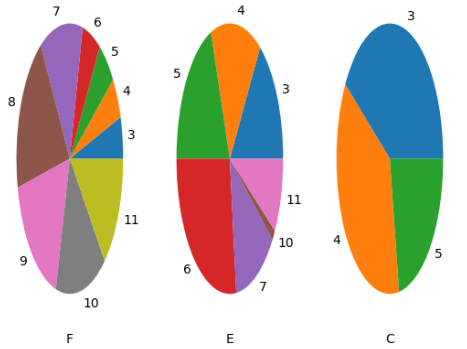


Figura 4.31: Representação da distribuição dos dias tipo relevantes, ao longo dos meses da janela temporal, no modo horário

Para o modo horário, segundo a figura 4.31, existe uma prevalência do dia 'F' de julho a novembro e 'E' de março a julho. 'C' apenas acontece entre os meses de março a maio.

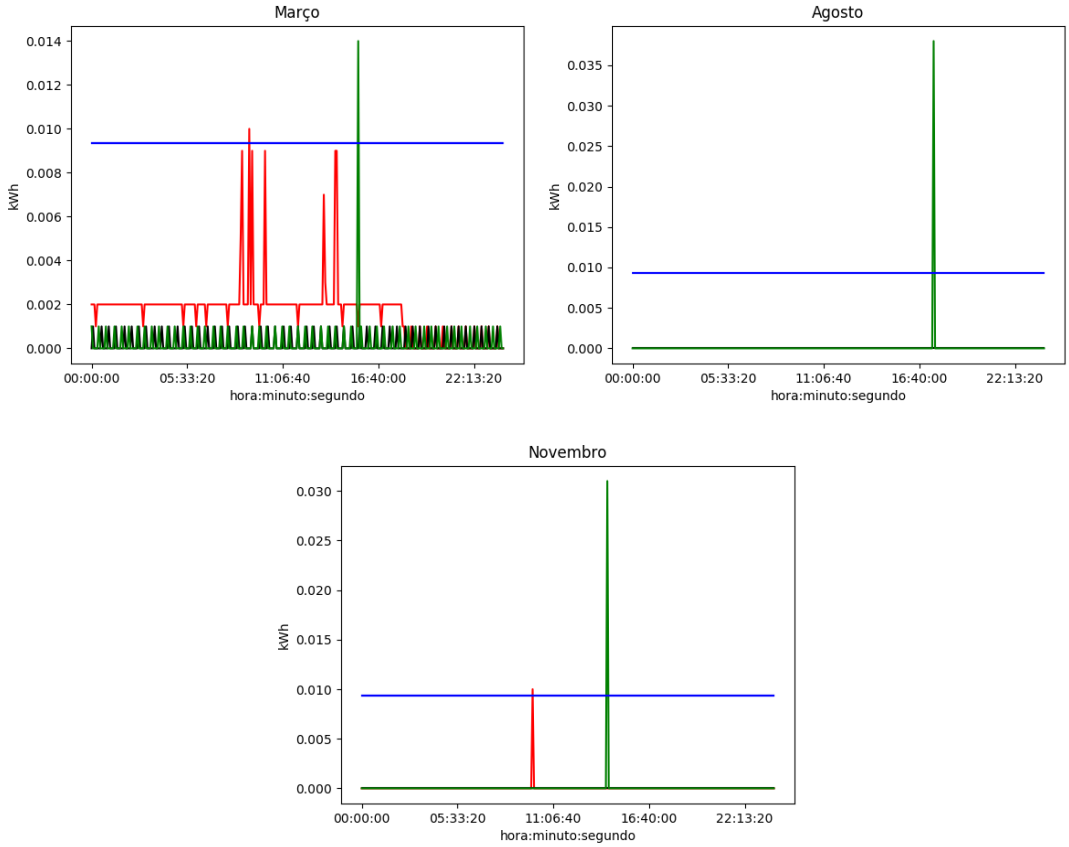


Figura 4.32: Representação, em modo base, para a máquina de café, dos dias tipo 'E' a verde, 'F' a vermelho e 'C' a preto.

Na figura 4.32, o dia tipo 'C', representado a preto, por ter um consumo nulo ao longo do tempo, está camuflado pelos restantes dias nos gráficos dos meses em que ocorre. Como já visto nas figuras 4.28 e 4.30, tem predominância de ocorrência aos fins-de-semana e nos meses de verão, mesmo durante a semana. Os dias 'F' caracterizam-se por terem uma predominância de ocorrência no período da manhã, e 'E', no período da tarde. Como verificado na figura 4.30, 'F' tem uma predominância de ocorrência de março a junho, e 'E', de setembro a novembro, em dias úteis. Isto leva à conclusão que existe uma tendência de substituição dos dias tipo 'F' por 'E', para os dias laborais. No mês de agosto, praticamente todos os dias foram classificados com 'C', o que leva a concluir que a máquina de café esteve desligada durante a maior parte deste mês.

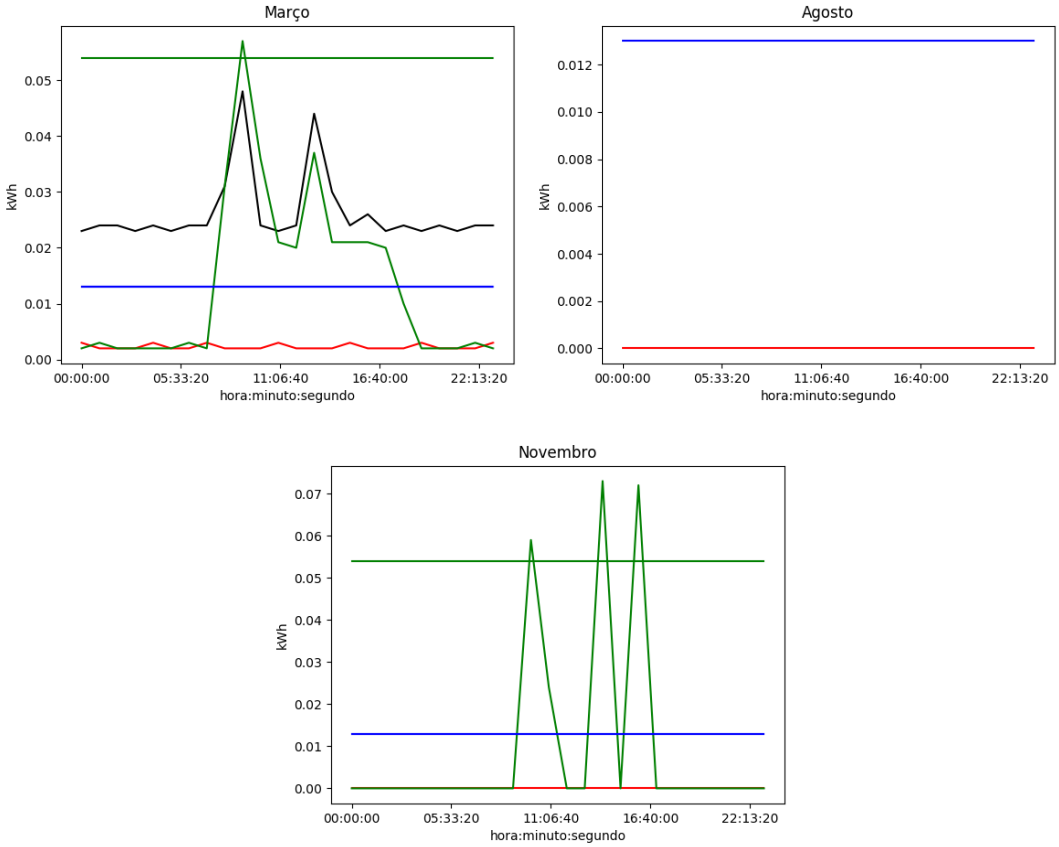


Figura 4.33: Representação, em modo horário, dos dias tipo 'F' a vermelho, 'E' a verde e 'C' a preto.

No modo horário, figura 4.33, é visível o dia 'F', a vermelho, que corresponde a dias sem consumos. Todos os dias de fim-de-semana são classificados com este dia tipo. Este dia tipo também tem predominância nos meses de verão, como visto na figura 4.31. Daqui se conclui que o dia tipo 'F' é equivalente ao dia tipo 'C', no modo base. Os dias tipo 'E' e 'C', classificadores de dias úteis, têm picos de consumos de igual amplitude e ocorrem aproximadamente nos mesmos horários, mas, o dia tipo 'C', que se verifica nos meses de março a maio, apresenta um consumo base muito superior à linha horizontal azul, em períodos em que não é expectável atividade num dia laboral, o que quer dizer que a máquina de café, nestes dias tipo, teve uma atividade constante

e de consumo considerável. Estes dias podem corresponder a dias em que a máquina de café foi deixada ligada. Quando isto acontece, a máquina fica constantemente a aquecer água. Este consumo representa um gasto que pode ser evitado. Este dia tipo só se verifica entre março a maio, portanto, é provável que a anomalia tenha sido detetada e corrigida.

É importante referir que, no modo horário, os picos de consumo, para os dias 'E' e 'C', apresenta-se ao inicio da manhã e ao início da tarde, como já referido, estas são as alturas em que é mais comum as pessoas beberem café.

Apesar do o modo de aquisição horário ter-se revelado fundamental, para este aparelho, na identificação de consumos desnecessários, não representativos de efetiva atividade do aparelho (dia tipo 'C', neste modo de aquisição com consumo de fundo bastante elevado), a análise segue-se apenas no modo base. Conclui-se assim que o modo base representa melhor o número real de arranques, representativos de efetiva atividade do aparelho, por ter um intervalo de tempo de registo próximo do funcionamento do aparelho.

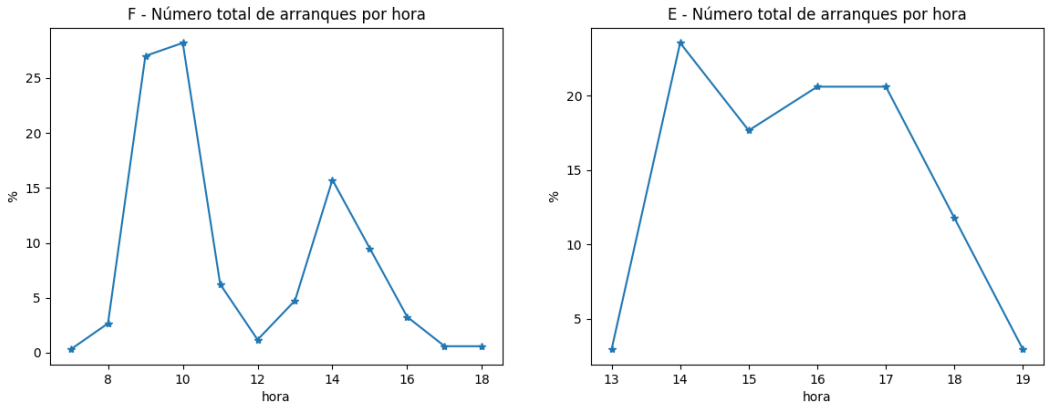


Figura 4.34: Distribuição do número de arranques, para cada hora, no modo base para a máquina de café.

A figura 4.34, para o modo base, representa o número de arranques para os dias tipo mais frequentes, com atividade, que são os dias 'F' e 'E'. O dia 'F' apresenta picos

de arranque entre as 9 e as 10 da manhã e logo após a hora de almoço, às 14 horas. Já os dias tipo 'E' apresentam um número de arranques uniforme no período da tarde.

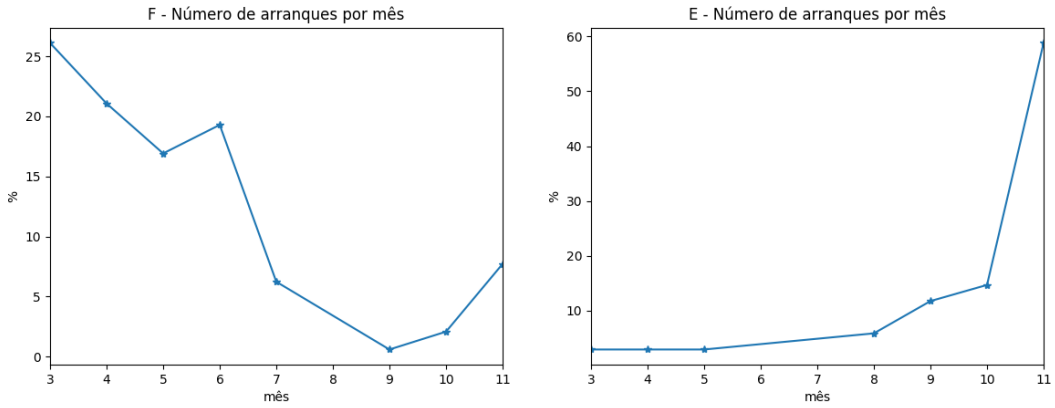


Figura 4.35: Distribuição do número de arranques, para cada mês, no modo base para a máquina de café.

O número total de arranques por mês, figura 4.35, para o modo base, para os dias tipo com atividade, é contrário entre os dias tipo 'F' e 'E', isto é, o número de arranques para os dias tipo 'F' diminui ao longo do avançar dos meses e 'E' aumenta. Tal significa que existe uma tendência, ao longo do tempo, para a máquina de café ser utilizada de manhã e tarde, para passar a ser utilizada apenas de tarde. Tal facto é sustentado pela figura 4.30, que mostra uma maior prevalência dos dias 'F' para os primeiros meses em análise e 'E' para os últimos.

Ora, adicionando esta informação à concluída para a figura 4.34, verifica-se ao longo do tempo, a tendência de utilização da máquina de café por curtos períodos de tempo no início da manhã e da tarde para passar a ser usada de forma mais uniforme no tempo, durante a tarde.

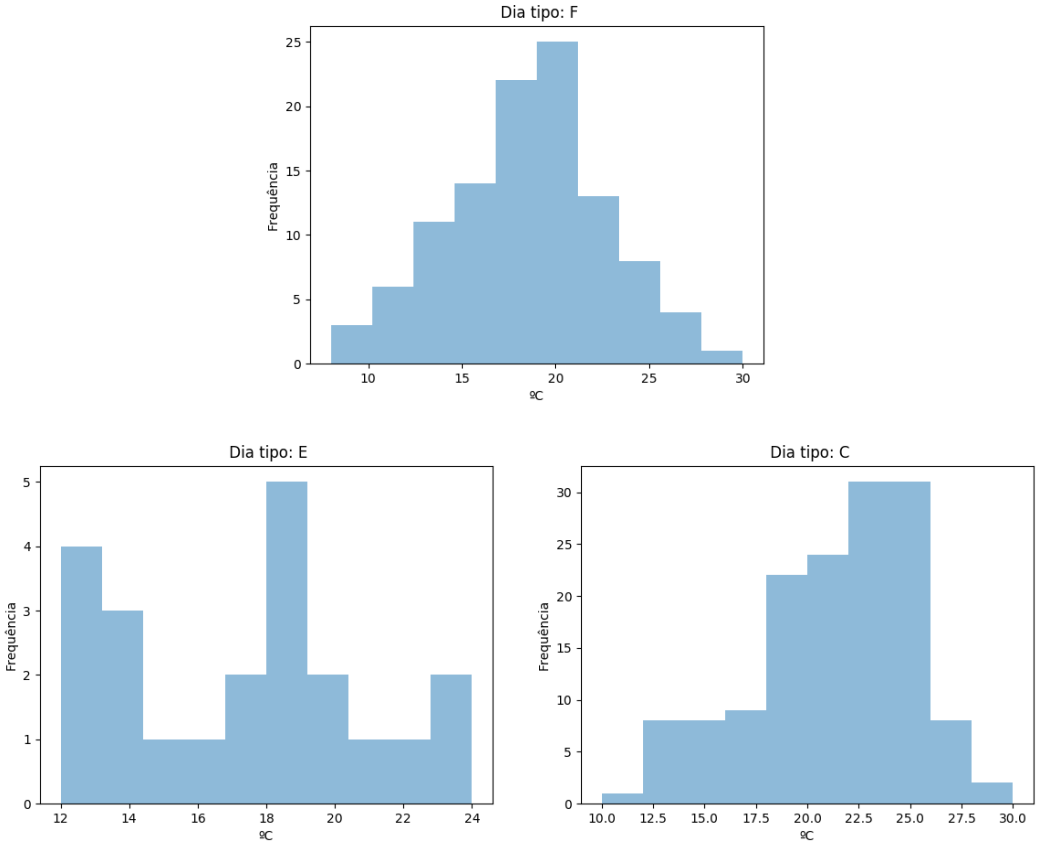


Figura 4.36: Distribuição das temperaturas máximas para a máquina de café, no modo de aquisição base.

Pela figura 4.36, e confrontando com a figura 4.8, verifica-se que as temperaturas máximas para o dia tipo 'F' situam-se maioritariamente entre os 17.5 e os 22.5ºC, para o dia 'E' entre os 12-14ºC e os 18-19ºC e para 'C' entre os 20 e os 25ºC. As temperaturas verificadas vão de encontro à figura 4.30. 'C' apresenta uma gama de temperaturas alta, pois, ocorre em maioria nos meses de Verão. O dia tipo 'F' revela-se para temperaturas ligeiramente mais baixas. Tal é consequência deste dia tipo se verificar com frequência considerável durante todo o ano, com uma baixa muito significativa nos meses de verão. O dia tipo 'E' apresenta também uma gama de temperaturas baixa por apresentar uma predominância de ocorrência nos meses de outubro e novembro.

	Duração (min)	# [C1,C2,C6]	Centro
7	5	0,0,1	A
8	6	0,0,9	A
9	6	1,0,34	A
10	8	0,1,13	A
11	5	0,0,2	A
12	5	0,0,1	A
13	5	0,0,15	A
14	6	0,0,8	A
15	5	0,0,2	A
17	5	0,0,1	A

a) Março;

	Novembro		
	Duração (min)	# [C4,C5,C6,C7]	Centro
9	5	0,0,3,0	A
10	6	2,1,0,1	B
11	7	2,0,3,1	A
14	6	2,1,0,1	B
15	8	1,0,1,0	B
16	6	4,1,2,0	B

b) Novembro;

Tabela 4.20: Caracterização dos dias tipo com atividade, para a máquina de café, em modo base, para o dia tipo 'F'.

	Duração (min)	# [C6]	Centro
15	5	1	A

a) Março;

	Duração (min)	# [C6]	Centro
17	5	2	A

b) Agosto;

	Duração (min)	# [C4,C5,C6,C7]	Centro
13	10	0,0,0,1	BA
14	6	3,2,2,0	B
15	6	1,1,1,1	B
16	6	4,2,0,0	B
17	8	0,0,2,0	AA

c) Novembro;

Tabela 4.21: Caracterização do dias tipo com atividade, para a máquina de café, em modo base para o dia tipo 'E'.

As tabelas 4.20 e 4.21 permitem analisar um dia tipo, em que é registada atividade, para os meses de março, agosto e novembro, no modo base.

Para o mês de março, para o dia tipo 'F' (tabela 4.20 a)), existe uma predominância de arranques entre as 8 e as 10 da manhã e entre as 13 e 14 horas.

O mês de agosto não é analisado, para o dia tipo 'F' (tabela 4.20), por não conter dias assim classificados.

Em 4.21 a) e b), março e agosto, verifica-se que a existência de consumos nos dias tipo 'E' é residual.

No mês de novembro verificam-se arranques, em número considerável, para os dias tipo 'F' e 'E'. Para os dias tipo 'F', os arranques são em menor número que no mês de março, e o pico, embora pouco evidente, é registado às 16 horas, revelando que também para este dia tipo se verifica uma tendência de prevalência de consumos durante a tarde, para os últimos meses da janela temporal. Para o dia tipo 'E', os consumos são distribuídos ao longo da tarde, entre as 14 e 16 horas. De notar que para ambos os dias tipo, no mês de novembro, os consumos têm maioritariamente centros horários de patamar mais elevado, isto é, ocorre maioritariamente o patamar B, ao contrario do A registado em março.

As médias de duração dos consumos são, para todas as horas, entre os 5 e os 10 minutos, o que vai de encontro a hipótese de que o estado ativo do equipamento é da ordem do intervalo de tempo de aquisição do modo base.

Refrigerador de Água

O refrigerador de água é nada mais do que um depósito de água, que é constantemente refrigerada, e se destina a consumo dos trabalhadores. O seu gasto energético é monitorizado por uma *smart plug*.

A divisão entre a atividade e não atividade do equipamento, é feita pelo cálculo do valor máximo registado na base de dados de consumos, não considerando *outliers*, e dividindo este máximo por dez, tanto no modo de aquisição horário como no base.

	Mínimo	Máximo	Média
ON	0.002	0.038	0.011
OFF	0.000	0.002	0.0003

a) Modo base, valores em kWh;

	Mínimo	Máximo	Média
ON	0.013	0.110	0.055
OFF	0.000	0.012	0.004

b) Modo horário, valores em kWh;

Tabela 4.22: Divisão dos valores de consumo bruto (3.1) em patamares de consumo ON e OFF (3.11), para o refrigerador de água.

Pela tabela 4.22, verifica-se que a gama de valores do patamar ON, no modo horário, é muito superior à gama detetada para o modo base. Tal pode significar que o refrigerador está ativo durante um período considerável, a cada hora em que é detetado estado ON.

	Mínimo	Máximo	Média
A	0.002	0.004	0.003
B	0.005	0.009	0.007
C	0.010	0.011	0.011
D	0.012	0.015	0.013
E	0.016	0.019	0.018
F	0.038	0.038	0.038

a) Modo base, valores em kWh;

	Mínimo	Máximo	Média
A	0.013	0.056	0.035
B	0.057	0.073	0.065
C	0.074	0.096	0.085
D	0.098	0.110	0.103

b) Modo horário, valores em kWh;

Tabela 4.23: Divisão dos valores de ON (3.11), em patamares de consumo, para o refrigerador de água.

	Centro de cada cluster*	
	Dados base	**Dados horários**
C1	AB^2 (13.954%)	BC (62.414%)
C2	$ABCB^6$ (83.363%)	A^2 (37.587%)
C3	AB^2CB^9A (2.683%)	-

Tabela 4.24: Centro de cada *cluster* e respetiva percentagem de ocorrência, para o refrigerador de água (# 1118-modo base; # 1011-modo horário).

A tabela 4.24 revela os centros das *clusters* mais frequentes. Para o modo base, a *cluster* mais frequente, C2, tem um centro constituído por 9 caracteres, ou seja, tem uma duração de 45 minutos, indo de encontro ao já suspeito de que o período de estado ativo do equipamento é mais próximo da ordem do modo de aquisição horário que do modo base. De notar que o centro de C2 monstra que se vai escalando nos patamares,

de A a C, e depois se volta para patamares mais baixos, B. Significa que o aparelho atinge um pico de forma gradual, o que pode ser originado por um consumo periódico.

No modo horário, os centros das duas *clusters* identificadas têm uma duração de duas horas.

Com isto se chega à conclusão que, ou o equipamento tem consumos longos, mas não apresenta muitos arranques em horas sucessivas, ou então o consumo é periódico, mas com picos de consumo, em amplitude, de duração de duas horas, sendo que, estes picos, por terem grande amplitude, camuflam os restantes consumos, acusando-os como consumos de fundo (OFF).

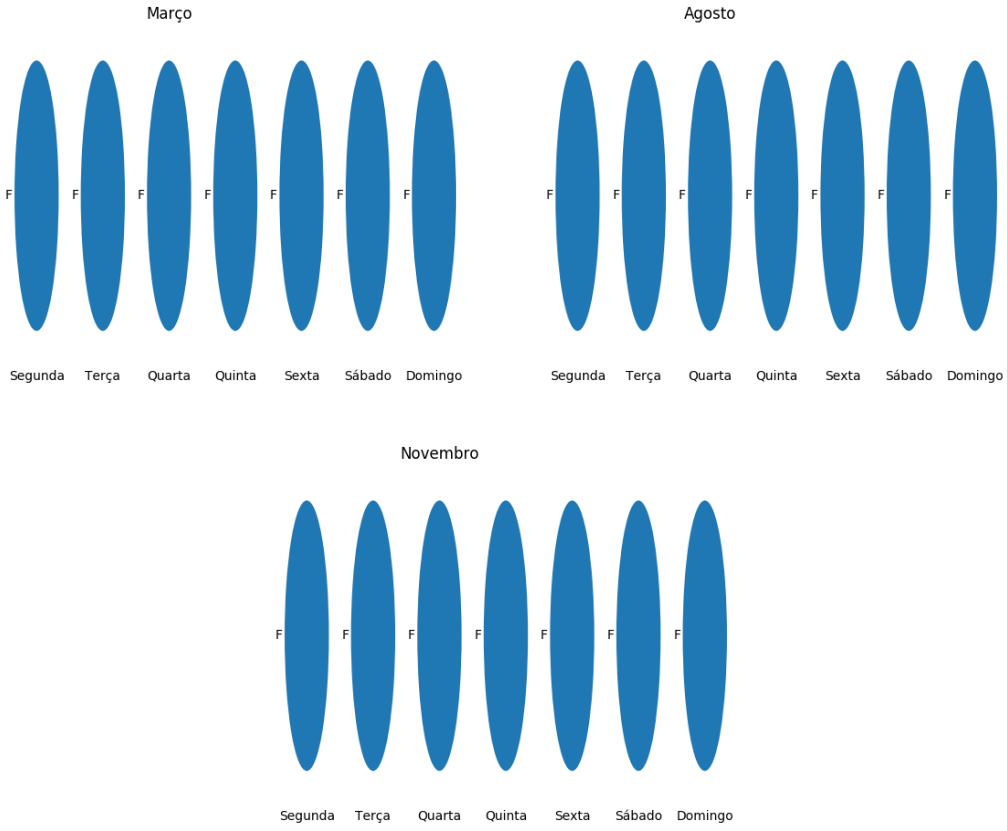


Figura 4.37: Representação da distribuição dos dias tipo, pelos dias da semana, no modo base

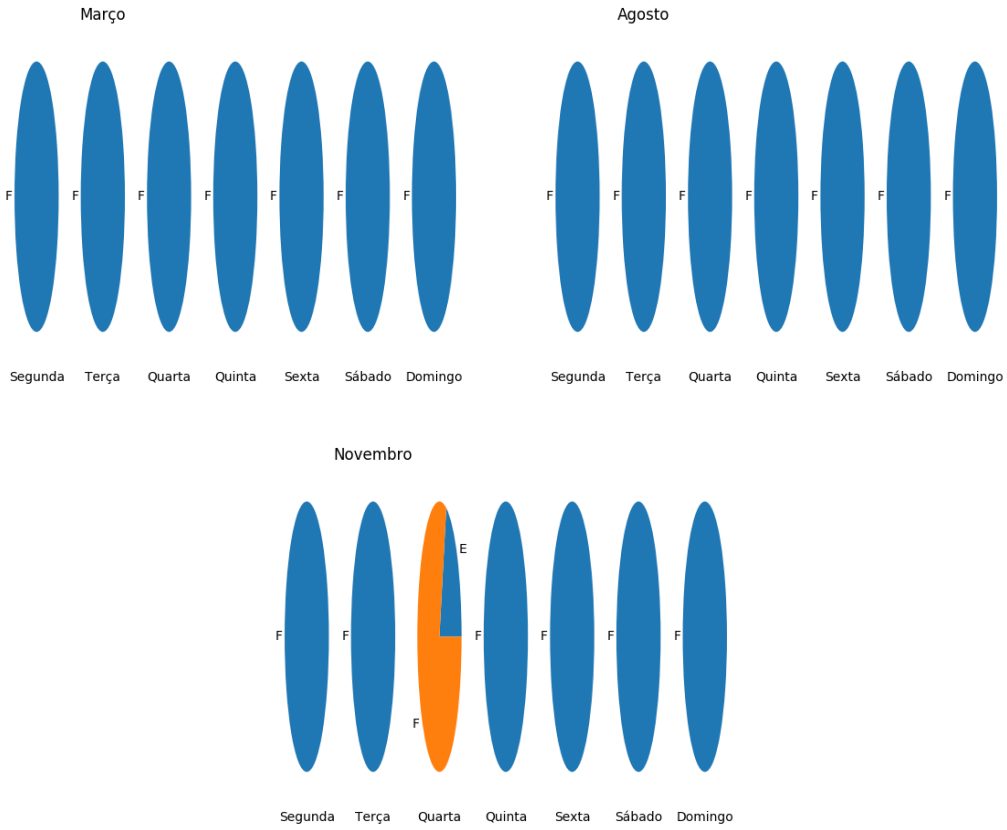


Figura 4.38: Representação da distribuição dos dias tipo, pelos dias da semana, no modo horário

Por análise das figuras 4.37 e 4.38, é notória a predominância do dia tipo 'F', para todos os dias analisados, em todos os meses, para ambos os modos de aquisição.

O facto de todos os dias se classificarem com o mesmo dia tipo, leva a concluir que existe uma forte chance da atividade do aparelho ser independente da atividade do edifício.

Como consequência de praticamente todos os dias serem classificados como 'F', a distribuição destes dias tipo, para o ambos os modos de aquisição, é uniforme para os meses da janela temporal em análise, como se pode verificar na figura 4.39.

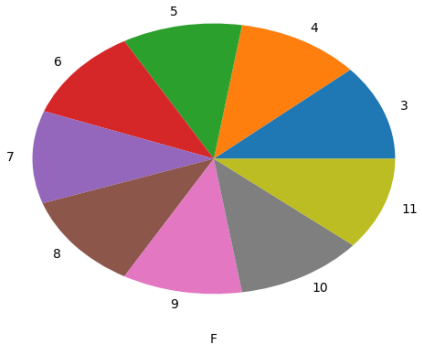


Figura 4.39: Representação da distribuição dos dias tipo relevantes, ao longo dos meses da janela temporal, no modo base.

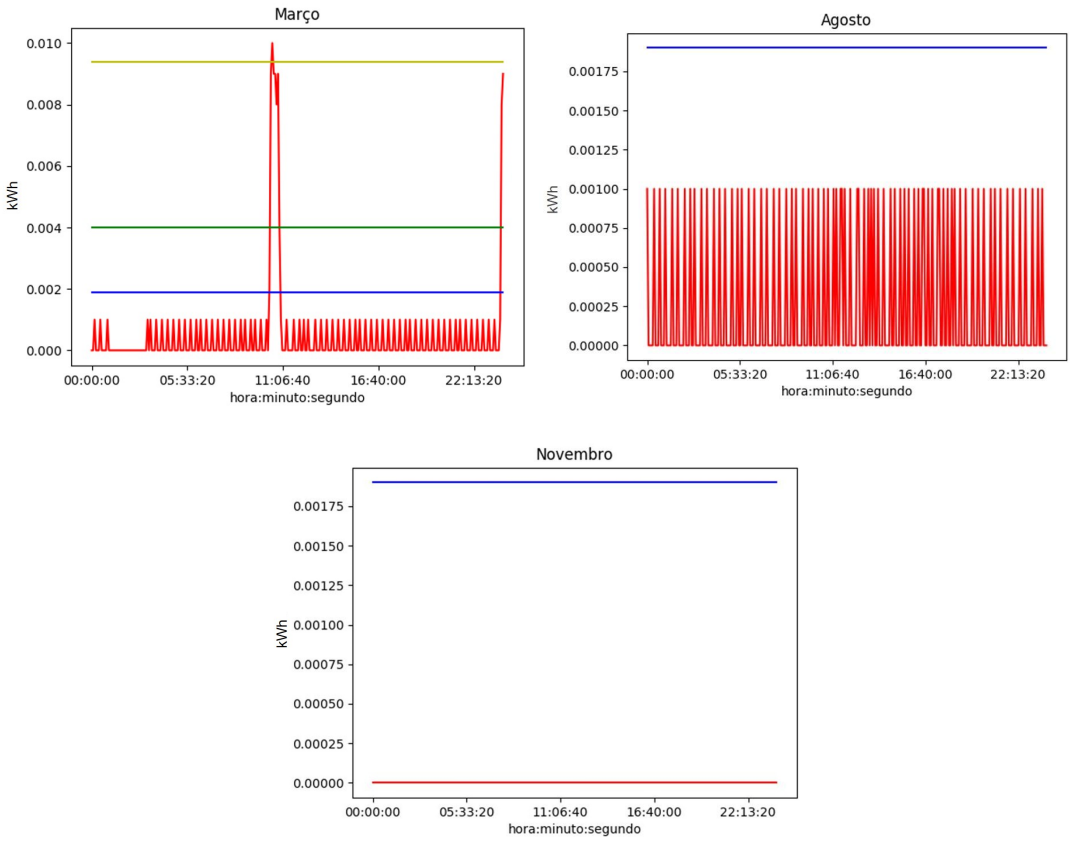


Figura 4.40: Representação, em modo base, para o refrigerador de água, dos dias tipo 'F' a vermelho.

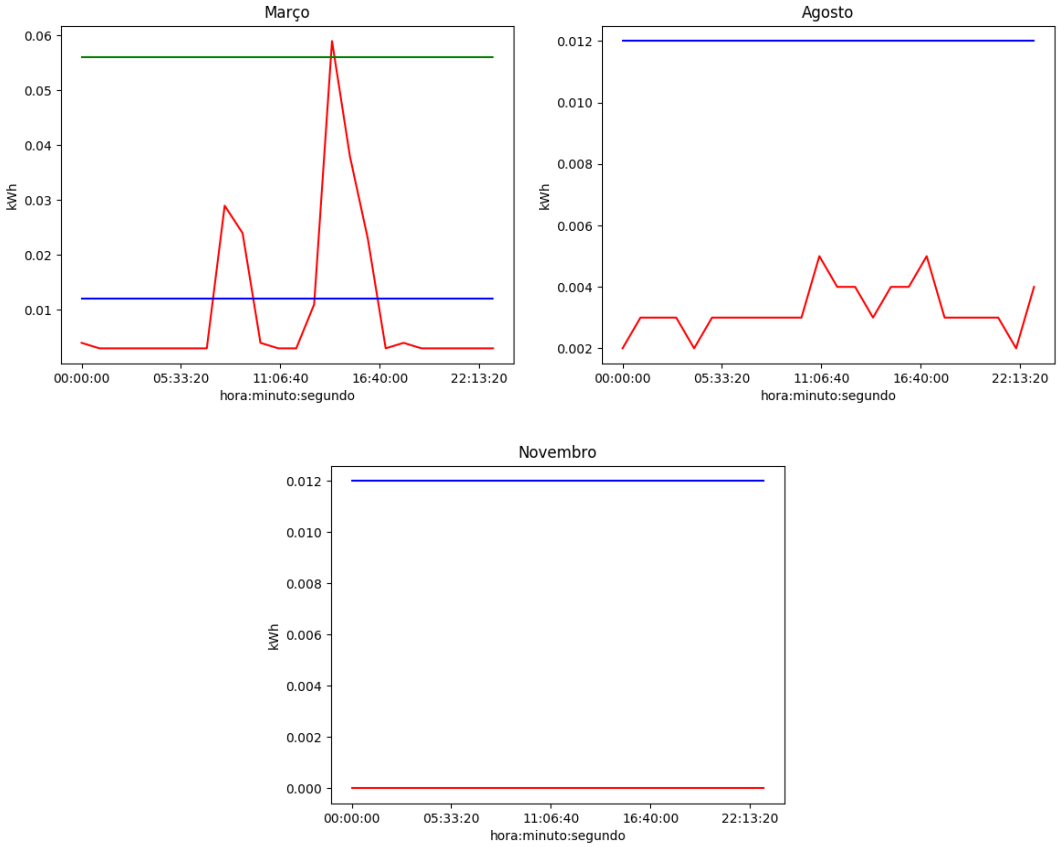


Figura 4.41: Representação, em modo horário, para o refrigerador de água, dos dias tipo 'F' a vermelho.

Pela análise das figuras 4.40 e 4.41, constata-se que por o refrigerador de água ter consumos a rondar uma hora, os gráficos base e horários revelam informações redundantes. O gráfico horário representa os dados de forma mais suave. No entanto, a ideia de que os consumos são periódicos, com alguns picos de amplitude ao longo do dia, com as restantes amplitudes camufladas (estado OFF) por estes picos, é sustentada com estas imagens. De relembra que a linha horizontal azul divide o estado ativo do inativo, no equipamento

Conclui-se que para o mês de março, um dia típico tem dois picos, um de manhã e um à tarde. No mês de agosto existe apenas um consumo residual (periódico) e em novembro não existe qualquer consumo, dando a ideia que o refrigerador esteve desli-

gado.

Por comparação com os restantes meses da janela temporal, não representados nas figuras 4.40 e 4.41, constata-se que o mês de agosto e novembro não são de todo regra, o comportamento mais habitual destes dois equipamentos é da existência de dois picos de consumo ao longo de um dia.

Por a figura 4.40 e 4.41 serem redundantes e o estado ativo do equipamento não atingir uma hora, as análises seguintes são feitas para o modo de aquisição base.

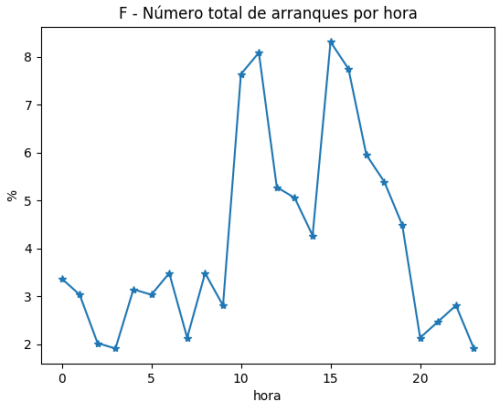


Figura 4.42: Distribuição do número de arranques, para o refrigerador de água, para cada hora, no modo base.

Na figura 4.42 são apresentados dois picos, aproximadamente entre as 11 e as 12 horas e entre as 16 e as 17 horas. Estes picos de consumo podem ter origem num maior uso do refrigerador nestas horas, que resulte num maior gasto energético, visto se apresentarem durante o horário laboral. No entanto, os arranques nas restantes horas não são desprezáveis.

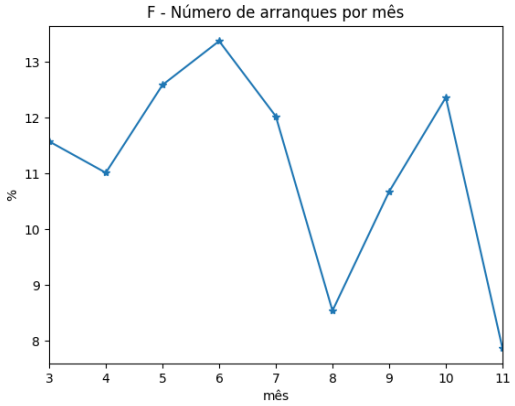


Figura 4.43: Distribuição do número de arranques, para o refrigerador de água, para cada mês, no modo base.

A figura 4.43 revela uma distribuição aproximadamente uniforme do número de arranques, ao longo dos meses, há um pequeno decréscimo em abril, um claro decréscimo no mês de agosto e um número de arranques mínimo no mês de novembro, que vai de encontro ao verificado nas figuras 4.40 e 4.41, em que não são registados consumos acima da linha horizontal azul para os meses de agosto e novembro.

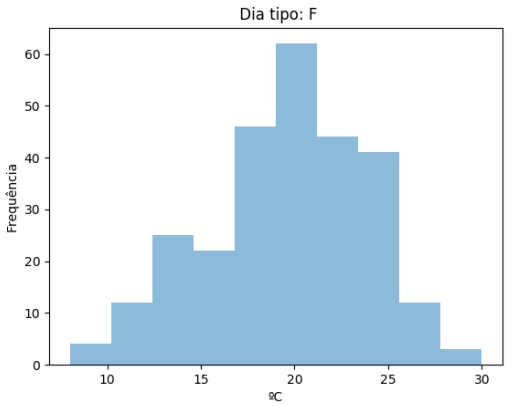


Figura 4.44: Distribuição da temperatura máxima, para o refrigerador de água, no modo base.

A distribuição de temperaturas representada na figura 4.44 apresenta a mesma distribuição que a apresentada na figura 4.8, o que não é de estranhar visto a base de

dados ser classificada com este dia tipo, de forma massiva. Portanto, este dia tipo não é sazonal.

	Março		
	Duração (min)	# [C1,C2,C3]	Centro
0	40	0,3,0	AB^2CB^4
1	32	0,2,0	B^2CB^2A
2	39	0,4,0	AB^7
3	35	0,1,0	BCB^4A
4	35	0,2,0	BCB^5
5	29	1,3,0	$ABCB^4$
6	37	0,3,0	B^7
7	40	0,2,0	BCB^5A
8	37	1,3,0	B^7A
9	43	0,3,0	B^2C^4A
10	47	0,2,0	$ABCB^5A$
11	47	0,3,0	B^3CB^6
12	39	1,7,1	AB^3CB^4
13	51	0,7,0	AB^7
14	48	0,3,0	ACB^7
15	44	0,4,0	BCB^2CB^4
16	30	0,5,0	BCB^5
17	46	2,5,0	$AB^2(CB)^2BA$
18	29	0,6,0	A
19	26	3,4,0	A
20	40	0,2,0	$A(BC)^2B^2A$
21	40	0,2,0	AB^3CB^2A
22	35	1,1,0	$ABCB^3$
23	35	0,2,0	BCB^5

Tabela 4.25: Caracterização do dia tipo 'F', em modo base para o mês de março.

	Agosto		
	Duração (min)	# [C1,C2,C3]	Centro
0	39	0,5,0	B^7A
1	35	0,3,0	BCB^5
3	35	0,3,0	BCB^4A
4	35	0,4,0	BCB^5
5	35	0,3,0	$(BC)^2B^3$
6	35	0,2,0	AB^3CB^2
8	33	0,3,0	BCB^4
9	35	0,2,0	BCB^5
10	50	0,5,0	B^2CB^5A
11	43	1,8,0	B^2CB^6
12	47	0,3,0	B^3CB^6
13	45	0,1,0	BCB^6A
14	42	0,3,0	B^2CB^5
15	35	1,4,0	$ABCB^4$
16	49	0,6,1	$(BC)^2B^5A$
17	42	1,6,0	$A(CB)^3BA$
18	45	0,3,0	$AB^2(CB)^2BA$
19	40	0,2,0	AB^2CB^4
20	35	0,3,0	B^7
21	35	0,1,0	B^3CB^3
23	35	0,2,0	AB^2CB^3

Tabela 4.26: Caracterização do dia tipo 'F', em modo base para o mês de agosto.

	Novembro		
	Duração (min)	# [C1,C2]	Centro
0	25	1,2	B^6
1	25	1,2	B^5A
2	35	0,2	AB^5A
3	30	0,1	B^2CB^3
4	30	2,2	B^5
5	30	0,2	BCB^4
6	32	0,2	$ABCB^3A$
8	30	0,3	B^6
9	50	0,2	AB^2CBCB^5
10	43	0,6	B^2CB^5
11	31	1,6	B^7
12	32	0,2	B^6
13	35	0,3	B^6
14	31	2,5	$ABCB^4$
15	38	0,5	B^3CB^3A
16	40	0,5	$ABCB^5$
17	33	0,3	CB^2CB^3
18	35	0,2	B^2CB^4
19	30	0,1	BCB^4
21	32	0,4	B^2CB^3
22	32	0,2	BCB^4
23	30	0,1	B^6

Tabela 4.27: Caracterização do dia tipo 'F', em modo base para o mês de novembro.

Nas tabelas 4.25:27, praticamente todos os arranques têm uma duração entre os 20 e os 50 minutos e estão bem divididos ao longo do tempo, com ligeiros picos no período da manhã e da tarde, não muito evidentes. De notar que, com o avançar dos meses, verifica-se um ligeiro decréscimo no número de arranques. De relembrar o concluído com a figura 4.43, o número de arranques é maior para o mês de março, diminuí em agosto e é mínimo em novembro.

Discussão

Para a empresa piloto em estudo, os períodos de estado ativo de um equipamento, que dependem do uso por parte de um utilizador (impressora, computadores e máquina de

café), vão de encontro à atividade expectável para um edifício tipo empresarial, com os dias e horas laborais típicas. O seu uso ocorre em dias úteis, segunda a sexta-feira, com a exceção de feriados, entre as 9 e as 19 horas, com uma quebra na hora de almoço. É evidente uma quebra de uso deste tipo de aparelho nos meses de verão, no número de arranques total, a iniciar-se no mês de julho e até setembro (com ligeiras variações mas ocorrendo essencialmente nestes meses). Mostram também uma quebra de consumo nos mês de abril. Abril, em 2017, teve um feriado, 25 de abril, a uma terça-feira, que tendencialmente leva os trabalhadores a fazerem fim-de-semana prolongado, é um mês de 30 dias e a Páscoa ocorreu neste mês. Estes podem ser os motivos do decréscimo.

Verifica-se uma certa tendência dos dias sem consumos a ocorrerem em dias quentes, provavelmente, devido ao número de arranques ser menor nos meses de verão, por consequência de uma menor atividade do edifício.

O refrigerador de água não depende da atividade do edifício, apresentando um consumo periódico. No entanto, apresenta picos de arranques, dois, ao longo de um dia, em períodos em que se estima haver atividade no edifício.

4.2.2 Casa Piloto

Para esta casa piloto assume-se (não sendo necessariamente verdade) que habita uma família de quatro pessoas, que estão ausentes do edifício entre as 9 e as 19 horas, aproximadamente, em dias úteis (segunda a sexta-feira).

Para a casa piloto são monitorizados um frigorífico e uma máquina de lavar, com recurso a *smart plugs*, e os consumos agregados da habitação, com um *smart meter*.

Frigorífico

Para o frigorífico, por análise direta da base de dados, concluí-se que os períodos em que não há atividade no frigorífico, o seu consumo é nulo. Portanto, a divisão, para ambos os modos de aquisição, resume-se a atribuir o patamar ON a consumos superiores a zero.

	Mínimo	Máximo	Média
ON	$1.429 * 10^{-4}$	0.112	0.020
OFF	0.000	0.000	0.000

a) Modo base, valores em kWh;

	Mínimo	Máximo	Média
ON	0.002	0.22	0.043
OFF	0.000	0.000	0.000

b) Modo horário, valores em kWh;

Tabela 4.28: Divisão dos valores de consumo bruto (3.1) em patamares de consumo ON e OFF (3.11), para o frigorífico.

Com a tabela 4.28, conclui-se que a gama de valores do período ON, no modo horário, é aproximadamente o dobro do modo base. É assim suspeito que exista consumos, a cada hora de consumo ativo, por um período de, aproximadamente, 5 a 15 minutos, não necessariamente sucessivos.

	Mínimo	Máximo	Média
A	$1.429 * 10^{-4}$	0.025	0.009
B	0.028	0.038	0.033
C	0.044	0.055	0.049
D	0.081	0.081	0.081
E	0.112	0.112	0.112

a) Modo base, valores em kWh;

	Mínimo	Máximo	Média
A	0.002	0.047	0.026
B	0.050	0.053	0.052
C	0.103	0.142	0.121
D	0.220	0.220	0.220

b) Modo horário, valores em kWh;

Tabela 4.29: Divisão dos valores de ON (3.11), em patamares de consumo, para o frigorífico.

Na tabela 4.29 estão representados os patamares de consumo pela qual a base de dados se divide. De notar que o patamar mais baixo de consumo ativo, A, apresenta uma média, aproximadamente, três vezes mais baixa para o consumo base, relativamente ao horário. De resto, os restantes patamares apresentam gamas de valores muito curtas, o que pode significar que classificam registos *outliers*.

	Centro de cada *cluster*	
	Dados base	**Dados horários**
C1	A^4 (99.892%)	A (83.333%)
C2	A^{24} (0.108%)	$A^{1519}BA^{1560}C^7$ (16.666%)

Tabela 4.30: Centro de cada *cluster* e respetiva percentagem de ocorrência, para o frigorífico (# 8316-modo base; # 12-modo horário).

A tabela 4.30 revela um número residual de arranques no modo horário, 12. Em

comparação com o número de arranques no modo base, 8316, é uma diferença muito significativa. Esta situação tem, provavelmente, origem num funcionamento cíclico do frigorífico, com períodos inferiores a uma hora. Por consequência, é quase sempre acusado o modo operacional do aparelho, no modo horário.

Para o modo de aquisição base, a *cluster* mais frequente é a C1, que tem um centro constituído por quatro A's sucessivos. Representa um consumo constante, de baixo patamar e com duração de 20 minutos.

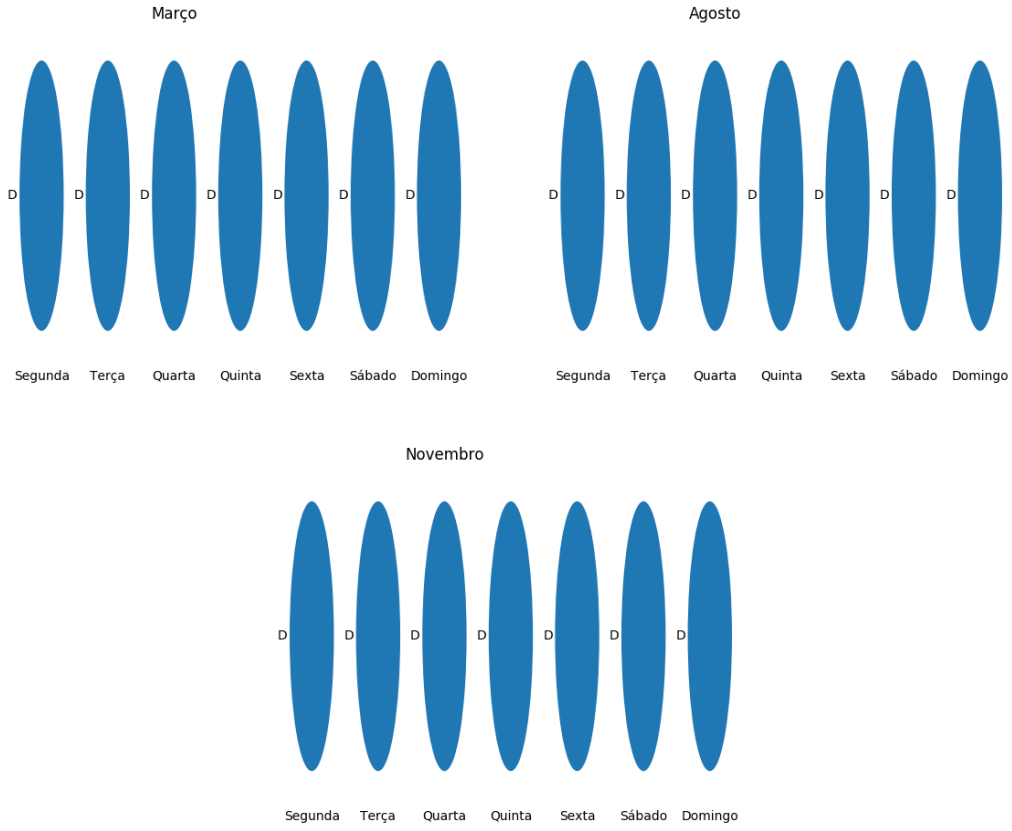


Figura 4.45: Representação da distribuição dos dias tipo, para o frigorífico, pelos dias da semana, no modo base.

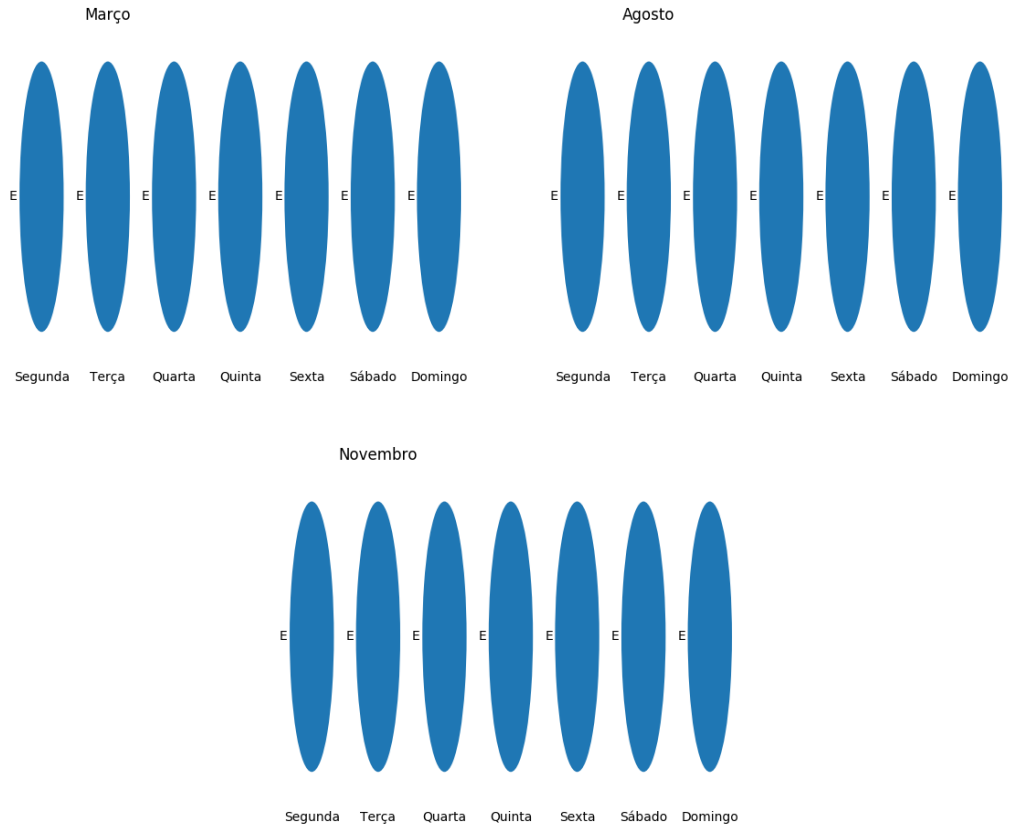


Figura 4.46: Representação da distribuição dos dias tipo, para o frigorífico, pelos dias da semana, no modo horário.

Como já tinha ocorrido, para o refrigerador de água, na análise à empresa piloto, quer no modo horário, quer no modo base, todos os dias são classificados com um único dia tipo. Para o modo base é atribuído o caracter 'D' e no modo horário o 'E'.

Visto os dias disponíveis nas base de dados serem classificados de forma massiva com os mesmos dias tipo, a distribuição destes é uniforme ao longo dos meses em análise, como se pode verificar na figura 4.47.

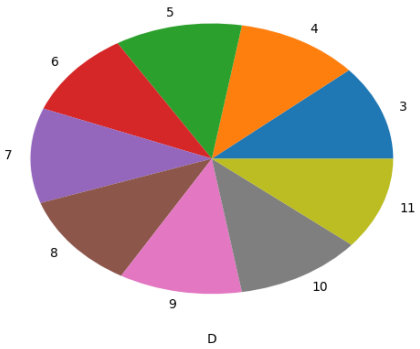


Figura 4.47: Representação da distribuição dos dias tipo relevantes, para o frigorífico, ao longo dos meses da janela temporal, no modo base.

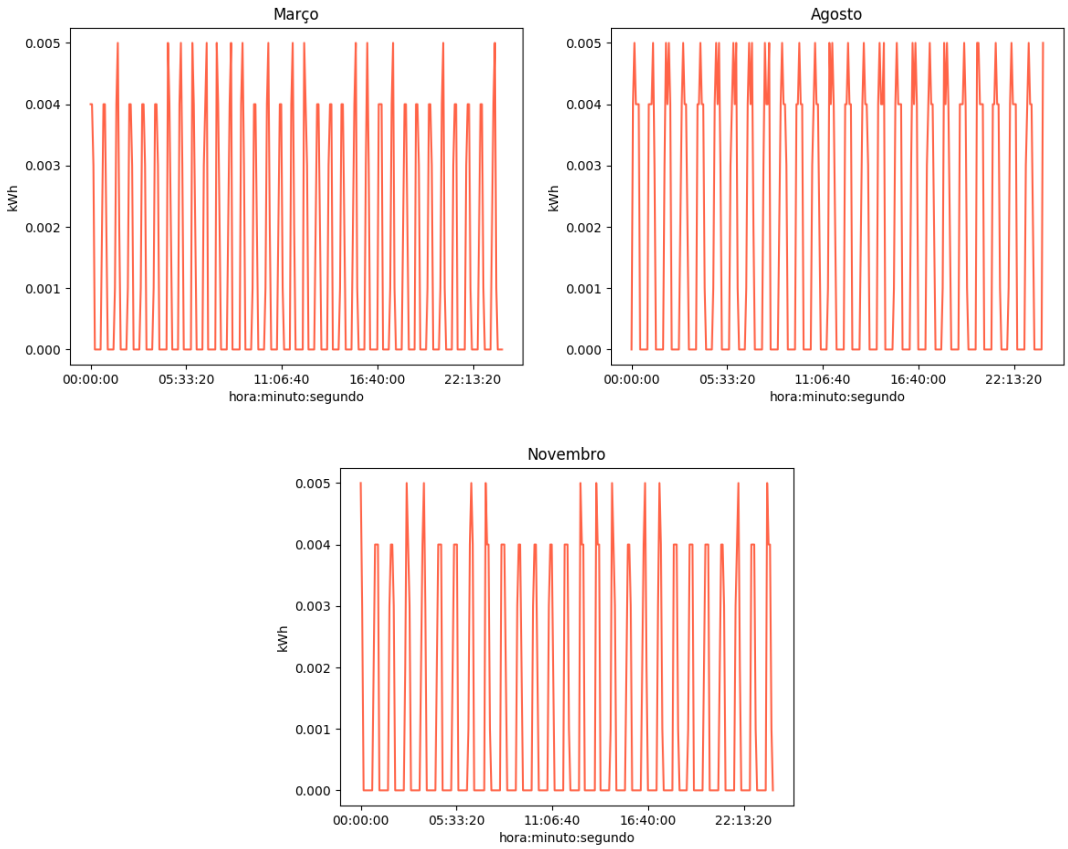


Figura 4.48: Representação, em modo base, para o frigorífico, do dia 'D' a vermelho.

Na figura 4.48, verificam-se picos de consumo em intervalos constantes e de forma recorrente. A figura mostra que o equipamento apresenta estados ativos de forma cíclica. Visto o centro da *cluster* mais frequente, para este modo de aquisição, ser constituído por quatro caracteres, espera-se que esses período de ativação do equipamento sejam dessa ordem. De notar que os picos de consumo variam a sua amplitude.

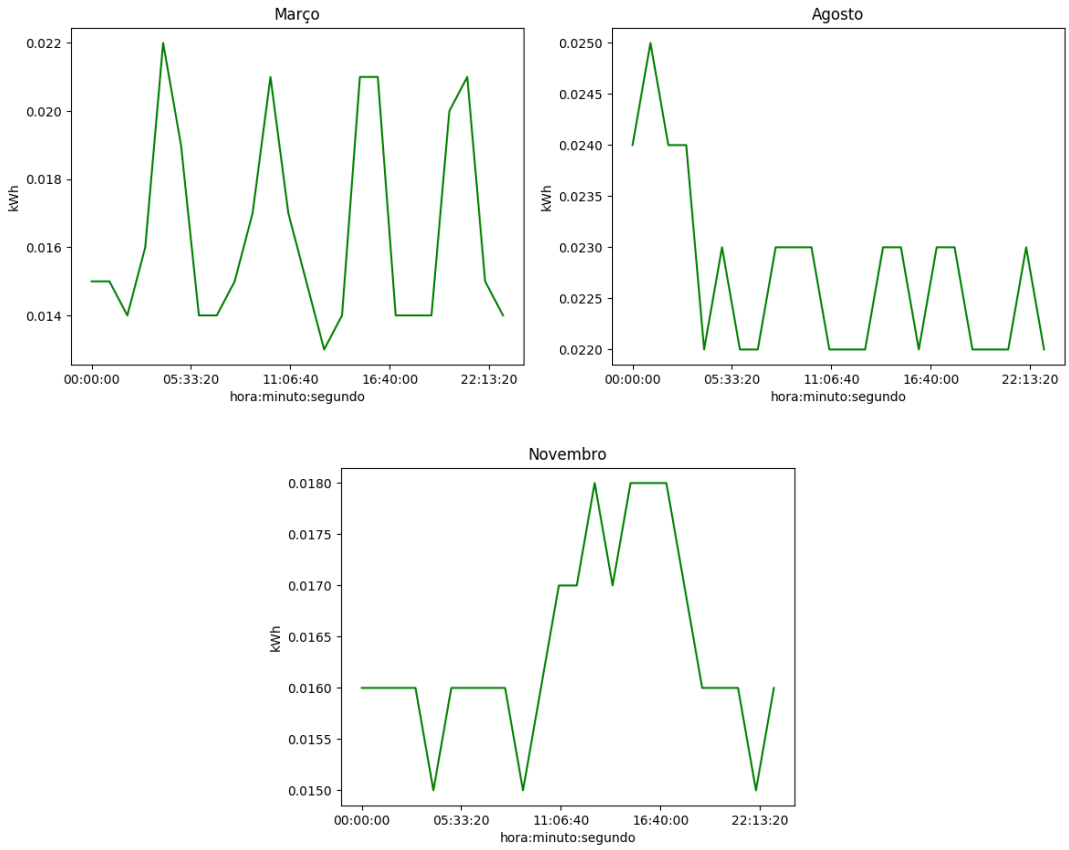


Figura 4.49: Representação, em modo horário, para o frigorífico, do dia 'E' a verde.

A figura 4.49 mostra um consumo sempre acima do patamar ON/OFF, que corresponde ao eixo das abcissas, consequência de se verificar, sempre, pelo menos um arranque a cada hora. No entanto, os picos horários alteram bastante a sua amplitude, ao longo do dia. Este último facto tem origem na variação dos picos de consumo ativo identificados para 4.48. A figura 4.49 acaba por representar uma modelagem dos picos

de consumo. É notório o aumento da amplitude dos picos no mês de agosto, estes picos também duram mais tempo para este mês e para o mês de novembro, isto é, a largura dos picos é maior.

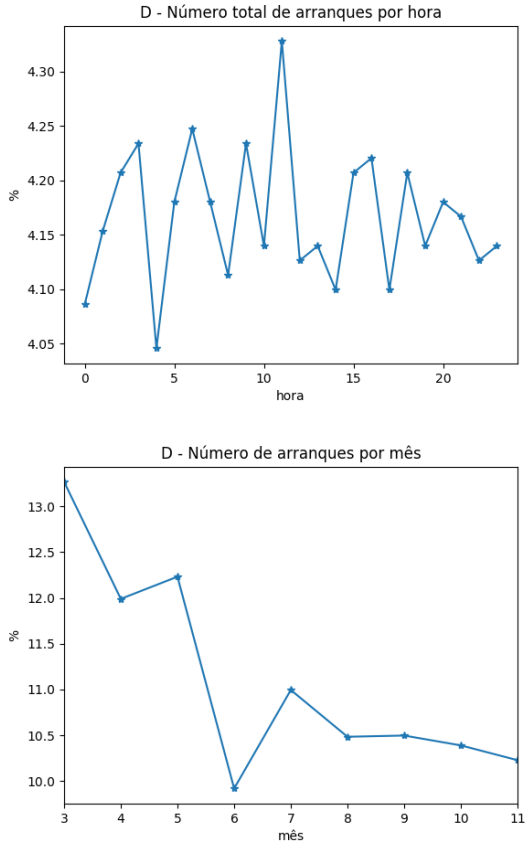


Figura 4.50: Distribuição do número de arranques, para cada hora e para cada mês, para os dias tipo 'D', no modo base.

A análise do número de arranques, para cada hora e para cada mês, é apresentada apenas no modo base, pois, apenas este modo apresenta o número real de arranques do frigorífico. Estes arranques são representados na figura 4.50. Ao longo do período de um dia, para a análise do número de arranques a cada hora, não se verifica nenhuma tendência. O número de arranques a cada hora ocorre aproximadamente com a mesma frequência, as probabilidades de arranque variam entre os 4% e os 4.3%, sendo que se

verifica uma ligeira tendência de aumento de probabilidade e sucessivo decréscimo na hora seguinte, de forma periódica. Tal facto reflete assim que, a densidade de arranques é periódica, contribuindo para a forma dos gráficos nas figuras 4.48.

Quanto à distribuição de arranques ao longo dos meses, a figura 4.50 também tem uma baixa gama de variação, isto é, a probabilidade de um arranque ocorrer em cada mês varia aproximadamente entre 10% e os 13%. No entanto, é notória uma quebra continua entre os primeiros quatro meses da janela temporal, apresentando um pico mínimo no mês de junho. A partir do mês referido, o número de arranques para cada mês é aproximadamente uniforme. Assumindo o funcionamento periódico do equipamento, esta variação pode ter origem num aumento do período de funcionamento do equipamento.

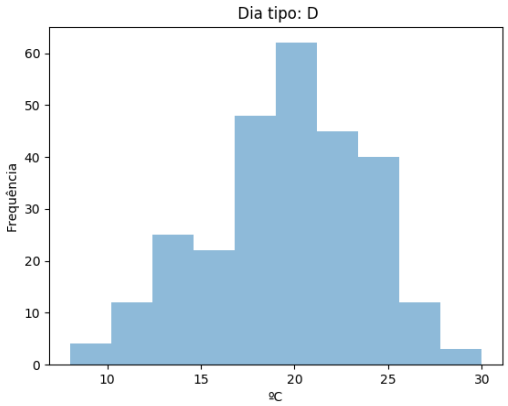


Figura 4.51: Distribuição da temperatura máxima, para o frigorífico, para os dias tipo 'D', no modo base.

Mais uma vez, e por consequência de praticamente todos os dias, no modo base, serem classificados como 'D', a figura 4.51 é em tudo semelhante à apresentada em 4.8, concluindo-se assim que não existe uma sazonalidade característica deste dia tipo.

	Março		
	Duração (min)	# [C1]	Centro
0	19	37	AAAA
1	19	37	AAAA
2	19	43	AAAA
3	19	45	AAAA
4	19	39	AAAA
5	19	41	AAAA
6	19	40	AAAA
7	19	39	AAAA
8	18	44	AAAA
9	18	40	AAAA
10	18	43	AAAA
11	19	41	AAAA
12	19	40	AAAA
13	19	44	AAAA
14	19	36	AAAA
15	19	42	AAAA
16	19	41	AAAA
17	19	39	AAAA
18	19	44	AAAA
19	19	40	AAAA
20	19	42	AAAA
21	19	40	AAAA
22	19	41	AAAA
23	20	45	AAAA

Tabela 4.31: Caracterização do dia tipo 'D', em modo base para o mês de março.

	Agosto		
	Duração (min)	# [C1,C2]	Centro
0	29	27,0	A^6
1	28	33,0	A^5
2	28	33,0	A^5
3	27	33,0	A^5
4	28	32,0	A^5
5	28	36,0	A^5
6	28	30,0	A^5
7	28	34,0	A^5
8	26	34,0	A^5
9	27	33,0	A^5
10	28	32,0	A^5
11	28	31,0	A^5
12	27	34,0	A^5
13	26	33,0	A^5
14	26	31,0	A^5
15	28	32,0	A^5
16	27	33,0	A^5
17	30	29,1	A^6
18	36	35,2	A^5
19	30	28,1	A^5
20	29	29,0	A^5
21	28	33,0	A^5
22	31	31,1	A^6
23	28	32,0	A^5

Tabela 4.32: Caracterização do dia tipo 'D', em modo base para o mês de agosto.

	Novembro		
	Duração (min)	# [C1,C2]	Centro
0	23	29,0	A^4
1	23	31,0	A^4
2	24	30,0	A^4
3	22	33,0	A^4
4	23	33,0	A^4
5	22	31,0	A^4
6	23	31,0	A^4
7	22	31,0	A^5
8	23	29,0	A^4
9	22	37,0	A^4
10	23	31,0	A^4
11	22	34,0	A^4
12	23	31,0	A^4
13	22	32,0	A^4
14	22	32,0	A^4
15	22	33,0	A^4
16	22	32,0	A^5
17	22	32,0	A^4
18	22	30,0	A^4
19	27	34,1	A^4
20	23	31,0	A^4
21	22	30,0	A^4
22	24	33,0	A^4
23	24	30,0	A^4

Tabela 4.33: Caracterização do dia tipo 'D', em modo base para o mês de novembro.

As tabelas 4.31:33 caracterizam os dias com atividade nos meses de março, agosto e novembro. Não se registam horas preferências para picos, o que sustenta a ideia do funcionamento do equipamento ser periódico. A duração dos arranques é superior em agosto e novembro, o que sustenta a hipótese apresentada para a figura 4.50. Se o período é maior nestes meses, é natural que exista uma ligeira quebra de arranques.

Máquina de Lavar Roupa

Igualmente ao ocorrido para o frigorífico, a máquina de lavar roupa presente na casa piloto, tem uma monitorização de consumos com recurso a uma *smart plug*. A divisão

do período ON e OFF é feita com o critério de que todos os consumos, superiores a zero kWh, são considerados estados ativos do aparelho, em ambos os modos de aquisição. Este critério foi averiguado por análise direta dos dados.

	Mínimo	Máximo	Média
ON	0.001	0.172	0.078
OFF	0.000	0.000	0.000

a) Modo base, valores em kWh;

	Mínimo	Máximo	Média
ON	0.001	1.421	0.154
OFF	0.000	0.000	0.000

b) Modo horário, valores em kWh;

Tabela 4.34: Divisão dos valores de consumo bruto (3.1) em patamares de consumo ON e OFF (3.11), para máquina de lavar roupa.

Com a tabela 4.34, verifica-se que o consumo médio e máximo, no modo horário, é uma ordem de grandeza superior ao mesmo consumo no modo base. Provavelmente, o intervalo de tempo em que a máquina de lavar roupa funciona, quando é detetado um estado ativo no modo de aquisição horário, numa certa hora, no intervalo de tempo correspondente, no modo base, é detetado estado ativo num período de tempo elevado desse intervalo.

	Mínimo	Máximo	Média
A	0.001	0.103	0.048
B	0.105	0.152	0.129
C	0.155	0.172	0.163

a) Modo instantâneo, valores em kWh;

	Mínimo	Máximo	Média
A	0.001	0.206	0.079
B	0.233	0.466	0.349
C	0.506	0.646	0.592
D	1.421	1.421	1.421

b) Modo horário, valores em kWh;

Tabela 4.35: Divisão dos valores de ON (3.11), em patamares de consumo, para a máquina de lavar roupa.

Na tabela 4.35, os patamares de estados ativos apresentam gamas de valores de amplitude consideráveis. Apenas no modo horário, o patamar D se limita a um valor possível, o que tudo indica ser um *outlier*. Portanto, é expectável haver alguma heterogeneidade de patamares nos padrões de consumo.

	Centro de cada *cluster*	
	Dados base	**Dados horários**
C1	A^7 (80.494%)	A^{18} (0.758%)
C2	ABA^{22} (19.506%)	BA^{10} (9.091%)
C3	-	A^4BA^4 (15.152%)
C4	-	A^3 (75.000%)

Tabela 4.36: Centro de cada *cluster* e respetiva percentagem de ocorrência, para a máquina de lavar (# 405 modo base; # 132-modo horário).

De acordo com a tabela 4.36, os centros de ambas as *clusters*, no modo base, têm probabilidades de ocorrência consideráveis. O centro de C1 tem uma duração de 35 minutos e C2 de duas horas, sendo que C1 é mais provável que C2. É assim sustentada a hipótese de que o estado ativo, em modo base, tem um funcionamento longo. No modo horário existem duas *clusters*, com probabilidades de ocorrência significativas, C3 e C4. C3 apresenta uma duração de 9 horas e C4, mais provável que C3, de 3 horas.

A máquina de lavar roupa apresenta um funcionamento muito pontual e aleatório no tempo, isto é, espera-se que uma máquina de lavar roupa trabalhe entre 2 a 5 vezes, para uma família de quatro elementos, durante uma semana, sem ser expectável que funcione sempre nos mesmos dias, às mesmas horas. Os dias tipo foram portanto classificados em apenas dois tipos. Dias em que é detetado funcionamento na máquina de lavar (representados pela letra 'E'), e dias em que não é detetado qualquer estado ativo no equipamento (representados pela letra 'F'). Os resultados das classificações encontram-se representados em 4.52 e 4.53.

As figuras 4.52 e 4.53 são equivalentes. É visível uma predominância dos dias tipo 'E', com consumos, aos fins-de-semana, embora também vá havendo alguns consumos ao longo dos restantes dias da semana.

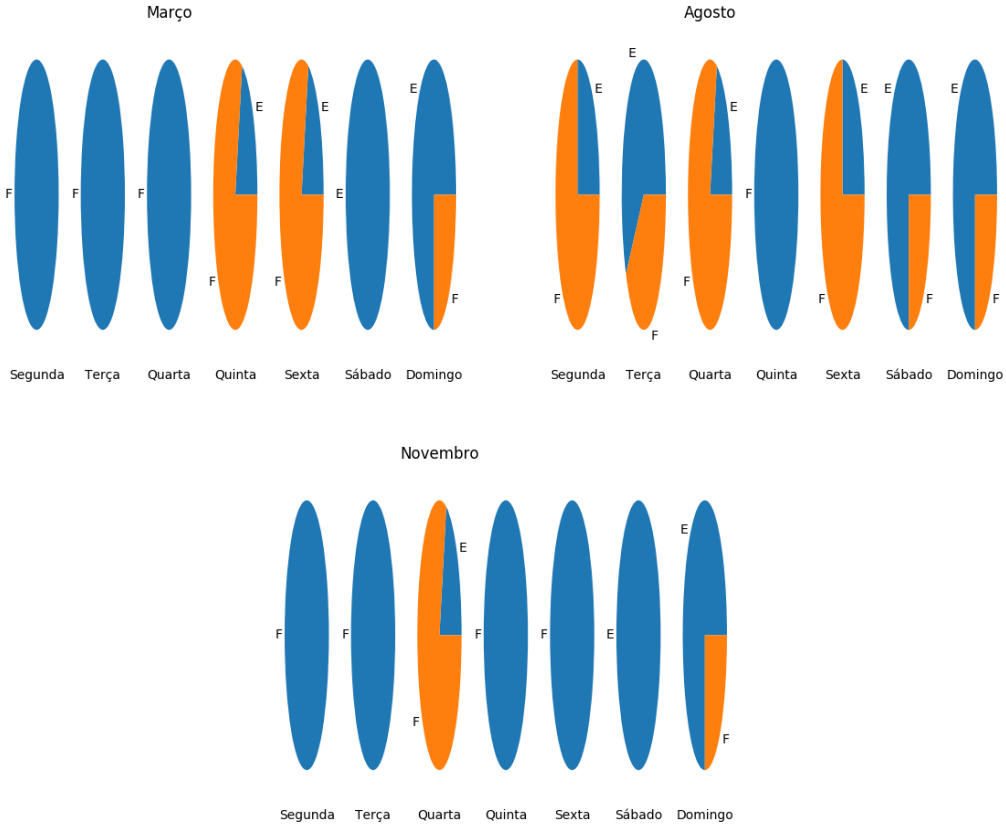


Figura 4.52: Representação da distribuição dos dias tipo, pelos dias da semana, no modo base, para a máquina de lavar roupa.

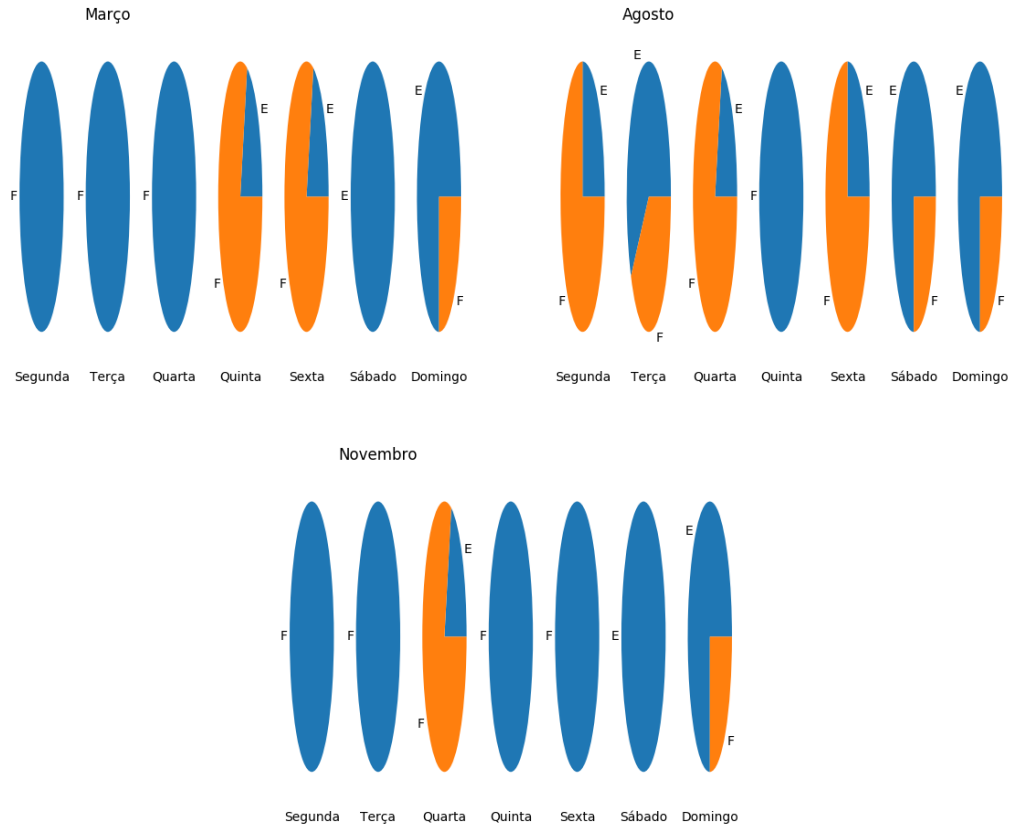


Figura 4.53: Representação da distribuição dos dias tipo, pelos dias da semana, no modo horário, para a máquina de lavar roupa.

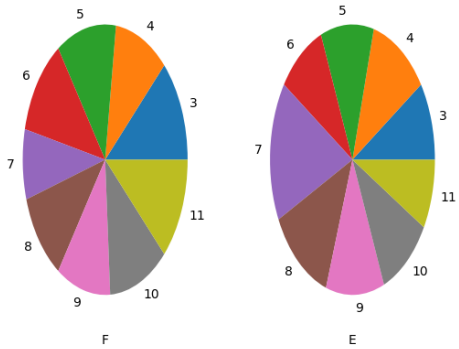


Figura 4.54: Representação da distribuição dos dias tipo relevantes, ao longo dos meses da janela temporal, no modo horário.

É revelado, na figura 4.54, que existe um maior número de dias tipo 'E' (dias com atividade do equipamento) no mês de julho, relativamente aos restantes meses. De resto, todos os meses parecem ter uma distribuição de dias tipo, aproximadamente, uniforme. A distribuição dos dias tipo ao longo dos meses, no modo base, não é apresentada, a distribuição é equivalente, para ambos os modos de aquisição, como visto nas figuras 4.52 e 4.53.

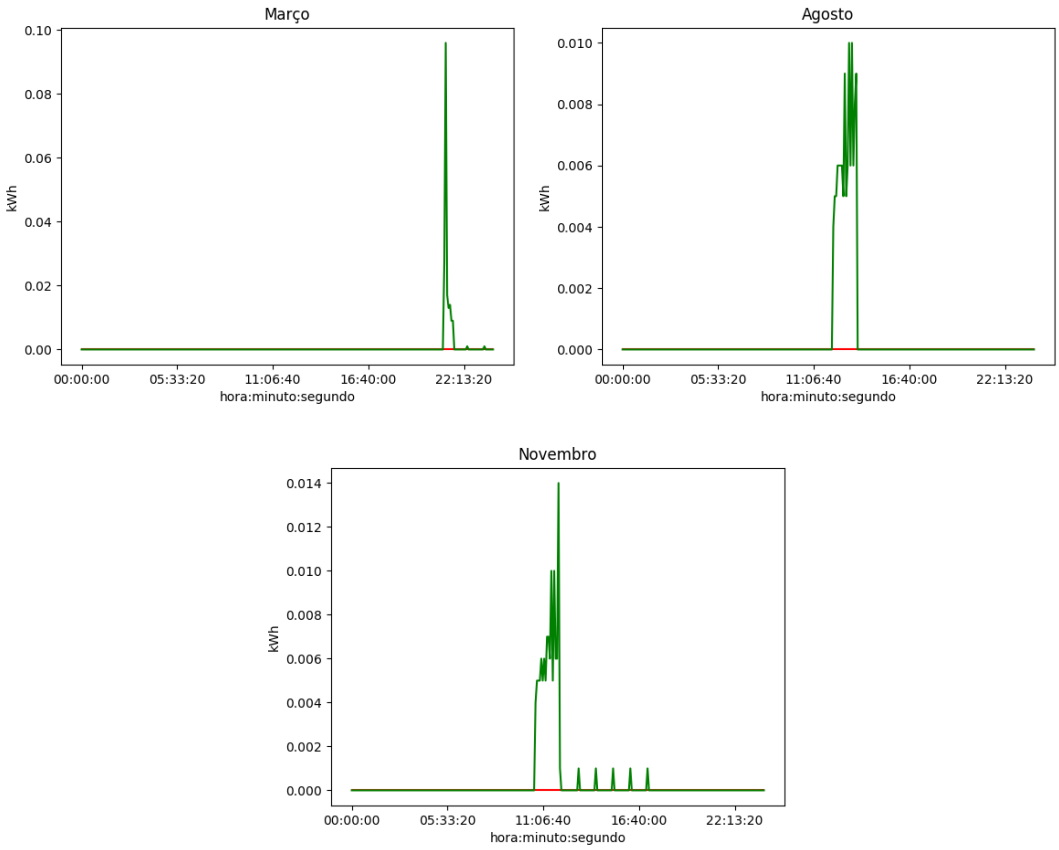


Figura 4.55: Representação, em modo base, dos dias tipo 'E' a verde e 'F' a vermelho, para a máquina de lavar roupa.

Por análise da figura 4.55, representativa dos perfis em modo base, conclui-se que um consumo da máquina de lavar roupa é tipicamente de um pico de aproximadamente uma ou duas horas de duração, sendo por vezes sucedido por vários picos de curta duração, de uma forma periódica, num período de tempo considerável.

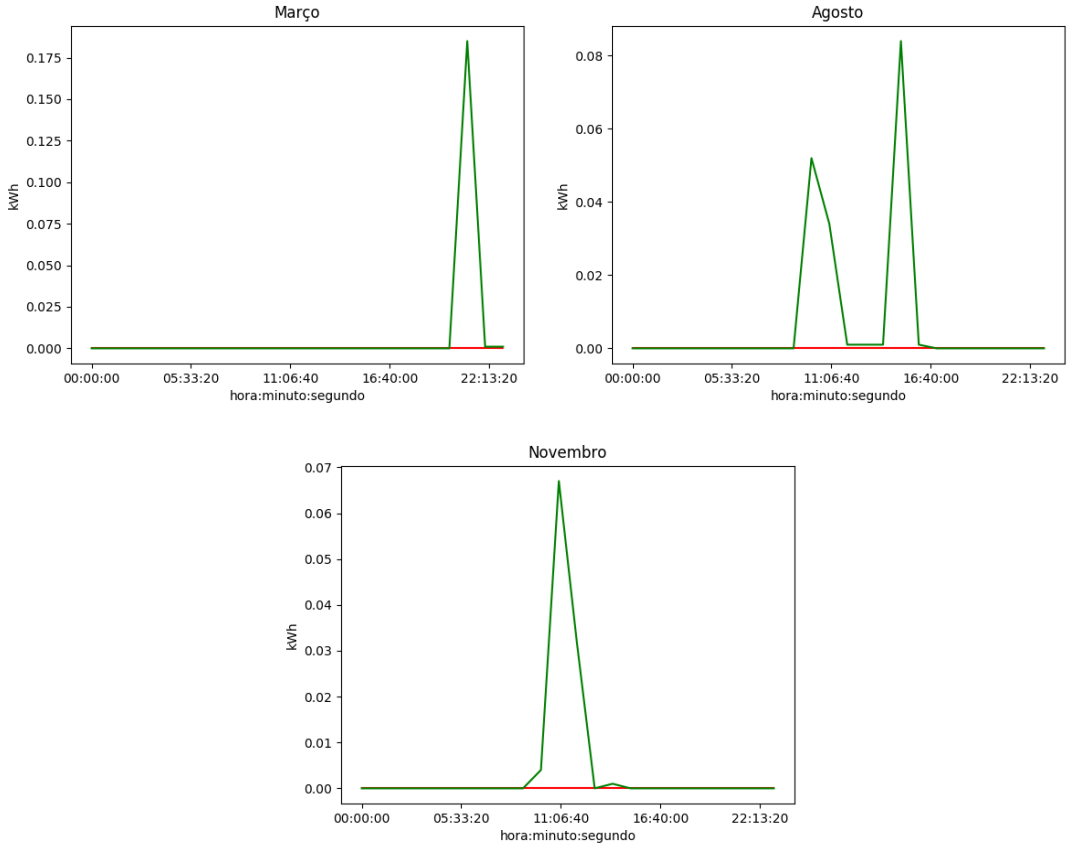


Figura 4.56: Representação, em modo horário, dos dias tipo 'E' a verde e 'F' a vermelho, para a máquina de lavar roupa.

O perfil diário dos dias com consumos, 'E', representado na figura 4.56, em modo horário, apresenta melhores resultados, mais claros, daquilo que é um consumo típico do aparelho, comparativamente à figura 4.55. Tal facto deve-se a no modo base se verificar constantes arranques sucessivos que, provavelmente, correspondem ao mesmo funcionamento da máquina de lavar roupa, dando a ideia que a máquina executou várias lavagens quando, na verdade, os sucessivos arranques correspondem todos à mesma.

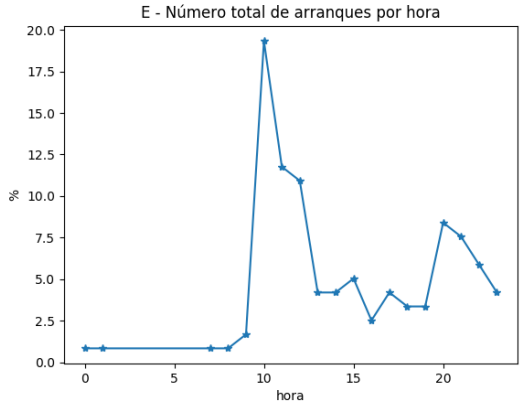


Figura 4.57: Distribuição do número de arranques, para cada hora, dos dias tipo 'E', no modo horário.

A figura 4.57 mostra a distribuição da frequência dos arranques, no modo horário, ao longo de um dia categorizado como 'E', um dia com consumos. A representação não é apresentada para o modo base, como já visto, o modo de aquisição que representa os arranques reais da máquina de lavar roupa é o modo horário. Conclui-se, por análise de 4.57, que os arranques são em maioria entre as 10-12 horas e entre as 20-21 horas sendo que este último horário tem menor probabilidade de ocorrência que o primeiro.

O pico de arranques identificado, de maior probabilidade, para as 10-12 horas pode dever-se, provavelmente, a um aproveitamento das horas de sol para secar a roupa ao ar livre. O pico também relevante, mas de menor probabilidade, é entre as 20-21 horas, que corresponde às horas em que se inicia a atividade numa habitação, após as horas laborais. Neste período há uma maior disponibilidade para as tarefas domésticas que no período da manhã, num dia laboral. Estas conclusões são tiradas por suposição, não sendo necessariamente verdade, de que os habitantes da casa piloto estão fora da habitação das 9 às 19 horas, num dia laboral.

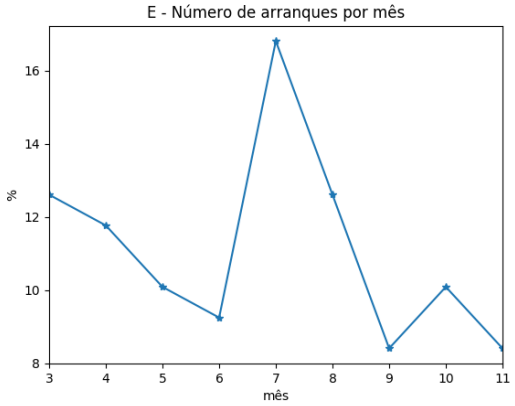


Figura 4.58: Distribuição do número de arranques, para cada mês, para os dias tipo 'E', no modo horário.

Igualmente ao realizado para a figura 4.57, e pelas mesmas razões enunciadas, apenas são representados, na figura 4.58, o número de arranques ao longo dos meses, no modo horário. É de destaque um pico de arranques nos meses de verão, com números de arranques máximo no mês de julho, que vai de encontro à distribuição apresentada na figura 4.54. Existe uma descida gradual, entre março e junho, do número de arranques. O mês de setembro e o mês de novembro apresentam uma frequência de arranques semelhante, havendo um ligeiro pico, por comparação com estes meses, no mês de outubro. O mês de outubro foi anormalmente quente no ano de 2017, podendo ter uma causa conjunta com os meses de verão para os picos de arranques.

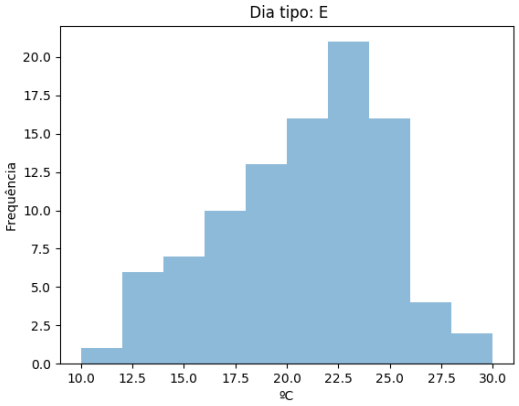


Figura 4.59: Distribuição da temperatura máxima, para os dias tipo 'E', no modo horário.

Por análise da figura 4.59, e sua comparação com 4.8, conclui-se que existe uma predominância do número de arranques em dias mais quentes, resultado do maior número de arranques em meses de verão, como já verificado na figura 4.58.

	Março		
	Duração (horas)	# [C2,C3,C4]	Centro
1	4	0,0,1	A^4
7	2	0,0,1	A^3
9	1	0,0,1	A
10	5	1,0,1	A^3
11	2	0,0,1	A^2
12	4	0,0,1	CA^3
13	6	1,0,1	BA^2
14	7	0,1,0	A^3BA^3
19	4	0,0,1	ADA^2
20	3	0,0,1	A^3
21	11	1,0,0	A^{11}

Tabela 4.37: Caracterização do dia tipo 'E', em modo horário para o mês de março.

	Agosto		
	Duração (horas)	# [C2,C3,C4]	Centro
8	2	0,0,1	A^2
10	8	1,2,1	A^5
11	2	0,0,1	A^2
12	2	0,0,1	A^2
15	6	0,0,1	A^6
16	2	0,0,1	ABA
17	1	0,0,1	A
20	2	0,0,1	A^2
22	1	0,0,2	A
23	2	0,0,1	A^2

Tabela 4.38: Caracterização do dia tipo 'E', em modo horário para o mês de agosto.

	Novembro		
	Duração (horas)	# [C3,C4]	Centro
9	2	0,1	A^2
10	7	3,1	A^8
11	7	1,0	A^7
12	2	0,1	A^2
14	1	0,1	A
18	2	0,1	A^2
19	2	0,1	A^2

Tabela 4.39: Caracterização do dia tipo 'E', em modo horário para o mês de novembro.

As tabelas 4.37:39 caracterizam um dia com atividade para os meses de março, agosto e novembro. De março para agosto verifica-se uma subida quase impercetível do número de arranques. De agosto para novembro nota-se uma queda deste número. Os arranques dão-se tipicamente no período da manhã, ao início da tarde ou ao início da noite. Como já referido, estas horas típicas de funcionamento podem ter origem num aproveitamento das horas de sol para secar a roupa, ou, de um dia típico laboral, em que há mais tempo para tratar de tarefas domésticas ao fim do dia. Os consumos ativos têm uma duração entre uma a duas horas, tipicamente.

Consumos de duração longa podem ter origem em lavagens consecutivas de roupa. Mas, quando este período de funcionamento é exagerado, pode ter origem num consumo

residual, após um pico de consumo, como acontece no mês de novembro, visível em 4.55.

Consumos Agregados

Nesta secção é feita a análise dos consumos agregados da casa piloto, isto é, os consumos totais, o somatório de todos os consumos elétricos, gerados por todos os equipamentos presentes na casa piloto que num dado momento estão a consumir energia.

Entenda-se como atividade do edifício os momentos em que se deteta o estado ON de equipamentos que dependem de fatores externos, como a hora do dia ou a temperatura registada, não funcionando se forma constante no tempo (como é o caso do frigorífico para a casa piloto), ou seja, não contribuindo para o consumo de fundo do edifício, destacando-se deste.

Tome-se então a distinção de consumos ON e OFF (3.11), para as séries de consumos agregados, como a classificação dos eventos (3.1) que correspondem à existência de atividade no edifício. Neste sentido, os arranques da habitação são referentes aos momentos em que se estima que se iniciou atividade.

Esta análise tem interesse para que se possa estudar os momentos em que, tipicamente, existe um maior consumo na habitação e os fatores externos de que dependem.

Para fazer a divisão do estado inativo para o estado ativo da casa, é utilizado o critério de 1/15 do valor máximo para o modo base e 1/10 do valor máximo para o modo horário, ambos desprezando *outliers*.

	Mínimo	**Máximo**	**Média**
ON	0.046	0.706	0.250
OFF	0.000	0.044	0.020

a) Modo base, valores em kWh;

	Mínimo	**Máximo**	**Média**
ON	0.320	4.293	1.253
OFF	0.000	0.317	0.226

b) Modo horário, valores em kWh;

Tabela 4.40: Divisão dos valores de consumo bruto (3.1) em patamares de consumo ON e OFF (3.11), para os consumos agregados.

Observando a tabela 4.40, verifica-se que a gama de valores de estado ON, no modo de aquisição horário, é bastante superior, uma ordem de grandeza, que no modo base. Tal pode significar que quando um consumo horário é identificado como ativo, ON, que

nesse período, no modo base, na maior parte do tempo, o aparelho foi identificado como ativo.

	Mínimo	**Máximo**	**Média**
A	0.046	0.143	0.093
B	0.145	0.244	0.193
C	0.246	0.474	0.347
D	0.490	0.706	0.559

a) Modo base, valores em kWh;

	Mínimo	**Máximo**	**Média**
A	0.320	3.468	1.244
B	4.128	4.293	4.220

b) Modo horário, valores em kWh;

Tabela 4.41: Divisão dos valores de ON (3.11), em patamares de consumo, para os consumos agregados.

A tabela 4.41 revela patamares de A, em ambos os modos de aquisição, com uma grande amplitude da gama de valores, portanto, provavelmente, os restantes patamares têm baixa ocorrência.

	Centro de cada cluster*	
	Dados base	**Dados horários**
C1	$*$ (0.010%)	A^{173} (0.114%)
C2	$AABAA$ (95.171%)	A^{89} (0.229%)
C3	$*$ (4.308%)	A^6BA^4 (85.943%)
C4	$*$ (0.511%)	A^{21} (12.343%)
C4	-	A^{48} (1.371%)

Tabela 4.42: Centro de cada *cluster* e respetiva percentagem de ocorrência, para os consumos agregados (# 9795-modo base; # 875-modo horário).

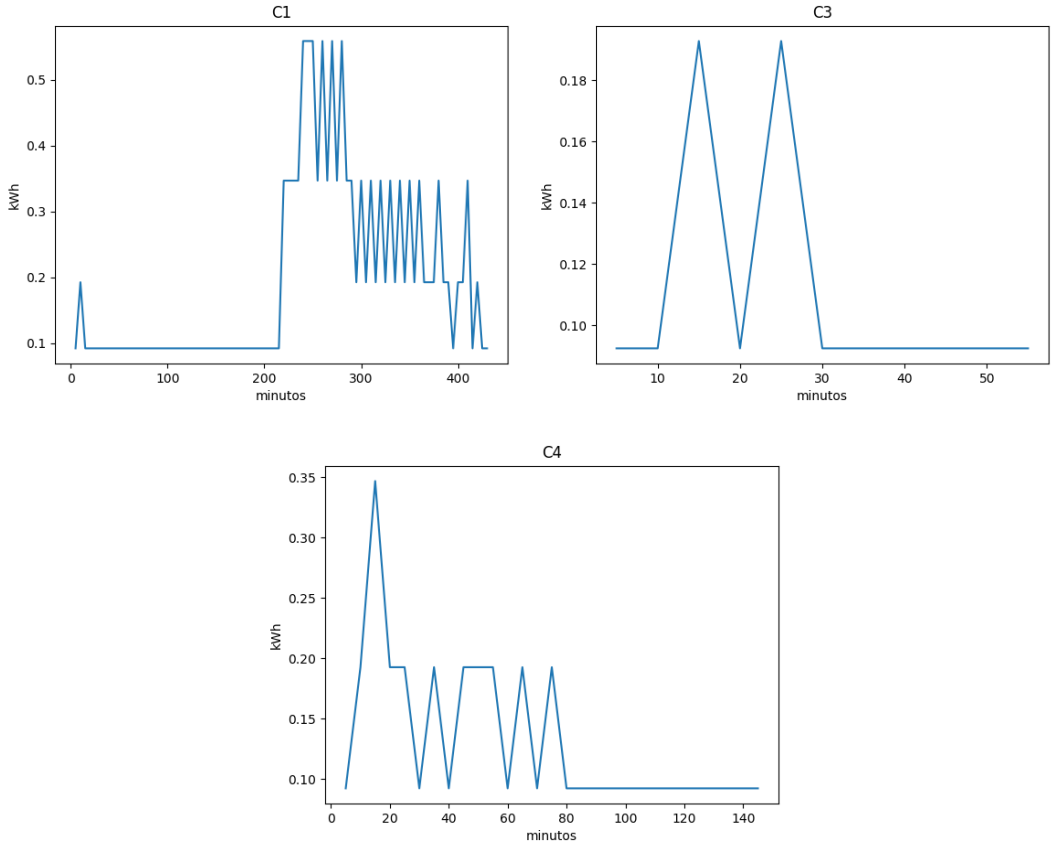


Figura 4.60: Representação da *cluster* C1, C3 e C4 no modo base, para os consumos agregados.

As *clusters* mais frequentes são C2, no modo de aquisição base, com uma duração do seu centro de 25 minutos e C3, no modo horário, com uma duração de 11 horas. Isto significa que os consumos agregados apresentam consumos durante 11 horas, embora provavelmente, não de forma ininterrupta. Estas conclusões foram obtidas por análise da tabela 4.42. As *clusters* identificadas com um $*$, por terem centros demasiado longos e heterogéneos, encontram-se representadas na figura 4.60.

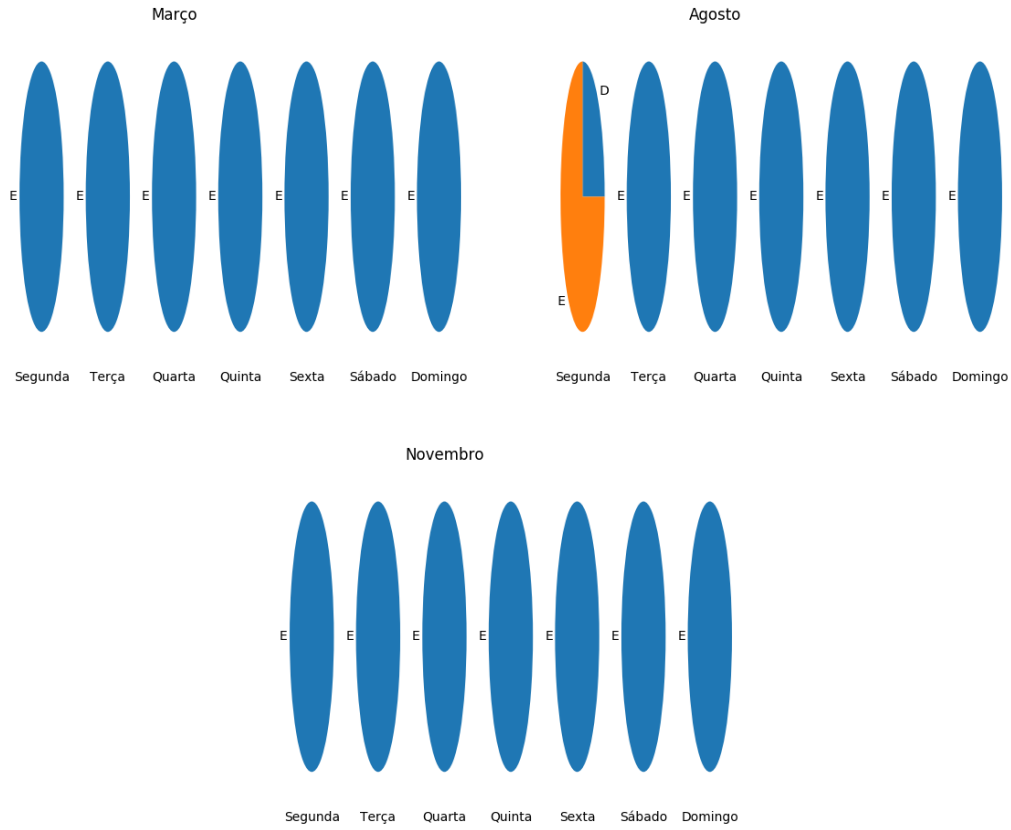


Figura 4.61: Representação da distribuição dos dias tipo, pelos dias da semana, no modo base, para os consumos agregados.

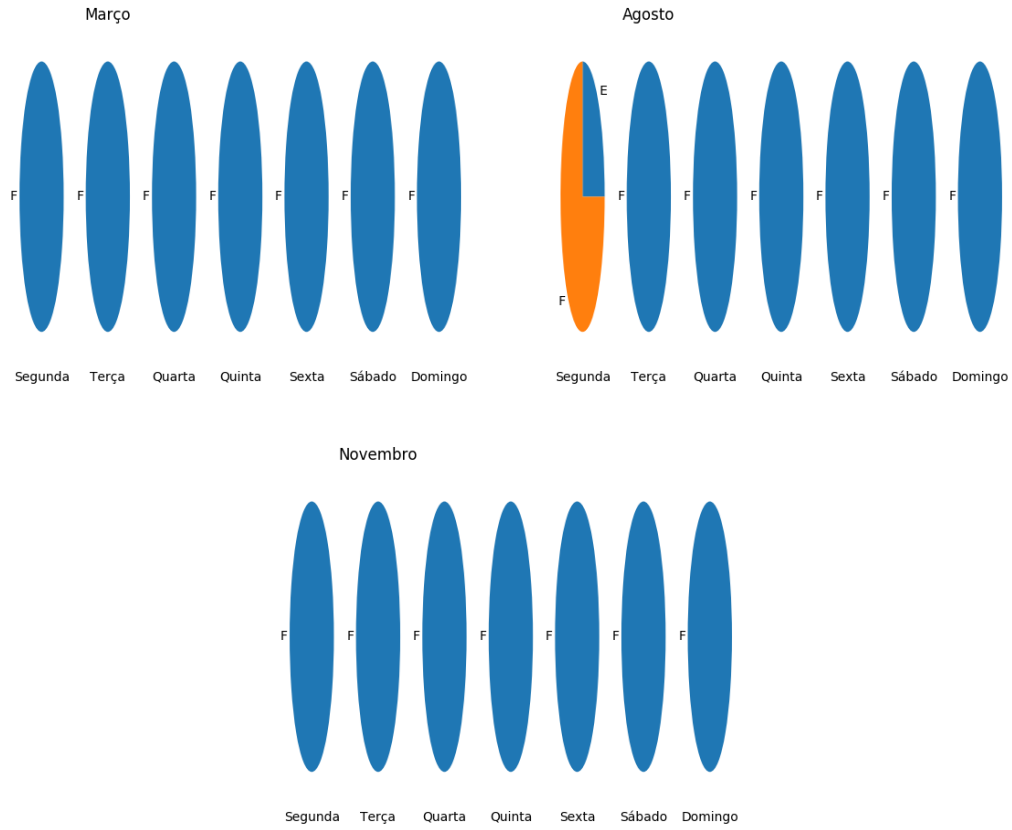


Figura 4.62: Representação da distribuição dos dias tipo, pelos dias da semana, no modo horário, para os consumos agregados.

Por análise das figuras 4.61 e 4.62, verifica-se que os dias são classificados de forma massiva com o dia tipo 'E', para o modo base, e, de forma equivalente, com o dia tipo 'F' para o modo horário.

Visto a classificação em modo horário e em modo base serem equivalentes, interessa apenas analisar uma delas. Na figura 4.63 encontra-se representada a distribuição do dia tipo 'E', no modo base. Verifica-se, como expectável, por os dias serem classificados em maioria com este dia tipo, que a distribuição é uniforme ao longo dos meses.

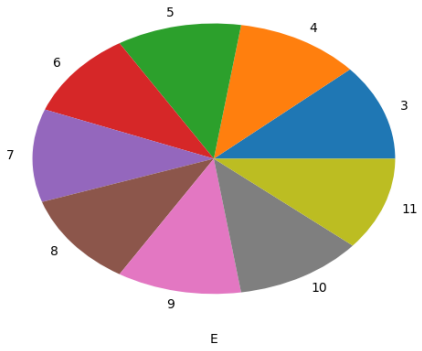


Figura 4.63: Representação da distribuição dos dias tipo relevantes, ao longo dos meses da janela temporal, no modo base, para os consumos agregados.

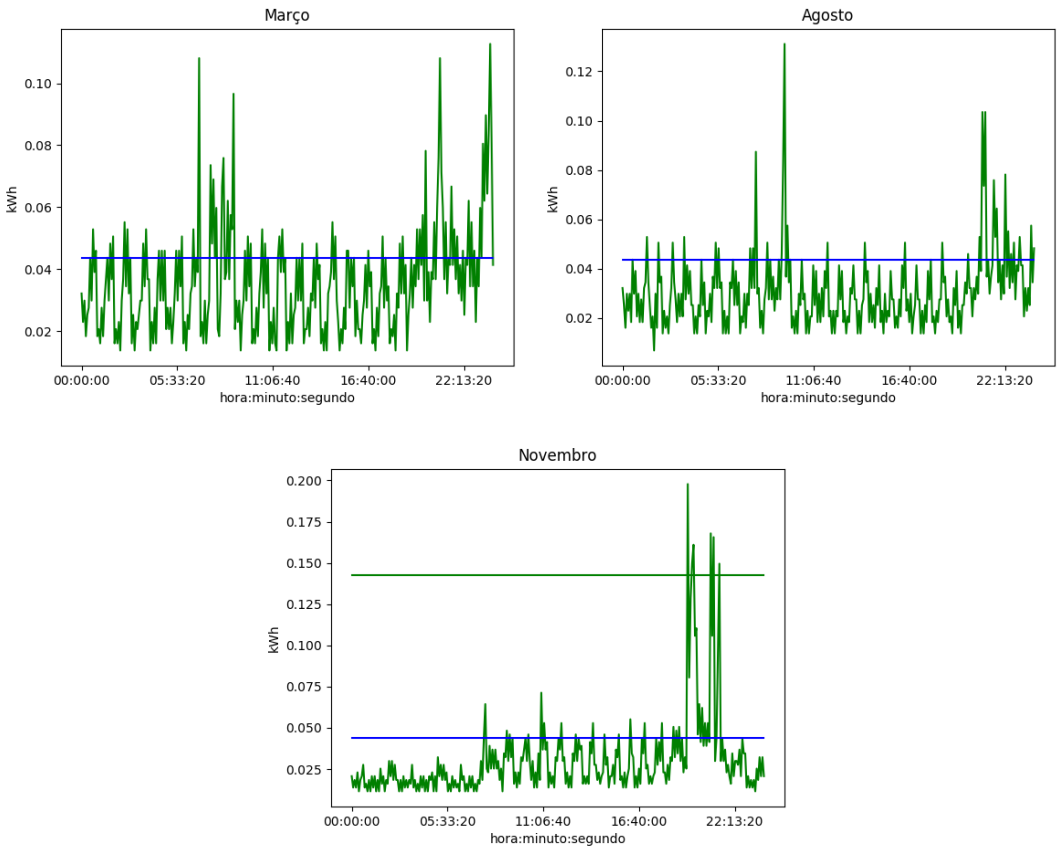


Figura 4.64: Representação, em modo base, dos dias tipo 'E' a verde, para os consumos agregados.

Para as figuras 4.64 e 4.65, a linha horizontal azul representa a divisão entre os consumos de fundo e os consumos representativos de atividade, no edifício habitacional. A linha horizontal verde representa o limite superior do patamar A.

Por observação da figura 4.64, verificam-se dois picos de consumo, um ao início da manhã e outro no período da noite. No mês de novembro, os picos que ocorrem durante a noite são de maior amplitude que os picos equivalentes em março e agosto.

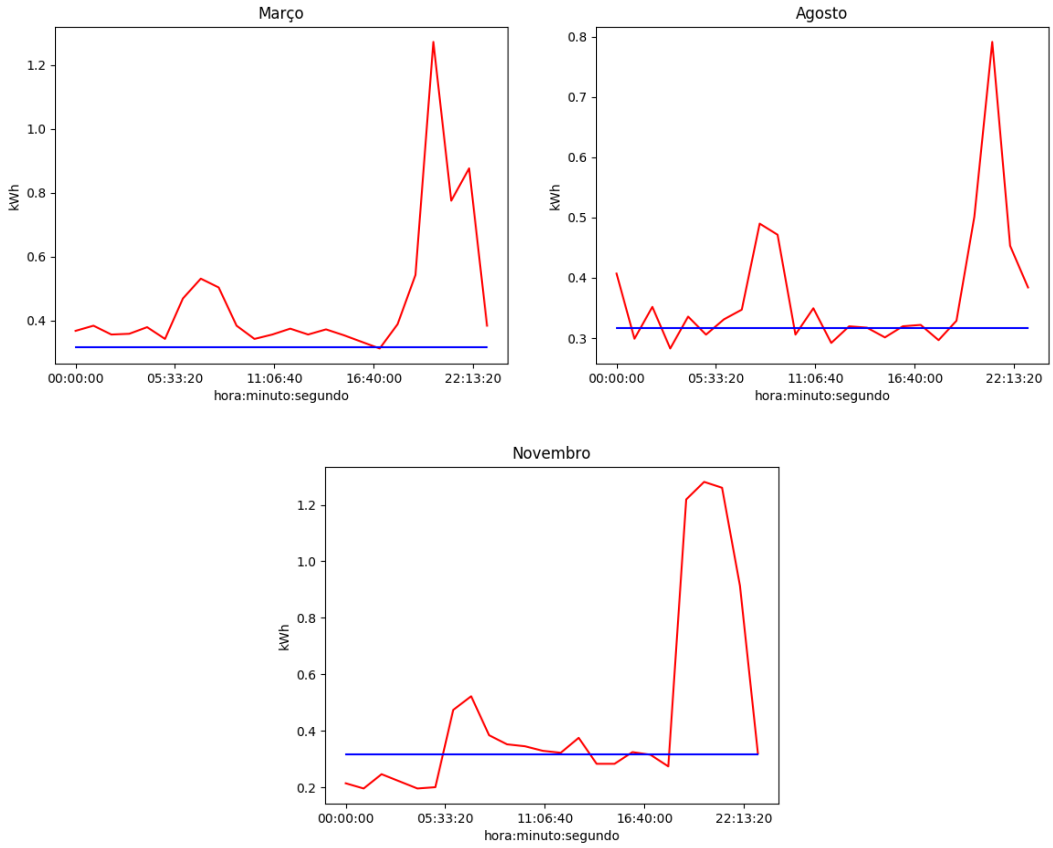


Figura 4.65: Representação, em modo horário, dos dias tipo 'F' a vermelho, para os consumos agregados.

No modo horário, representado na figura 4.65, conclui-se que existe uma tendência de interceção da linha azul, ou uma tendência de se manter abaixo desta, por consumos de fundo para os meses quentes, e uma maior tendência para se manter acima desta

linha, nos meses mais frios. Para este modo de aquisição torna-se claro os picos de consumo durante a manhã e durante a noite. Ocorrem aproximadamente entre as 7 e as 8 horas e as 18 e as 22 horas.

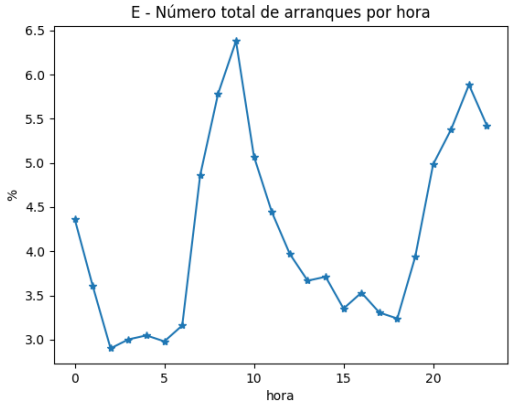


Figura 4.66: Distribuição do número de arranques, para cada hora, para os dias tipo 'E', no modo base.

Na figura 4.66, é perfeitamente observável os picos de consumo entre as 8 e as 10 horas e entre as 20 horas e a meia-noite. Não é analisado o número de arranques no modo horário, porque neste modo de aquisição o consumo resultante de atividade no edifico confunde-se muitas vezes com o consumo de fundo da habitação, ficando a linha de consumos por longos períodos de tempo acima da linha horizontal azul (figura 4.65). No modo base, a intersecção da linha horizontal azul (figura 4.64), pelos consumos de fundo, acaba por se difundir nos arranques reais, de consumo por atividade no edifício, e estes últimos, por serem em muito maior número, se evidenciarem.

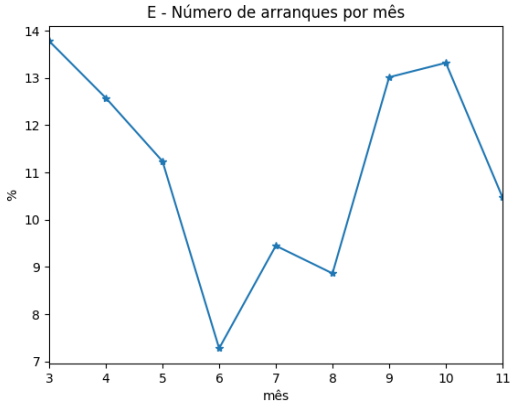


Figura 4.67: Distribuição do número de arranques, para cada mês, para os dias tipo 'E', no modo base.

A figura 4.67 mostra a distribuição das percentagens de arranques ao longo dos meses em análise. A percentagem de arranques é aproximadamente uniforme, a menos dos meses de verão e do mês de novembro, em que existe um decréscimo acentuado de arranques. Nos meses de verão é expectável que uma família passe mais tempo fora de casa, resultando em menos arranques de consumos na habitação.

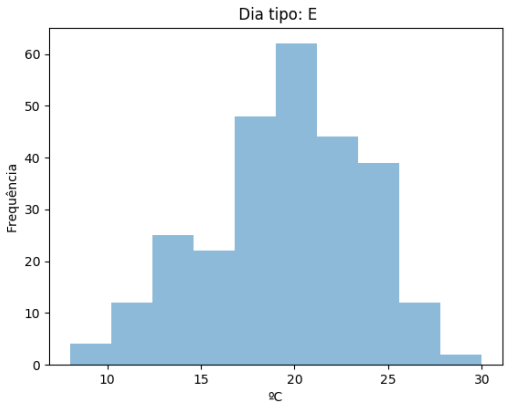


Figura 4.68: Distribuição da temperatura máxima, para os dias tipo 'E', no modo base.

Por comparação da figura 4.67 com 4.8, conclui-se que não existe sazonalidade para o dia tipo 'E', no modo base. Tal facto é expectável visto praticamente todos os dias presentes na base de dados serem classificados com o mesmo dia tipo.

	Março		
	Duração (min)	# [C2,C3,C4]	Centro
0	7	33,0,0	*A*
1	6	34,0,0	*A*
2	5	34,0,0	*A*
3	5	34,0,0	*A*
4	5	37,0,0	*A*
5	5	32,0,0	*A*
6	7	48,0,0	*A*
7	7	72,1,0	*A*
8	8	89,1,0	*A*
9	10	68,3,0	A
10	8	58,2,0	*A*
11	14	61,2,1	*A*
12	14	41,0,2	*A*
13	9	45,2,0	*A*
14	7	41,1,0	*A*
15	12	39,3,0	*A*
16	9	48,1,3	*A*
17	11	34,3,0	*A*
18	6	49,0,0	*A*
19	20	62,9,2	*A*
20	25	42,17,0	*A*
21	17	37,6,0	*A*
22	15	60,5,1	*A*
23	14	55,5,1	*A*

Tabela 4.43: Caracterização do dia tipo 'E', em modo base para o mês de março.

	Agosto		
	Duração (min)	# [C2,C3,C4]	Centro
0	8	24,0,0	*A*
1	8	26,1,0	*A*
2	5	19,0,0	*A*
3	5	20,0,0	*A*
4	5	21,0,0	*A*
5	5	21,0,0	*A*
7	5	37,0,0	*A*
8	7	28,1,0	*A*
9	8	54,0,0	*A*
10	9	46,0,1	*A*
11	7	36,1,0	*A*
12	16	32,4,0	*A*
13	18	27,3,0	*A*
14	9	38,1,0	*A*
15	14	31,2,0	*A*
16	11	36,1,0	*A*
17	14	32,2,0	*A*
18	13	27,1,1	*A*
19	5	29,0,0	A
20	13	40,3,0	*A*
21	27	37,5,3	*A*
22	10	41,2,0	*A*
23	11	38,2,0	*A*

Tabela 4.44: Caracterização do dia tipo 'E', em modo base para o mês de agosto.

	Novembro		
	Duração (min)	# [C2,C3,C4]	Centro
0	5	21,0,0	A
1	5	16,0,0	A
2	5	25,0,0	A
3	5	17,0,0	A
4	5	27,0,0	A
5	5	20,0,0	A
6	6	40,0,0	A
7	6	63,0,0	A
8	7	65,0,0	A
9	7	61,1,0	A
10	11	55,5,0	A
11	7	48,1,0	A
12	13	39,4,1	A
13	7	46,1,0	A
14	21	38,3,2	A
15	6	29,0,0	A
16	12	26,0,1	A
17	9	40,2,0	A
18	21	30,0,1	A
19	33	38,14,3	A
20	27	26,9,0	A
21	15	37,5,0	A
22	13	30,4,0	A
23	7	30,1,0	A

Tabela 4.45: Caracterização do dia tipo 'E', em modo base para o mês de novembro.

As tabelas 4.43:45 caracterizam um dia com consumos, dia tipo 'E', no modo base. Para os três meses em análise, verifica-se uma tendência do aumento da duração dos consumos ativos ao longo do avançar das horas, ao longo de um dia. Comparativamente ao mês de março, em agosto e novembro ocorre um menor número de arranques, no geral. Como já visto, os picos de arranques dão-se no início da manhã e no fim do dia. Os centros destes picos são, para o mês de Março e Novembro, às 8 e 19 horas e para o mês de agosto às 9 e 20 horas. Por no mês de agosto anoitecer mais tarde, comparativamente a março e novembro, poderá haver uma tendência de deitar e acordar ligeiramente mais tarde.

Discussão

Com toda a informação reunida, para as monitorizações realizadas na casa piloto, é possível tirar algumas conclusões gerais.

O frigorífico não apresenta nenhuma relação evidente com uma possível atividade no edifício, nem com as estações do ano. Revela-se um consumo periódico, com uma tendência a aumentar o período de atividade com o curso do tempo.

Para a máquina de lavar roupa, é notória uma tendência de uso nos meses de verão, mas com um destaque de uso no mês de outubro, relativamente aos meses de setembro e novembro, que foi quente para o ano em análise. Existe uma preferência de uso deste equipamento no período da manhã, início da tarde e início da noite. O seu uso é predominante ao fim-de-semana. Embora pontualmente também se verifiquem consumos durante a semana, dias úteis. É suspeito que os consumos ao final do dia sejam realizados em dias úteis, e as restantes horas típicas de uso em fim-de-semana. No entanto, visto a classificação dos dias tipo ser atribuída com o critério de existir ou não consumos para um dia, com a informação obtida para este equipamento, não se obtém conhecimento que sustente ou contrarie esta hipótese . Um consumo típico deste aparelho dura entre uma a duas horas.

Por último é analisada a monitorização efetuada aos consumos agregados. Nesta monitorização, que apresenta o somatório de todos os consumos efetuados na habitação, durante o intervalo de tempo de aquisição, é notório um pico de consumo, de curta duração, ao início da manhã e outro ao inicio da noite. Verifica-se também uma tendência a um menor número de arranques nos meses de verão. Nestes meses, tipicamente, as pessoas tendem a passar menos tempo em casa, tendem a passear mais e tirar férias.

Portanto, devidas às características das três monitorizações, os consumos monitorizados não apresentam grande correlação entre si. Isto é, apesar dos consumos agregados conterem todos os consumos do edifício, o consumo do frigorífico é periódico e o da máquina de lavar roupa muito aleatório no tempo, não permitindo grandes associações entre o conhecimento obtido pelas suas análises de consumos.

4.3 Associações entre Valores de Variáveis Distintas

Nesta secção são apresentadas relações existentes entre valores de variáveis distintas, para cada uma das bases de dados, empresa e casa piloto, por aplicação dos conceitos descritos na secção 3.6.

As associações entre valores de variáveis das bases de dados em causa, são determinadas apenas para os instantes de estados ON (3.11), dos consumos de cada monitorização, efetuada em cada edifício. Poderia-se realizar um cálculo contando com a hora de funcionamento, o mês e a temperatura máxima para o registo em que foi acusada associação, mas, iria causar muito ruído na análise. É preferencial calcular as associações entre estados ativos nas monitorizações, e posto isto, verificar se existe alguma relação com os restantes valores das variáveis referidas.

É necessário estabelecer alguns valores *threshold* para que as associações sejam calculadas. Estabeleceu-se $lift = 1$ (3.20) e $suporte = 0.001$ (3.18).

Não se justifica incluir, nas relações de associação, aparelhos de uso contínuo ou periódico. Visto este tipo de aparelho não ter nenhuma relação causa-efeito com fatores externos. Na base de dados da empresa piloto fica, então, excluído da análise o refrigerador de água. Da casa piloto exclui-se o frigorífico. Desta última base de dados sobra apenas as monitorizações efetuadas à máquina de lavar roupa e aos consumos agregados. Ora, não faz sentido verificar uma relação de associação entre os estados ativos das duas monitorizações, visto que, os consumos agregados incluem a monitorização da máquina de lavar roupa. Não resta assim nenhum par de antecessor e sucessores possível de formar, para a casa piloto. A casa piloto fica portanto excluída desta secção, apenas são analisadas associações para a empresa piloto. A avaliação de associações entre uma monitorização individual e a hora de uso, mês e temperatura máxima não tem interesse, pois, essa análise já foi efetuada na secção 4.2.1.

4.3.1 Empresa Piloto

Para identificar associações, de estados ativos (3.11), entre os equipamentos, aplica-se então o algoritmo exposto na secção 3.6 às monitorizações dos estados ON (3.11) para a impressora, ilha de computadores e máquina de café, no modo de aquisição base, com o *suporte* e o *lift* mínimos enunciados.

Antecessores	**Sucessores**	***Suporte***	***Confiança***	***lift***
$\{1,3\}$	$\{2\}$	0.003	1	1.011
$\{3,2\}$	$\{1\}$	0.011	0.305	1.095
$\{1,2\}$	$\{3\}$	0.269	0.013	1.035
$\{3\}$	$\{1,2\}$	0.012	0.278	1.035
$\{1\}$	$\{2,3\}$	0.278	0.012	1.095
$\{2\}$	$\{1,3\}$	0.990	0.003	1.010

Tabela 4.46: Resultado da aplicação do algoritmo de associação às monitorizações: 1-Impressora, 2-Ilha de computadores, 3-Máquina de café;

A tabela 4.46 revela as seguintes associações, com uma *confiança* (3.19) considerável:

- $\{1,3\} \rightarrow \{2\}$: O estado ativo da impressora e da máquina de café como antecessor, e a ilha de computadores como sucessor;
- $\{3,2\} \rightarrow \{1\}$: O estado ativo da ilha de computadores e da máquina de café como antecessor, e a impressora como sucessor;
- $\{3\} \rightarrow \{1,2\}$: O estado ativo da máquina de café como antecessor, e a impressora e ilha de computadores como sucessores;

De notar que as associações com uma *confiança* relevante, têm sempre incluídas o estado ativo da ilha de computadores. Não é de estranhar, de acordo com as conclusões tiradas na secção 4.2.1, a impressora e a máquina de café funcionarem durante curtos instantes, mas com alguma frequência, durante o período laboral, e a ilha de computadores tem um estado ativo durante todo o período laboral. Quanto ao uso da impressora e da máquina de café, é natural que ocorra alguma intersecção dos seus estados ativos por terem um intervalo de funcionamento, expectável, semelhante, sem

que esse uso esteja diretamente relacionado. Os equipamentos podem ser usados por diferentes empregados, em simultâneo. Mas, pode acontecer o fenómeno de um trabalhador ir tirar um café enquanto espera que os documentos acabem de imprimir.

Não é interessante a análise destas associações com a temperatura máxima, os meses e a hora em que ocorrem. É óbvia a intersecção de horas típicas de funcionamento, para um dia laboral, neste edifício e para estes equipamentos na secção 4.2.1. Análises adicionais a estas variáveis seriam redundantes.

4.4 Consumos Anormais

Os consumos anormais são detetados com recurso à metodologia descrita na secção 3.7. São detetados dias atípicos para o dia da semana em questão, consumos que ocorram em horas atípicas, com uma duração fora do padrão ou com um padrão atípico. Os dias de referência para se detetar os consumos anormais são os dias da semana equivalentes, entre os 30 dias que antecedem cada dia em análise. Os consumos com início a cada hora são comparados com os consumos com início nas mesmas horas, nos dias em comparação.

Para esta análise são usados, novamente, os meses representativos de cada estação do ano, presentes na base de dados disponível, março, agosto e novembro.

Não são analisados consumos anormais para equipamentos que apresentam um funcionamento periódico. A análise não é efetuada a equipamentos periódicos, pois, o algoritmo desenvolvido está preparado para analisar equipamentos que dependem da atividade do edifício, com horas típicas de funcionamento, sendo apenas estudado o estado ativo do equipamento. Para um equipamento periódico, a comparação teria de ser efetuada por análise dos períodos dos últimos 30 dias, construindo um modelo periódico torna-se possível a comparação entre o consumo registado e o previsto pelo modelo. Caso se desvie, em demasia, do modelo é reportada anomalia.

Também não são analisados consumos agregados devido à heterogeneidade deste tipo de monitorização. Um consumo anormal pode ser confundido com a introdução

de um novo equipamento, ou, um equipamento de uso muito pontual. Ou até, apenas com o uso simultâneo de vários aparelhos. Esta análise apenas se revela útil na deteção de estados de atividade, ou pelo contrário, inatividade do edifício em horas não desejadas. Para o efeito, o algoritmo de deteção de anomalias teria de ser aprimorado nesse sentido. Como se encontra, não se revela conclusivo para avaliar estados de atividade.

4.4.1 Empresa Piloto

Para a análise de consumos anormais, para a base de dados da empresa piloto, tem-se disponível as monitorizações efetuadas à impressora, ilha de computadores e máquina de café. Esta análise não é feita para o refrigerador de água por apresentar um funcionamento periódico.

A análise é feita, no modo base, para a impressora e para a máquina de café, e no modo horário para a ilha de computadores. A escolha é feita por identificação do intervalo de tempo de aquisição mais conveniente para o período de funcionamento típico de cada equipamento.

As tabelas 4.47:49 apresentam os resultados dos algoritmos de deteção de consumos anormais aplicados às monitorizações referidas. Os índices destas tabelas representam os dias em questão, para cada mês em análise. Caso não seja indicado qual o dia da semana a que pertencem, assume-se que se trata de um dia entre segunda a sexta-feira.

Para se entender os consumos anormais apresentados, a tabela 4.47 deve ser comparada com 4.8, 4.14, 4.20 a) e 4.21 a), 4.48 com 4.9, 4.15 e 4.21 b) e 4.49 com 4.10, 4.16, 4.20 b) e 4.21 c).

	Março		
	Hora	**Monitorização**	**Falha**
2	14	3	Padrão: AA
3	8	3	Padrão: AA
6	9	3	Padrão: AA
7	9	3	Padrão: AA
	0	2	Duração: 8h
8	13	3	Duração: 10 min
9	10	1	Duração: 55 min
12-Domingo	-	3	Dia: E
20	9	3	Padrão: AA
	17	1	Duração: 90 min
21	14	1	Duração: 110 min
22	9	3	Padrão: AA
24	9	3	Duração: 15 min
	10	2	Duração: 14h
27	12	1	Duração: 85 min
	10	3	Duração: 25 min
28	0	2	Duração: 24h
	10	3	Padrão: AA
29	10	3	Padrão: AA
	0	2	Duração: 24h
30	0	2	Duração: 23h
	9	3	Padrão: AA
31	0	2	Duração: 21h
	19	1	Duração: 35 min

Tabela 4.47: Consumos anormais detetados em março para: 1-Impressora; 2-Ilha de computadores; 3-Máquina de café;

	Agosto		
	Hora	**Monitorização**	**Falha**
3	10	2	Duração: 11h
14	14	1	Duração: 75 min
	10	2	Duração: 9h
15-Feriado	-	1	Dia: F
	-	2	Dia: F
18	-	3	Dia: E
22	13	1	Duração: 80 min
	18	1	Duração: 60 min
24	2	2	Duração: 22h
	-	3	Dia: E
25	15	1	Duração: 140 min
	0	2	Duração: 120 min
31	10	2	Duração: 14h

Tabela 4.48: Consumos anormais detetados em agosto para: 1-Impressora; 2-Ilha de computadores; 3-Máquina de café;

	Novembro		
	Hora	**Monitorização**	**Falha**
1-Feriado	-	2	Dia: F
	-	1	Dia: F
7	17	3	Padrão: AA
8	10	1	Duração: 80 min
	-	3	Dia: E
13	-	3	Dia: E
14	1	1	Duração: 170 min
16	-	3	Dia: F
17	-	3	Dia: E
21	16	1	Duração: 155 min
22	11	1	Duração: 135 min
24	11	3	Padrão: AA
28	17	1	Duração: 80 min
	15	3	Padrão: AA
30	14	1	Duração: 55 min

Tabela 4.49: Consumos anormais detetados em novembro para: 1-Impressora; 2-Ilha de computadores; 3-Máquina de café;

Por análise das tabelas 4.47:49, salienta-se que o padrão AA, para a máquina de café, é frequentemente classificado como anormal. Isto acontece, porque, em maioria, o padrão registado é A, ou seja, um consumo isolado de baixo patamar. Durante o mês de agosto são raros os consumos, portanto, quando a máquina de café está ativa é considerado um consumo anormal, em comparação com os dias antecedentes.

As restantes anomalias dão-se por os aparelhos apresentarem consumos muito longos no tempo e/ou fora de horas típicas de uso. De notar que, para a ilha de computadores, tipicamente, se inicia o estado anormal à meia-noite. Isto acontece por defeito, o algoritmo apenas analisa um dia de forma individual. Se um consumo se iniciou às 15h:30m de um dia, e se prolongar até as 23:00 do dia seguinte, por defeito, este algoritmo assume que o consumo se inicia à meia-noite. Daí, todos os consumos anómalos que ocorrem durante a madrugada, com início no dia anterior, apresentam um início à meia-noite do próprio dia. De notar que este tipo de anomalia acontece em maioria no mês de março, portanto, ou foi corrigida ou se tornou uma constante, sendo bastante mais provável, pela informação recolhida na secção 4.2.1, a primeira hipótese.

Os feriados que se dão entre segunda a sexta-feira são considerados dias anormais.

4.4.2 Casa Piloto

Para a casa piloto, a monitorização sujeita a análise de falhas é a monitorização efetuada à máquina de lavar roupa, no modo horário. Para esta monitorização, o algoritmo sofreu uma pequena alteração. Visto a máquina de lavar roupa apresentar um consumo muito aleatório no tempo, não faz sentido a comparação dos arranques apenas com os arranques da mesma hora, dos dias disponíveis para comparação. Portanto, é comparado com todos os arranques registados nos trinta dias antecedentes.

	Hora	Falha
2/3-Quinta-Feira	-	Dia: E
17/3-Sexta-Feira	-	Dia: E
27/8-Domingo	-	Dia: E
30/8-Terça-feira	-	Dia: F
18/11-Sábado	10	Duração: 8h

Tabela 4.50: Consumos anormais detetados para a máquina de lavar roupa. O índice indica o dia/mês.

Como já visto anteriormente, para a máquina de lavar roupa, um dia tipo 'E' corresponde a um dia com consumos e 'F' sem consumos. Cada série de estados ON (3.14) da máquina de lavar roupa tem uma duração aproximada de três horas. Portanto, um consumo que dure oito horas é claramente um consumo anormal. Os dias tipo classificados com 'E' ou 'F', que são considerados anormais, são classificados como anomalias porque não é comum nesse dia da semana haver essa atribuição de dia tipo, tendo sempre como referência os trinta dias antecedentes.

4.5 Desagregação de Consumos Elétricos Agregados

Uma forma mais detalhada de obter informação dos consumos da casa piloto, de forma pouco intrusiva, passa por, aos consumos agregados, aplicar algoritmos de desagregação.

Como exposto na secção 3.8, a desagregação para ser efetuada necessita de alguma informação adicional sobre os equipamentos.

Seguidamente são testadas duas metodologias, uma que usa uma assinatura de mudanças de estado sucessivas, registadas nos consumos agregados, para uma série de consumos ativos (3.14) do equipamento que se pretende identificar, nos consumos agregados. A segunda metodologia necessita de uma base de dados de teste para o equipamento que se pretende identificar, em que se conheça os instantes que o equipamento esteve ON ou OFF (3.11), com a informação da mudança de estado verificada nos consumos agregados associada (3.21), para cada instante. Esta última metodologia faz uso

de cadeias de *Markov* escondidas para caracterizar as mudanças de estado dos consumos agregados, estados observáveis, com um estado ON ou OFF (3.11) do equipamento associado, estados escondidos. Com esta cadeia caracterizada, na base de dados de teste, é possível determinar a sequência de estados escondidos associada à sequência de mudanças de estado dos consumos agregados, para outras janelas temporais que não a de teste.

Para teste dos algoritmos de desagregação, é utilizada a monitorização efetuada à máquina de lavar roupa como equipamento a identificar nos consumos agregados. São aplicadas as duas metodologias enunciadas para desagregação dos consumos da máquina de lavar roupa nos consumos totais, agregados, e comparados os seus resultados. As séries de dados consideradas são as adquiridas no modo base.

4.5.1 Mudanças de Estado Detetadas

Antes de se seguir para a desagregação de consumos totais propriamente dita, é necessário associar aos eventos de consumos elétricos registados, na monitorização dos consumos agregados (3.1), a diferença de consumos, entre o consumo elétrico agregado do instante do evento considerado, e o seguinte (3.21), chamada mudanças de estado ou $step_{t,d}$.

Estes estados, para facilitar a análise, são divididos em patamares e devidamente classificados. De notar que a divisão por patamares é simétrica, relativamente à diferença nula.

A tabela 4.51 apresenta os resultados da divisão por patamares (3.24) dos valores da variável *step* (3.22).

	Mínimo	**Máximo**	**Média**
O	-0.340	-0.368	-0.411
N	-0.250	-0.277	-0.319
M	-0.158	-0.201	-0.248
L	-0.046	-0.096	-0.156
$*$	0.043	0.000	-0.043
A	0.156	0.096	0.046
B	0.248	0.201	0.158
C	0.319	0.277	0.250
D	0.411	0.368	0.340

Tabela 4.51: Divisão dos valores das mudanças de estado (*step*) (3.22), em patamares de consumo, para os consumos agregados

O patamar que inclui a mudança de estado nula é classificado com o símbolo vazio, $*$, pois este não corresponde a uma mudança de estado real de consumos, mas sim, originada por flutuações no consumo de fundo. Classificando este patamar com um caracter vazio ($*$), ao aplicar o modelo do pacote de palavras a séries de estados ativos (3.14) dos consumos agregados com $step_p$ (3.24) associados, as mudanças de estado provocados por ruído de fundo não se manifestam. Cada palavra obtida por este modelo, representa então uma mudança sucessiva de estados nos consumos agregados relevantes, associadas às séries de consumos ativos (3.14).

	Centro de cada *cluster*
	Dados Base
C1	$*$ (0.048%)
C2	A^2M (78.794%)
C3	$A(LA)^2M$ (15.558%)
C4	$*$ (5.601%)

Tabela 4.52: Centros de cada *cluster*, representativas das mudanças de estado nos consumos agregados e respetivas percentagens de ocorrências (# 2089-número total de arranques).

As *clusters* assinalados com $*$, na tabela 4.52, visto terem padrões representativos do seu centro heterogéneos a nível de patamares e ocorrerem entre longos períodos, são representadas na figura 4.69.

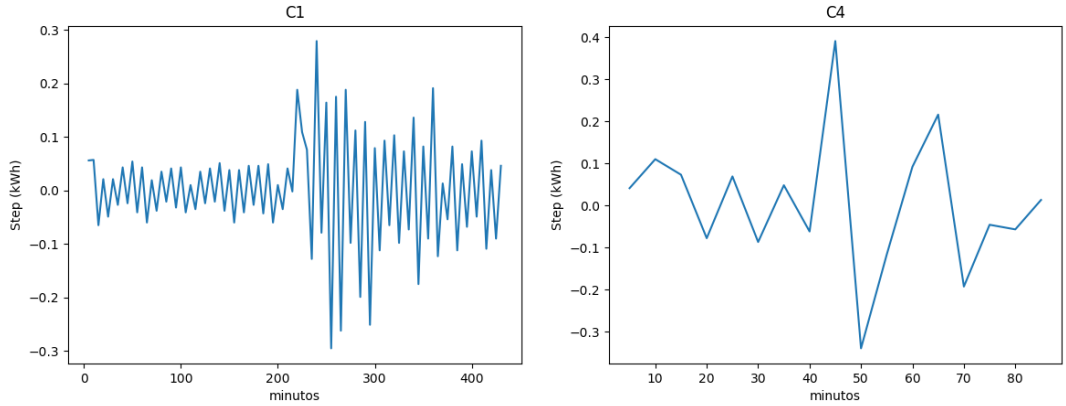


Figura 4.69: Representação das *clusters* C1 e C4 no modo base, para as mudanças de estado nos consumos agregados.

Com a tabela 4.52 e a figura 4.69, tem-se então assim caracterizadas as mudanças de estado típicas, associadas a uma série de estados ativos dos consumos agregados.

De notar que o centro da sucessão de estados mais frequente é caracterizada por duas subidas consecutivas nos consumos, classificadas com o patamar A, seguida de uma descida, classificada com o patamar M. A análise dos centros frequentes é útil no sentido que pode dar pistas ao cliente, de que equipamentos ativos estão associados ao padrões.

Tendo as mudanças de estados associadas aos consumos agregados, é possível prosseguir para a aplicação dos algoritmos de desagregação de cargas.

4.5.2 Assinatura dos Equipamentos

A assinatura do equipamento que se pretende identificar, na base de dados dos consumos agregados da casa piloto, máquina de lavar roupa, é conseguida pela identificação dos instantes em que ocorre uma série aleatória de estados ativos do equipamento (3.14) que se pretende desagregar, identificação dos eventos (3.1), para esses instantes, dos consumos agregados e posterior cálculo das variáveis *step* (3.21). Associando os patamares correspondentes e construindo uma palavra representativa das mudanças de estado,

tem-se assim uma assinatura do equipamento, isto é, a assinatura das mudanças de estado que gera nos consumos agregados, quando este está num estado de atividade.

Para a máquina de lavar foi associada a seguinte sucessão de patamares, num estado ativo aleatório:

$$\textbf{LALL}: * \rightarrow * \rightarrow * \rightarrow \mathrm{L} \rightarrow \mathrm{A} \rightarrow \mathrm{L} \rightarrow * \rightarrow \mathrm{L}$$

Identificada a assinatura do equipamento, LALL, com a metodologia apresentada na secção 3.8.1, por aplicação do algoritmo 3.4, são assim previstos os instantes em que se estima que a máquina de lavar roupa esteve operacional e calculada a *confusion matrix* (3.33), que se resume na tabela 3.2.

Real / Previsão	**Positivo**	**Negativo**
Positivo	205	1537
Negativo	2424	83097

Tabela 4.53: *Confusion Matrix* de 2x2, positivo=ON e negativo=OFF.

A partir da tabela 4.53, que apresenta a *Confusion Matrix* calculada para a previsão em causa, pode-se calcular os seguintes parâmetros de avaliação da previsão:

- $f_{VP} = \dfrac{205}{205 + 2424} = 7.798\%$ (3.34)
- $f_{VN} = \dfrac{83097}{1537 + 83097} = 98.184\%$ (3.35)
- $accuracy = \dfrac{205 + 83097}{205 + 2424 + 1537 + 83097} = 95.461\%$ (3.36)

Com estes parâmetros enunciados, verifica-se que o algoritmo apresenta uma grande percentagem de acertos nos valores negativos (OFF), avaliada pela equação (3.35). Este algoritmo não se revela tão forte nos acertos em positivos (ON) (3.34).

Como o número de ocorrências de valores OFF é extremamente superior (83097+1537) ao número de valores ON (205+2424), este primeiro acaba por mascarar a falta de acertos nos valores ON (3.35) e concluir uma exatidão dos resultados alta (3.36).

4.5.3 *Hidden Markov Model*

Para efetuar a previsão com recurso a cadeias escondidas de *Markov*, divide-se as séries de dados numa fase de treino e numa fase de testes. Na primeira fase deve ser atribuída 70% da janela temporal disponível e na segunda os restantes 30%, sendo que, cada excerto das série de dados deve conter instantes consecutivos.

Na fase de treino pretende-se caracterizar a cadeia escondida de *Markov*.

Para se caracterizar as cadeias de *Markov* escondidas, é necessário, primeiramente, indicar quais são os estados escondidos (3.25) e os observáveis (3.30).

Os estados observáveis correspondem aos valores que a variável de mudança de estado (3.22), traduzidos por patamares (3.24), podem tomar na base de dados dos consumos agregados, e são eles $L = \left\{A, L, B, M, C, N, D, O, *\right\}$ (3.30).

Os estados escondidos correspondem aos estados ativos ou inativos da máquina de lavar roupa (3.11), associados aos consumos brutos (3.1). Esta variável apenas pode tomar dois valores, esses dois valores constituem os estados escondidos e são eles $V = \left\{ON, OFF\right\}$ (3.25).

Posto isto, por análise das sequências de estados observáveis e dos estados escondidos associados na janela temporal de treino, concluiu-se as seguintes matrizes caracterizantes da cadeia escondida de *Markov*, apresentadas na forma de tabela em 4.54:56.

,

	p_i
ON	0.004
OFF	0.996

Tabela 4.54: Probabilidade, p_i (3.29), de um estado escondido (3.25) ocorrer no início da cadeia de *Markov*.

	ON	**OFF**
ON	0.845	0.155
OFF	0.005	0.995

Tabela 4.55: Probabilidades de transição (3.27).

	ON	OFF
A	0.152	0.033
L	0.175	0.034
B	0.022	0.003
M	0.013	0.002
C	0.003	$2.6627 * 10^{-4}$
N	$8.295 * 10^{-4}$	$1.708 * 10^{-4}$
D	0.305	0.445
O	0.000	$1.314 * 10^{-5}$
"	0.329	0.482

Tabela 4.56: Probabilidades de emissão (3.32).

Com as probabilidades apresentadas nas tabelas 4.54:56 tem-se a cadeia de *Markov* escondida caracterizada. Segue-se o cálculo da sequência mais provável de ocorrer, de estados escondidos associados aos estados observáveis, na janela temporal de testes, por aplicação do algoritmo 3.5. Por comparação com a sequência conhecida de estados escondidos, pode-se prosseguir para o cálculo da *confusion matrix* que permite concluir a exatidão das previsões calculadas e avaliar a eficiência do algoritmo.

Real / Previsão	**Positivo**	**Negativo**
Positivo	73	908
Negativo	145	7514

Tabela 4.57: *Confusion Matrix* de 2x2, positivo=ON e negativo=OFF.

A partir da *Confusion Matrix*, apresentada em 4.57, é permitido calcular os seguintes parâmetros:

- $f_{VP} = \dfrac{73}{73 + 145} = 33.486\%$ (3.34)
- $f_{VN} = \dfrac{7514}{7514 + 908} = 89.219\%$ (3.35)
- $accuracy = \dfrac{73 + 7514}{73 + 145 + 7514 + 908} = 87.813\%$ (3.36)

De notar que o número total de previsões apresentadas na tabela 4.57 é muito menor que as apresentadas na tabela 4.53. Tal facto é resultado de para as cadeias escondidas

de *Markov* necessitarem de parte da série de dados para treinar o algoritmo 3.5.

Por análise dos parâmetros calculados a partir da *Confusion Matrix*, verifica-se novamente, como já ocorrido para a metodologia anterior, a percentagem de acertos no estado OFF (3.35) do aparelho é bastante superior à de estado ON (3.34), o que não é surpreendente visto as ocorrências do valor OFF serem em número muito superior às de valor ON.

4.5.4 Discussão

Com a determinação das *Confusion Matrix* dos dois algoritmos de identificação do estado ativo da máquina de lavar roupa, pode-se comparar a eficiência destes.

Conclui-se que o algoritmo da assinatura dos equipamentos é mais eficiente na deteção de estados OFF que o algoritmo de cadeias escondidas de *Markov*, e que este último é mais eficiente na deteção de estados ON em comparação com o primeiro. No entanto, em ambos os algoritmos, a percentagem de acertos é bastante elevada na previsão do estado OFF da máquina de lavar roupa. Este facto tem origem provável no funcionamento muito pontual deste equipamento. Tipicamente tem estados ativos de duração entre uma a duas horas, 2 a 5 vezes por semana, aproximadamente. Sendo pouco frequente o estado ON, é expectável uma tendência dos algoritmos a prever o estado OFF.

Apesar das boas frequências de acertos no estado OFF, os acertos no estado ON não vão além dos 33.486%, ou seja, o objetivo final da identificação do estado ativo do equipamento não é bem sucedido.

Esta falta de eficiência dos algoritmos poderá ser resultado do intervalo de tempo de aquisição ser muito elevado, cinco minutos, para que se consiga identificar mudanças reais de estado nos consumos agregados, e também da dimensão da base de dados de treino ser baixa.

A desagregação, com recurso à assinatura das mudanças de estado associadas ao equipamento, não é tolerante a mudanças de estado causadas por outros equipamen-

tos no decorrer do seu funcionamento, e não tolera padrões de funcionamento muito dispares do usado para a assinatura. No entanto, é exigida muito pouca informação adicional para efetuar a desagregação.

Já a desagregação com recurso a cadeias escondidas de *Markov*, por usar uma larga base de dados de teste, é tolerante a interferências de outros equipamentos, desde que essa interferência seja comum. Também é flexível a vários padrões de funcionamento do equipamento, desde que a sua frequência seja relevante. No entanto, é necessária uma larga dimensão de dados de treino para que o algoritmo aprenda e possa tirar conclusões de novos dados.

O ideal seria aplicar um algoritmo de *machine learning*, que fosse independente de qualquer informação sobre o edifício para que se pretende desagregar os consumos. Para isso seria necessário um elevado histórico de consumos, e a reunião de *datasheets** de um número bastante elevado de aparelhos.

Capítulo 5

Conclusão

Com a aplicação dos algoritmos desenvolvidos, no âmbito deste trabalho, às séries de dados de consumo de uma empresa e casa piloto, obteve-se conhecimento da forma como a energia é gasta nestes edifícios. Descobriu-se e analisou-se padrões de consumo, rotinas, o número de arranques dos equipamentos ao longo do tempo, relações entre as rotinas e a temperatura exterior registada e consumos anormais para um equipamento. Também se analisou associações entre o uso de equipamentos distintos e desagregou-se uma série de dados obtida por um *smart meter*, que passou pela identificação dos momentos em que uma máquina de lavar roupa trabalhou, por análise dos consumos agregados. Por generalização, este algoritmo poderia ser usado para identificar todos os equipamentos presentes na casa piloto, desagregando os consumos por completo, caso se tivesse informações úteis sobre os restantes equipamentos presentes no edifício.

A descoberta de padrões e rotinas nos consumos elétricos, para as bases de dados disponíveis, foi conseguida com sucesso. Obteve-se conhecimento que vai ao encontro daquilo que é expectável para os equipamentos em análise e para o tipo de edifícios em questão. Esta informação, descoberta com a aplicação dos algoritmos desenvolvidos, permite assim saber a forma como a energia é gasta e os fatores externos que influenciam o seu consumo. Serve assim de suporte para sugestões de deslocação de consumos elétricos, quando estes são feitos em alturas de sobrecarga da rede, sem afetar o conforto

e necessidades dos clientes. Serve também para se ter um conhecimento detalhado do uso da energia e assim perceber quais são os consumos que podem ser reduzidos ou evitados.

No decorrer deste projeto, foram também determinadas associações entre os consumos dos equipamentos para a empresa piloto, para um intervalo de tempo de aquisição de cinco minutos. A determinação de associações entre consumos é importante para que se possa conhecer os equipamentos que são usados em conjunto e a sua associação a atividades desempenhadas no edifício. Assim, ao sugerir sugestões de poupança e deslocamento de carga, esta informação é tida em conta para que as necessidades e conforto dos ocupantes do edifício não sejam comprometidas. Para equipamentos com um funcionamento característico inferior ao intervalo de tempo de aquisição, o estado ativo acaba por ser acusado para todo esse intervalo de tempo, não estando o equipamento a funcionar durante todo esse período, apenas funcionando numa fração desse tempo. Portanto, pode conduzir a um acusar de uso simultâneo de equipamentos, quando estes não estão a funcionar ao mesmo tempo, apenas funcionam no intervalo de tempo referido. Assumindo que o problema não se coloca para a base de dados em questão, este algoritmo permite descobrir relações entre o uso de vários equipamentos. No entanto, para a empresa piloto, não são reveladas associações relevantes.

Tinha-se também como objetivo a deteção e diagnóstico de falhas dos equipamentos, mas, devida às características dos dados disponíveis, são adquiridos em intervalos de cinco minutos ou de uma hora e correspondem à energia acumulada consumida nesses intervalos. Com dados de consumo acumulado de cinco minutos (que é o intervalo de aquisição mais baixo) para um equipamento, torna-se difícil avaliar o funcionamento do equipamento e comparar com o seu traço comum. O ideal seria um registo ao segundo. A juntar a este facto, não se possui qualquer informação adicional sobre o equipamento, especificações do equipamento, ou um histórico de dados com indicação dos momentos de falha e seu tipo. No entanto, seria possível detetar falhas apenas com uma dimensão considerável do histórico de dados, adquiridos num curto intervalo de tempo. O seu diagnóstico é que caberia ao cliente. Portanto, nas condições indicadas, a tarefa

de descoberta de falhas fica muito dificultada. Por comparação dos consumos com o histórico de dados, desenvolveu-se um algoritmo para detetar consumos anormais. Este algoritmo, por confronto de um consumo com os realizados para as mesmas horas, para os mesmos dias tipo, nos trinta dias antecedentes, apura se o consumo é expectável ou não. Cabe ao cliente, após a identificação de anomalia, identificar a causa. Tem-se assim o reporte ao cliente de consumos anormais, sem causas associadas. O cliente é informado e assim pode tomar as ações que achar necessárias face às anomalias detetadas.

Por último, desenvolveu-se algoritmos de desagregação. Novamente, o intervalo de tempo de aquisição mais baixo, cinco minutos, revelou-se alto demais para o sucesso da tarefa de desagregação. Portanto, a exatidão dos algoritmos desenvolvidos não é muito alta, não tendo necessariamente causa na qualidade destes, mas sim nas características da base de dados. Com os algoritmos de desagregação pretende-se caminhar para que a tarefa de monitorização seja o menos intrusiva e dispendiosa possível, havendo cada vez mais edifícios com monitorização dos seus consumos.

Pode-se assim concluir que as limitações deste projeto têm origem nos dados, na baixa frequência de aquisição, e não nos algoritmos desenvolvidos. Esta limitação não se revela na deteção de padrões e rotinas e consumos anormais, ou seja, nos algoritmos de *data mining*, apenas se revela nos algoritmos de previsão, de *machine learning*.

Com o conhecimento obtido no decorrer deste projeto, tem-se assim um perfil dos consumos nos edifícios em análise. Conhecendo as necessidades da rede, as alturas do dia em que é necessário aliviar os picos de carga, as alturas em que a rede está mais livre e confrontando com as necessidades do cliente, podem assim ser efetuadas sugestões de deslocamento de consumos do cliente, sem comprometer o seu conforto e necessidades.

Com o conhecimento aprofundado da forma como a energia é gasta, os fatores externos que influenciam o consumo e a deteção de anomalias, o cliente tem toda informação necessária para poder eliminar desperdícios energéticos e consumir a energia de forma mais eficiente.

Este projeto cumpre assim o objetivo de contribuir para um mundo mais verde,

por construção de algoritmos possíveis de usar num serviço de eficiência e poupança energética, tornando também este serviço mais cómodo e acessível possível.

Referências Bibliográficas

[1] M.-S. Chen, J. Han e P.S. Yu. «Data Mining: An Overview from a Database Perspective». Em: *IEEE Transactions on knowledge and data engineering* 8.6 (Dezembro de 1998), pp. 866–833.

[2] J.G. Carbonell, R.S. Michalski e T.M. Mitchell. «Capítulo 1: An Overview Of Machine Learning». Em: *Machine Learning: An Artificial Intelligence Approach.* California: TIOGA Publishing co., 1983, pp. 3–23.

[3] P. Palensky e D. Dietrich. «Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads». Em: *IEEE Transactions on Industrial Informatics* 7.3 (Agosto de 2011), pp. 381–388.

[4] B. DuCharme. *data science glossary.* Acedido 6 março de 2018. URL: `http://www.datascienceglossary.org/#feature`.

[5] J. Lin et al. «Chapter 1: Pattern Recognition in Time Series». Em: *Advances in Machine Learning and Data Mining for Astronomy* (2012).

[6] B. Lkhagua, Y. Suzuki e K. Kawagoe. «Extended SAX: Extension of Symbolic Aggregate Approximation for Financial Time Series Data Representation». Em: *Proceeding of IEICE the 17th Data Engineering Workshop* (jan. de 2006).

[7] C. Chen, D.J. Cook e Crandall A.S. «The user side of sustainability: Modeling behavior and energy usage in the home». Em: *Pervasive and Mobile Computing* 9.1 (2013), pp. 161–175.

[8] A.K. Join, M.N. Murty e P.J. Flynn. «Abstract». Em: *Data Clustering: A Review.*

[9] U. Habib e G. Zucker. «Finding the Different Patterns in Buildings Data Using Bag of Words Representation with Clustering». Em: *2015 13th International Conference on Frontiers of Information Technology (FIT)* (2015), pp. 303–308.

[10] J.E. Seem. «Pattern recognition algorithm for determining days of the week with similar energy consumption profiles». Em: *Energy and Buildings* 37.2 (2005), pp. 127–139.

[11] J. M. Abreu, F.C. Pereira e P. Ferrão. «Using pattern recognition to identify habitual behavior in residential electricity consumption». Em: *Energy and Buildings* 49 (2012), pp. 479–487.

[12] L. Hernández et al. «Classification and Clustering of Electricity Demand Patterns in Industrial Parks». Em: *Energies* 5 (2012), pp. 5215–5228.

[13] S. Rollins e Nilanjan B. «Using rule mining to understand appliance energy consumption patterns». Em: *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (2014), pp. 29–37.

[14] X. Dai e Z. Gao. «From Model, Signal to Knowledge: A Data-Driven Perspective of Fault Detection and Diagnosis». Em: *IEEE Transactions on Industrial Informatics* 9.4 (2013), pp. 2226–2238.

[15] C. Kwan et al. «A novel approach to fault diagnostics and prognostics». Em: vol. 1. 2003, pp. 604–609.

[16] L. Farinaccio e R. Zmeureanu. «Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses». Em: *Energy and Buildings* 30.3 (Agosto de 1999), pp. 245–259.

[17] O. Parson et al. «Non-Intrusive Load Monitoring Using Prior Models of General Appliance Types». Em: *Proceedings of the National Conference on Artificial Intelligence* 1 (jan. de 2012).

[18] Acedido 9 março de 2018. URL: `https://www.wunderground.com/weather/api/d/docs`.

[19] E.M. Knorr e R.T. Ng. «Algorithms for Mining Distance-Based Datasets Outliers in Large Datasets». Em: *Proceedings of the 24rd International Conference on Very Large Data Bases* (1998), pp. 392–403.

[20] Michael Cox e David Ellsworth. «Managing big data for scientific visualization». Em: *ACM Siggraph* 97 (Agosto de 1997), pp. 21–38.

[21] S. Kotsiantis e D. Kanellopoulos. «Discretization Techniques: A recent survey». Em: *GESTS International Transactions on Computer Science and Engineering* 32 (2006), pp. 47–58.

[22] T. Ganu et al. «nPlug: A smart Plug for Alleviating Peak Loads». Em: *2012 Third International Conference on Future Systems: Where Energy, Computing and Communication Meet (e-Energy)*. 2012, pp. 1–10.

[23] S. Garcia-Vallvé e P. Puigbò. *DendroUPGMA tutorial*. URL: `http://genomes.urv.es/UPGMA/`.

[24] O. Maimon e L. Rokach. «Capítulo 6: Discretization Methods». Em: *Data mining and knowledge discovery handbook*. 2ª ed. Springer, 2010, pp. 101–116.

[25] C.D. Manning, P. Raghavan e H. Schütze. «Capítulo 17: Hierarchical clustering». Em: *Introduction to Information Retrieval*. Cambridge University Press, Online edition, 2009, pp. 377–401.

[26] S. Salvador e P. Chan. «Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms». Em: *16th IEEE International Conference on Tools with Artificial Intelligence*. 2004, pp. 576–584.

[27] S. Schimke, C. Vielhauer e J. Dittmann. «Using adapted Levenshtein distance for on-line signature authentication». Em: *Proceedings of the 17th International Conference on Pattern Recognition* 2 (2004), pp. 931–934.

[28] A. Pande e M. Abdel-Aty. «Market basket analysis of crash data from large jurisdictions and its potential as a decision support tool». Em: *Safety Science* 47.1 (2009), pp. 145–154.

[29] D.A. Kelly. «Disaggregating Smart Meter Readings using Device Signatures». Submitted in partial fulfilment of the requirements for the MSc Degree in Computing Science of Imperial College London. Imperial College London, Department of Computing, set. de 2011.

[30] A. Zoha et al. «Non-Intrusive Load Monitoring Approaches for Disaggregated Energy Sensing: A Survey». Em: *Sensors* 12.12 (Dezembro de 2012), pp. 16838–16866.

[31] D. Jurafsky e J.H. Martin. «Capítulo 9: Hidden Markov Models». Em: *Speech and Language Processing*. 3ª ed. 2006.

[32] B. Christopher. *Introduction to Hidden Markov Models with Python Networkx and Sklearn*. Acedido em 8 março de 2018. URL: `http://www.blackarbs.com/blog/introduction-hidden-markov-models-python-networkx-sklearn/2/9/2017`.

[33] M. Story e R.G. Congalton. «Accuracy Assessment: A User's Perspective». Em: *American Society for Photogrammetry and Remote Sensing* 52.3 (mar. de 1986), pp. 397–399.

[34] T. Fawcett. «An introduction to ROC analysis». Em: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874.

[35] A. Kashif, J. Dugdale e S. Ploix. «Simulating Occupants Behavior For Energy Waste Reduction In Dwellings: A Multiagent Methodology». Em: *Advances in Complex Systems* 16.04n05 (Maio de 2013), 1350022 (37 páginas).