UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Análise e Previsão de Anomalias de Consumo

Oana Raluca Pascu

Mestrado em Matemática Aplicada à Economia e Gestão

Trabalho de Projeto orientado por:
Maria Teresa Alpuim
Pedro Alexandre Silva Felício

2018

Agradecimentos

Este trabalho simboliza o realizar de mais uma meta, de mais um degrau alcançado na escada da vida académica, pelo que tenho de deixar o meu agradecimento a algumas pessoas que contribuíram para que isto fosse possível.

Quero agradecer aos meus pais, Nicolae e Mariana, por me terem incutido desde cedo que é imperativo a palavra “desistir” nunca fazer parte do meu leque de escolhas. Isso permitiu que eu tivesse sempre foco nas metas e objetivos aos quais me propus, o que me levou a chegar até aqui. À minha irmã, Nicole, porque palavras nunca são necessárias, basta o olhar de uma criança para nunca duvidarmos do orgulho, admiração e reconhecimento que têm por nós.

À minha orientadora, professora Teresa Alpuim, por me ter impulsionado a aprofundar os meus conhecimentos, a procurar o desconhecido e a não me conformar apenas com o que sei. Por todo o apoio, disponibilidade e ideias que trocámos ao longo da realização deste projeto e por ser sempre incansável em todas as sessões de esclarecimentos.

Ao meu orientador, Pedro Felício, que nunca duvidou das minhas capacidades e tentou ver sempre além das minhas ideias, para atingir a excelência. À empresa EDP Distribuição e à minha equipa de trabalho, pelo apoio que me prestaram e por toda a compreensão ao longo do mestrado.

Por fim, não poderia deixar de agradecer às minhas colegas de faculdade, que me acompanham desde o início desta etapa académica e que nunca me deixaram esquecer que o apoio dos amigos é incondicional.

Resumo

Atualmente a energia elétrica é considerada um bem com uma fundamental aplicação no quotidiano, pois é utilizado para variados fins e sem a eletricidade o ser humano sente-se condicionado. A dependência dos consumidores face à energia elétrica leva a que esta tenha de ser alvo de estudos cada vez mais aprofundados, para que a utilização deste bem não seja ultrapassada pelos avanços tecnológicos da era em que vivemos e, portanto, os consumidores cada vez melhor informados e mais exigentes, não vejam a utilização da energia elétrica como algo limitado.

No presente projeto é efetuada uma contextualização ao tema da energia elétrica, com principal foco no consumo de eletricidade por parte de clientes cujas instalações possuem determinadas características. É também abordado o tema da qualidade dos dados em estudo, para que se consiga determinar quando esses dados são considerados anómalos ou, por outras palavras, não tipificados. Este estudo é feito com base na análise de *clusters*, utilizando dados de consumo históricos, em que as instalações são agregadas em grupos homogéneos. Aplicam-se dois métodos, o de Ward e o de *k*-means, para se observar a distribuição dos objetos pelos vários grupos. O intuito da análise de *clusters* é criar grupos com consumos padrão, bem como grupos com consumos considerados suspeitos, para uma imediata comparação entre si.

Para complementar a análise de *clusters* são criados intervalos de 95% de confiança para o valor médio dos consumos padrão, com vista a observar se os grupos suspeitos apresentam consumos considerados estatisticamente tipificados, ou seja, até que ponto as instalações com consumo suspeito podem afinal enquadrar-se nos grupos de referência.

Por fim, para compreender qual dos métodos utilizados na análise de *clusters* é mais eficiente e robusto em termos da identificação de objetos com consumo efetivamente anómalo, é efetuada a comparação entre os resultados obtidos por ambos e outras fontes de informação acerca dos dados.

Palavras-Chave: Análise de *clusters*, Ward, *k*-means, Intervalos de confiança, Consumo de energia elétrica.

Abstract

Nowadays the electric energy is fundamental to daily life as it is the basis for many usages in such a way that a power shortage causes tremendous impact. As the consumers become more active players in the energy market, more well informed and more demanding, it is of utmost importance to deeply study this market behavior, so that electric energy services keep up to date with technological advances in all human areas.

This project presents a brief background to the electric energy theme, focusing on electric usage by costumers with some specific characteristics, with the aim of studying and understanding when abnormal data are present, in other words, identifying non-typified data. Using historical consumption data, the proposed method to identify abnormal data starts by dividing the observations into homogeneous groups with the help of two different methods: Ward and *k*-means. This allows to observe the similarity between the variables classified in the same groups as well as the differences among these groups. The goal of the cluster analysis is precisely to group similar consumption locations, targeting reference (standard) patterns, as well as grouping objects with non-typified consumption for an immediate comparison.

After dividing the observations into groups, 95% confidence intervals for the means of the standard consumption are built to supplement the cluster analysis and to confirm the non-typified consumption objects grouped before. Finally, both methods, Ward and *k*-means, are compared looking at the grouping of objects with non-typified consumption in an attempt to conclude which method is better.

Keywords: Cluster analysis, Ward, *k*-means, Confidence interval, Electric energy consumption.

Índice

Índice de Figuras

Índice de Tabelas

Lista de Siglas e Abreviaturas

AT	Alta Tensão
BT	Baixa Tensão
BTE	Baixa Tensão Especial
BTN	Baixa Tensão Normal
CPE	Código de Ponto de Entrega
DGE	Direção de Gestão de Energia
EDP	Energias de Portugal
EDPD	EDP Distribuição – Energia S.A.
EDP SU	EDP Serviço Universal, S.A.
ERSE	Entidade Reguladora dos Serviços Energéticos
GMLDD	Guia de Medição, Leitura e Disponibilização de Dados
IC	Intervalo de Confiança
IP	Iluminação Pública
MAT	Muito Alta Tensão
ML	Mercado Livre
MR	Mercado Regulado
MT	Média Tensão
ORD	Operador de Rede de Distribuição
ORT	Operador de Rede de Transporte
PC	Potência Contratada
REN	Redes Energéticas Nacionais
RND	Rede Nacional de Distribuição
RNT	Rede Nacional de Transporte
RRC	Regulamento de Relações Comerciais
SE	Setor Elétrico

1. Introdução

Atualmente associa-se a energia elétrica a pequenas ações como ligar e desligar um interruptor, apesar de este bem ir muito além disso. A descoberta da eletricidade levou a uma evolução significativa da qualidade de vida do ser humano, tendo começado a ser investigada na Grécia Antiga, a partir do momento em que Thales de Mileto observou o primeiro fenómeno de eletricidade estática [1]. Este matemático verificou que a fricção de âmbar na pele de um carneiro provocava a atração de pequenos fragmentos de palha. A partir deste instante, a eletricidade passou a ser conhecida como "elektron" em grego, cujo significado é âmbar e que originou posteriormente a palavra "electricus" em latim, de onde deriva a palavra "elétrico" em português. Depois deste marco na história da energia elétrica, esta foi sendo alvo de estudos e investigações por ser uma área pouco explorada nessa época. Aos poucos, a informação acerca deste bem foi-se desenvolvendo cada vez mais, após a energia elétrica ter sido estudada e relacionada com várias áreas, até ao momento em que o cientista William Gilbert, em 1600, concluiu que a Terra é magnética, motivo pelo qual as bússolas apontam para Norte. Depois de 72 anos da descoberta de Gilbert, Otto Von Guericke, correspondente de Leibnitz, implementa a primeira máquina eletrostática, com base em eletricidade gerada através de fricção [1]. Posteriormente, vários cientistas foram contribuindo para o desenvolvimento do estudo da energia elétrica e do conhecimento cada vez mais aprofundado deste tema, chegando atualmente ao estatuto de um recurso facilmente acessível, utilizado para variados fins e com uma versatilidade ilimitada. Ainda mais recentemente, o uso consciente da energia elétrica tem ajudado a que esta seja consumida de um modo menos prejudicial ao ambiente.

Em Portugal, este bem começou a ser explorado a 30 de outubro de 1878, momento no qual surgiram seis candeeiros Jablochkoff no Chiado. Um marco muito importante para Portugal foi o início da produção de energia elétrica, no início da década de 50, dado pela produção hídrica. Contudo, foi no fim da década de 80 que a eletricidade passou a fazer parte de todos os recantos do país [2]. Atualmente, pode afirmar-se que cada vez mais alguns fatores como a consciencialização ambiental e as flutuações dos preços têm levado a um investimento acrescido nas energias renováveis [3].

O grupo Energias de Portugal (EDP) é um dos maiores operadores europeus no setor da energia elétrica, considerando-se o maior grupo industrial português nesta área. Este grupo é responsável pela maior fatia correspondente à produção, comercialização e distribuição de energia elétrica no país [4]. O grupo está dividido em várias empresas, sendo duas delas a EDP Serviço Universal, S.A. (EDP SU) e a EDP Distribuição – Energia S.A. (EDPD). A EDP SU é uma das maiores empresas na área da comercialização de energia elétrica em Portugal, enquanto a EDPD está responsável pela distribuição deste bem no território nacional continental.

Existe uma crescente preocupação para que a geração da eletricidade em Portugal seja cada vez menos prejudicial à vida do ser humano e a tudo o que o rodeia, por isso a EDP SU fornece energia elétrica aos consumidores, maioritariamente, a partir de fontes renováveis. Esta característica permite uma diminuição da poluição ambiental havendo, no entanto, o problema de não poder ser utilizada continuamente devido à sua dependência relativamente a condições atmosféricas [5]. Apesar de a EDP procurar uma utilização cada vez maior deste tipo de fontes, é natural que as energias não renováveis ainda tenham um papel relevante, dado que não dependem de fatores climatéricos, tais como o vento ou a radiação solar, para que possam ser utilizadas quando a produção de energia elétrica através de fontes renováveis não está disponível, de forma a que o consumo seja sempre satisfeito. Segundo

informação disponível da EDP SU, a evolução mensal de energia elétrica consumida por tecnologia, entre julho de 2017 e junho de 2018 apresenta-se da seguinte forma:

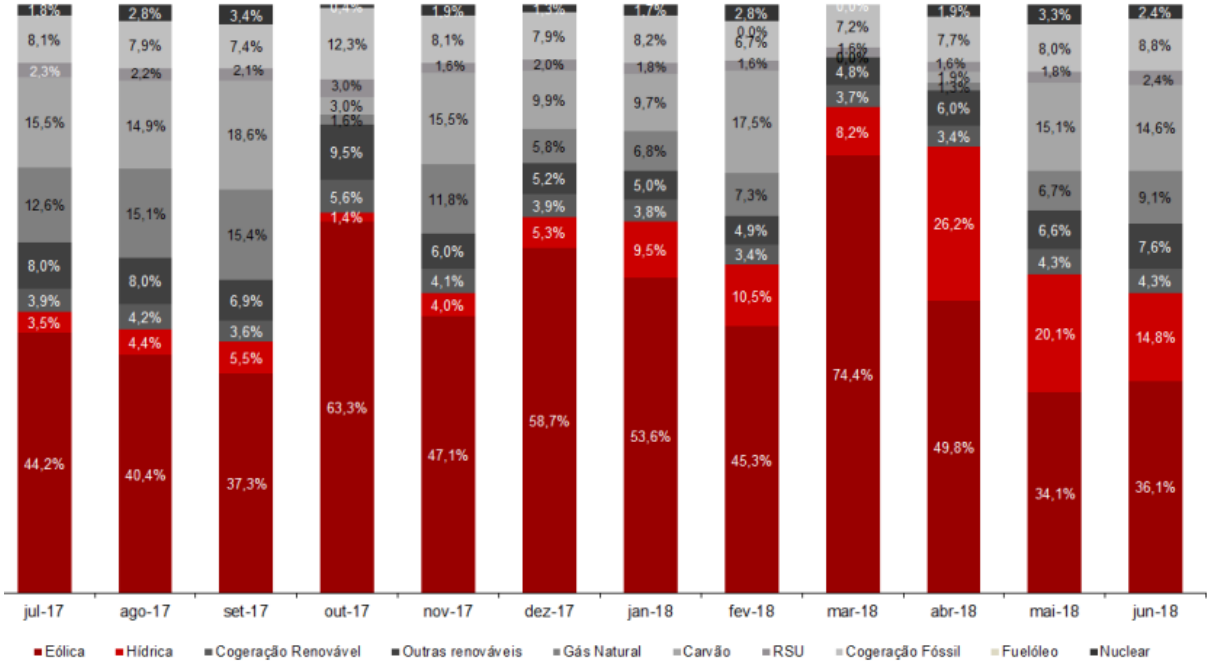

Figura 1.1: Evolução mensal da energia consumida por tecnologia, na EDP SU [5].

O consumo de energia elétrica é um tema que está em constante mudança devido aos avanços tecnológicos, entre outros. É cada vez mais relevante descobrir padrões de consumo, para que se possam efetuar estudos de previsibilidade e análises mais complexas acerca da qualidade dos dados de energia elétrica medidos. No presente estudo foram utilizados dados de consumo de energia elétrica históricos, associados a instalações com certas características, especificadas nos capítulos 2 e 3. O objetivo deste estudo foi a análise e previsão de anomalias de consumo, isto é, a análise dos dados para verificação de consumos não tipificados, ou anómalos, e posterior estudo dos conjuntos de instalações com este tipo de consumo. O estudo foi maioritariamente baseado na análise de *clusters*, com foco nos métodos de classificação hierárquica e não hierárquica, de Ward e *k*-means, respetivamente. Estes algoritmos foram aplicados com o intuito de agrupar as instalações de acordo com os respetivos consumos, para se averiguar quais os grupos possivelmente anómalos. Para esta análise foi utilizado o programa R, o qual permitiu aplicar os algoritmos referidos, com o objetivo de determinar a melhor forma de categorizar os *clusters* em conjuntos de instalações com consumo padrão e conjuntos com consumo não tipificado.

A previsibilidade de anomalias estudada baseia-se na construção de intervalos de 95% de confiança calculados para os grupos com consumo tipificado, de modo a observar se é previsível, ou não, que algum dos conjuntos de instalações com consumo anómalo se deva classificar afinal como *cluster* cujo consumo tem um comportamento padrão. Os intervalos de confiança (IC) foram criados com recurso ao programa Microsoft Excel, obtendo-se assim os resultados apresentados nos capítulos 5.1 e 5.2. No final do relatório é efetuada a comparação entre as duas metodologias utilizadas, de Ward e de *k*-means, de modo a verificar qual o algoritmo mais adequado para a análise e previsão de anomalias de consumo. Por conseguinte, isto permitirá concluir qual a metodologia mais robusta para encontrar conjuntos de instalações cujo consumo é não tipificado.

2. Contextualização

A energia elétrica tem-se tornado um bem cada vez mais essencial, de tal forma que é imprescindível para a vida, sendo necessária em praticamente todas as atividades quotidianas. Para permitir que qualquer pessoa tenha acesso a este bem, existe toda uma cadeia de valor onde estão inseridos os Produtores, os Operadores de Rede e os Comercializadores, de modo a que a energia elétrica chegue ao consumidor final. Em Portugal Continental existem dois grandes operadores de rede, estando um responsável pelo transporte e outro pela distribuição de energia elétrica. A Redes Energéticas Nacionais (REN) é o Operador da Rede de Transporte (ORT) e está responsável pela Rede Nacional de Transporte (RNT) [6] que opera em Muito Alta Tensão (MAT). Por sua vez, a EDPD é um dos Operadores da Rede de Distribuição (ORD) e está responsável pela Rede Nacional de Distribuição (RND) que opera em Alta Tensão (AT), Média Tensão (MT) e Baixa Tensão (BT).

Atualmente parte do Setor Elétrico (SE) é liberalizado para promover uma maior concorrência, de forma a que o consumidor consiga uma melhor qualidade no serviço e no preço da fatura energética. Na comercialização de energia elétrica existem dois tipos de mercado, denominados Mercado Livre (ML) e Mercado Regulado (MR) [7]. O ML é caracterizado pela possibilidade de os comercializadores deste setor concorrerem uns com os outros relativamente aos preços praticados pelo consumo de eletricidade, bem como atuarem sobre as suas próprias regras de comercialização de energia elétrica, dentro do previsto pelo Regulamento de Relações Comerciais (RRC). Por outro lado, o MR é supervisionado pela Entidade Reguladora dos Serviços Energéticos (ERSE) relativamente aos preços de venda de energia elétrica a pagar pelos consumidores finais, entre outros temas. Assim, os consumidores podem escolher o comercializador com o qual pretendem estabelecer um acordo comercial de fornecimento de energia elétrica.

Os consumidores finais são aqueles que recebem a energia elétrica para consumo próprio e podem caracterizar-se de acordo com o Código do Ponto de Entrega (CPE), o contador associado, a Potência Contratada (PC) e o nível de tensão para a respetiva instalação, entre outros. O CPE é um código associado ao ponto de entrega da instalação que, segundo o Guia de Medição, Leitura e Disponibilização de Dados (GMLDD), é o "ponto de rede onde se faz a entrega ou receção de energia elétrica à instalação do cliente, produtor ou outra rede, localizado nos terminais, do lado da rede, do órgão de corte, que separa as instalações" [7]. Por sua vez, um contador de eletricidade é o equipamento de medição do consumo ou da produção de energia elétrica do consumidor.

A PC é a potência que o indivíduo contrata para ter no seu ponto de entrega ou, segundo a ERSE, é a "potência que o distribuidor vinculado coloca, em termos contratuais, à disposição do cliente" [7]. Este valor, para consumidores do segmento Baixa Tensão Normal (BTN), é definido em escalões que variam entre os 1.15 kVA e os 41.4 kVA, num total de 13 escalões. Os escalões de 1.15 kVA e 2.3 kVA são contratados para instalações em que os equipamentos elétricos ligados são de baixa potência, como por exemplo iluminação, e estas instalações podem ser garagens, arrecadações e quiosques. Os escalões 3.45 kVA a 10.35 kVA são usualmente contratados para instalações em que os equipamentos ligados são de maior potência e podem estar ligados em simultâneo. Estas instalações podem ser do tipo residencial ou lojas. Os escalões 13.8 kVA a 41.4 kVA são utilizados em instalações nas quais os equipamentos ligados são de potência elevada e estão ligados em simultâneo. Estas instalações podem ser do tipo comércio ou indústria. [8]. Para resumir a informação anteriormente dada, observam-se na tabela 2.1 os 13 escalões de PC existentes no segmento BTN:

Tabela 2.1: Escalões de potência contratada (kVA) no nível Baixa Tensão Normal [8].

1.15	2.3	3.45	4.6	5.75	6.9	10.35	13.8	17.25	20.7	27.6	34.5	41.4

O nível de tensão da instalação é a diferença de potencial elétrico que alimenta a instalação e varia entre a BT e a MAT, sendo dado pelo valor da tensão entre fases. Pode também ser classificado como o valor máximo de energia elétrica que passa pelas linhas de eletricidade. Os diferentes níveis de tensão encontram-se caracterizados na tabela 2.2:

Tabela 2.2: Caracterização dos níveis de tensão [9], [10].

Nível de Tensão		**Tensão entre Fases (kV)**	**Tensão de Exploração (kV)**	**Potência Contratada (kVA)**
MAT		> 110	150; 220; 400	Ilimitada
AT		45 – 110	60	
MT		1 – 45	10; 15; 30	
BT	E	≤ 1	0,4/0,23	> 41.4
	N			≤ 41.4

Tal como é apresentado na tabela acima, existem quatro tipos de classificação relativos aos níveis de tensão: MAT, AT, MT e BT. Cada um destes tem associado um intervalo de tensões entre fases admitidas, bem como valores de tensão em exploração pelos operadores de rede. Pode ser observado que para os níveis de tensão MT, AT e MAT não é definido nenhum valor mínimo ou máximo de potência contratada, enquanto que para as instalações BT existem valores definidos por regulamentação. O nível de tensão BT divide-se em dois segmentos: Baixa Tensão Especial (BTE) e Baixa Tensão Normal. A diferença entre estes segmentos é verificada no valor da PC pelo consumidor final, ou seja, se este apresentar um valor inferior a 41.4 kVA, a instalação terá um nível de tensão BTN, caso contrário, o consumidor final estará associado a uma instalação do tipo BTE.

Como referido anteriormente, a distribuição de energia elétrica para os níveis de tensão compreendidos entre AT e BT é da responsabilidade da EDPD. Para conseguir efetuar a entrega de energia aos consumidores finais esta empresa está organizada em várias direções, cada qual com o seu âmbito de atuação para cumprimento do objetivo da empresa, tendo sido este projeto realizado com a colaboração do Departamento de Gestão de Dados, área de Validação e Certificação de Dados, inserida na Direção de Gestão de Energia (DGE). Esta direção tem sobre a sua alçada, entre outros, o controlo da qualidade dos dados de energia, identificando, prevenindo e corrigindo situações anómalas seguindo sempre as regras acordadas com a ERSE.

O objeto de trabalho da DGE são os dados de energia, dados esses medidos por contadores colocados nas instalações dos clientes. A sua correta medição é possível se forem garantidas duas condições: conformidade metrológica do contador instalado e conformidade das ligações entre a instalação e o contador. Por vezes de forma fortuita, ou não, estas condições não são garantidas e daí surge o conceito de anomalia de consumo. Quando tal sucede, os dados recolhidos não estão corretos, pelo que é necessário regularizá-los. Esta tarefa consiste na determinação do período temporal em que a anomalia afeta os dados e em que medida esses dados são afetados. É então necessário quantificar a energia que não foi medida pelo contador para que a energia medida anteriormente seja regularizada. As anomalias de consumo dividem-se em anomalias de medição e leitura, procedimento fraudulento e furto, sendo regularizadas de acordo com o estipulado do GMLDD elaborado pela ERSE [6]. De acordo com o GMLDD, uma anomalia de medição e leitura caracteriza-se entre anomalia tipificada e não tipificada. Se for tipificada pode tratar-se de um erro de medição, de configuração, de leitura por acesso local ou de comunicação de dados por acesso remoto. Contrariamente, uma anomalia é não tipificada se não está inserida em nenhum dos grupos anteriores ou, em alternativa, se provoca a

carência de dados de medição que exceda em 10% a energia elétrica associada ao período de faturação anterior. Para as anomalias tipificadas, tem-se então:

- **Erros de medição**: erros de ligação do equipamento, disparidade entre as relações de transformação dos transformadores de medida, defeito de funcionamento ou desregulação do equipamento;
- **Erros de configuração**: incorreta parametrização do equipamento ou dos sistemas de informação associados à medição;
- **Erros de leitura por acesso local**: enganos manuais dos agentes de leitura aquando da observação, recolha ou registo dos valores do equipamento de medição;
- **Erros de comunicação de dados por acesso remoto:** falhas dos processos automáticos, provocando falta de dados.

Quando o cliente possui um contrato de fornecimento de energia elétrica e atua com o objetivo de falsificar o normal funcionamento do equipamento de medição, as suas ações enquadram-se no denominado procedimento fraudulento. A caracterização do procedimento fraudulento é efetuada da seguinte forma:

- **Ligações diretas**: ligações da energia elétrica sem passagem pelo equipamento de medição, ocorrendo furto da quantidade de energia elétrica consumida;
- **Manipulação de equipamento de medição**;
- **Manipulação dos dispositivos de controlo de potência**: manejo de modo a modificar as funções dos equipamentos ou a ter acesso a uma potência que exceda a contratada;
- **Veiculação de energia elétrica entre instalações distintas**:
- **Manipulação ou violação de selagem/fechaduras.**

O último registo de anomalia determinado pelo GMLDD é o furto, sendo caracterizado por ocorrência de utilização ilegal de energia elétrica quando não existe contrato entre o cliente e o comercializador.

3. Descrição do Projeto

O presente projeto é realizado utilizando uma amostra composta por 2682 instalações BTN, de modo a averiguar a presença e previsibilidade de anomalias nos dados correspondentes. Primeiramente, as instalações são agrupadas de acordo com o valor da PC associada, ficando-se assim com os dados correspondentes repartidos em 12 grupos. Uma das potências contratadas referidas no capítulo 2 não é considerada, pois o estudo da previsibilidade de anomalias de consumo é efetuado com o pressuposto de que cada grupo deve conter pelo menos vinte instalações. Como esta PC abrange menos que vinte pontos de medida e, tendo em conta que este valor representa menos de 1% da amostra, não se considera este escalão de potência representativo para o estudo. Procede-se à agregação das instalações com o objetivo de obter grupos com consumos de energia semelhantes entre si, tendo por base a ideia de que a PC da instalação é uma variável que influencia fortemente o padrão de consumo da mesma. Um dos fatores a ter em conta aquando das conclusões retiradas é o número de instalações por grupo de potência contratada, pois a quantidade de pontos de entrega por grupo varia.

A fim de discriminar as instalações cujo consumo de energia elétrica é anómalo, realiza-se uma análise de *clusters* através dos métodos de Ward e de *k*-means. Como se explica no capítulo 4, as duas metodologias aplicadas são diferentes pois o método de classificação hierárquica – método de Ward – baseia-se num processo de agrupamento dos objetos em que, no início, cada grupo é constituído por uma única observação e, em cada passo, se vão juntando observações ou classes até que façam parte do mesmo *cluster*. Por outro lado, o método de classificação não hierárquica – método de *k*-means – prende-se pela escolha prévia do número de *clusters* e, após cada iteração do algoritmo é comparada a distância de cada elemento dos grupos aos respetivos centroides, isto é, aos vetores nos quais cada componente corresponde à média das mesmas componentes dos elementos do grupo. Após a comparação, se se verificar que existe algum elemento cuja distância ao centroide é inferior, face à média do grupo ao qual o objeto pertence, este passa a fazer parte do *cluster* cuja distância entre si e o respetivo centroide é mais pequena. Note-se ainda que nesta metodologia os elementos podem mudar de grupo ao longo do algoritmo, na medida em que uma observação estar contida num grupo não significa necessariamente que no final do procedimento esteja inserida nesse mesmo *cluster*. Em ambos os métodos, a junção dos grupos é feita tendo por base a sua distância, que pode ser definida de várias formas, mas verificando determinadas regras, descritas pormenorizadamente no capítulo 4. A análise de *clusters* permite dividir um conjunto inicial de observações multivariadas em grupos homogéneos e com características semelhantes.

Neste trabalho utiliza-se a análise de *clusters* com o objetivo de uniformizar os padrões de consumo de energia elétrica, o que facilita a verificação da ocorrência de anomalias. Note-se que o agrupamento pelo método de Ward tem por base a matriz de distâncias *D*, cujas entradas correspondem ao quadrado das distâncias euclidianas entre cada grupo, calculadas com base nos consumos elétricos de cada ponto de medida (observação). Os métodos utilizados, de Ward e de *k*-means, são aplicados separadamente aos valores de consumo de três em três meses, isto é, consideram-se os dados de janeiro a março, de abril a junho, de julho a setembro e, finalmente, de outubro a dezembro. Esta divisão efetua-se devido à grande quantidade de dados que a amostra contém. Para os meses de 31 dias existem 2976 quartos de hora, para fevereiro há 2688 valores para o consumo de 15 em 15 minutos e para os restantes meses verificam-se, no máximo, 2880 valores para cada ponto de medida. Deste modo, a matriz de distâncias inicial é uma matriz quadrada de dimensão 2682 x 2682, já que este corresponde ao número de instalações em análise. Assim, na primeira iteração, a entrada (i,j) da matriz de distâncias corresponde à distância euclidiana entre a instalação i e a instalação j em

que cada componente do vetor é o consumo correspondente a cada quarto de hora. Como já foi referido, o número de quartos de hora varia com o trimestre em análise.

A divisão das observações por trimestre é efetuada tendo em atenção razões de natureza diversa. Em primeiro lugar, uma vez que os consumos de energia elétrica apresentam variabilidade diária e semanal, este procedimento procura eliminar a variabilidade anual. Com efeito, as variações sazonais consideram-se bastante acentuadas, podendo confundir o processo de identificação das variações devidas tanto a padrões diferentes de consumo como a observações anómalas. Procura-se, portanto, eliminar uma fonte de variação que é diferente da que se pretende identificar. Por outro lado, as séries de consumos em cada instalação, o número de quartos de hora em análise, não deve ser demasiado grande pois a variabilidade que naturalmente acontece entre um número muito grande de pontos torna difícil fazer o agrupamento em grupos de dimensão razoável, resultado que se pretende obter para tipificar os consumos. Isto significa que duas instalações podem ter um padrão de consumo muito semelhante no inverno, mas serem bastante diferentes no verão. Assim, é possível que pertençam a um mesmo grupo quando considerado apenas o primeiro trimestre, mas estejam contidas em grupos diferentes se se considerar o ano completo.

Após averiguar quais os grupos de instalações com consumos médios possivelmente não tipificados, estes são analisados com mais pormenor através de intervalos de 95% confiança para o valor médio. O objetivo desta análise é verificar rigorosamente se as instalações contidas nos grupos candidatos a anómalos podem, ou não, ser inseridas nos *clusters* com consumo médio padrão. Por conseguinte, para que um grupo com consumo suspeito possa fazer parte de um grupo padrão, espera-se que pelo menos 95% das médias dos consumos em cada quarto de hora correspondentes estejam contidos nos IC construídos com base nos consumos dos *clusters* de referência. Caso isso não aconteça, estatisticamente pode-se concluir que os *clusters* em estudo são de facto anómalos e o seu consumo médio é considerado não tipificado.

Uma vez que o número de amostras para o qual se obtém um intervalo de confiança para a média é igual ao número de quartos de hora do trimestre em análise, isto é, um número que varia entre 8637 e 8833, torna-se impossível na prática proceder à análise e ajustamento de uma distribuição para cada uma destas populações, ou seja, para o consumo num determinado quarto de hora. No entanto, como o número de instalações em cada grupo de referência é bastante grande, utiliza-se o método de construção de um intervalo de confiança para o valor médio de uma população normal. Este procedimento é particularmente robusto, pois em consequência do Teorema do Limite Central e da Lei dos Grandes Números, a distribuição de probabilidade da variável fulcral na qual o método se baseia tem uma distribuição que é bem aproximada pela normal (ou t-Student, já que para grandes valores do número de graus de liberdade esta distribuição se aproxima da normal) para valores não muito grandes da dimensão da amostra [11].

Como a variância é desconhecida e se utiliza o pressuposto de que a amostra segue, aproximadamente, uma distribuição Normal com valor médio μ e desvio padrão σ, o intervalo de $100(1-\alpha)\% = 100(1-0.05)\% = 95\%$ de confiança calculado para o valor médio é dado por

$$\left(\bar{x} - t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}};\ \bar{x} - t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) \tag{3.1}$$

onde $\bar{x}$ representa o consumo médio no respetivo quarto de hora para o grupo em estudo, *s* o desvio padrão correspondente, *n* a dimensão do *cluster* e *t* o quantil de probabilidade $1-\frac{\alpha}{2}$ da distribuição *t*-Student com $n-1$ graus de liberdade. Neste intervalo de confiança estão representados os limites inferior e superior a considerar, respetivamente, para se averiguar se o consumo médio em estudo está contido nos mesmos. Caso isto aconteça, significa que o quarto de hora respetivo está contido no

intervalo de 95% de confiança, considerando-se assim um período de 15 minutos para o qual o consumo médio não é anómalo. Por fim comparam-se dos dois métodos utilizados, método de Ward e de *k*-means, de forma a concluir qual deles é mais robusto e eficiente na identificação de *clusters* anómalos.

4. Análise de *clusters*

A análise de *clusters* ou análise classificatória é um critério de agregação de variáveis em grupos mutuamente exclusivos (cada elemento pertence a um e um só grupo), exaustivos (cada elemento pertence obrigatoriamente a um dos grupos) e homogéneos (os elementos do grupo devem ser mais uniformes entre si, face aos elementos dos restantes grupos). Em geral, cada elemento ou observação denomina-se por objeto, constituído por um determinado número de variáveis, ou seja, é multidimensional e é esta observação que constitui o *cluster*, cuja análise pode ser efetuada através de métodos hierárquicos ou não hierárquicos [12].

Nos métodos hierárquicos os grupos são constituídos passo a passo de tal modo que:

- Quando um elemento é incluído num *cluster*, já não pode ser excluído do mesmo;
- O número de grupos é desconhecido *a priori;*

Para a formação dos grupos podem ser empregues metodologias aglomerativas ou divisivas. Nos métodos aglomerativos, o procedimento consiste em juntar grupos em cada etapa até que haja apenas um *cluster* com todos os elementos da amostra. Por outro lado, os métodos divisivos têm por base a construção de um *cluster* inicial com todos os elementos incluídos no mesmo, sendo estes separados ao longo da execução do algoritmo. Por conseguinte, no final do procedimento cada variável é considerada como um grupo [13].

Os métodos não hierárquicos consistem na definição de uma partição inicial das observações, tendo por base o princípio de que os elementos podem mudar de grupo durante o algoritmo utilizado. Deste modo, o objeto pode mudar de *cluster* acabando por fazer parte, no final do procedimento, do grupo que apresenta uma melhoria no algoritmo efetuado, ou seja, que reflete um decréscimo na variabilidade. No presente estudo, será utilizada a análise de *clusters* com base em ambos os métodos hierárquico e não hierárquico, o de Ward e o de *k*-means, respetivamente.

Para a execução da análise de *clusters* é necessário construir a matriz de distâncias [14]. A dissemelhança entre dois grupos é uma medida da sua homogeneidade e exprime-se através de uma função real não negativa, ou seja, associando a cada par de objetos com *p* variáveis um valor não negativo:

$$\begin{aligned} d: \mathbb{R}^p \text{x } \mathbb{R}^p &\rightarrow \mathbb{R}_0^+ \\ (\boldsymbol{x}, \boldsymbol{y}) &\rightarrow d(\boldsymbol{x}, \boldsymbol{y}) \geq 0 \end{aligned} \tag{4.1}$$

A função *d* pode ser denominada de formas diferentes consoante as suas propriedades:

1. É um índice de dissemelhança se
$$d(\boldsymbol{x}, \boldsymbol{x}) = 0, \forall \boldsymbol{x} \in \mathbb{R}^p \quad \textbf{e} \quad d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}), \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^p \text{x } \mathbb{R}^p;$$
2. É um índice de distância se verifica a propriedade anterior e, além disso,
$$d(\boldsymbol{x}, \boldsymbol{y}) = 0 \implies \boldsymbol{x} = \boldsymbol{y}, \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^p;$$
3. É uma métrica se verifica as propriedades anteriores e
$$d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y}), \forall\, \boldsymbol{x}, \boldsymbol{y} \text{ e } \boldsymbol{z} \in \mathbb{R}^p;$$
4. É uma distância ultramétrica se verifica as propriedades acima e
$$d(\boldsymbol{x}, \boldsymbol{y}) \leq \max\{d(\boldsymbol{x}, \boldsymbol{z}), d(\boldsymbol{z}, \boldsymbol{y})\}, \forall\, \boldsymbol{x}, \boldsymbol{y} \text{ e } \boldsymbol{z} \in \mathbb{R}^p.$$

Para além das propriedades citadas acima, existem ainda alguma notas importantes acerca da função *d*, as quais são apresentadas de seguida:

1. Se *d* for uma distância ultramétrica, também se considera como uma distância ou métrica;
2. A soma de duas métricas é uma métrica;

3. O produto de duas métricas não é sempre uma métrica;
4. As medidas de dissemelhança mais utilizadas na análise de *clusters* são a distância euclidiana, a distância de Minkowski e a distância euclidiana padronizada ou de Karl Pearson.

4.1. Método de Ward

No presente estudo interessa, em particular, tipificar o consumo médio de instalações com base em grupos de variabilidade reduzida em torno dessa média. Além disso, o método de formação de grupos será posteriormente completado com a construção de intervalos de confiança para o valor médio da distribuição normal. Por estas razões, o método hierárquico mais adequado ao problema seria o método de Ward que, de uma forma resumida, pode afirmar-se que se baseia no quadrado da distância euclidiana entre dois objetos, dado por:

$$d^2(\boldsymbol{x_r}, \boldsymbol{x_s}) = \sum_{j=1}^{p} \left(\boldsymbol{x_{r_j}} - \boldsymbol{x_{s_j}} \right)^2 \tag{4.2}$$

onde $\boldsymbol{x_{r_j}}$ e $\boldsymbol{x_{s_j}}$ correspondem aos valores da *j*-ésima variável para os objetos *r* e *s*, respetivamente. Deste modo, $d^2(\boldsymbol{x_r}, \boldsymbol{x_s})$ pode ser vista como a distância física entre os dois objetos, $\boldsymbol{x_r'} = (x_{r_1}, x_{r_2}, \dots, x_{r_p})$ e $\boldsymbol{x_s'} = (x_{s_1}, x_{s_2}, \dots, x_{s_p})$, no espaço Euclidiano *p*-dimensional [15]. O cálculo desta distância é uma consequência do Teorema de Pitágoras, no qual se estabelece que a soma do quadrado dos catetos é igual ao quadrado da hipotenusa (ver figura 4.1).

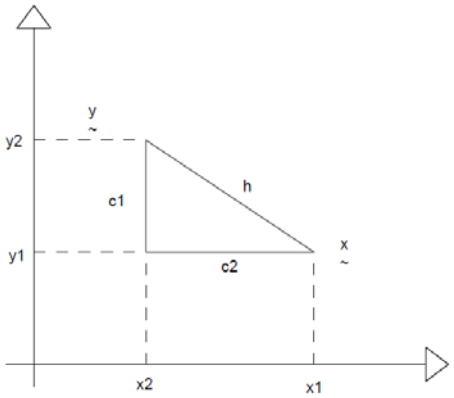

Figura 4.1: Exemplo utilizado para o Teorema de Pitágoras.

.

$$c_1^2 + c_2^2 = h^2 \tag{4.3}$$

A distância euclidiana e o Teorema de Pitágoras estão intimamente relacionados pois este teorema pode ser aplicado no campo multidimensional, tendo em conta que o quadrado da distância entre dois vetores corresponde à soma dos quadrados das distâncias entre as respetivas coordenadas [16].

O método de Ward utiliza, em cada passo, uma generalização da distância euclidiana, em que cada ponto é um grupo de objetos e em que a sua distância se define como a distância entre os vetores de médias de cada grupo. Neste caso, o quadrado da distância euclidiana visa medir a distância entre dois centroides designando-se, mais uma vez, por centroide de um *cluster* o vetor no qual cada componente corresponde à média das mesmas componentes dos elementos do

grupo. Este algoritmo utiliza como distância entre *clusters* a medida baseada na média das dissemelhanças entre os centroides das classes, dada por:

$$d_{\ell,k} = \frac{n_\ell n_k}{n_\ell + n_k} d^2(\overline{\boldsymbol{x}}_\ell, \overline{\boldsymbol{x}}_k) \tag{4.4}$$

Defende-se a adequabilidade deste método para o estudo apresentado porque a expressão (4.4) utiliza o quadrado da distância euclidiana entre os centroides das classes C_ℓ e C_k, respetivamente, $\overline{\boldsymbol{x}}_\ell$ e $\overline{\boldsymbol{x}}_k$, sendo que estes valores também servirão para a análise posterior, na qual se analisarão os grupos cujo consumo médio de energia elétrica é anómalo. Note-se que (4.4) é calculada para decidir sobre a junção das classes C_ℓ e C_k e que n_ℓ e n_k correspondem ao número de elementos dentro das classes referidas, como se esquematiza na figura 4.2.

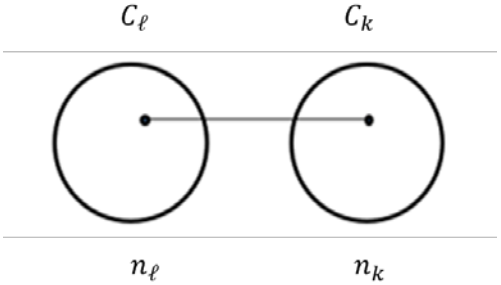

Figura 4.2: Junção de dois objetos das classes C_ℓ e C_k, respetivamente.

Como apresentado em seguida, a expressão (4.4) é igual ao incremento da soma dos quadrados das distâncias dos objetos das classes aos respetivos centroides quando C_ℓ e C_k se fundem. O método de Ward assume que os objetos podem ser representados geometricamente no espaço Euclidiano podendo, no entanto, o resultado ser influenciado por observações muito discordantes. Esta abordagem tem como objetivo, em cada passo, minimizar a variabilidade dentro os grupos. Para melhor compreender este método representa-se por $x_{\ell ij}$ o valor observado da variável x_j, $j = 1, \dots, p$, para o objeto i, $i = 1, \dots, n_\ell$ pertencente ao *cluster* ℓ, $\ell = 1, \dots, N$. Em seguida demonstra-se que o método de Ward, ao juntar os grupos com a menor distância entre si, com base na distância definida pela expressão (4.4), permite o agrupamento dos dois grupos que provocam um menor aumento da variabilidade dentro do novo grupo.

Para ver que assim é, considere-se $\bar{x}_{\ell j}$ a média no *cluster* ℓ para a variável *j*,

$$\bar{x}_{\ell j} = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} x_{\ell ij}, \tag{4.5}$$

e $\bar{x}_{kj}$ a média no *cluster k* para a variável *j*,

$$\bar{x}_{kj} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{kij}. \tag{4.6}$$

A variabilidade do *cluster* ℓ, com a dimensão de n_ℓ objetos, é dada por

$$E_\ell = \sum_{i \in C_\ell} \sum_{j=1}^{p} \left(x_{\ell ij} - \bar{x}_{\ell j}\right)^2 \tag{4.7}$$

e, analogamente, a variabilidade da classe *k*, com dimensão de n_k objetos, é:

$$E_k = \sum_{i \in C_k} \sum_{j=1}^{p} \left(x_{kij} - \bar{x}_{kj}\right)^2 \tag{4.8}$$

Note-se que as expressões (4.7) e (4.8) representam o quadrado da soma dos erros em cada *cluster*, isto é, a soma dos quadrados das distâncias dos objetos do grupo ℓ ou k ao vetor de médias do mesmo [13].

Aquando da junção das classes ℓ e k a variabilidade do *cluster* resultante, indexado por q, representa-se por:

$$E_{\ell \cup k} = E_q \tag{4.9}$$

Seguindo o raciocínio utilizado nas expressões acima, a média do grupo q para a variável j é dada por:

$$\bar{x}_{qj} = \frac{1}{n_\ell + n_k}\left(\sum_{i=1}^{n_\ell} x_{\ell ij} + \sum_{i=1}^{n_k} x_{kij}\right) = \frac{n_\ell \bar{x}_{\ell j} + n_k \bar{x}_{kj}}{n_\ell + n_k} \tag{4.10}$$

Por conseguinte, a variabilidade da classe q calcula-se da seguinte forma:

$$E_q = \sum_{i \in C_\ell} \sum_{j=1}^{p} \left(x_{\ell ij} - \bar{x}_{qj}\right)^2 + \sum_{i \in C_k} \sum_{j=1}^{p} \left(x_{kij} - \bar{x}_{qj}\right)^2 \Leftrightarrow$$

$$\Leftrightarrow E_q = \sum_{i \in C_\ell} \sum_{j=1}^{p} \left(x_{\ell ij} - \bar{x}_{\ell j} + \bar{x}_{\ell j} - \bar{x}_{qj}\right)^2 + \sum_{i \in C_k} \sum_{j=1}^{p} \left(x_{kij} - \bar{x}_{kj} + \bar{x}_{kj} - \bar{x}_{qj}\right)^2 \Leftrightarrow$$

$$\Leftrightarrow E_q = \sum_{i \in C_\ell} \sum_{j=1}^{p} \left(x_{\ell ij} - \bar{x}_{\ell j}\right)^2 + n_\ell \sum_{j=1}^{p} \left(\bar{x}_{\ell j} - \bar{x}_{qj}\right)^2 + 2 \sum_{i \in C_\ell} \sum_{j=1}^{p} \left(x_{\ell ij} - \bar{x}_{\ell j}\right)\left(\bar{x}_{\ell j} - \bar{x}_{qj}\right) +$$

$$+ \sum_{i \in C_k} \sum_{j=1}^{p} \left(x_{kij} - \bar{x}_{kj}\right)^2 + n_k \sum_{j=1}^{p} \left(\bar{x}_{kj} - \bar{x}_{qj}\right)^2 + 2 \sum_{i \in C_k} \sum_{j=1}^{p} \left(x_{kij} - \bar{x}_{kj}\right)\left(\bar{x}_{kj} - \bar{x}_{qj}\right)$$

Contudo, sabe-se que:

$$\sum_{j=1}^{p} \left(\left(\bar{x}_{\ell j} - \bar{x}_{qj}\right) \sum_{i=1}^{n_\ell} \left(x_{\ell ij} - \bar{x}_{\ell j}\right)\right) = 0 \tag{4.11}$$

e

$$\sum_{j=1}^{p} \left(\left(\bar{x}_{kj} - \bar{x}_{qj}\right) \sum_{i=1}^{n_k} \left(x_{kij} - \bar{x}_{kj}\right)\right) = 0 \tag{4.12}$$

Deste modo:

$$E_q = \sum_{i \in C_\ell} \sum_{j=1}^{p} \left(x_{\ell ij} - \bar{x}_{\ell j}\right)^2 + n_\ell \sum_{j=1}^{p} \left(\bar{x}_{\ell j} - \bar{x}_{qj}\right)^2 + \sum_{i \in C_k} \sum_{j=1}^{p} \left(x_{kij} - \bar{x}_{kj}\right)^2 + n_k \sum_{j=1}^{p} \left(\bar{x}_{kj} - \bar{x}_{qj}\right)^2 \Leftrightarrow$$

$$\Leftrightarrow E_q - E_\ell - E_k = n_\ell \sum_{j=1}^{p} \left(\bar{x}_{\ell j} - \bar{x}_{qj}\right)^2 + n_k \sum_{j=1}^{p} \left(\bar{x}_{kj} - \bar{x}_{qj}\right)^2 \tag{4.13}$$

Resta verificar que a expressão

$$n_\ell \sum_{j=1}^{p} \left(\bar{x}_{\ell j} - \bar{x}_{qj}\right)^2 + n_k \sum_{j=1}^{p} \left(\bar{x}_{kj} - \bar{x}_{qj}\right)^2 \tag{4.14}$$

equivale à distância observada entre os grupos ℓ e k, ou seja, que é igual a

$$d(c_\ell, c_k) = \frac{n_\ell n_k}{n_\ell + n_k} d^2(\bar{\boldsymbol{x}}_\ell, \bar{\boldsymbol{x}}_k) = \frac{n_\ell n_k}{n_\ell + n_k} \sum_{j=1}^{p} \left(\bar{x}_{\ell j} - \bar{x}_{kj}\right)^2 \tag{4.15}$$

Tem-se então que

$$n_\ell \sum_{j=1}^{p} \left(\bar{x}_{\ell j} - \bar{x}_{qj}\right)^2 + n_k \sum_{j=1}^{p} \left(\bar{x}_{kj} - \bar{x}_{qj}\right)^2 =$$

$$= \sum_{j=1}^{p} \left(n_\ell \bar{x}_{\ell j}^2 - 2 n_\ell \bar{x}_{\ell j} \bar{x}_{qj} + n_\ell \bar{x}_{qj}^2\right) + \sum_{j=1}^{p} \left(n_k \bar{x}_{kj}^2 - 2 n_k \bar{x}_{kj} \bar{x}_{qj} + n_k \bar{x}_{qj}^2\right)$$

$$= \sum_{j=1}^{p} \left(n_\ell \bar{x}_{\ell j}^2 + n_k \bar{x}_{kj}^2 - 2 \bar{x}_{qj}\left(n_\ell \bar{x}_{\ell j} + n_k \bar{x}_{kj}\right) + (n_\ell + n_k) \bar{x}_{qj}^2\right) =$$

$$= \sum_{j=1}^{p} \left(n_\ell \bar{x}_{\ell j}^2 + n_k \bar{x}_{kj}^2 - (n_\ell + n_k) \bar{x}_{qj}^2\right) \tag{4.16}$$

Sabendo que a dimensão do *cluster* resultante da agregação das classes ℓ e k é representada por $n_q = n_\ell + n_k$, para qualquer j:

$$n_q\bar{x}_{qj} = n_\ell\bar{x}_{\ell j} + n_k\bar{x}_{kj} \Leftrightarrow$$
$$\Leftrightarrow n_q^2\bar{x}_{qj}^2 = n_\ell^2\bar{x}_{\ell j}^2 + n_k^2\bar{x}_{kj}^2 + 2n_\ell\bar{x}_{\ell j}n_k\bar{x}_{kj} \Leftrightarrow$$
$$\Leftrightarrow n_q^2\bar{x}_{qj}^2 = n_\ell^2\bar{x}_{\ell j}^2 + n_k^2\bar{x}_{kj}^2 + 2n_\ell\bar{x}_{\ell j}n_k\bar{x}_{kj} - n_\ell n_k\bar{x}_{\ell j}^2 + n_\ell n_k\bar{x}_{\ell j}^2 - n_\ell n_k\bar{x}_{kj}^2 + n_\ell n_k\bar{x}_{kj}^2 \Leftrightarrow$$
$$\Leftrightarrow n_q^2\bar{x}_{qj}^2 = n_\ell(n_\ell + n_k)\bar{x}_{\ell j}^2 + n_k(n_\ell + n_k)\bar{x}_{kj}^2 - n_\ell n_k\left(\bar{x}_{\ell j}^2 + \bar{x}_{kj}^2 - 2\bar{x}_{\ell j}\bar{x}_{kj}\right) \Leftrightarrow$$
$$\Leftrightarrow n_q^2\bar{x}_{qj}^2 = n_\ell n_q\bar{x}_{\ell j}^2 + n_k n_q\bar{x}_{kj}^2 - n_\ell n_k\left(\bar{x}_{\ell j}^2 - \bar{x}_{kj}^2\right) \Leftrightarrow$$
$$\Leftrightarrow n_q\bar{x}_{qj}^2 = n_\ell\bar{x}_{\ell j}^2 + n_k\bar{x}_{kj}^2 - \frac{n_\ell n_k}{n_q}\left(\bar{x}_{\ell j}^2 - \bar{x}_{kj}^2\right) \quad (4.17)$$

Assim, substituindo a expressão anterior em (4.16) vem que:

$$\sum_{j=1}^{p} n_\ell\left(\bar{x}_{\ell j} - \bar{x}_{qj}\right)^2 + \sum_{j=1}^{p} n_k\left(\bar{x}_{kj} - \bar{x}_{qj}\right)^2 = \sum_{j=1}^{p}\left(n_\ell\bar{x}_{\ell j}^2 + n_k\bar{x}_{kj}^2 - \left(n_\ell\bar{x}_{\ell j}^2 + n_k\bar{x}_{kj}^2 - \frac{n_\ell n_k}{n_q}\left(\bar{x}_{\ell j}^2 - \bar{x}_{kj}^2\right)\right)\right) \Leftrightarrow$$
$$\Leftrightarrow \sum_{j=1}^{p} n_\ell\left(\bar{x}_{\ell j} - \bar{x}_{qj}\right)^2 + \sum_{j=1}^{p} n_k\left(\bar{x}_{kj} - \bar{x}_{qj}\right)^2 = \frac{n_\ell n_k}{n_\ell + n_k}\sum_{j=1}^{p}\left(\bar{x}_{\ell j} - \bar{x}_{kj}\right)^2 \quad (4.18)$$

Conclui-se assim que, a medida do incremento da variabilidade resultante da junção de dois *clusters* é dada por:

$$E_q - E_\ell - E_k = \frac{n_\ell n_k}{n_\ell + n_k}\sum_{j=1}^{p}\left(\bar{x}_{\ell j} - \bar{x}_{kj}\right)^2, \quad (4.19)$$

expressão utilizada para o cálculo da distância entre classes. Por conseguinte, pode afirmar-se que o quadrado da distância euclidiana entre os centroides dos *clusters* agrupados é proporcional ao aumento da variabilidade do novo *cluster* [13].

O algoritmo inicia-se com *n* grupos, número correspondente à dimensão da amostra em estudo. Assim, cada objeto constitui um *cluster* e a distância da mesma à média do respetivo grupo é nula, levando a que $E_\ell = 0$ para cada grupo $\ell = 1, \ldots, N$. Como referido anteriormente, este método visa a minimização do aumento da variabilidade dentro dos grupos, em cada etapa do algoritmo.

O método de Ward baseia-se nos seguintes passos:

1. Construir a matriz de dissemelhanças com base no cálculo da expressão (4.4);
2. Agrupar os *clusters* com menor distância entre si;
3. Atualizar a matriz *D;*
4. Repetir os passos 1 a 3 para os próximos $n - 1$ ciclos;

De acordo com (Anderberg, 1973) [17], as entradas da matriz *D* podem ser obtidas de várias formas: dando pesos às variáveis, com recurso aos cálculos de componentes principais ou a relações não lineares entre as variáveis. Estes factos levam a que o método de Ward seja considerado flexível para a análise de *clusters*. Para que melhor se entenda a aplicação deste método, apresenta-se de seguida um exemplo retirado de (Anderberg, 1973, pg. 43) com a respetiva resolução.

Exemplo 1:

Supõe-se a existência de uma amostra, composta pelos elementos 1, 2, 5, 7, 9 e 10. Dado que a dimensão da amostra é de $n = 6$, o algoritmo corre em seis etapas. Os resultados para cada iteração estão expostos na tabela 4.1 e os cálculos de apenas algumas das distâncias entre os elementos é apresentado de seguida, dado que estes cálculos são análogos para todos os elementos ou grupos.

Tabela 4.1: Primeiro exemplo de utilização do método de Ward.

Iteração	Número de grupos	Grupos
1ª.	6	(1)(2)(5)(7)(9)(10)
2ª.	5	(1,2)(5)(7)(9)(10)
3ª.	4	(1,2)(5)(7)(9,10)
4ª.	3	(1,2)(5,7)(9,10)
5ª.	2	(1,2)(5,7,9,10)
6ª.	1	(1,2,5,7,9,10)

Passo 1:

Grupos (1)(2)(5)(7)(9)(10);

Passo 2:

Cálculo de distâncias e respetiva matriz *D*:

$$d_{12} = d_{21} = \frac{n_1 n_2}{n_1+n_2} d^2(\overline{\boldsymbol{x}}_2, \overline{\boldsymbol{x}}_1) = \frac{1\text{x}1}{1+1}(2-1)^2 = 0.5$$

$$d_{15} = d_{51} = \frac{n_1 n_5}{n_1+n_5} d^2(\overline{\boldsymbol{x}}_5, \overline{\boldsymbol{x}}_1) = \frac{1\text{x}1}{1+1}(5-1)^2 = \frac{16}{2} = 8.0$$

$$d_{17} = d_{71} = \frac{n_1 n_7}{n_1+n_7} d^2(\overline{\boldsymbol{x}}_7, \overline{\boldsymbol{x}}_1) = \frac{1\text{x}1}{1+1}(7-1)^2 = \frac{36}{2} = 18.0$$

$$d_{19} = d_{91} = \frac{n_1 n_9}{n_1+n_9} d^2(\overline{\boldsymbol{x}}_9, \overline{\boldsymbol{x}}_1) = \frac{1\text{x}1}{1+1}(9-1)^2 = \frac{64}{2} = 32.0$$

$$d_{110} = d_{101} = \frac{n_1 n_{10}}{n_1+n_{10}} d^2(\overline{\boldsymbol{x}}_{10}, \overline{\boldsymbol{x}}_1) = \frac{1\text{x}1}{1+1}(10-1)^2 = 40.5$$

D	***1***	***2***	***5***	***7***	***9***	***10***
1	0	0.5	8.0	18.0	32.0	40.5
2	**0.5**	0	4.5	12.5	24.5	32.0
5	8.0	4.5	0	2.0	8.0	12.5
7	18.0	12.5	2.0	0	2.0	4.5
9	32.0	24.5	8.0	2.0	0	0.5
10	40.5	32.0	12.5	4.5	**0.5**	0

Nesta iteração é indiferente agrupar os valores 1 e 2 ou 9 e 10. Desta forma, opta-se pela construção de um *cluster* constituído pelos elementos 1 e 2.

Resultado:

Grupos (1,2)(5)(7)(9)(10);

Passo 3:

$$d_{12,5} = d_{5,12} = \frac{n_5 n_{12}}{n_5+n_{12}} d^2(\overline{\boldsymbol{x}}_{12}, \overline{\boldsymbol{x}}_5) = \frac{1\text{x}2}{1+2}\left(\frac{3}{2}-5\right)^2 = \frac{2}{3}\left(-\frac{7}{2}\right)^2 = \frac{2}{3}\text{x}\frac{49}{4} \simeq 8.2$$

$$d_{12,7} = d_{7,12} = \frac{n_7 n_{12}}{n_7+n_{12}} d^2(\overline{\boldsymbol{x}}_{12}, \overline{\boldsymbol{x}}_7) = \frac{1\text{x}2}{1+2}\left(\frac{3}{2}-7\right)^2 = \frac{2}{3}\left(-\frac{11}{2}\right)^2 = \frac{2}{3}\text{x}\frac{121}{4} \simeq 20.2$$

D	***(1,2)***	***5***	***7***	***9***	***10***
(1,2)	0	8.2	20.2	37.5	48.2
5	8.2	0	2.0	8.0	12.5
7	20.2	2.0	0	2.0	4.5
9	37.5	8.0	2.0	0	0.5
10	48.2	12.5	4.5	**0.5**	0

Neste passo agrupam-se os valores 9 e 10 num *cluster*.

Resultado:
Grupos (1,2)(5)(7)(9,10);

Passo 4:

$$d_{12,910} = d_{910,12} = \frac{n_{910}n_{12}}{n_{910}+n_{12}} d^2(\overline{\boldsymbol{x}}_{12}, \overline{\boldsymbol{x}}_{910}) = \frac{2\text{x}2}{2+2}\left(\frac{3}{2} - \frac{19}{2}\right)^2 = \left(-\frac{16}{2}\right)^2 = \frac{256}{4} = 64.0$$

$$d_{5,910} = d_{910,5} = \frac{n_5 n_{910}}{n_5 + n_{910}} d^2(\overline{\boldsymbol{x}}_{910}, \overline{\boldsymbol{x}}_5) = \frac{1\text{x}2}{1+2}\left(\frac{19}{2} - 5\right)^2 = \frac{2}{3}\left(\frac{9}{2}\right)^2 = \frac{2}{3}\text{x}\frac{81}{4} = 13.5$$

$$d_{7,910} = d_{910,7} = \frac{n_7 n_{910}}{n_7+n_{910}} d^2(\overline{\boldsymbol{x}}_{910}, \overline{\boldsymbol{x}}_7) = \frac{1\text{x}2}{1+2}\left(\frac{19}{2} - 7\right)^2 = \frac{2}{3}\left(\frac{5}{2}\right)^2 = \frac{2}{3}\text{x}\frac{25}{4} \simeq 4.2$$

D	***(1,2)***	***5***	***7***	***(9,10)***
(1,2)	0	8.2	20.2	64.0
5	8.2	0	2.0	13.5
7	20.2	**2.0**	0	4.2
(9,10)	64.0	13.5	4.2	0

A matriz de distâncias revela que o próximo grupo será formado pelos elementos 5 e 7.
Resultado:
Grupos (1,2)(5,7)(9,10);

Passo 5:

$$d_{12,57} = d_{57,12} = \frac{n_{57}n_{12}}{n_{57}+n_{12}} d^2(\overline{\boldsymbol{x}}_{12}, \overline{\boldsymbol{x}}_{57}) = \frac{2\text{x}2}{2+2}\left(\frac{3}{2} - \frac{12}{2}\right)^2 = \left(-\frac{9}{2}\right)^2 \simeq 20.3$$

$$d_{12,910} = d_{910,12} = \frac{n_{910}n_{12}}{n_{910}+n_{12}} d^2(\overline{\boldsymbol{x}}_{12}, \overline{\boldsymbol{x}}_{910}) = \frac{2\text{x}2}{2+2}\left(\frac{3}{2} - \frac{19}{2}\right)^2 = \left(-\frac{16}{2}\right)^2 = \frac{256}{4} = 64.0$$

$$d_{57,910} = d_{910,57} = \frac{n_{910}n_{57}}{n_{910}+n_{57}} d^2(\overline{\boldsymbol{x}}_{57}, \overline{\boldsymbol{x}}_{910}) = \frac{2\text{x}2}{2+2}\left(\frac{12}{2} - \frac{19}{2}\right)^2 = \left(-\frac{7}{2}\right)^2 \simeq 12.3$$

D	***(1,2)***	***(5,7)***	***(9,10)***
(1,2)	0	20.3	64.0
(5,7)	20.3	0	12.3
(9,10)	64.0	**12.3**	0

A distância entre os grupos (5,7) e (9,10) é inferior à distância entre os grupos (1,2) e (5,7) ou (1,2) e (9,10). Assim, agregam-se os *clusters* (5,7) e (9,10), obtendo-se por fim dois grupos (1,2) e (5,7,9,10).

Resultado:
Grupos (1,2)(5,7,9,10);

Passo 6:

$$d_{12,57,910} = d_{57,910,12} = \frac{n_{12}n_{57910}}{n_{12}+n_{57910}} d^2(\overline{\boldsymbol{x}}_{12}, \overline{\boldsymbol{x}}_{57910}) = \frac{2x4}{2+4}\left(\frac{3}{2} - \frac{31}{2}\right)^2 = \frac{4}{3}(-14)^2 \simeq 261.3$$

D	*(1,2)*	*(5,7,9,10)*
(1,2)	0	261.3
(5,7,9,10)	261.3	0

Resta juntar os últimos dois *clusters*, (1,2) e (5,7,9,10), para que o algoritmo termine.

Resultado:
Grupos (1,2,5,7,9,10);

Segue um segundo exemplo para a agregação de grupos formados inicialmente por observações de duas variáveis.

Exemplo 2:

Considere-se agora um conjunto de quatro objetos constituídos por observações de duas variáveis: {(1,3)}, {(4,5)}, {(7,8)} e {(9,11)}.

Tabela 4.2: Segundo exemplo de utilização do método de Ward.

Iteração	Número de grupos	Grupos
1ª.	4	{(1,3)}{(4,5)}{(7,8)}{(9,11)}
2ª.	3	{(1,3),(4,5)}{(7,8)}{(9,11)}
3ª.	2	{(1,3),(4,5)}{(7,8),(9,11)}
4ª.	1	{(1,3),(4,5),(7,8),(9,11)}

Passo 1:
Grupos: {(1,3)}, {(4,5)}, {(7,8)}, {(9,11)};

Passo 2:
Matriz de distâncias:

$$d_{13,45} = d_{45,13} = (1-4)^2 + (3-5)^2 = 13.0$$
$$d_{13,78} = d_{78,13} = (1-7)^2 + (3-8)^2 = 61.0$$
$$d_{13,911} = d_{911,13} = (1-9)^2 + (3-11)^2 = 128.0$$
$$d_{45,78} = d_{78,45} = (4-7)^2 + (5-8)^2 = 18.0$$
$$d_{45,911} = d_{911,45} = (4-9)^2 + (5-11)^2 = 61.0$$
$$d_{78,911} = d_{911,78} = (7-9)^2 + (8-11)^2 = 13.0$$

D	*{(1,3)}*	*{(4,5)}*	*{(7,8)}*	*{(9,11)}*
{(1,3)}	0	13.0	61.0	128.0
{(4,5)}	**13.0**	0	18.0	61.0
{(7,8)}	61.0	18.0	0	13.0
{(9,11)}	128.0	61.0	13.0	0

Nesta iteração juntam-se os dois grupos, {(1,3)} e {(4,5)}.

Resultado:
Grupos {(1,3),(4,5)}{(7,8)}{(9,11)};

Passo 3:
Matriz de distâncias:

$$d_{1345,78} = d_{78,1345} = \frac{n_{1345}n_{78}}{n_{1345}+n_{78}} d^2(\overline{\boldsymbol{x}}_{1345}, \overline{\boldsymbol{x}}_{78}) = \frac{2\text{x}1}{2+1}\left(\left(\frac{5}{2}-7\right)^2 + (4-8)^2\right) = \frac{2}{3}\left(\left(\frac{9}{2}\right)^2 + 16\right) \simeq 24.2$$

$$d_{1345,911} = d_{911,1345} = \frac{n_{1345}n_{911}}{n_{1345}+n_{911}} d^2(\overline{\boldsymbol{x}}_{1345}, \overline{\boldsymbol{x}}_{911}) = \frac{2\text{x}1}{2+1}\left(\left(\frac{5}{2}-9\right)^2 + (4-11)^2\right) = \frac{2}{3}\left(\left(\frac{13}{2}\right)^2 + 49\right) \simeq 60.8$$

$$d_{78,911} = d_{911,78} = (7-9)^2 + (8-11)^2 = 13.0$$

D	***{(1,3),(4,5)}***	***{(7,8)}***	***{(9,11)}***
{(1,3),(4,5)}	0	24.2	60.8
{(7,8)}	24.2	0	13.0
{(9,11)}	60.8	**13.0**	0

Após a agregação dos grupos {(1,3)} e {(4,5)}, segue-se a junção dos grupos {(7,8)} e {(9,11)}.

Resultado:
Grupos {(1,3),(4,5)} e {(7,8),(9,11)}.

Passo 4:
Matriz de distâncias:

$$d_{1345,78911} = d_{78911,1345} = \frac{n_{1345}n_{78911}}{n_{1345}+n_{78911}} d^2(\overline{\boldsymbol{x}}_{1345}, \overline{\boldsymbol{x}}_{78911}) = \frac{2x2}{2+2}\left(\left(\frac{5}{2}-8\right)^2 + \left(4-\frac{19}{2}\right)^2\right) = 60.5$$

D	***{(1,3),(4,5)}***	***{(7,8),(9,11)}***
{(1,3),(4,5)}	0	60.5
{(7,8),(9,11)}	60.5	0

Nesta última iteração, juntam-se os *clusters* {(1,3),(4,5)} e {(7,8),(9,11)}. Posto isto, termina o algoritmo.

Resultado:
Grupos {(1,3),(4,5),(7,8),(9,11)}.

4.2. Método de *k*-means

O método de *k*-means é um método de classificação não hierárquica, ou de partição, que leva à criação de grupos disjuntos cuja união é o conjunto inicial dos objetos. Este tipo de método é conhecido por operar sobre uma matriz de dados, exigir a prévia escolha do número de grupos e apenas classificar objetos. A escolha da partição inicial das classes pode ser feita de três formas: com base no problema em estudo, de acordo com o resultado de um estudo prévio ou ao acaso [18].

Esta metodologia utiliza-se com o objetivo de minimizar a soma de todas as distâncias euclidianas entre cada objeto e o seu centroide [19]. Como vantagem, apresenta-se a sua grande eficácia na formação de grupos com grande dimensão, sendo o algoritmo dado como concluído apenas quando a distância entre cada objeto e o centroide do seu grupo é a menor. Em cada iteração, os grupos são alterados de modo a que a sua variabilidade interna diminua. Se tal não for possível o algoritmo termina. Por esta razão e porque o número possível de grupos a formar é finito, o algoritmo é obrigatoriamente convergente. Deste modo, cada objeto é deslocado de um grupo para outro, até que se encontre contido naquele cujo centroide está mais próximo de si [20]. A formação dos grupos é efetuada de acordo com a decomposição da variabilidade total, dada pela soma entre a variabilidade dentro de cada grupo com a variabilidade entre os grupos:

$$\boldsymbol{T} = \boldsymbol{W} + \boldsymbol{B}, \tag{4.20}$$

em que $\boldsymbol{T}$ é a variabilidade total, dada por

$$\boldsymbol{T} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{x_{ij}} - \overline{\boldsymbol{x}})(\boldsymbol{x_{ij}} - \overline{\boldsymbol{x}})', \tag{4.21}$$

$\boldsymbol{W}$ representa a variabilidade dentro de cada grupo,

$$\boldsymbol{W} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{x_{ij}} - \overline{\boldsymbol{x}}_{\boldsymbol{i}})(\boldsymbol{x_{ij}} - \overline{\boldsymbol{x}}_{\boldsymbol{i}})', \tag{4.22}$$

e **B** corresponde à variabilidade entre grupos, e é dado pela expressão:

$$\boldsymbol{B} = \sum_{i=1}^{k} n_i(\overline{\boldsymbol{x}}_{\boldsymbol{i}} - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_{\boldsymbol{i}} - \overline{\boldsymbol{x}})', \tag{4.23}$$

com *k* o número de *clusters* escolhido, n_i o número de objetos pertencentes ao grupo *i* e $\boldsymbol{x_{ij}}$ o vetor de observações do objeto *j* no grupo *i*. Note-se ainda que a média do grupo *i* corresponde a:

$$\overline{\boldsymbol{x}}_{\boldsymbol{i}} = \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x_{ij}}, \tag{4.24}$$

o número total de objetos em estudo, *n*, é dado por:

$$n = \sum_{i=1}^{k} n_i \tag{4.25}$$

e a média total calcula-se da seguinte forma:

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\boldsymbol{x_{ij}} \tag{4.26}$$

Tal como referido anteriormente, o método de classificação não hierárquica *k*-means visa minimizar a soma dos quadrados das distâncias euclidianas entre os objetos e as médias dos respetivos *clusters*, representada por tr(W) [18]. Esta expressão é dada por:

$$tr(W) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{x_{ij}} - \overline{\boldsymbol{x}}_{\boldsymbol{i}})'(\boldsymbol{x_{ij}} - \overline{\boldsymbol{x}}_{\boldsymbol{i}}) \Leftrightarrow$$
$$\Leftrightarrow tr(W) = \sum_{l=1}^{p}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\boldsymbol{x_{ijl}} - \overline{\boldsymbol{x}}_{\boldsymbol{il}}\right)^2 \Leftrightarrow$$
$$\Leftrightarrow tr(W) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} d_{ij,i}^2 \quad (4.27)$$

A escolha prévia do número de grupos no início do algoritmo é baseada no cálculo da seguinte expressão:

$$R_k^2 = 1 - \frac{tr(W)}{tr(T)} \quad (4.28)$$

Esta escolha é frequentemente efetuada com base na representação gráfica de R^2_k ou através do método de Elbow:

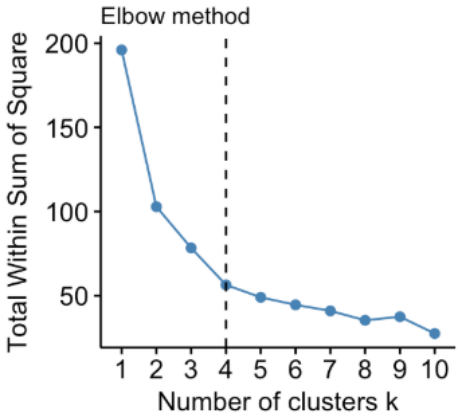

Figura 4.3: Exemplo de gráfico utilizado para escolher o número ótimo de *clusters*, no método de *k*-means [21].

A figura 4.3 é uma representação gráfica que permite a decisão da escolha do melhor número de grupos para se iniciar o algoritmo *k*-means, visto que possibilita a escolha do número ótimo de *clusters*, tendo em conta a minimização da variabilidade dentro dos grupos. Considera-se que o melhor número de grupos para a aplicação deste método é aquele a partir do qual a variabilidade dentro dos grupos começa a estabilizar. Na figura 4.3 o número de *clusters* escolhido deve ser 4 porque a partição das observações por mais do que 4 *clusters* começa a não provocar um decréscimo muito grande na variabilidade, face à escolha de agrupamento das observações em 1, 2 ou 3 grupos. Esta escolha é efetuada com base no método de Elbow, que permite visualizar a percentagem de variabilidade explicada em função do número de *clusters*, *k*. No exemplo da figura 4.3, esta ideia leva á conclusão de que a escolha de 5 ou mais *clusters* não melhora o resultado do algoritmo [22]. Após a seleção do número de *clusters* inicial, *k*, o algoritmo é caracterizado pelos seguintes passos:

1. Avaliação de todas as deslocações de cada objeto do seu grupo para cada um dos restantes grupos;
2. Registo das alterações produzidas no critério de formação de grupos registados, isto é, das distâncias dos elementos aos centroides;
3. Deslocação do elemento, de acordo com a mudança que provoca um maior valor da melhoria verificada para o critério de formação de grupos;
4. Repetição dos primeiros 3 passos até se verificar que a deslocação de qualquer objeto não produz melhoria no critério de formação de grupos.

Para que melhor se entenda como funciona este método, segue abaixo um exemplo de aplicação retirado dos apontamentos da Unidade Curricular Análise Exploratória de Dados Multivariados [18].

Exemplo:

Neste exemplo escolhem-se previamente dois grupos para a agregação de cinco indivíduos: A, B, C, D e E. Os dados do problema são os seguintes:

Tabela 4.3: Dados correspondentes ao exemplo utilizado para aplicar o método de *k*-means.

	$\mathbf{X_1}$	$\mathbf{X_2}$
A	2.0	8.0
B	5.0	1.0
C	4.0	12.0
D	15.0	4.0
E	16.0	5.0

Passo 1:

Clusters: {A, B}, {C, D, E}

Passo 2:

Cálculo dos centroides:

$$\frac{1}{2}(\boldsymbol{x_A} + \boldsymbol{x_B}) = \overline{\boldsymbol{x}}_1 = \frac{1}{2}\left(\begin{bmatrix}2\\8\end{bmatrix} + \begin{bmatrix}5\\1\end{bmatrix}\right) = \begin{bmatrix}3.5\\4.5\end{bmatrix}$$

$$\frac{1}{2}(\boldsymbol{x_C} + \boldsymbol{x_D} + \boldsymbol{x_E}) = \overline{\boldsymbol{x}}_2 = \frac{1}{3}\left(\begin{bmatrix}4\\12\end{bmatrix} + \begin{bmatrix}15\\4\end{bmatrix} + \begin{bmatrix}16\\5\end{bmatrix}\right) = \begin{bmatrix}11.67\\7.0\end{bmatrix}$$

Distâncias dos elementos aos seus centroides:

$$d^2(A, \overline{\boldsymbol{x}}_1) = (2.0 - 3.5)^2 + (8.0 - 4.5)^2 = 14.5$$

$$d^2(A, \overline{\boldsymbol{x}}_2) = (2.0 - 11.7)^2 + (8.0 - 7.0)^2 \approx 94.5$$

$$d^2(B, \overline{\boldsymbol{x}}_1) = (5.0 - 3.5)^2 + (1.0 - 4.5)^2 = 14.5$$

$$d^2(B, \overline{\boldsymbol{x}}_2) = (5.0 - 11.7)^2 + (1.0 - 7.0)^2 \approx 80.5$$

O cálculo das distâncias dos elementos C e D aos respetivos centroides é análogo. Os resultados correspondentes a esta iteração encontram-se abaixo:

Tabela 4.4: Resultado do segundo passo tomado no método de *k*-means.

	Centroides		**Distância (d^2) do elemento ao centroide**				
Grupo *i*	$\overline{\boldsymbol{x}}_{i1}$	$\overline{\boldsymbol{x}}_{i2}$	A	B	C	D	E
{A, B}	3.5	4.5	14.5	14.5	56.5	132.8	156.5
{C, D, E}	11.7	7	94.5	80.5	83.8	20.1	22.8

Tendo em conta o pressuposto de que se os pontos estão contidos no grupo correto, a sua distância ao centroide do respetivo grupo vai ser a mais baixa, considera-se que os indivíduos A e B pertencem à classe correta, pois as suas distâncias ao centroide do primeiro grupo são inferiores (14.5 para os dois elementos), face à distância apresentada de A e B ao grupo {C, D, E}, com valores de 94.5 e 80.5, respetivamente. No entanto, o objeto C está mais próximo do centroide associado ao *cluster* {A, B} do que daquele no qual está inserido. Procede-se então à deslocação deste elemento para o primeiro grupo apresentado. Finalmente, pode concluir-se que os elementos D e E estão bem posicionados porque o centroide do respetivo *cluster* é aquele que se encontra mais perto dos indivíduos referidos.

Clusters: {A, B, C}, {D, E}

Passo 3:

Cálculo dos centroides:

$$\frac{1}{3}(\boldsymbol{x_A} + \boldsymbol{x_B} + \boldsymbol{x_C}) = \overline{\boldsymbol{x}}_1 = \frac{1}{3}\left(\begin{bmatrix}2\\8\end{bmatrix} + \begin{bmatrix}5\\1\end{bmatrix} + \begin{bmatrix}4\\12\end{bmatrix}\right) \approx \begin{bmatrix}3.7\\7.0\end{bmatrix}$$

$$\frac{1}{2}(\boldsymbol{x_D} + \boldsymbol{x_E}) = \overline{\boldsymbol{x}}_2 = \frac{1}{2}\left(\begin{bmatrix}15\\4\end{bmatrix} + \begin{bmatrix}16\\5\end{bmatrix}\right) = \begin{bmatrix}15.5\\4.5\end{bmatrix}$$

Distâncias dos elementos aos seus centroides:

$$d^2(A, \overline{\boldsymbol{x}}_1) = (2.0 - 3.7)^2 + (8.0 - 7.0)^2 \approx 3.8$$
$$d^2(A, \overline{\boldsymbol{x}}_2) = (2.0 - 15.5)^2 + (8.0 - 4.5)^2 = 194.5$$
$$d^2(B, \overline{\boldsymbol{x}}_1) = (5.0 - 3.7)^2 + (1.0 - 7.0)^2 \approx 37.8$$
$$d^2(B, \overline{\boldsymbol{x}}_2) = (5.0 - 15.5)^2 + (1.0 - 4.5)^2 = 122.5$$

Tal como no passo anterior, o cálculo das distâncias dos elementos C e D aos respetivos centroides é análoga, face aos cálculos apresentados acima. Os resultados correspondentes a este passo encontram-se abaixo:

Tabela 4.5: Resultado do terceiro passo tomado no método de *k*-means.

	Centroides		**Distância (d^2) do elemento ao centroide**				
Grupo *i*	$\overline{\boldsymbol{x}}_{\boldsymbol{i1}}$	$\overline{\boldsymbol{x}}_{\boldsymbol{i2}}$	A	B	C	D	E
{A, B, C}	3.7	7.0	3.8	37.8	25.1	737.8	156.0
{D, E}	15.5	4.5	194.5	122.5	188.8	0.5	0.5

Conclusão: observando os resultados obtidos, todos os indivíduos apresentam menor distância ao centroide do grupo onde estão contidos, face ao centroide do outro *cluster* em análise. Isto significa que não deve ser efetuada mais nenhuma deslocação, pois o resultado que provoca uma melhoria no critério de formação de grupos é dado pelo agrupamento dos 5 objetos da seguinte forma: {A, B, C}, {D, E}.

4.3. Intervalos de confiança

Um IC é construído com base numa amostra, sendo um intervalo de valores aleatórios, isto é, que dependem da amostra ao qual se atribui um grau de confiança em que o respetivo intervalo contém o verdadeiro valor em estudo do parâmetro desconhecido. Pode também dizer-se, ao calcular-se um intervalo de confiança, que se obtém uma estimativa intervalar do valor do parâmetro desconhecido. Se a amostra aleatória em estudo for $(X_1, \dots, X_n)$, proveniente de uma população com distribuição $F(x;\ \theta)$, tem-se sempre que:

$$E(\bar{X}) = \mu, \tag{4.31}$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} \tag{4.32}$$

para os quais

$$\mu = E(X_i), \forall i = 1, \dots, n \tag{4.33}$$

$$\sigma^2 = Var(X_i), \forall i = 1, \dots, n \tag{4.34}$$

A variável aleatória $S(X_1, \dots, X_n; \theta)$ é função da amostra aleatória e de θ e é denominada variável fulcral se a distribuição de probabilidade correspondente não depender do parâmetro θ. O nível de confiança é percentual e é dado por $100(1-\alpha)\%$, no qual α toma, em geral, os valores 0.01, 0.05 ou 0.1, isto é, o intervalo terá um grau de 99%, 95% ou 90% de confiança correspondente à sua amplitude que, por sua vez, está associada à incerteza face ao parâmetro em estudo [23]. O valor que é frequentemente utilizado para α é 0.05. Deste modo, no presente relatório os intervalos criados têm um grau de 95% de confiança.

Existem vários tipos de IC, dos quais aquele que será utilizado para o presente relatório é um IC para o valor médio, ou seja, o parâmetro em estudo é $\theta = \mu$ e o IC para μ será (L_1,L_2) tal que:

$$P(L_1 < \mu < L_2) = 1 - \alpha \tag{4.35}$$

Ao calcular-se o intervalo com $100(1-\alpha)\%$ de confiança para o valor médio, μ, com base na amostra aleatória recolhida da respetiva população o intervalo obtido contém, ou não, o verdadeiro valor do parâmetro. Este tipo de intervalo baseia-se no pressuposto de que a variável aleatória *X* segue uma distribuição Gaussiana, com valor médio μ e desvio padrão σ, $X \cap N(\mu, \sigma)$ [23]. Neste caso, ambos os parâmetros são desconhecidos e o desvio padrão é substituído pelo seu estimador:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{4.36}$$

Neste caso, a variável fulcral é dada por:

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \tag{4.37}$$

que tem uma distribuição t-Student com *n-1* graus de liberdade, isto é

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \cap t_{n-1} \tag{4.38}$$

Para o cálculo do IC deve fixar-se previamente o nível de confiança *1-α* com $0 < \alpha < 1$, de tal modo que:

$$P\left(-t_{1-\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < t_{1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow \quad (4.39)$$

$$\Leftrightarrow P\left(\bar{X} - t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha \quad (4.40)$$

Obtém-se assim o estimador intervalar:

$$\left(\bar{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}; \bar{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) \quad (4.41)$$

Conclui-se que, para a amostra em estudo, o intervalo de 95% de confiança para o valor médio μ com σ desconhecido é dado por:

$$\left(\bar{x} - t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}; \bar{x} + t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) \quad (4.42)$$

Este IC baseia-se no pressuposto de que a população da qual foi recolhida a amostra tem distribuição Gaussiana [11]. No caso em estudo, este método será utilizado para verificar se a média de grupos com pequena dimensão está razoavelmente próxima dos intervalos de confiança construídos para a média dos grupos cujo consumo foi considerado padrão, grupos esses de dimensão considerável. Uma vez que este processo se repetirá para cada quarto de hora, num total de 8637 e 8833 consoante o trimestre, torna-se praticamente impossível verificar qual a distribuição de probabilidade correspondente aos grupos de consumo padrão. No entanto, esta metodologia tem uma grande vantagem que é a sua robustez, isto é, pode ser utilizada em amostras provenientes de populações com qualquer distribuição de probabilidade desde que essas amostras tenham uma dimensão razoavelmente grande, como é o caso dos grupos de consumo padrão. Estas boas propriedades devem-se ao Teorema do Limite Central e à Lei dos Grandes Números que permitem que, para valores da dimensão da amostra não muito pequenos, o IC seja válido para qualquer distribuição de probabilidade da população em estudo [11].

5. Apresentação de Resultados

Neste capítulo, apresentam-se os resultados da aplicação dos procedimentos já resumidamente descritos nos capítulos 4.1 e 4.2 aos consumos de energia da amostra de instalações em estudo. Assim, iniciou-se o processo através de uma análise de *clusters* que foi aplicada a cada uma das 12 potências contratadas referidas no capítulo 2. Os métodos de agrupamento utilizados foram o método de Ward, pois pareceu o mais consentâneo com a ideia base de procurar grupos com médias semelhantes e variâncias reduzidas, e o método de *k*-means, que se baseia numa lógica semelhante mas que segue um método não hierárquico. Pretende-se também comparar os resultados obtidos pelos dois métodos com o objetivo de perceber qual o mais apropriado para a identificação de observações anómalas.

Para cada PC o método de classificação hierárquica produziu grupos homogéneos com um número razoável de observações e alguns grupos com menos observações e bem diferenciados relativamente aos outros grupos. Este tipo de resultado é o ideal para a identificação de observações candidatas a serem consideradas anómalas, uma vez que estas tendem a ter um padrão próprio e distinto dos grupos de maior dimensão, que refletem os padrões de consumo dos diversos tipos de consumidores. Relativamente ao método de classificação não hierárquica, para cada PC e para cada grupo de três meses, foi criado o mesmo número de *clusters* do que o obtido pelo método de Ward, de modo a que se consigam comparar os resultados produzidos por cada um dos algoritmos. Neste capítulo apresentam-se as conclusões resultantes da aplicação do método de Ward, seguido pelo método de *k*-means e, por fim, a comparação dos dois algoritmos referidos. Pretende-se assim perceber se os resultados para ambos os métodos coincidem, provando a eficiência dos agrupamentos obtidos.

5.1. Resultados pelo método de Ward

A aplicação do método de Ward tem por base o cálculo do quadrado da distância euclidiana, tal como foi mencionado no capítulo 4, para o posterior agrupamento das instalações em grupos homogéneos. O processo de agregação pode ser analisado através do dendrograma, representação gráfica que permite a visualização dos grupos obtidos pela aplicação do algoritmo associado ao referido método [14]. Os grupos estão representados no eixo horizontal do gráfico e as distâncias euclidianas entre os objetos apresentam-se no eixo vertical. Para que os grupos sejam mais homogéneos dividiram-se os dados por trimestre, o que originou um dendrograma para cada trimestre. A apresentação de resultados efetuada de seguida será para a PC E porque é uma das potências que contém mais instalações e, por isso, pode apresentar mais variabilidade nos consumos quartos-horário. Nos anexos A.1, A.2 e A.3 podem observar-se os gráficos resultantes da aplicação do método de Ward, para as PC A, B e J, tendo sido ocultadas as figuras referentes às 8 restantes potências contratadas, dado que o procedimento foi análogo.

Para a utilização da classificação hierárquica, isto é, através do algoritmo implementado para o método de Ward, apresentam-se de seguida os resultados do agrupamento das observações, nos quais é efetuada a análise pormenorizada por trimestre. Assim, a análise inicia-se para os meses de janeiro a março, através das figuras 5.1 a 5.3.

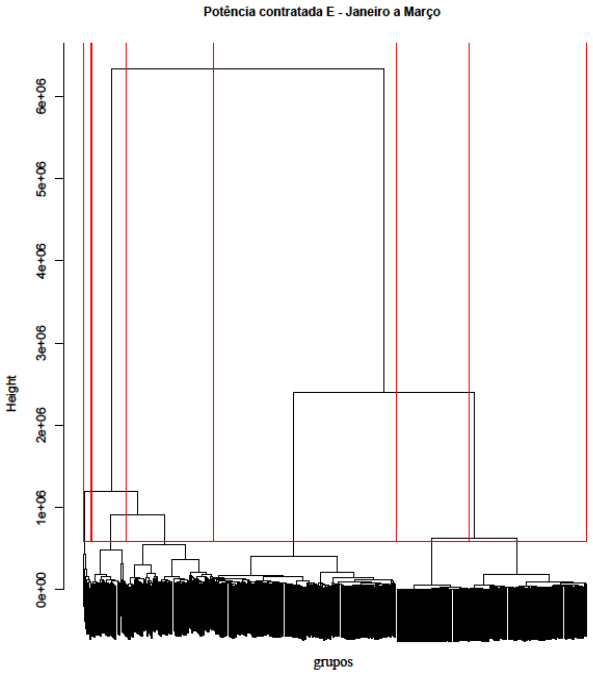

Figura 5.1: Dendrograma associado à PC E, para o primeiro trimestre de 2017, pelo método de Ward.

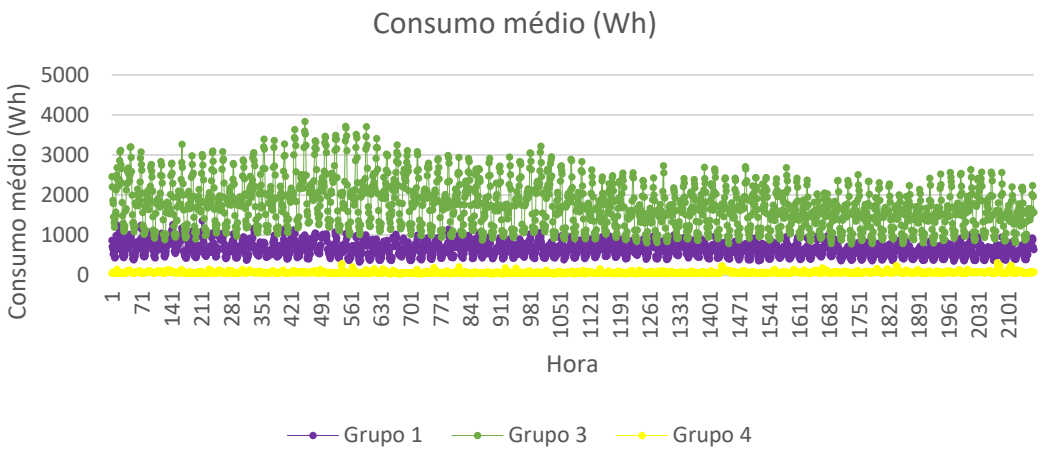

Figura 5.2: Consumo médio (Wh) para os grupos 1, 3 e 4 no primeiro trimestre de 2017, pelo método de Ward.

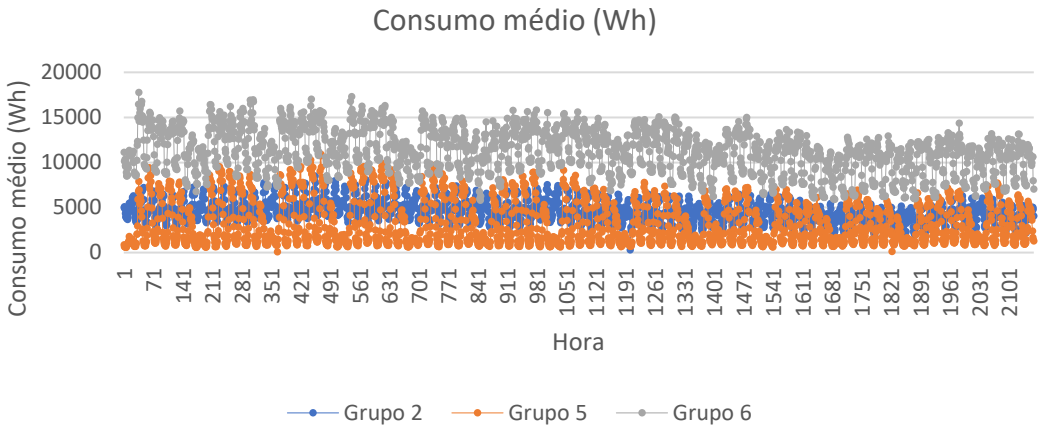

Figura 5.3: Consumo médio (Wh) para os grupos 2, 5 e 6 no primeiro trimestre de 2017, pelo método de Ward.

No primeiro trimestre de 2017, a aplicação do método de Ward originou seis conjuntos de instalações que podem ser visualizados no dendrograma apresentado na figura 5.1. A escolha deste número de grupos resulta de um compromisso entre a estrutura visual do dendrograma e a análise do eixo vertical, isto é, uma escolha de modo a que a distância entre elementos ou subgrupos de um mesmo *cluster* não seja demasiado grande. Pretendem-se construir grupos relativamente homogéneos cuja variabilidade interna não atinja valores demasiado grandes. Por exemplo, se for escolhida uma divisão em dois grupos, isso significa que os elementos de um mesmo grupo podem distar entre si por um valor de cerca de $2.5x10^6$ W por trimestre, ou seja, uma distância média de aproximadamente 312.5 W por quarto de hora, o que parece exagerado.

Através da figura 5.1 observa-se que, de janeiro a março, os grupos com menos instalações são os dois primeiros, podendo ser considerados candidatos a anómalos devido ao facto de a distância entre os seus consumos quartos-horário e os consumos correspondentes aos restantes pontos de medida ser considerada grande, de acordo com o algoritmo utilizado. Os restantes quatro grupos revelam uma dimensão superior e são os possíveis candidatos a *clusters* de referência para a comparação dos consumos de energia entre as instalações dos vários grupos.

As figuras 5.2 e 5.3 refletem o consumo médio (Wh) dos grupos construídos com base no método de classificação hierárquica, para a PC E, nos primeiros três meses de 2017. Embora a análise de *clusters* tenha sido feita utilizando os consumos em cada quarto de hora, optou-se por representar graficamente as médias de cada grupo correspondentes aos consumos horários, pois caso contrário os gráficos tornar-se-iam demasiado pesados e complexos, dificultando a visualização do comportamento dos diferentes grupos. O eixo das ordenadas representa o consumo médio e o das abcissas as horas do trimestre em causa, tendo em conta que este trimestre contém 2159 horas. Para uma melhor visualização dos resultados obtidos criaram-se dois gráficos – figuras 5.2 e 5.3 – porque a colocação dos seis conjuntos de instalações apenas numa figura tornaria a análise menos precisa, devido à escala e aos consumos médios que cada grupo atinge serem bastante diferentes. Para a análise dos restantes trimestres de 2017 o raciocínio de construção dos gráficos associados aos consumos médios foi análogo.

Foram criados seis *clusters* cuja energia elétrica despendida, em média, varia entre os 0 Wh e os 17500 Wh. Devido ao facto de os grupos serem diferentes entre si, é necessário escolher aqueles que contêm maior número de instalações e cujo consumo parece enquadrar-se melhor neste escalão de PC para que sejam tidos como *clusters* com consumo padrão. Pela visualização das figuras 5.2 e 5.3 observa-se que os grupos 1, 2, 3, e 4 apresentam consumos médios de energia elétrica na ordem dos 600 Wh, 4500 Wh, 1700 Wh e 70 Wh, respetivamente, sendo também estes os *clusters* que contêm um maior número de instalações, o que permite a escolha dos mesmos como conjuntos de instalações com consumo de referência. Por conseguinte, os *clusters* 5 e 6 apresentam menor dimensão e um valor de energia elétrica média gasta de, aproximadamente, 3000 Wh e 11000 Wh, respetivamente. Comparando os valores máximos que estes dois grupos atingem no primeiro trimestre com os valores dos grupos de referência, os *clusters* 5 e 6 são então considerados como grupos suspeitos em relação ao consumo médio de energia elétrica, isto porque os valores de energia elétrica média consumida por estes dois grupos não estão de acordo com as características desta PC. No final da análise deste método, os grupos tidos como suspeitos nestes primeiros três meses – *clusters* 5 e 6 – serão comparados com os grupos formados nos restantes trimestres de 2017, de modo a averiguar se ao longo do ano as instalações inseridas nestes dois grupos continuam a ser consideradas anómalas.

Em suma, de acordo com os resultados obtidos para os meses de janeiro a março, é necessário reter as seguintes conclusões:

- Criação de seis grupos;

- Grupos 1, 2, 3 e 4 são considerados como conjuntos de instalações com consumo médio de referência:
  - Grande dimensão;
- Grupos 5 e 6 apresentam consumos médios possivelmente não tipificados:
  - Menor dimensão, face aos grupos de referência;
  - Consumos máximos atingidos não vão de encontro às características deste escalão de PC.

As figuras 5.4, 5.5 e 5.6 indicam os resultados obtidos com a aplicação do método de Ward para a PC E, com os dados do segundo trimestre do ano em estudo . É efetuada de seguida uma análise aos resultados identificados, de modo a inferir quais os *clusters* cujo consumo médio de energia elétrica é possivelmente não tipificado.

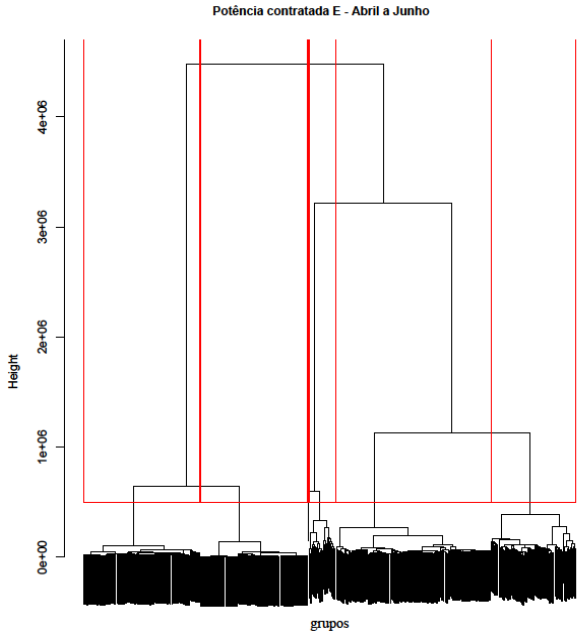

Figura 5.4: Dendrograma associado à PC E, para o segundo trimestre de 2017, pelo método de Ward.

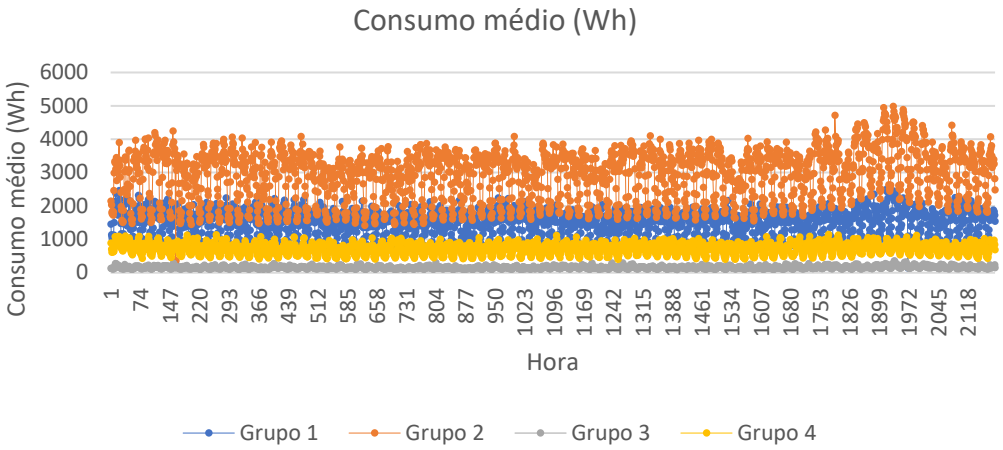

Figura 5.5: Consumo médio (Wh) para os grupos 1, 2, 3 e 4 no segundo trimestre de 2017, pelo método de Ward.

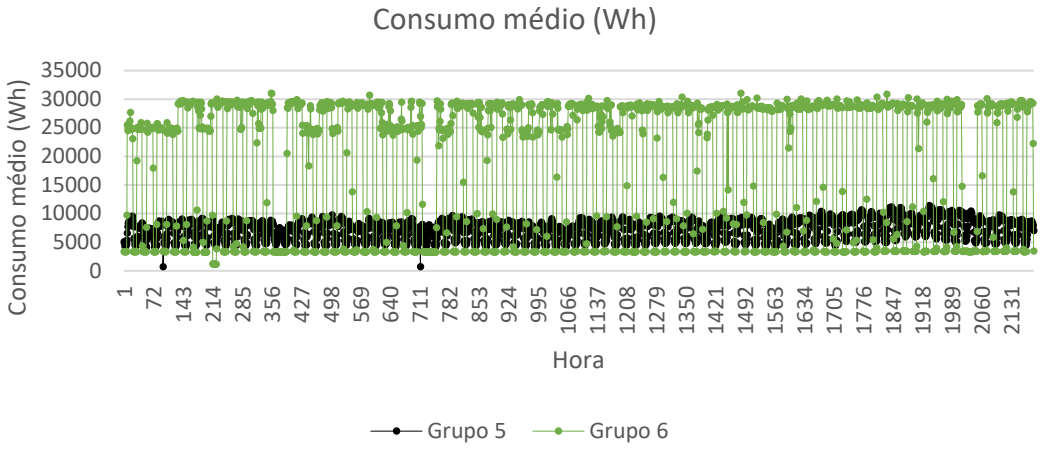

Figura 5.6: Consumo médio (Wh) para os grupos 5 e 6 no segundo trimestre de 2017, pelo método de Ward.

No segundo trimestre as instalações desta PC foram agregadas em seis *clusters*, dos quais apenas o terceiro e quarto grupos apresentados na figura 5.4 são considerados conjuntos de instalações candidatos a ter um consumo anómalo. Mais uma vez, o princípio utilizado para a caracterização dos grupos foi a dimensão de cada *cluster* pois espera-se que para o número considerável de instalações que são abrangidas por este escalão de PC, os grupos com maior dimensão sejam aqueles cujo consumo é tido como tipificado.

De abril a junho, com base neste algoritmo, podem observar-se nas figuras 5.5 e 5.6 os consumos médios horários dos seis grupos construídos. Ao contrário do que foi observado para o trimestre anterior, verifica-se agora uma variação do consumo elétrico médio entre os 0 Wh e os 30000 Wh. Mais uma vez, são considerados *clusters* suspeitos aqueles cuja eletricidade horária consumida, em média, é grande quando verificadas as características da PC em estudo. Os grupos 1, 2, 3 e 4 apresentam um consumo médio mais baixo, na ordem dos 1400 Wh, 3000 Wh, 160 Wh e 700 Wh, respetivamente. Os *clusters* 5 e 6 revelam um consumo médio mais significativo de, aproximadamente, 7000 Wh e 18000 Wh. Ao saber-se também que estes dois têm menor dimensão, serão considerados conjuntos de pontos de entrega com um consumo de energia elétrica possivelmente não tipificado, necessitando assim de uma análise mais detalhada. Esta análise permitirá averiguar até que ponto estes podem, ou não, ser considerados *clusters* com consumo de eletricidade anómalo e será baseada na construção de intervalos de 95% de confiança.

Entre abril e junho de 2017, é importante reter as seguintes ideias:

- Criação de seis grupos;
- Grupos 1, 2, 3 e 4 apresentam consumos médios possivelmente tipificados:
  - Maior número de instalações inseridas nos grupos;
- Grupos 5 e 6 revelam consumos médios possivelmente não tipificados:
  - Menor dimensão, face aos grupos de referência;
  - Consumos máximos atingidos não vão de encontro às características deste escalão de PC.

Os resultados obtidos para o terceiro trimestre de 2017 são apresentados de seguida através das figuras 5.7 a 5.10, que permitirão inferir acerca de quais os grupos com consumo possivelmente não tipificado durante estes meses.

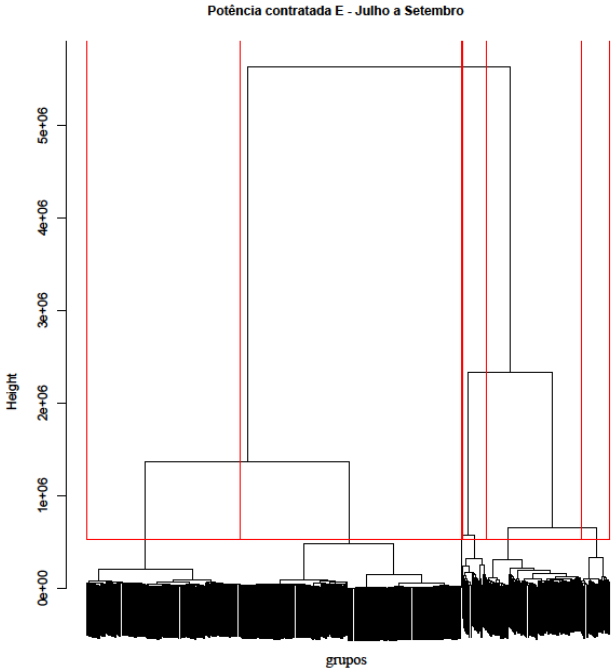

Figura 5.7: Dendrograma associado à PC E, para o terceiro trimestre de 2017, pelo método de Ward.

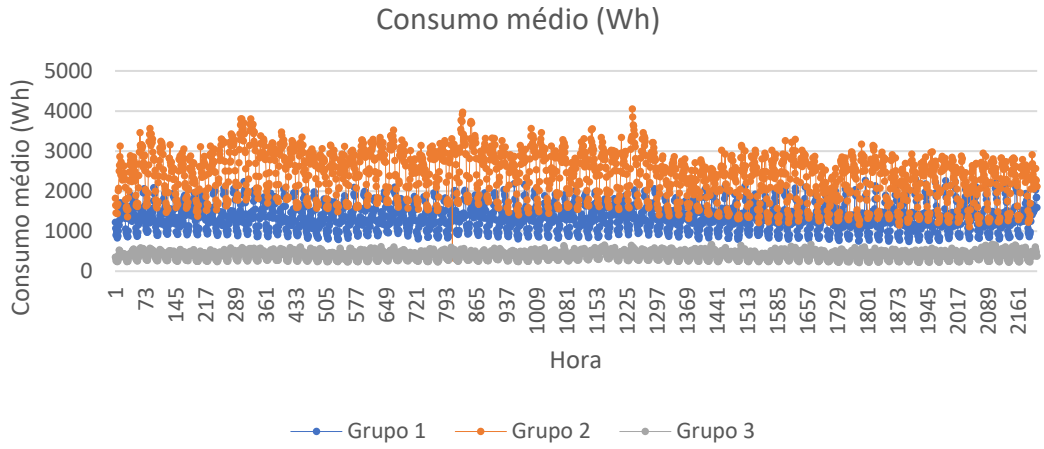

Figura 5.8: Consumo médio (Wh) para os grupos 1, 2 e 3 no terceiro trimestre de 2017, pelo método de Ward.

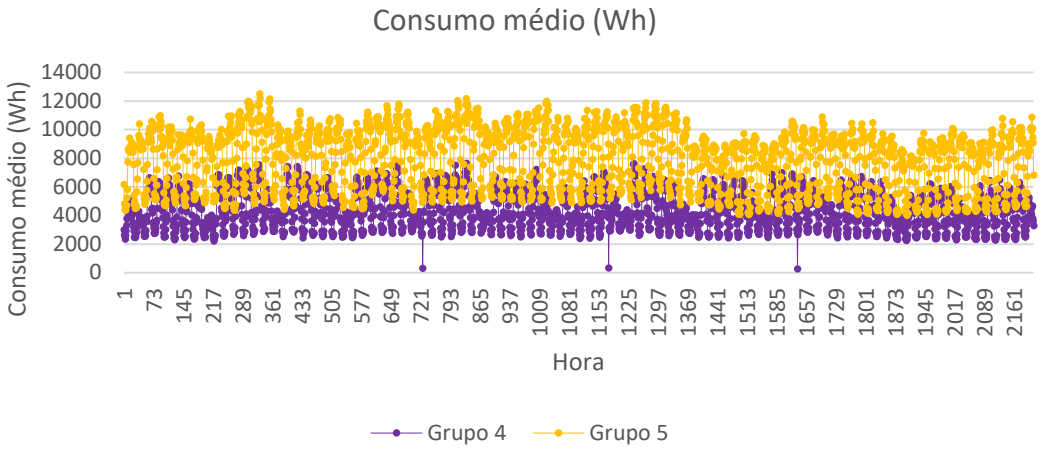

Figura 5.9: Consumo médio (Wh) para os grupos 4 e 5 no terceiro trimestre de 2017, pelo método de Ward.

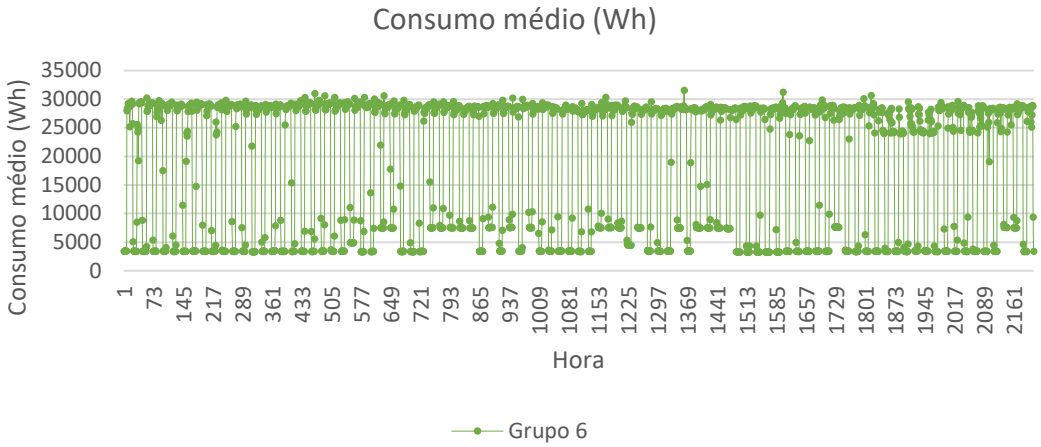

Figura 5.10: Consumo médio (Wh) para o grupo 6 no terceiro trimestre de 2017, aplicando o método de Ward.

De julho a setembro, pela figura 5.7 verifica-se a existência de três grupos com menor dimensão, o terceiro, quarto e sexto, estudados com mais pormenor através dos consumos médios (Wh) para cada *cluster*. A maior parte das instalações pertencentes à PC E apresentam um comportamento diferente no terceiro trimestre de 2017, face aos meses anteriores, relativamente ao consumo médio horário de eletricidade. As figuras 5.8, 5.9 e 5.10 apresentam o consumo médio horário (Wh) dos grupos construídos a partir da aplicação do método de Ward, no terceiro trimestre de 2017. O grupo 3 revela um consumo médio na ordem dos 370 Wh, valor considerado baixo relativamente aos consumos dos restantes grupos, cujo valor de energia elétrica despendida em média é de, aproximadamente, 1200 Wh, 2200 Wh, 4000 Wh, 7350 Wh e 19000 Wh para os grupos 1, 2, 4, 5 e 6, respetivamente. Pelos valores referidos anteriormente pode observar-se que o *cluster* 6 reflete um consumo médio horário de eletricidade muito superior aos restantes conjuntos de instalações. Este grupo tem também uma menor dimensão, a par dos *clusters* 4 e 5, o que leva a que se espere que estes três conjuntos de instalações tenham o seu consumo médio horário de energia elétrica caracterizado como não tipificado. Deste modo, os *clusters* considerados de referência são todos aqueles cujo consumo médio de energia elétrica por hora se podem visualizar na figura 5.8.

Os pontos chave da análise do terceiro trimestre são os seguintes:

- Criação de seis grupos;
- Grupos 1, 2 e 3 apresentam consumos médios de referência:
  - Contêm uma grande parte das instalações com este escalão de PC;
- Grupos 4, 5 e 6 revelam consumos médios suspeitos:
  - Menor dimensão, face aos restantes *clusters*;
  - Consumos máximos atingidos não estão de acordo com as características deste escalão de PC.

Para que a análise de *clusters* pelo método de classificação hierárquica esteja completa, segue abaixo a informação obtida para os últimos três meses de 2017, na qual constam os grupos formados e a explicação de quais destes apresentam características de um consumo não tipificado.

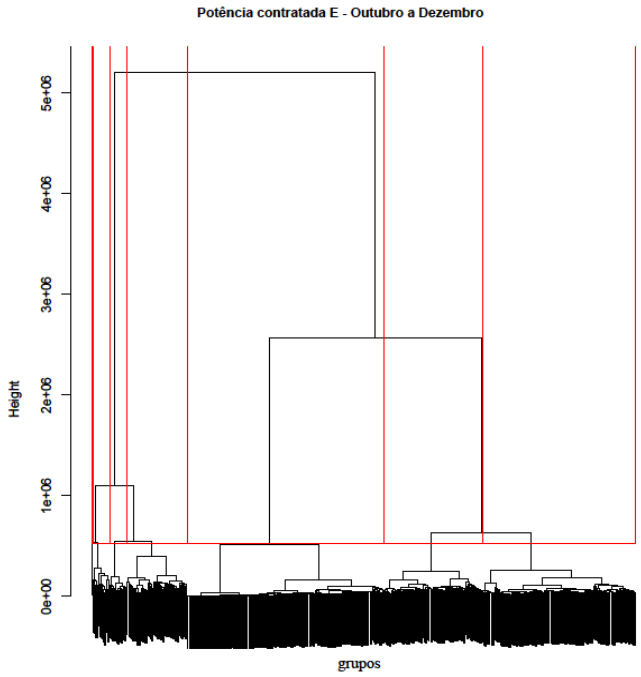

Figura 5.11: Dendrograma associado à PC E, para o quarto trimestre de 2017, pelo método de Ward.

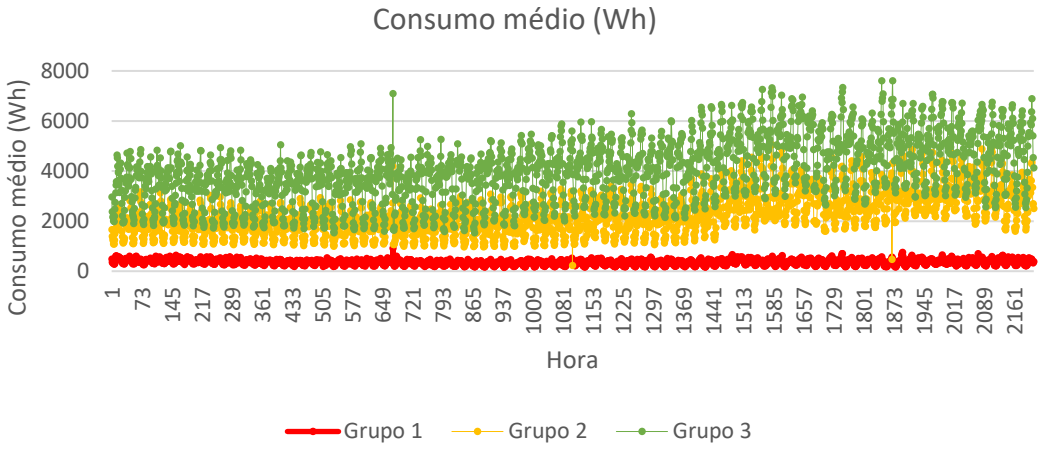

Figura 5.12: Consumo médio (Wh) para os grupos 1, 3 e 4 no quarto trimestre de 2017, pelo método de Ward.

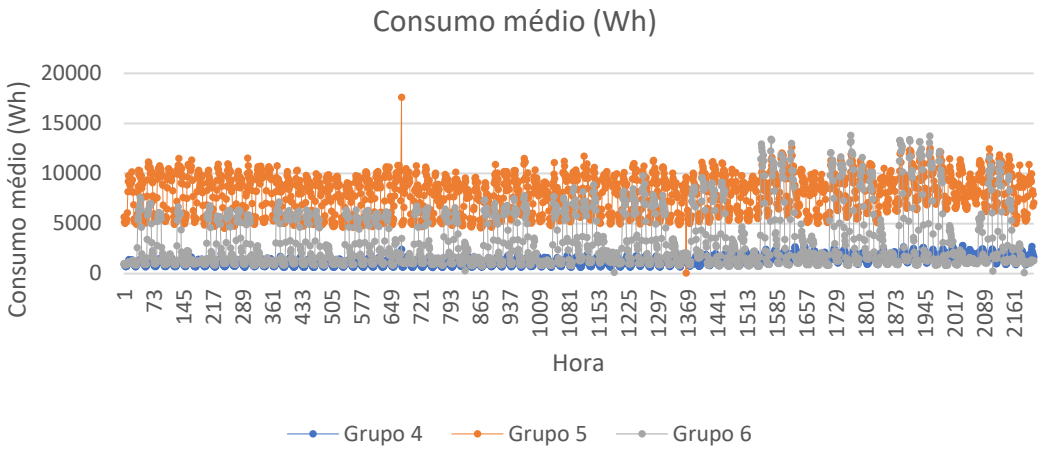

Figura 5.13: Consumo médio (Wh) para os grupos 2, 5 e 6 no quarto trimestre de 2017, pelo método de Ward.

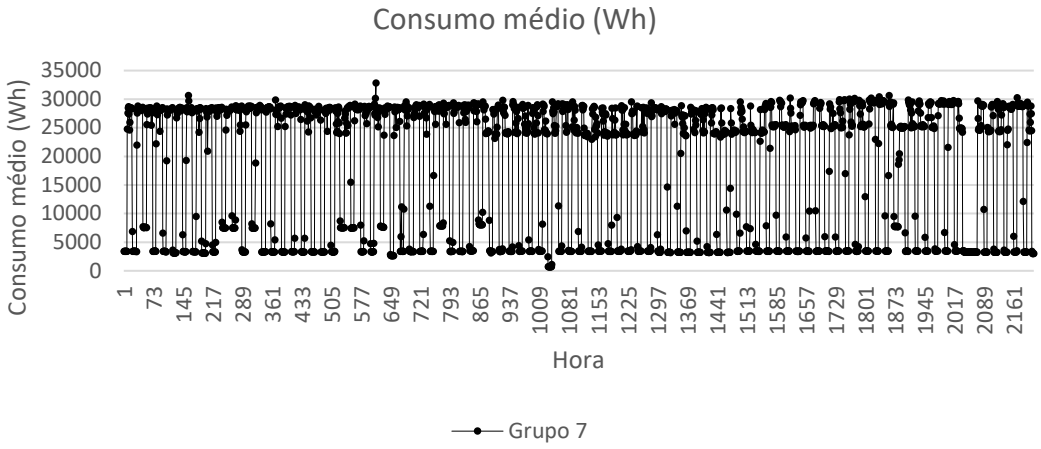

Figura 5.14: Consumo médio (Wh) para o grupo 7 no quarto trimestre de 2017, pelo método de Ward.

Contrariamente aos três trimestres anteriores, para os quais foram criados seis grupos, de outubro a dezembro as instalações pertencentes a este escalão de PC foram agregadas em sete clusters, como se pode observar na figura 5.11. Dos conjuntos de instalações formados, aqueles que contêm menos pontos de entrega são os primeiros três. Estes grupos serão estudados mais aprofundadamente para se averiguar a existência de comportamentos não tipificados relativamente aos seus consumos de energia elétrica no último trimestre de 2017.

Os consumos médios horários de energia elétrica de cada grupo, entre outubro e dezembro de 2017, estão representados graficamente nas figuras 5.12, 5.13 e 5.14. Observa-se através das figuras mencionadas que os *clusters* 2 e 3 revelam um comportamento semelhante, com consumo médio de energia elétrica horário tendencialmente maior, que se pode verificar pelos valores obtidos a partir de novembro. Relativamente aos grupos 1, 2, 3, 4, 5, e 6, o consumo médio de energia elétrica por hora é de cerca de 350 Wh, 2000 Wh, 3900 Wh, 1200 Wh, 7800 Wh e 3100 Wh, respetivamente. Destes, aqueles que contêm um maior número de instalações são os *clusters* 1, 2, 3 e 4, o que leva a que sejam considerados grupos de referência, para efeitos de comparação. O *cluster* 5 revela um comportamento semelhante aos já referidos, nestes três meses em estudo, com a diferença de que apresenta um pico do seu consumo médio horário de quase 18000 Wh, quando em média as instalações pertencentes a este grupo apresentam um consumo de 7800 Wh. Posto isto, o *cluster* 5 é considerado um grupo com consumo médio elétrico candidato a anómalo, facto também suportado pelos valores apresentados na figura 5.13 para este grupo serem bastante elevados, tendo em conta os consumos admitidos na PC E. O grupo 6 apresenta picos de consumo médio que atingem os 13000 Wh no último mês do ano, quando este conjunto de instalações tem consumos na ordem dos 8000 Wh entre outubro e novembro. Deste modo, este grupo apresenta um consumo médio de energia elétrica possivelmente não tipificado. Por fim, o *cluster* 7 é aquele cujas instalações apresentam um consumo médio de energia elétrica que varia entre os 4500 Wh e os 30000 Wh. Pela figura 5.14 pode verificar-se esta variação havendo a ocorrência de um pico de quase 33000 Wh e uma cava de 700 Wh, isto é, o valor máximo observado para o consumo médio deste grupo é de 33000 Wh e o valor mínimo de 700 Wh. Como este *cluster* tem menor dimensão e apresenta um consumo médio horário bastante elevado, é também considerado candidato a anómalo.

De acordo com os resultados obtidos de outubro a dezembro, segue o resumo das conclusões retiradas:

- Formação de sete *clusters*;
- Grupos 1, 2, 3 e 4 apresentam consumos médios padrão:
  - Contêm a maior parte das instalações pertencentes a esta PC;
- Grupos 5, 6 e 7 revelam consumos médios suspeitos:
  - Menor dimensão, face aos restantes *clusters*;
  - Consumos máximos atingidos não estão de acordo com as características deste escalão de PC.

Tal como referido aquando da análise dos resultados obtidos para o primeiro trimestre de 2017, na tabela 5.1 pode visualizar-se como estão agora distribuídas as instalações pertencentes aos grupos suspeitos de janeiro a março – *clusters* 5 e 6.

Tabela 5.1: Percentagem de distribuição dos grupos suspeitos no primeiro trimestre pelos restantes grupos ao longo do ano (Ward).

	Segundo Trimestre	**Terceiro Trimestre**	**Quarto Trimestre**
Grupo 5	Padrão: 1, 2, 3, 4 (100%)	Padrão: 1, 2, 3 (72.7%) Suspeitos: 4 (27.3%)	Padrão: 1, 2, 3, 4 (60.6%) Suspeitos: 6 (39.4%)
Grupo 6	Padrão: 1, 2 (25%) Suspeitos: 5, 6 (75%)	Padrão: 3 (25%) Suspeitos: 4, 5, 6 (75%)	Padrão: 1, 3 (25%) Suspeitos: 5, 7 (75%)

Para as instalações pertencentes ao grupo 5, pode verificar-se de acordo com os valores relativos apresentados que nos restantes meses do ano estas estão distribuídas maioritariamente por *clusters* com consumo de referência. Relativamente ao grupo 6, as respetivas instalações apresentam-se divididas em dois grupos padrão e dois suspeitos no segundo e quarto trimestres de 2017, contrariamente ao terceiro trimestre, no qual estes pontos de medida parecem ser maioritariamente suspeitos. Contudo, a percentagem de instalações pertencentes ao grupo 6 que se encontram distribuídas por grupos padrão nos restantes trimestres é sempre de 25%. Para se efetuar uma melhor inferência face aos *clusters* considerados suspeitos de janeiro a março efetuar-se-á de seguida uma análise detalhada com base em intervalos de 95% de confiança.

Dada a análise anterior, são construídos de seguida intervalos de confiança para o valor médio dos consumos padrão com o objetivo de verificar se, para um nível de 95% de confiança, os consumos médios quartos-horário para os grupos candidatos a anómalos podem ser considerados consumos padrão. Esta análise foi mencionada anteriormente, aquando da explicação dos resultados obtidos em cada trimestre do ano e serve para se averiguar se, estatisticamente, os grupos caracterizados como suspeitos devem de facto ser, ou não, considerados como tal. Não serão apresentadas as representações gráficas correspondentes às bandas de confiança calculadas para os grupos de referência, visto que existem entre 8637 e 8833 quartos de hora ao longo dos trimestres e a sua visualização não seria possível devido à grande escala dos eixos das figuras. Para que um grupo suspeito seja classificado, estatisticamente, como um grupo de referência, espera-se que num intervalo de 95% de confiança pelo menos 95% das suas observações estejam contidas nas bandas de confiança. Seguem abaixo as tabelas com a percentagem de quartos de hora contidos nos IC, para cada grupo candidato a anómalo e em cada trimestre de 2017. Assim, poder-se-á inferir quais os grupos cujo consumo considerado não

tipificado mais se aproxima de um consumo padrão, ao longo do ano em estudo. É de notar que as instalações contidas no grupo 1 no primeiro trimestre de 2017 podem não ser as mesmas que estão associadas ao mesmo grupo, nos restantes meses do ano e o mesmo se aplica aos restantes grupos resultantes deste método. Esta análise serve para complementar as conclusões retiradas das figuras acima (figuras 5.1 a 5.14), do presente capítulo, nomeadamente se os grupos apontados como suspeitos apresentam consumos efetivamente anómalos.

Tabela 5.2: Percentagem de períodos anómalos contidos nos IC dos grupos padrão do primeiro trimestre (Ward).

	Grupos suspeitos	
Grupos padrão	**Grupo 5**	**Grupo 6**
Grupo 1	3.9%	0.0%
Grupo 2	9.8%	0.0%
Grupo 3	18.6%	0.0%
Grupo 4	0.0%	0.0%

Na tabela 5.2 estão apresentados os resultados correspondentes à quantidade, em termos relativos, de consumos médios quartos-horário de cada grupo suspeito contidos nos IC construídos para cada grupo de referência, no primeiro trimestre de 2017. Neste trimestre, os grupos candidatos a anómalos são o 5 e o 6, sendo os restantes quatro considerados *clusters* com consumo médio padrão. Relativamente ao grupo 6, de acordo com a tabela 5.2, nenhum dos seus consumos se insere nos intervalos de 95% de confiança construídos para cada *cluster* de referência, apresentando assim o valor de 0.0% de consumos contidos em cada IC. Por outro lado, o grupo 5 tem 3.9%, 9.8%, 18.6% e 0.0% dos seus consumos médios quartos-horário contidos nos IC dos grupos 1, 2, 3 e 4, respetivamente. Observando os valores referidos, conclui-se que estes dois conjuntos de instalações são de facto *clusters* com consumos médios anómalos, tendo em conta que em nenhum dos IC estão contidos pelo menos 95% dos consumos médios de energia elétrica correspondentes aos períodos de 15 minutos.

Tabela 5.3: Percentagem de períodos anómalos contidos nos IC dos grupos padrão do segundo trimestre (Ward).

	Grupos suspeitos	
Grupos padrão	**Grupo 5**	**Grupo 6**
Grupo 1	0.0%	0.1%
Grupo 2	0.0%	6.9%
Grupo 3	0.0%	0.0%
Grupo 4	0.0%	0.1%

A tabela 5.3 diz respeito à percentagem de períodos de 15 minutos anómalos contidos nos intervalos de 95% de confiança correspondentes aos *clusters* de referência, entre abril e junho de 2017. Os grupos 1, 2, 3 e 4 têm consumo médio padrão, sendo os restantes dois candidatos a anómalos. Para o grupo 5, pode observar-se que a percentagem de consumos médios quartos-horário inseridos nos IC de cada *cluster* de referência é 0.0%, contrariamente ao grupo 6 que, apesar de apresentar percentagens baixas, apenas para o IC construído com os dados do *cluster* 3 é que nenhum dos consumos quartos-horário é abrangido. Observando os restantes valores, a percentagem de períodos anómalos contidos nos IC dos grupos 1 e 4 é de 0.1% e para o grupo 2 este valor aumenta, chegando aos 6.9%. Mais uma vez pode concluir-se, pela verificação da tabela 5.3, que os *clusters* 5 e 6 são efetivamente grupos com consumo médio elétrico anómalo, tendo em conta que as percentagens de consumos quartos-horário contidos nos IC construídos são muito baixas para o nível de confiança escolhido.

Tabela 5.4: Percentagem de períodos anómalos contidos nos IC dos grupos padrão do terceiro trimestre (Ward).

	Grupos suspeitos		
Grupos padrão	**Grupo 4**	**Grupo 5**	**Grupo 6**
Grupo 1	0.0%	0.0%	0.0%
Grupo 2	1.7%	0.0%	1.1%
Grupo 3	0.0%	0.0%	0.0%

Na tabela 5.4 estão representados os valores associados ao número de consumos médios quartos-horário dos grupos suspeitos, em termos relativos, contidos nos IC construídos para os *clusters* de referência, no terceiro trimestre de 2017. Mais uma vez, pode observar-se que as percentagens de períodos de 15 minutos anómalos contidos nestes IC continuam a ser baixas, sendo que os intervalos associados aos *clusters* de referência 1 e 3 não contêm nenhum dos consumos médios quartos-horário dos grupos suspeitos. Por sua vez, os IC construídos para o *cluster* 2 contêm cerca de 1.7% dos períodos anómalos do grupo 4 e 1.1% dos períodos do grupo 6. Em suma, pode afirmar-se que os *clusters* 4, 5 e 6 são efetivamente conjuntos cujas instalações apresentam um consumo médio de energia elétrica anómalo.

Tabela 5.5: Percentagem de períodos anómalos contidos nos IC dos grupos padrão do quarto trimestre (Ward).

	Grupos suspeitos		
Grupos padrão	**Grupo 5**	**Grupo 6**	**Grupo 7**
Grupo 1	0.0%	0.0%	0.0%
Grupo 2	0.0%	20.4%	4.2%
Grupo 3	1.4%	5.7%	15.8%
Grupo 4	0.0%	26.1%	0.3%

Foram criados, para os últimos três meses de 2017, IC correspondentes ao consumo médio dos grupos padrão, 1, 2, 3 e 4, para que se conclua se os restantes conjuntos de instalações podem, ou não, ser considerados também grupos com consumo de referência. Estes resultados estão resumidos na tabela 5.5, onde se pode verificar que a percentagem de períodos de 15 minutos anómalos de cada grupo suspeito inseridos nos IC associados ao *cluster* 1 é 0.0%. Relativamente aos intervalos de 95% de confiança associados ao consumo médio do grupo 2, estes contêm 0.0%, 20.4% e 4.2% dos períodos anómalos dos *clusters* 5, 6 e 7, respetivamente. Seguindo o mesmo raciocínio, verifica-se que 1.4%, 5.7% e 15.8% dos consumos médios quartos-horário correspondentes aos grupos 5, 6 e 7, respetivamente, estão contidos nos IC associados ao grupo 3. Por fim, observa-se que 26.1% e 0.3% dos períodos anómalos quartos-horário dos conjuntos de instalações 6 e 7, respetivamente, estão contidos nos IC construídos com base nos dados do *cluster* 4. Os valores apresentados na tabela 5.5 são muito baixos para que qualquer um dos conjuntos de instalações suspeitas seja considerado um grupo de referência.

Com base nos resultados apresentados nas tabelas 5.2 a 5.5, conclui-se assim que, todos os grupos criados através da aplicação do método de Ward e considerados *clusters* com consumo possivelmente não tipificado são, estatisticamente, conjuntos cujas instalações têm consumos médios de energia elétrica anómalos.

Segue abaixo a tabela resumo com a caracterização de todos os grupos, de acordo com o trimestre e a PC em estudo, pela aplicação do método de Ward.

Tabela 5.6: Caracterização de todos os grupos, para cada PC, pelo método de Ward.

	Trimestre			
PC	**Primeiro**	**Segundo**	**Terceiro**	**Quarto**
A	Padrão: 1, 2 Suspeitos: 3, 4, 5	Padrão: 1, 3, 4 Suspeitos: 2, 5, 6	Padrão: 1, 2 Suspeitos: 3, 4, 5, 6	Padrão: 1, 2, 3, 4 Suspeitos: 5, 6
B	Padrão: 1, 2, 3, 4 Suspeitos: 5	Padrão: 1, 2, 4, 5 Suspeitos: 3, 6, 7	Padrão: 1, 3, 4, 5 Suspeitos: 2, 6, 7	Padrão: 1, 3, 4, 5 Suspeitos: 2, 6, 7
C	Padrão: 1, 2 Suspeitos: 3, 4, 5	Padrão: 2, 3 Suspeitos: 1, 4, 5, 7	Padrão: 1, 2 Suspeitos: 3, 4, 5	Padrão: 1, 2 Suspeitos: 3, 4, 5
D	Padrão: 1, 2 Suspeitos: 3, 4, 5	Padrão: 1, 2 Suspeitos: 3	Padrão: 1, 2 Suspeitos: 3, 4	Padrão: 1, 2 Suspeitos: 3, 4, 5
F	Padrão: 1, 2, 3, 4 Suspeitos: 5, 6	Padrão: 1, 2, 3, 5 Suspeitos: 4, 6	Padrão: 1, 2, 3, 4 Suspeitos: 5, 6	Padrão: 1, 2, 4 Suspeitos: 3, 5, 6
G	Padrão: 1, 2, 3 Suspeitos: 4, 5, 6	Padrão: 1, 2, 3 Suspeitos: 4, 5, 6	Padrão: 1, 2, 4, 5 Suspeitos: 3, 6	Padrão: 1, 2, 3, 5 Suspeitos: 4, 6
H	Padrão: 1, 2, 4 Suspeitos: 3, 5, 6	Padrão: 1, 2, 3, 4 Suspeitos: 5, 6	Padrão: 1, 3, 4, 5 Suspeitos: 2, 6	Padrão: 1, 5 Suspeitos: 2, 3, 4, 6
J	Padrão: 1, 2, 3, 5 Suspeitos: 4, 6	Padrão: 1, 2, 3 Suspeitos: 4, 5, 6	Padrão: 1, 2, 3 Suspeitos: 4, 5	Padrão: 1, 3 Suspeitos: 2, 4, 5
L	Padrão: 1, 3 Suspeitos: 2, 4, 5, 6, 7	Padrão: 1, 4 Suspeitos: 2, 3, 5, 6, 7	Padrão: 1, 4 Suspeitos: 2, 3, 5, 6, 7	Padrão: 1, 2, 4 Suspeitos: 3, 5, 6, 7
M	Padrão: 1, 3 Suspeitos: 2, 4, 5, 6	Padrão: 1, 2 Suspeitos: 3, 4, 5, 6	Padrão: 1, 2 Suspeitos: 3, 4, 5, 6, 7, 8	Padrão: 1, 2 Suspeitos: 3, 4, 5, 6
N	Padrão: 1, 2 Suspeitos: 3, 4, 5, 6, 7	Padrão: 1, 2 Suspeitos: 3, 4, 5, 6, 7	Padrão: 2, 3 Suspeitos: 1, 4, 5, 6, 7	Padrão: 2, 5 Suspeitos: 1, 3, 4, 6

5.2. Resultados pelo método de *k*-means

O método de classificação não hierárquica aqui apresentado, *k*-means, baseia os seus resultados no cálculo da distância euclidiana para encontrar a melhor solução de agrupamento das instalações em conjuntos que permitam ter a menor distância entre os seus objetos e os respetivos centroides. Para existir termo de comparação entre este procedimento e o método de Ward, criaram-se os *clusters* para todas as potências contratadas, em cada trimestre de 2017, com base no número de grupos resultantes da aplicação do método de Ward. Os anexos B.1, B.2 e B.3 apresentam os consumos médios horários, para todos os grupos das PC A, B e J, respetivamente. Tendo sido o processo de agrupamento análogo para as restantes PC, não se encontram presentes neste relatório as respetivas representações gráficas, mas podem encontrar-se os resultados obtidos pela aplicação do algoritmo na tabela 5.11.

O algoritmo aplicado para a PC E permitiu a criação de seis *clusters* nos primeiros três trimestres de 2017 e de sete grupos nos restantes meses do ano referido, tal como no método apresentado anteriormente. As figuras abaixo apresentam esses resultados, nas quais se pode verificar o comportamento de cada grupo em cada trimestre, de acordo com o seu consumo médio de energia elétrica.

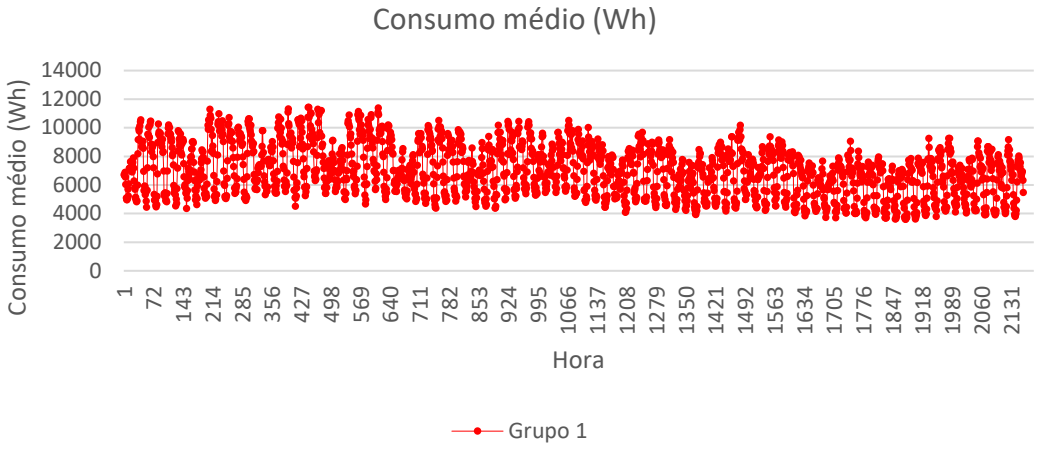

Figura 5.15: Consumo médio (Wh) para o grupo 1 no primeiro trimestre de 2017, aplicando o método de *k*-means.

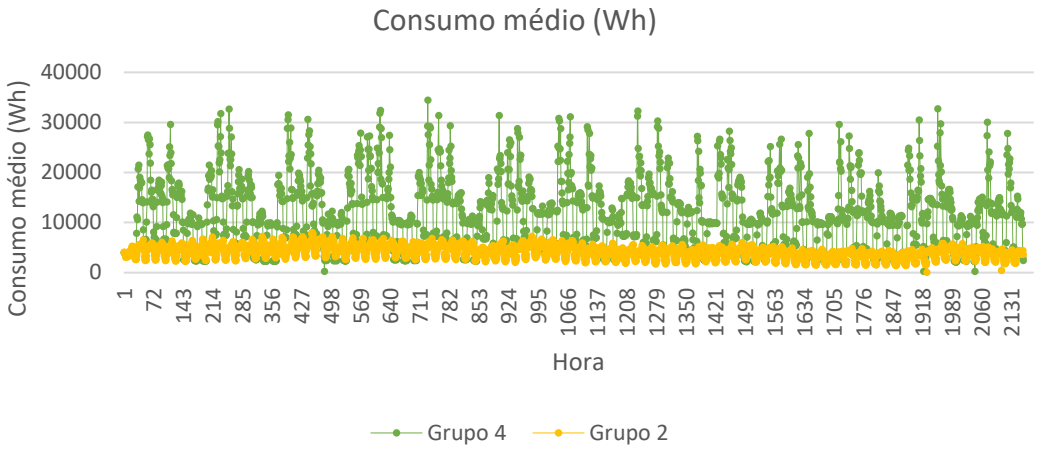

Figura 5.16: Consumo médio (Wh) para os grupos 2 e 4 no primeiro trimestre de 2017, pelo método de *k*-means.

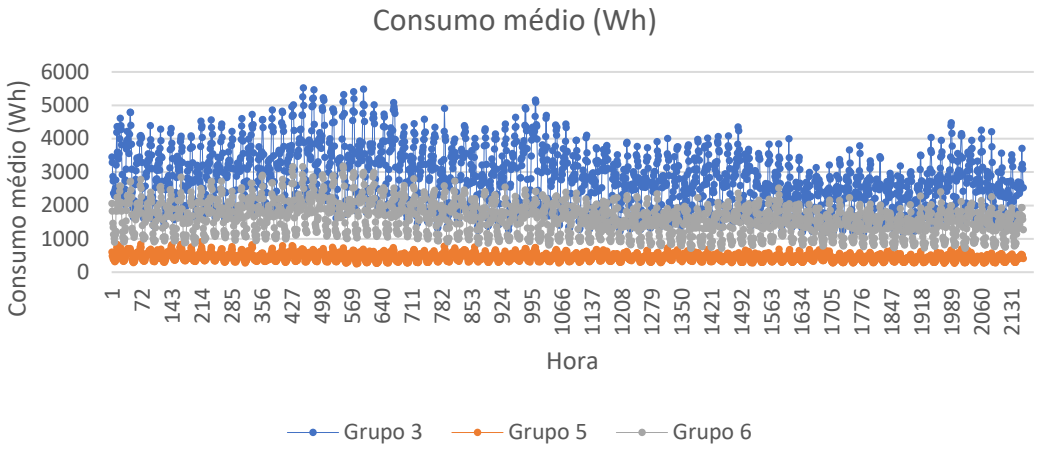

Figura 5.17: Consumo médio (Wh) para os grupos 3, 5 e 6 no primeiro trimestre de 2017, pelo método de *k*-means.

As figuras 5.15 a 5.17 indicam os grupos resultantes da aplicação do método de *k*-means para o primeiro trimestre de 2017. Foram criados três gráficos porque a colocação de todas as instalações numa só figura iria impedir a clara visualização dos consumos médios associados devido às diferentes escalas. Nas figuras 5.15 a 5.17 pode observar-se que os *clusters* com maior consumo médio de energia elétrica são o 1 e o 4, com valores na ordem dos 10000 Wh e 30000 Wh, respetivamente. Dado

que estes consumos são muito elevados para as características deste escalão de PC e que têm uma dimensão pequena, são considerados como grupos candidatos a anómalos neste trimestre. Os conjuntos de instalações 2, 3, 5 e 6 são grupos com consumo médio horário padrão, com valores na ordem dos 5000 Wh, 4000 Wh, 500 Wh e 2000 Wh, respetivamente. Pela visualização dos consumos médios referentes às instalações dos grupos de referência, pode concluir-se que o grupo 3 é aquele com comportamento menos consistente ao longo dos primeiros três meses de 2017 pois o seu consumo oscila bastante entre os 0 Wh e os 5500 Wh.

Com a aplicação do método de *k*-means para o primeiro trimestre de 2017, importa reter o seguinte:

- Formação de seis *clusters*;
- Grupos 2, 3, 5 e 6 com consumos médios padrão:
  - Grande dimensão;
  - Consumo médio horário enquadra-se neste escalão de PC;
- Grupos 1 e 4 com consumos médios suspeitos:
  - Menor dimensão, face aos restantes grupos;
  - Consumos médios não estão de acordo com as características deste escalão de PC.

Para que se consiga continuar a inferir sobre o comportamento destas instalações, indicam-se abaixo os resultados obtidos para a PC E com base nos dados de abril a junho. Caracterizar-se-ão então os grupos de acordo com os seus consumos médios diários.

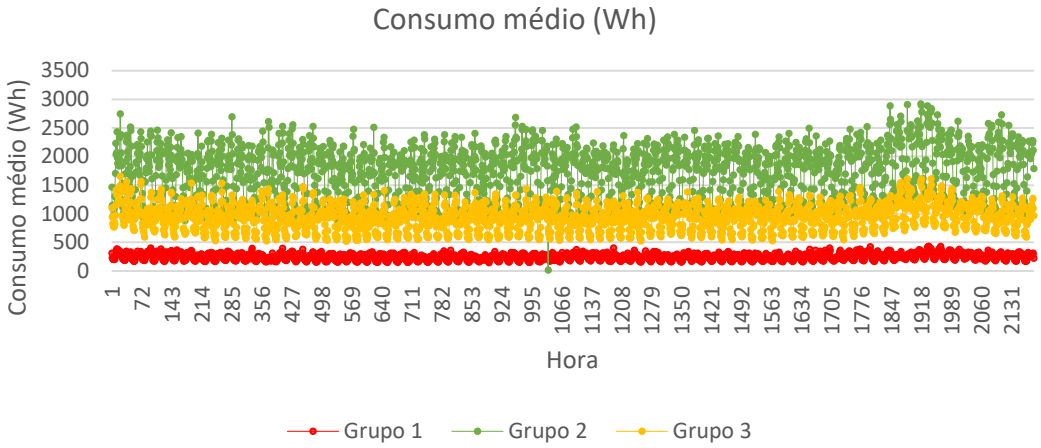

Figura 5.18 - Consumo médio (Wh) para os grupos 1, 2 e 3 no segundo trimestre de 2017, pelo método de *k*-means.

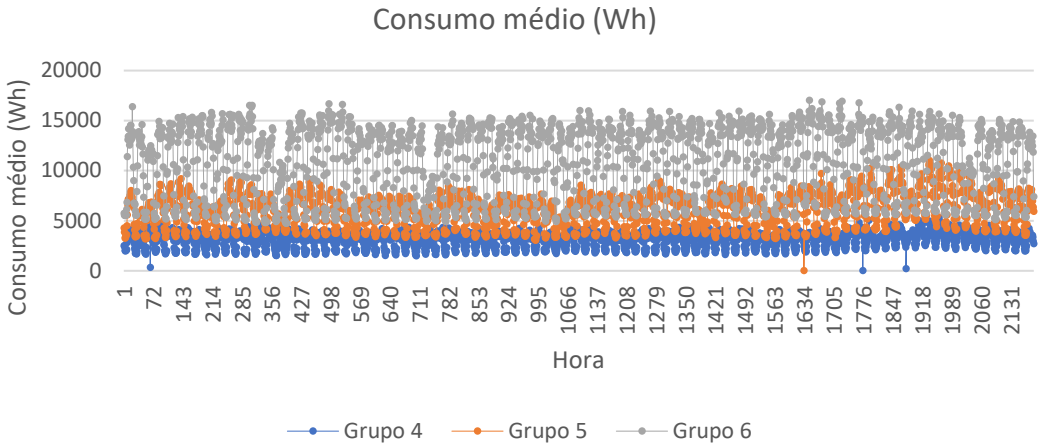

Figura 5.19 - Consumo médio (Wh) para os grupos 4, 5 e 6 no segundo trimestre de 2017, pelo método de *k*-means.

As figuras 5.18 e 5.19 refletem os consumos médios horários de energia dos seis grupos criados a partir da aplicação do método de classificação não hierárquica no segundo trimestre de 2017. Os *clusters* 1, 2, 3 e 4 são aqueles que revelam um menor consumo médio, com valores na ordem dos 250 Wh, 1000 Wh, 2250 Wh e 3000 Wh, respetivamente. Estes *clusters* contêm um elevado número de instalações e o consumo médio observado não parece suspeito, de acordo com as características da PC E, o que leva a que sejam escolhidos como grupos padrão. Por outro lado, os grupos 5 e 6 têm menor dimensão e exibem um consumo médio de eletricidade entre os 7000 Wh e os 14000 Wh, respetivamente. Apesar de estes dois grupos apresentarem um comportamento consistente ao longo dos meses de abril a junho, os valores dos seus consumos não estão de acordo com as características deste escalão de PC. Deste modo, serão considerados grupos com consumo médio considerado possivelmente não tipificado, sendo posteriormente analisados com mais detalhe através de intervalos de 95% de confiança.

Neste trimestre as conclusões associadas aos resultados obtidos foram as seguintes

- Formação de seis *clusters*;
- Grupos 1, 2, 3 e 4 apresentam consumos médios padrão:
  - Grande dimensão;
- Grupos 5 e 6 revelam consumos médios suspeitos:
  - Pequena dimensão quando comparados com os restantes grupos;
  - Consumos médios ultrapassam o que é previso para este escalão de PC;

De forma a averiguar a presença de instalações com consumo possivelmente não tipificado de julho a setembro, seguem os resultados obtidos para estes meses, nos quais foram formados seis grupos de instalações.

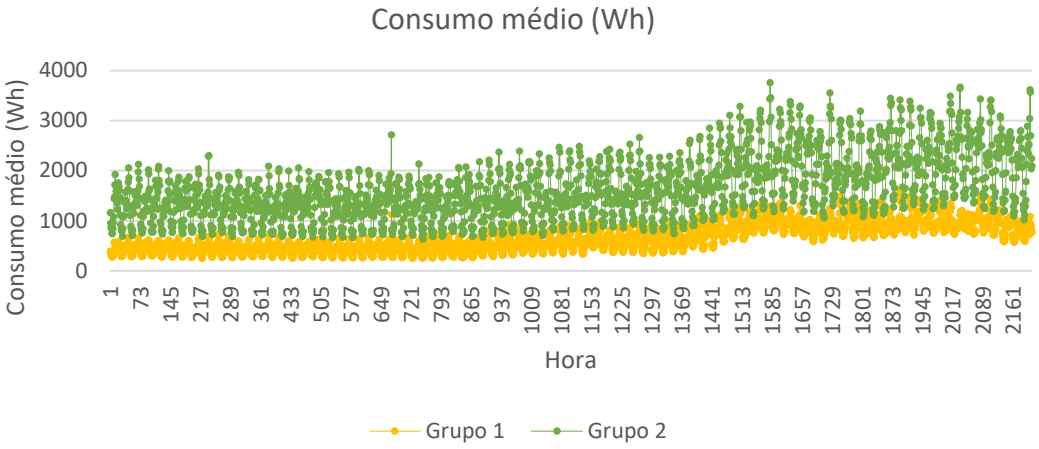

Figura 5.20 - Consumo médio (Wh) para os grupos 1 e 2 no terceiro trimestre de 2017, pelo método de *k*-means.

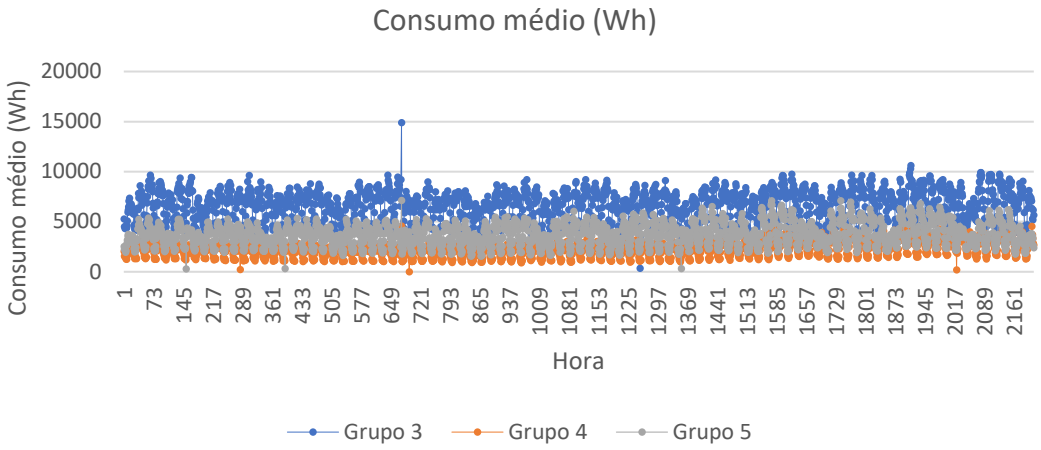

Figura 5.21: Consumo médio (Wh) para os grupos 3, 4 e 5 no terceiro trimestre de 2017, pelo método de *k*-means.

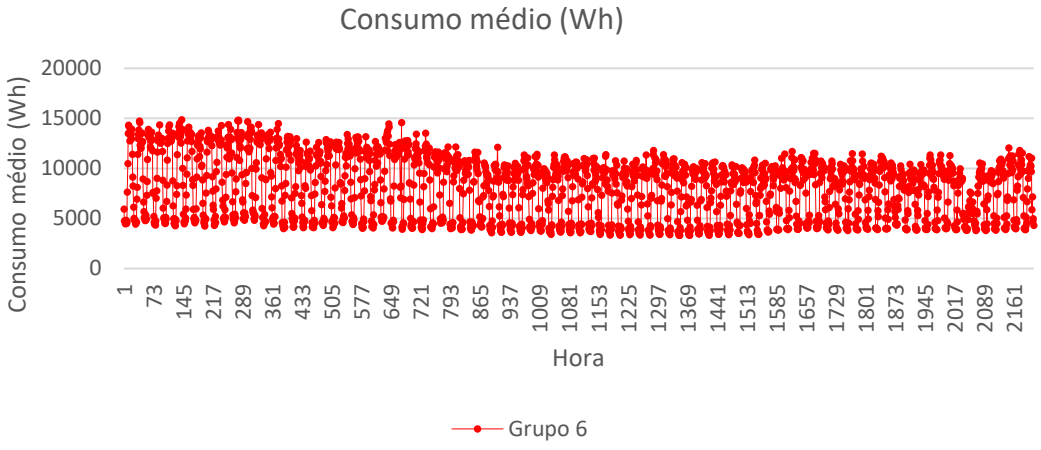

Figura 5.22: Consumo médio (Wh) para o grupo 6 no terceiro trimestre de 2017, aplicando o método de *k*-means.

As três figuras anteriores, 5.20 a 5.22, resultam da aplicação do método de *k*-means aos consumos de julho a setembro de 2017, através do qual se criaram seis grupos de instalações. Como foi referido anteriormente, os grupos com menor dimensão e cujas instalações apresentam um consumo médio horário suspeito para o tipo de PC em análise, são *clusters* candidatos a anómalos. Neste trimestre, os

clusters 3 e 6 são aqueles que revelam as características referidas, com consumos médios de eletricidade na ordem dos 8000 Wh e dos 10000 Wh, respetivamente, podendo ainda observar-se que o grupo 3 apresenta um pico de eletricidade média que ultrapassa os 14000 Wh. Por outro lado, os grupos 1, 2, 4 e 5 são aqueles cujo consumo médio de eletricidade é considerado padrão. Os *clusters* 1 e 2 revelam um comportamento semelhante quanto à tendência dos seus consumos, com a diferença de que o consumo médio de energia elétrica é de 500 Wh, aproximadamente, para o grupo 1 e para o grupo 2 este valor atinge o triplo do que é apresentado para o *cluster* 1. Por sua vez, o *cluster* 4 também parece ter um consumo médio de energia elétrica na ordem dos 2000 Wh e, por fim, o grupo 5 revela consumos na ordem dos 4000 Wh.

A aplicação do método de *k*-means para os consumos associados aos meses de julho a setembro, permitiu concluir o seguinte:

- Formação de seis *clusters*;
- Grupos 1, 2, 4 e 5 têm consumos médios padrão:
  - Grande dimensão;
- Grupos 3 e 6 apresentam consumos médios suspeitos:
  - Menor dimensão quando comparados com os restantes grupos;
  - Consumos médios ultrapassam o que é previso para este escalão de PC;

De seguida são apresentados os resultados obtidos com a aplicação do método não hierárquico, *k*-means, nos últimos três meses de 2017. Neste trimestre formaram-se sete grupos, cujo comportamento dos seus consumos se pode verificar de seguida.

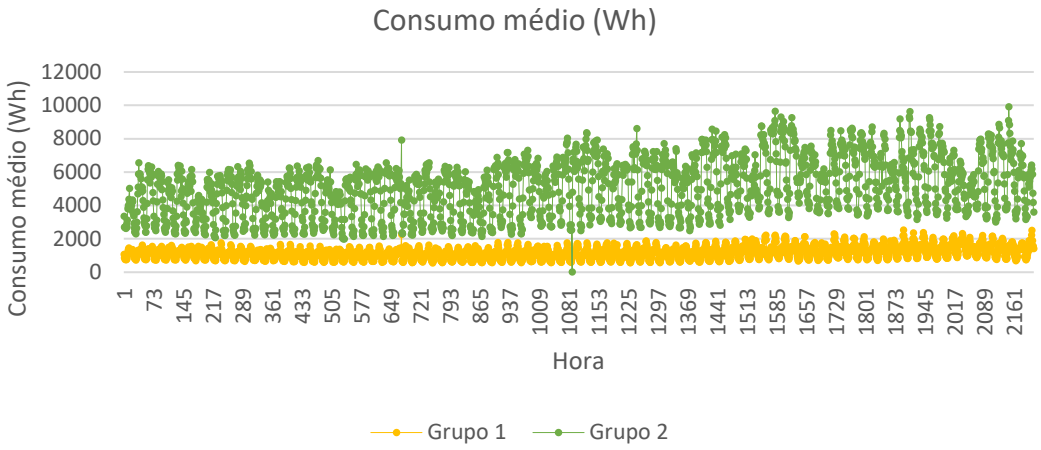

Figura 5.23: Consumo médio (Wh) para os grupos 1 e 2 no quarto trimestre de 2017, pelo método de *k*-means.

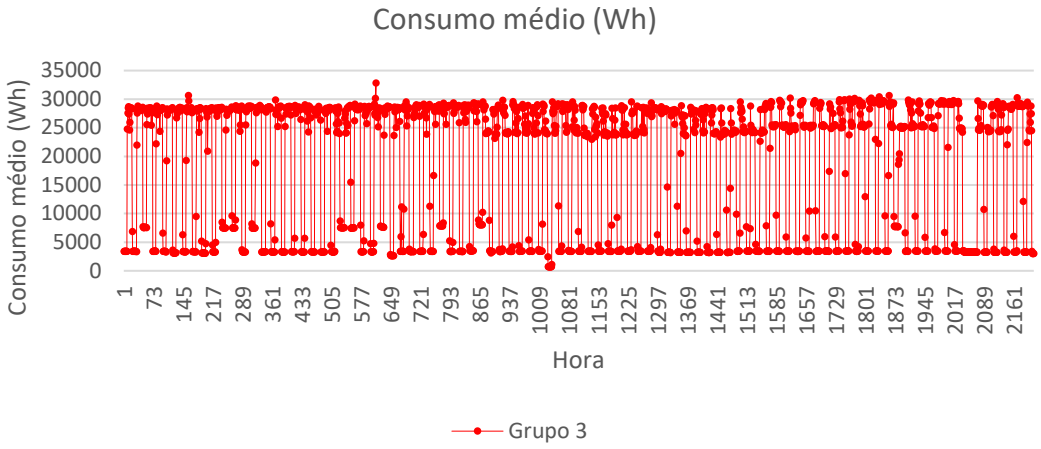

Figura 5.24: Consumo médio (Wh) para o grupo 3 no quarto trimestre de 2017, aplicando o método de *k*-means.

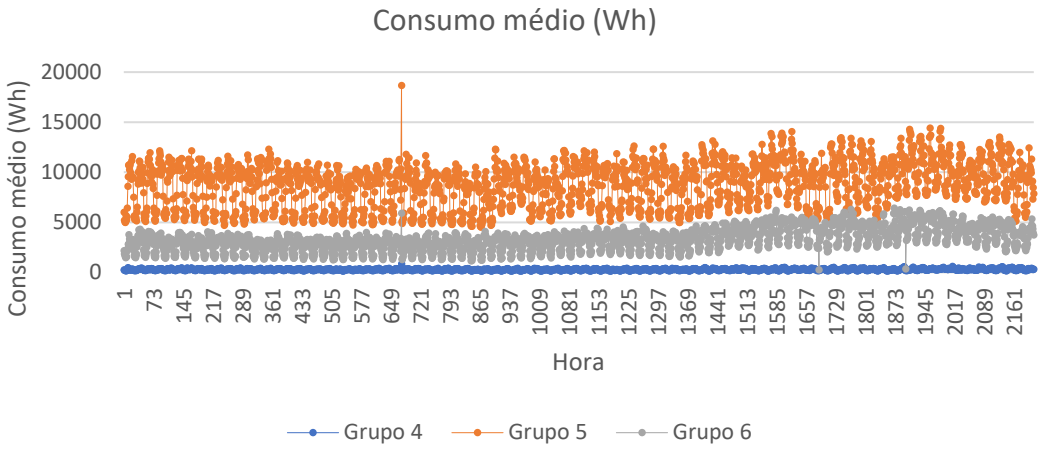

Figura 5.25: Consumo médio (Wh) para os grupos 4, 5 e 6 no quarto trimestre de 2017, pelo método de *k*-means.

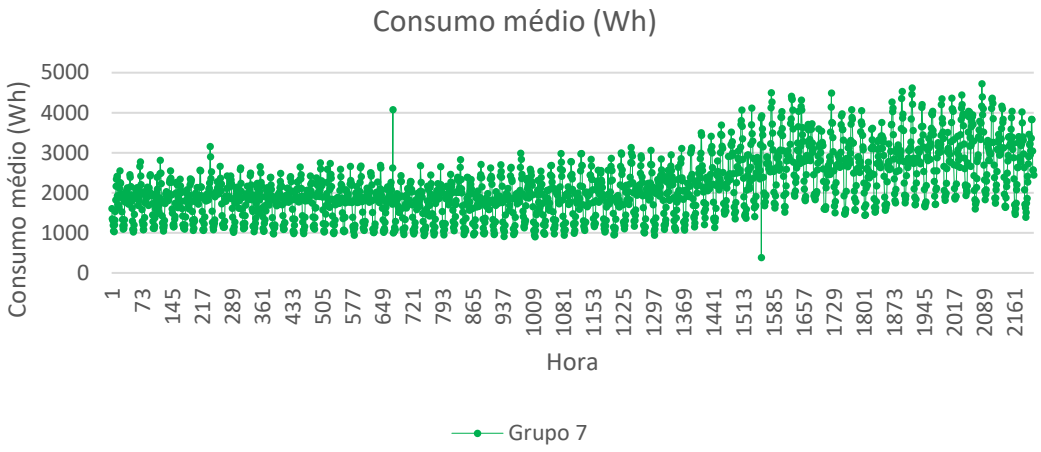

Figura 5.26: Consumo médio (Wh) para o grupo 7 no quarto trimestre de 2017, aplicando o método de *k*-means.

Para o último trimestre de 2017 foram construídos sete grupos cujos consumos médios de energia elétrica estão refletidos nas figuras 5.23 a 5.26. Tal como nos resultados comentados para os três primeiros trimestres de 2017, os grupos com menor dimensão e cujo consumo horário médio de energia elétrica não se enquadra nas características da PC E serão tidos como grupos candidatos a

anómalos. Assim, os *clusters* que se encontram nesta situação são conjuntos de instalações 2, 3 e 5, com consumos médios na ordem dos 6000 Wh, 29000 Wh e 11000 Wh, respetivamente. Estes *clusters* revelam um pico no seu consumo médio horário, que atinge um valor próximo de 9928 Wh, 32856 Wh e 18674 Wh, respetivamente. Por conseguinte, sabe-se que os grupos 1, 4, 6 e 7 são tidos como grupos de referência, com consumos médios de energia elétrica na ordem dos 1000 Wh, 500 Wh, 4000 Wh e 3000 Wh, respetivamente, dos quais o *cluster* 7 apresenta um pico que atinge os 4726 Wh.

Para os resultados obtidos no final do ano 2017, com o método de *k*-means, as ideias mais relevantes são as seguintes:

- Formação de sete *clusters*;
- Grupos 1, 4, 6 e 7 apresentam consumos médios padrão:
  - Grande dimensão;
  - Grupo 7 apresenta um pico no seu consumo médio, que atinge os 4726 Wh;
- Grupos 2, 3 e 5 revelam consumos médios suspeitos:
  - Menor dimensão quando comparados com os restantes grupos;
  - Apresentam um pico no seu consumo, na ordem dos 9928 Wh, 32856 Wh e 18674 Wh, respetivamente,

Analogamente ao método de Ward, segue uma análise aos grupos suspeitos no primeiro trimestre, 1 e 4, para se verificar a distribuição das suas instalações pelos respetivos grupos nos restantes meses de 2017.

Tabela 5.7: Percentagem da distribuição dos grupos suspeitos no primeiro trimestre pelos restantes grupos ao longo do ano (*k*-means).

	Segundo Trimestre	**Terceiro Trimestre**	**Quarto Trimestre**
Grupo 1	Padrão: 2, 3, 4 (44.8%) Suspeitos: 5, 6 (55.2%)	Padrão: 1, 2, 4, 5 (55.2%) Suspeitos: 3, 6 (44.8%)	Padrão: 1, 4, 6, 7 (20.7%) Suspeitos: 2, 5 (79.3%)
Grupo 4	Padrão: 4 (33.3%) Suspeitos: 5, 6 (66.7%)	Padrão: 1, 5 (66.7%) Suspeitos: 6 (33.3%)	Padrão: 1 (33.3%) Suspeitos: 3, 5 (66.7%)

Pela observação da tabela 5.7 pode verificar-se de que modo os grupos 1 e 4, cujos consumos são suspeitos no primeiro trimestre, têm as suas instalações espalhadas pelos *clusters* formados nos restantes nove meses de 2017. O grupo 1 apresenta as suas instalações maioritariamente contidas em grupos suspeitos no segundo e quarto trimestres. Relativamente ao grupo 4, as suas instalações estão divididas de igual forma no segundo e quarto trimestres, sendo no terceiro trimestre que 66.7% das respetivas instalações parecem ter um consumo menos suspeito. Para que melhor se entenda se estes dois *clusters* podem, ou não, ser considerados de referência, a par dos grupos suspeitos nos outros trimestres, far-se-á de seguida uma análise com base em intervalos de 95% de confiança.

Após efetuada a análise exaustiva aos consumos médios horários de cada *cluster* para a PC E, em cada trimestre, segue-se a análise baseada nos intervalos de 95% de confiança para o valor médio dos grupos de referência. As tabelas 5.8 a 5.11 ilustram ainda algumas conclusões visíveis relativamente aos grupos suspeitos. Mais uma vez, é importante mencionar o que foi referido no capítulo anterior, que as instalações constantes no grupo 1 do primeiro trimestre de 2017 podem não ser as mesmas que estão associadas ao grupo 1 dos restantes meses do ano, aplicando-se o mesmo para os restantes *clusters* formados.

Tabela 5.8: Percentagem de períodos anómalos contidos nos IC dos grupos padrão do primeiro trimestre (*k*-means).

	Grupos suspeitos	
Grupos padrão	Grupo 1	Grupo 4
Grupo 2	6.6%	11.6%
Grupo 3	0.0%	6.4%
Grupo 5	0.0%	0.0%
Grupo 6	0.0%	0.5%

Para o primeiro trimestre de 2017, de acordo com o algoritmo de *k*-means, apenas os *clusters* 1 e 4 apresentam consumos médios de energia elétrica suspeitos. Estes são avaliados na tabela 5.8 com base nos IC construídos para os grupos de referência. Relativamente ao *cluster* 1, 6.6% dos seus consumos médios estão contidos nos IC do grupo 2 e 0.0% nos restantes. Para o grupo 4, a percentagem de consumos quartos-horário contidos nos IC construídos é de 11.6%, 6.4%, 0.0% e 0.5% para os intervalos dos grupos 2, 3, 5 e 6, respetivamente. Conclui-se através desta comparação que os *clusters* 1 e 4 são efetivamente conjuntos de instalações com consumo médio não tipificado, sendo considerados estatisticamente grupos anómalos.

Tabela 5.9: Percentagem de períodos anómalos contidos nos IC dos grupos padrão do segundo trimestre (*k*-means).

	Grupos suspeitos	
Grupos padrão	Grupo 5	Grupo 6
Grupo 1	0.0%	0.0%
Grupo 2	0.0%	0.0%
Grupo 3	0.0%	0.0%
Grupo 4	0.5%	0.0%

Para o segundo trimestre de 2017 foram criados intervalos de 95% de confiança associados aos primeiros quatro conjuntos de instalações, tendo em conta que estes foram considerados *clusters* com consumo padrão. Por conseguinte, os *clusters* 5 e 6 são suspeitos e os seus períodos quartos-horário são analisados com base nos IC calculados. Verifica-se pela tabela 5.9, que nenhum dos consumos médios correspondentes aos grupos 5 e 6 se inserem nos IC construídos, à exceção de as instalações associadas ao *cluster* 5 terem 0.5% dos seus consumos médios contidos nos IC do grupo 4. Estes valores são muito baixos para que os *clusters* 5 e 6 sejam considerados conjuntos de instalações com consumo padrão, visto que para serem caracterizados como tal seria necessário que pelo menos 95% dos seus consumos se inserissem em algum dos IC construídos. Assim, os grupos 5 e 6 são estatisticamente anómalos.

Tabela 5.10: Percentagem de períodos anómalos contidos nos IC dos grupos padrão do terceiro trimestre (*k*-means).

	Grupos suspeitos	
Grupos padrão	Grupo 3	Grupo 6
Grupo 1	0.0%	0.0%
Grupo 2	0.0%	0.0%
Grupo 4	0.0%	0.0%
Grupo 5	3.4%	1.5%

A tabela 5.10 indica a percentagem de consumos médios quartos-horário correspondentes às instalações dos *clusters* 3 e 6, contidos nos IC construídos para os grupos padrão. Nenhum dos consumos referidos está contido nos IC associados aos *clusters* 1, 2 e 4. Contudo, o único conjunto de instalações cujo IC abrange consumos médios quartos-horário relativos aos grupos suspeitos é o *cluster* 5, cujo intervalo de 95% de confiança contém 3.4% e 1.5% dos consumos dos grupos 3 e 6, respetivamente. Tendo em conta que nenhum destes intervalos contém pelo menos 95% dos consumos

por períodos de 15 minutos, os grupos 3 e 6 são efetivamente considerados conjuntos de instalações com consumos não tipificados, ou seja, estatisticamente anómalos.

Tabela 5.11: Percentagem de períodos anómalos contidos nos IC dos grupos padrão do quarto trimestre (*k*-means).

	Grupos suspeitos		
Grupos padrão	**Grupo 2**	**Grupo 3**	**Grupo 5**
Grupo 1	0.0%	0.2%	0.0%
Grupo 4	0.0%	0.0%	0.0%
Grupo 6	14.9%	13.4%	0.1%
Grupo 7	0.2%	3.1%	0.0%

Para os últimos três meses de 2017 foram criados sete conjuntos de instalações, sendo os *clusters* 2, 3 e 5 analisados ao pormenor, de modo a verificar se os valores correspondentes à energia elétrica média consumida são de facto anómalos. Deste modo, foram construídos IC para cada grupo com consumo médio padrão – grupos 1, 4, 6 e 7 - para se averiguar se pelo menos 95% dos consumos médios quartos-horário dos grupos suspeitos estão contidos nos IC referidos. De facto, pela observação da tabela 5.11, pode concluir-se que os grupos 2, 3 e 5 têm efetivamente consumos médios não tipificados, tendo em conta que a quantidade de períodos suspeitos contidos nos IC de cada grupo padrão, em termos relativos, é muito inferior a 95%.

Em suma, pela observação dos resultados indicados nas tabelas 5.8 a 5.11, pode afirmar-se que qualquer um dos *clusters* obtidos pela aplicação do método de *k*-means é considerado estatisticamente um conjunto de instalações com consumo anómalo.

Segue abaixo a tabela resumo com a caracterização de todos os grupos, de acordo com o trimestre e a PC em estudo, pela aplicação do algoritmo correspondente ao método de *k*-means.

Tabela 5.12: Caracterização de todos os grupos, para cada PC, pelo método de *k*-means.

	Trimestre			
PC	**Primeiro**	**Segundo**	**Terceiro**	**Quarto**
A	Padrão: 1, 2, 3 Suspeitos: 4, 5	Padrão: 2, 3, 4, 5 Suspeitos: 1, 6	Padrão: 4, 5, 6 Suspeitos: 1, 2, 3	Padrão: 1, 2, 4, 6 Suspeitos: 3, 5
B	Padrão: 1, 5 Suspeitos: 2, 3, 4	Padrão: 1, 2, 3, 5 Suspeitos: 4, 6, 7	Padrão: 2, 4, 6, 7 Suspeitos: 1, 3, 5	Padrão: 2, 4, 5 Suspeitos: 1, 3, 6, 7
C	Padrão: 2, 3 Suspeitos: 1, 4, 5	Padrão: 3, 4, 5 Suspeitos: 1, 2, 6	Padrão: 1, 4, 5 Suspeitos: 2, 3	Padrão: 3, 4 Suspeitos: 1, 2, 5
D	Padrão: 2, 5 Suspeitos: 1, 3, 4	Padrão: 1 Suspeitos: 2, 3	Padrão: 2, 3 Suspeitos: 1, 4	Padrão: 1, 3 Suspeitos: 2, 4, 5
F	Padrão: 1, 4, 5, 6 Suspeitos: 2, 3	Padrão: 1, 2, 4, 5 Suspeitos: 3, 6	Padrão: 1, 5, 6 Suspeitos: 2, 3, 4	Padrão: 1, 2, 3 Suspeitos: 4, 5, 6
G	Padrão: 3, 4, 5, 6 Suspeitos: 1, 2	Padrão: 4, 5, 6 Suspeitos: 1, 2, 3	Padrão: 2, 4, 5 Suspeitos: 1, 3, 6	Padrão: 2, 4, 5 Suspeitos: 1, 3, 6
H	Padrão: 1, 2, 3, 4, 5 Suspeitos: 6	Padrão: 1, 2, 3, 5, 6 Suspeitos: 4	Padrão: 1, 2, 3, 6 Suspeitos: 4, 5	Padrão: 1, 2, 3, 4, 6 Suspeitos: 5
J	Padrão: 1, 4, 6 Suspeitos: 2, 3, 5	Padrão: 2, 3, 4 Suspeitos: 1, 5	Padrão: 2, 3, 5 Suspeitos: 1, 4	Padrão: 2, 4 Suspeitos: 1, 3, 5
L	Padrão: 3, 5 Suspeitos: 1, 2, 4, 6, 7	Padrão: 3, 4, 7 Suspeitos: 1, 2, 5, 6	Padrão: 6, 7 Suspeitos: 1, 2, 3, 4, 5	Padrão: 1, 7 Suspeitos: 2, 3, 4, 5, 6
M	Padrão: 3, 4 Suspeitos: 1, 2, 5, 6	Padrão: 1, 4, 6 Suspeitos: 2, 3, 5	Padrão: 2, 3, 5, 6, 8 Suspeitos: 1, 4, 7	Padrão: 2, 3, 6 Suspeitos: 1, 4, 5
N	Padrão: 1, 2, 4 Suspeitos: 3, 5, 6, 7	Padrão: 2, 3, 6 Suspeitos: 1, 4, 5, 7	Padrão: 4, 6, 7 Suspeitos: 1, 2, 3, 5	Padrão: 3, 6 Suspeitos: 1, 2, 4, 5

5.3. Comparação dos métodos utilizados

Os métodos utilizados na análise de *clusters*, método de Ward e de *k*-means, têm características diferentes, como referido nos capítulos 4.1 e 4.2, respetivamente, levando a que os resultados da aplicação dos algoritmos correspondentes aos mesmos sejam diferentes. O método de Ward é caracterizado como aglomerativo e, em cada iteração, após um elemento ser inserido num grupo não pode ser eliminado do mesmo. No entanto, no método de *k*-means é definido o número de *clusters a priori* e os elementos podem saltar de um grupo para outro, se isso justificar a diminuição da variabilidade dentro dos grupos. Nos capítulos 5.1 e 5.2 foram apresentados e explicados os resultados da aplicação dos métodos de classificação hierárquica e não hierárquica, respetivamente, para a PC E. Nestes capítulos concluiu-se que, nos quatro trimestres de 2017, os grupos com consumos suspeitos são considerados, estatisticamente, como grupos com consumo efetivamente anómalo para ambos os métodos. No entanto, os *clusters* anómalos podem não conter as mesmas instalações quando comparados os resultados obtidos pelo método de Ward com os resultados apresentados através do método de *k*-means.

Neste capítulo são comparadas as duas metodologias utilizadas para a análise e previsão de anomalias de consumo, de modo a averiguar qual o método mais eficiente e robusto para este tipo de análise. Espera-se que esta comparação permita escolher a melhor metodologia a aplicar para encontrar instalações com consumo não tipificado. Os resultados apresentados de seguida estão associados à PC E, tendo sido análogo o raciocínio para os restantes escalões de PC.

Tabela 5.13: Percentagem de instalações com consumo considerado anómalo através do método de Ward.

Trimestre	Instalações anómalas
Primeiro	52.6%
Segundo	51.9%
Terceiro	46.9%
Quarto	58.1%

Na tabela 5.13 é apresentada, em termos relativos, a quantidade de instalações que têm efetivamente consumo anómalo de entre as instalações pertencentes aos grupos suspeitos, pela aplicação do método de Ward, em cada trimestre de 2017. Pode observar-se que, das instalações contidas nos grupos com consumos não tipificados entre janeiro e março, 52.6% são pontos de medida cujo consumo, em algum momento desse período, se distancia daquele que é considerado como padrão nesse mesmo período. Por sua vez, esta percentagem diminui nos seis meses seguintes, com 51.9% e 46.9% de pontos de medida com consumo não tipificado, no segundo e terceiro trimestres, respetivamente. Nos últimos três meses do ano em estudo esta percentagem aumenta, sendo o trimestre no qual se verificam mais instalações com consumo suspeito, alcançando uma percentagem de 58.1%.

Tabela 5.14: Percentagem de instalações com consumo considerado anómalo através do método de *k*-means.

Trimestre	Instalações anómalas
Primeiro	68.8%
Segundo	53.3%
Terceiro	54.2%
Quarto	59.0%

A tabela 5.14 diz respeito à percentagem de instalações com consumo efetivamente anómalo, daquelas que estão inseridas nos grupos considerados suspeitos através do método de *k*-means, em cada um dos trimestres. Pela observação da tabela pode concluir-se que, o primeiro trimestre de 2017 é aquele cuja percentagem de pontos de medida anómalos é superior, alcançando os 68.8%. Esta percentagem diminui para os 53.3% no segundo trimestre e volta a aumentar, chegando a um total de 54.2% instalações com consumo não tipificado nos meses de julho a setembro de 2017. Nos últimos três meses do ano, a percentagem de instalações cujo consumo é anómalo chega aos 59.0%.

Tabela 5.15: Percentagem de instalações efetivamente anómalas consoante os métodos aplicados.

	Total Ward	**Total *k*-means**	**Só Ward**	**Só *k*-means**	**Ambos**	**Ward ou *k*-means**
%	52.6%	64.0%	45.0%	75.0%	60.5%	55.2%

Ao comparar os resultados obtidos pelas tabelas 5.13 e 5.14 verifica-se que, em todos os trimestres de 2017, o método de *k*-means parece ser melhor quando se trata de encontrar instalações com consumo efetivamente anómalo. Isto deve-se ao facto de a percentagem de instalações com consumo considerado anómalo pelo método de Ward ser sempre inferior, independentemente do trimestre em estudo. Contudo, é importante referir que nem todas as instalações consideradas anómalas pelo método de Ward são as mesmas caracterizadas como instalações com consumo não tipificado pelo método de *k*-means. Pode afirmar-se, pela tabela 5.15, que aplicando cada um dos algoritmos estudados, de todas as instalações pertencentes aos grupos suspeitos criados pelo método de Ward, 52.6% são efetivamente anómalas. Por sua vez, esta percentagem é superior quando se verificam os resultados dados pelo método de *k*-means, pois 64.0% das instalações assinaladas como anómalas apresentam realmente consumo não tipificado em algum instante do ano de 2017. Não se deve esquecer que as duas metodologias utilizadas criaram grupos cujos conjuntos de instalações, por vezes, estão inseridas em ambos os métodos de Ward e *k*-means.

Se neste estudo fosse aplicado apenas o método de classificação hierárquica, este iria permitir a sinalização de 45.0% dos pontos de entrega com consumo não tipificado e, ao empregar o método de classificação não hierárquica, iriam ser detetados 75.0% dos casos. Apesar de estas percentagens serem muito diferentes, mais uma vez aparentemente o método de *k*-means revela melhores resultados, o que não é verdade. Os valores indicados na tabela 5.15 são em termos relativos e, uma vez que o método de Ward em ambas as situações descritas (Total Ward e Só Ward) assinala grupos suspeitos com maior dimensão, face aos *clusters* suspeitos pelo método de *k*-means, o universo de conjuntos de instalações suspeitos em estudo é diferente para cada método. Isto significa que, na realidade, o método de Ward é mais eficiente pois o total de instalações possivelmente anómalas é superior logo, por exemplo, se for aplicado apenas este método 45.0% das instalações assinaladas representa uma quantidade absoluta superior, face aos 75.0% dos pontos de entrega dados pelo método de *k*-means. Deste modo, o método de Ward é mais eficiente quando utilizado individualmente.

O estudo focou-se na aplicação de dois métodos, um de classificação hierárquica e um de classificação não hierárquica, para que fossem comparados os resultados obtidos por ambas as metodologias. Nas últimas duas colunas da tabela 5.15 indicam-se as percentagens de instalações que foram indicadas como efetivamente anómalas de entre todas as que foram consideradas suspeitas simultaneamente pelos dois métodos (Ambos) ou por pelo menos um deles (Ward ou *k*-means). Então, das instalações dadas como suspeitas pela utilização simultânea dos dois métodos 60.5% apresentam consumo não tipificado em algum momento do ano. Se forem empregues o método de Ward, o de *k*-means ou ambos 55.2% dos pontos de medida assinalados como possivelmente anómalos efetivamente apresentam consumo não tipificado. Mais uma vez, esta última percentagem é inferior àquela que é

indicada pela aplicação da interseção das duas metodologias, com um total de 60.5%. Contudo, os 55.2% pontos de medida com consumo não tipificado apresentam, em termos absolutos, um valor superior face aos 60.5%.

Conclui-se assim que, para o estudo da análise e previsão de anomalias de consumo, quando se trata da identificação de grupos com consumo efetivamente não tipificado, o mais eficiente será utilizar pelo menos um dos dois métodos. Isto porque o universo de instalações suspeitas em estudo será maior e assim aumenta a probabilidade de que haja mais instalações efetivamente anómalas contidas nestes *clusters*. Optando por aplicar apenas uma destas metodologias, aquela que apresenta uma maior taxa de sucesso neste estudo é o método de Ward pois para o mesmo número de grupos suspeitos, este identifica mais instalações com consumo efetivamente anómalo. Por sua vez, o método de *k*-means caracteriza muito menos instalações como suspeitas, apesar de haver uma maior probabilidade desta metodologia encontrar pontos de medida com consumo efetivamente anómalo.

6. Conclusão

A análise de *clusters* efetuada com base nos dois métodos, Ward e *k*-means, permitiu identificar quais os conjuntos de instalações com consumo possivelmente não tipificado, em cada trimestre de 2017. A agregação das instalações em grupos, com base nos respetivos consumos, facilitou a análise dos resultados obtidos, tendo sido mais acessível a observação dos diferentes comportamentos destes grupos.

Os resultados obtidos através dos métodos de Ward e de *k*-means foram alvo de um estudo ainda mais aprofundado, com o intuito de se eliminar a possibilidade dos respetivos algoritmos considerarem como suspeitos os grupos cujo consumo deve, de facto, ser tomado como referência. Esta análise baseou-se na construção de intervalos de 95% de confiança para o valor médio do consumo dos grupos padrão. O objetivo deste procedimento foi identificar grupos que, apesar de serem considerados suspeitos pelo resultado da análise de *clusters*, poderiam inserir-se nos grupos de consumo padrão. Concluiu-se que nenhum dos intervalos de confiança continha pelo menos 95% dos consumos quartos-horário dos grupos suspeitos, independentemente do trimestre em estudo. Posto isto e tendo em conta que as percentagens de consumos médios por quarto de hora contidos em qualquer um dos IC atingiram no máximo 26.1%, concluiu-se que todos os *clusters* caracterizados como anómalos pelos dois métodos apresentam, estatisticamente, consumos não tipificados.

Após a validação de todos os *clusters* cujos consumos foram estatisticamente considerados como anómalos, a comparação efetuada entre os resultados obtidos através dos dois métodos levou a concluir qual o mais robusto e eficiente, quando se trata de encontrar instalações com consumo anómalo. Se apenas se puder aplicar uma das metodologias estudadas, deverá escolher-se o método de Ward porque forma grupos suspeitos com maior dimensão, o que aumenta a probabilidade de acerto relativamente às instalações com consumo realmente não tipificado ao longo do ano. Contudo, neste relatório utilizaram-se tanto o método de Ward como o de *k*-means, dando origem a que a melhor opção para caracterizar instalações suspeitas seja a utilização dos dois métodos, visto que o universo de grupos suspeitos em estudo será mais diversificado e abrangerá mais instalações, levando a que o resultado do número de pontos de medida anómalos seja mais próximo da realidade.

6.1. Trabalhos futuros

O tema da análise e previsão de anomalias de consumo na área da energia elétrica tem sido pouco estudado. Assim, parece que o método apresentado neste relatório, isto é, a análise de *clusters* e posterior construção de intervalos de confiança é uma boa abordagem inicial a este problema. Contudo, podem ser realizados outros estudos e análises para aperfeiçoar esta metodologia ou para encontrar métodos alternativos.

Como é demonstrado e concluído nos capítulos 5.3 e 6, ambas as metodologias aplicadas apresentam resultados positivos, sendo que a forma mais eficiente de identificar as instalações anómalas é através da utilização dos dois métodos, de Ward e de *k*-means. Conclui-se ainda que, individualmente, o método de classificação hierárquica é mais robusto pois agrega em *clusters* suspeitos um maior número de instalações, o que provoca uma maior taxa de sucesso. Pode ainda afirmar-se que para um mesmo número de *clusters*, o método de Ward identifica mais pontos de entrega efetivamente anómalos. Contudo, apesar de o método de *k*-means não ser tão eficiente quando

utilizado individualmente, espera-se que a criação de um maior número de *clusters* através deste método leve a uma maior percentagem de acerto, pois verifica-se que esta metodologia assinala menos instalações como suspeitas, mas desse universo o número absoluto de pontos de medida realmente anómalos é superior, face ao método de Ward. Assim, uma boa abordagem pode ser a agregação das observações em mais grupos, isto é, efetuar este estudo com o pressuposto de criar um número maior de *clusters*.

Uma forma possível de melhorar a metodologia apresentada é a de construir intervalos de confiança para quantis elevados, em alternativa aos intervalos de confiança para as médias de cada 15 minutos de cada grupo, uma vez que se pretende encontrar um limite para a observação e não para a sua média. Outro aspeto importante que não foi considerado neste trabalho é a procura de anomalias que ocorrem num curto espaço de tempo. Isto pode ser feito, por exemplo, através da utilização de modelos de séries temporais que apresentem um padrão específico de consumo. Subsequentemente, podem ser construídos intervalos de previsão para verificar se os eventuais picos e/ou cavas de consumo do grupo ao qual pertence essa mesma instalação se inserem nos referidos intervalos. Efetivamente existem várias abordagens que podem ser efetuadas a partir do estudo descrito neste relatório assinalando-se, entre outros, a extensão da análise de *clusters* e a possível implementação de uma mistura dos métodos utilizados, o de Ward e o de *k*-means, para averiguar até que ponto as características de cada algoritmo podem, ou não, influenciar a caracterização das instalações efetivamente anómalas.

Referências bibliográficas

[1] Indusmelec, "História da eletricidade." [Online]. Available: http://www.indusmelec.pt/newsletter/05/historia_electricidade.pdf. [Accessed: 01-Sep-2018].

[2] LojaLuz, "História do mercado de electricidade em Portugal." [Online]. Available: https://lojaluz.com/historia-mercado-eletricidade-portugal. [Accessed: 05-Sep-2018].

[3] F. Alves, "As Energias Renováveis em Portugal – Ponto da situação." [Online]. Available: http://naturlink.pt/article.aspx?menuid=5&cid=10094&bl=1&viewall=true. [Accessed: 31-Aug-2018].

[4] COTEC Portugal, "EDP - Energias de Portugal, SA." [Online]. Available: http://www.cotecportugal.pt/pt/quem-somos/associados/edp-energias-de-portugal-sa. [Accessed: 07-Sep-2018].

[5] EDP SU, "Origens da Eletricidade," 2018. [Online]. Available: http://www.edpsu.pt/pt/origemdaenergia/Pages/OrigensdaEnergia.aspx. [Accessed: 05-Sep-2018].

[6] ERSE, "Guia de Medição de Leituras e Disponibilização de Dados," 2016. [Online]. Available: http://www.erse.pt/pt/electricidade/regulamentos/relacoescomerciais/Documents/SubRegulamentação/GMLDD_2016.pdf. [Accessed: 01-Jul-2018].

[7] ERSE, "Glossário." [Online]. Available: http://www.erse.pt/pt/glossario/Paginas/glossario.aspx?folder=baeaae46-4f3f-401d-91ff-668518dd41e8. [Accessed: 01-Jun-2018].

[8] ERSE, "Tarifas transitórias," 2018. [Online]. Available: http://www.erse.pt/consumidor/electricidade/querosercliente/tenholigacaoarede/Documents/Documento_TVCF_Electricidade_T12018.pdf. [Accessed: 10-May-2018].

[9] ERSE, "Redes de transporte." [Online]. Available: http://www.erse.pt/pt/electricidade/actividadesdosector/transporte/Paginas/default.aspx. [Accessed: 03-Aug-2018].

[10] ERSE, "Redes de distribuição." [Online]. Available: http://www.erse.pt/pt/electricidade/actividadesdosector/distribuicao/Paginas/default.aspx?master=ErsePrint.master. [Accessed: 01-Jul-2018].

[11] T. Alpuim, "Probabilidade e Estatística."

[12] A. C. Estatística, "O que é Análise de Clusters?," 2018. [Online]. Available: http://www.abgconsultoria.com.br/blog/o-que-e-analise-de-cluster/. [Accessed: 20-Feb-2018].

[13] D. S. Brian S. Everitt, Sabine Landau, Morven Leese, *Cluster Analysis*, 5th ed. 2011.

[14] Técnico Lisboa, "Introdução à análise de clusters." [Online]. Available: https://fenix.tecnico.ulisboa.pt/downloadFile/3779579704252/SlidesACluster.pdf. [Accessed: 02-Apr-2018].

[15] R. Primicerio, M. Greenacre, "Measures of distance between samples: Euclidean," in *Multivariate Analysis of Ecological Data*, Fundación BBVA, Ed. 2013.

[16] M. das Graças, "Análise de Agrupamento Hierárquico Aglomerativo aplicada à Ecologia," 2017.

[17] M. R. Anderberg, *Cluster analysis for applications*, Elsevier, Ed. 1973.

[18] M. M. T. D. M. Leal, “Apontamentos da Unidade Curricular Análise Exploratória de Dados Multivariados.” 2016.

[19] Técnico Lisboa, “K-médias.” [Online]. Available: http://web.tecnico.ulisboa.pt/ana.freitas/bioinformatics.ath.cx/bioinformatics.ath.cx/index651a.html?id=147. [Accessed: 16-Jun-2018].

[20] ANACOM, “Análise de clusters - método K-means.” [Online]. Available: https://www.anacom.pt/streaming/anexo2_analise_clusters.pdf?contentId=1070331&field=ATTACHED_FILE. [Accessed: 18-Jun-2018].

[21] STHDA, “Cluster Validation Essentials,” 2017.

[22] P. Bholowalia and A. Kumar, “EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN,” *Int. J. Comput. Appl.*, vol. 105, No. 9, 2014.

[23] C. M. T. S. Rocha, “Apontamentos da Unidade Curricular Estatística Paramétrica.” Faculdade de Ciências da Universidade de Lisboa, 2015.

Anexos

Anexo A.1. Potência Contratada A

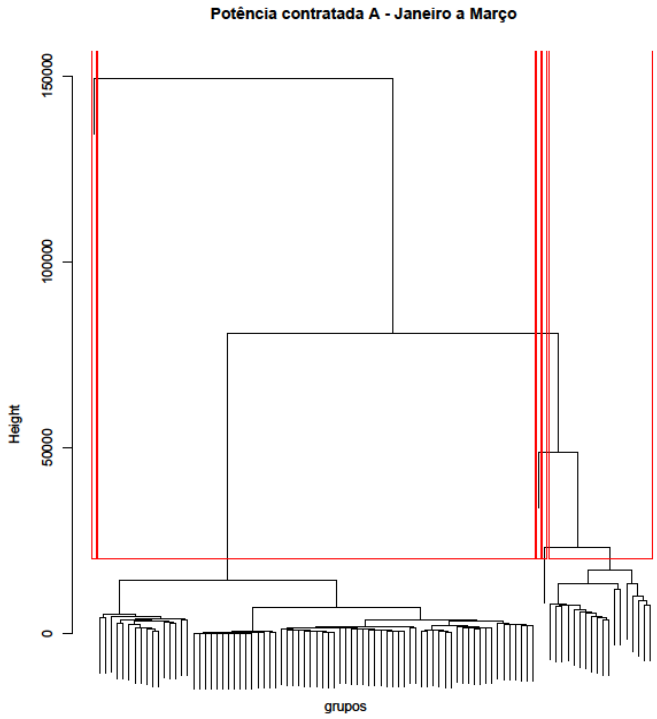

Figura A.1.1: Dendrograma associado à PC A, para o primeiro trimestre de 2017, pelo método de Ward.

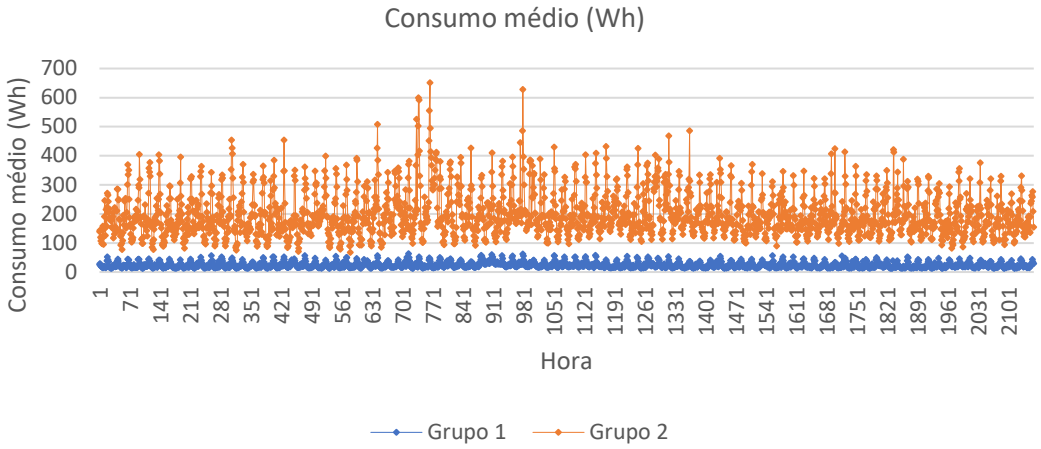

Figura A.1.2: Consumo médio (Wh) para os grupos 1 e 2, no primeiro trimestre de 2017, aplicando o método de Ward.

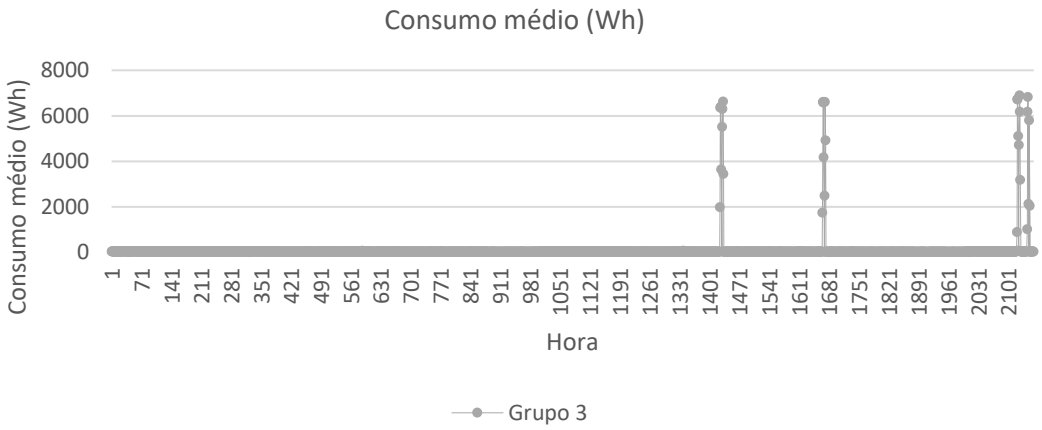

Figura A.1.3: Consumo médio (Wh) para o grupo 3, no primeiro trimestre de 2017, aplicando o método de Ward.

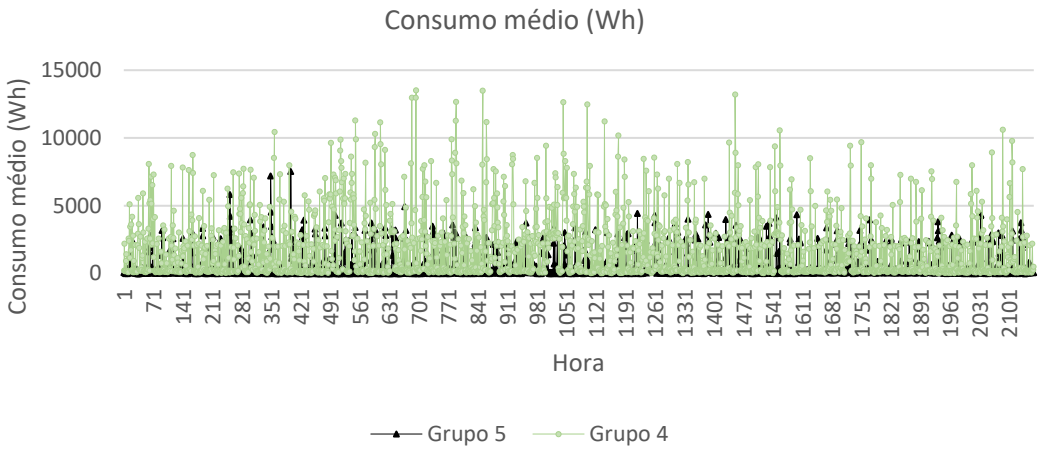

Figura A.1.4: Consumo médio (Wh) para os grupos 4 e 5, no primeiro trimestre de 2017, aplicando o método de Ward.

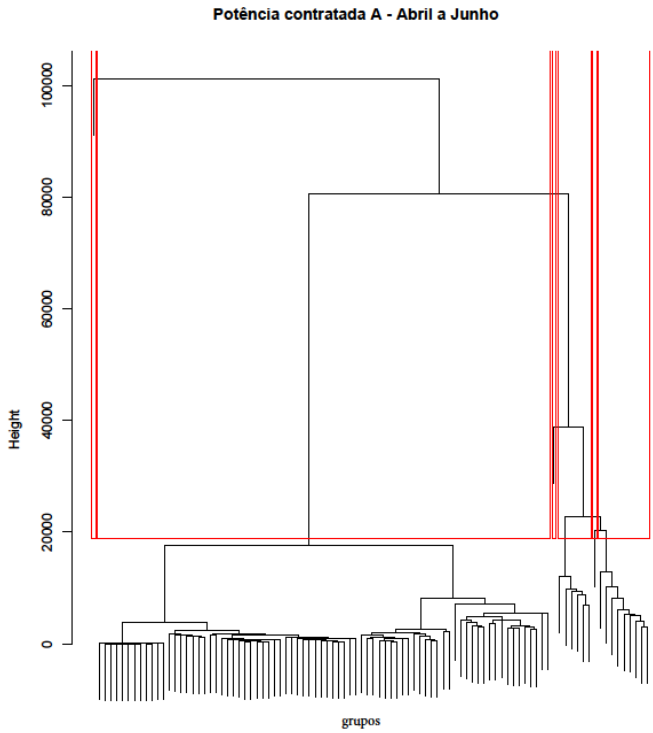

Figura A.1.5: Dendrograma associado à PC A, para o segundo trimestre de 2017, pelo método de Ward.

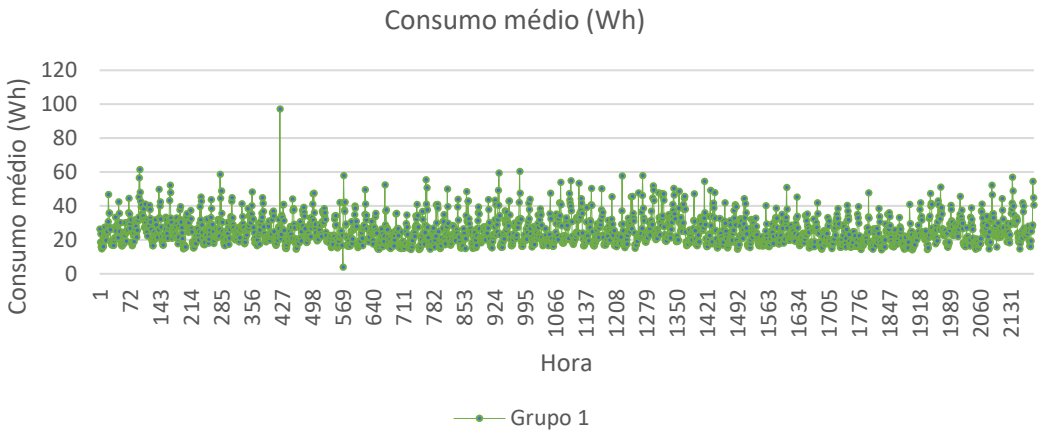

Figura A.1.6: Consumo médio (Wh) para o grupo 1, no segundo trimestre de 2017, aplicando o método de Ward.

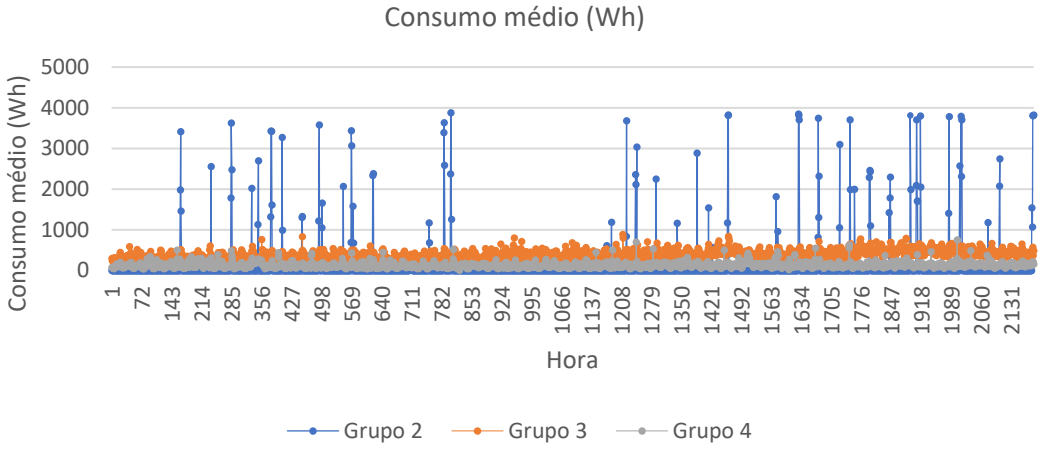

Figura A.1.7: Consumo médio (Wh) para os grupos 2, 3 e 4 no segundo trimestre de 2017, pelo método de Ward.

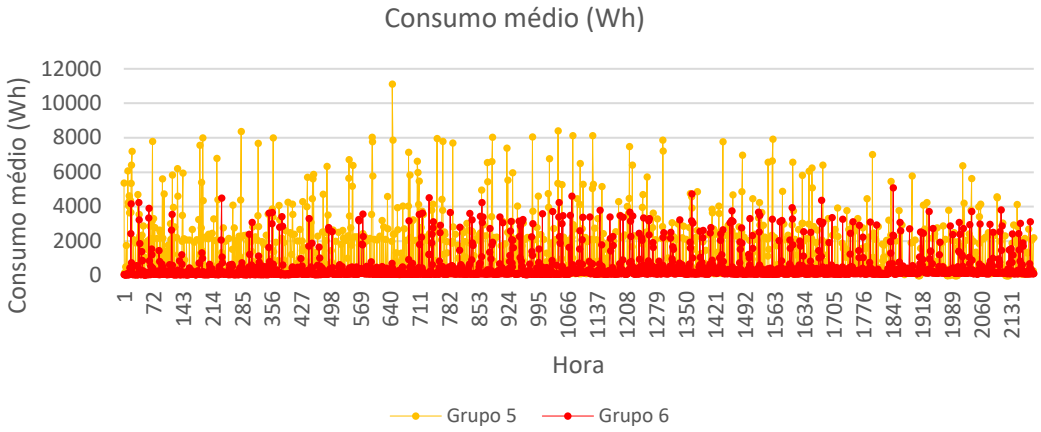

Figura A.1.8: Consumo médio (Wh) para os grupos 5 e 6, no segundo trimestre de 2017, pelo método de Ward.

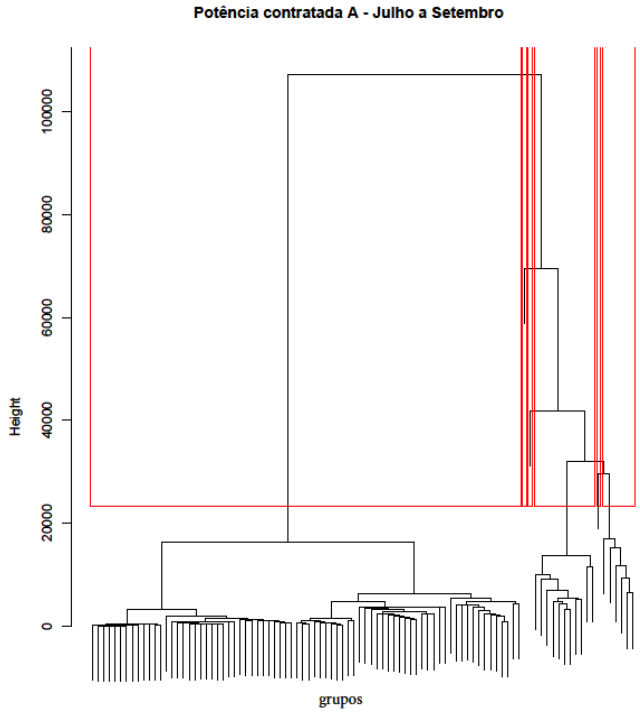

Figura A.1.9: Dendrograma associado à PC A, para o terceiro trimestre de 2017, pelo método de Ward.

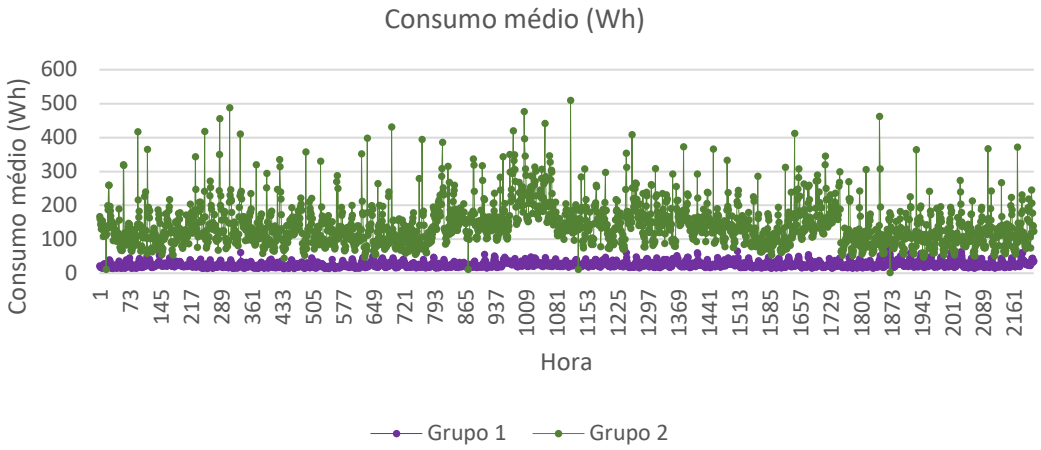

Figura A.1.10: Consumo médio (Wh) para os grupos 1 e 2, no terceiro trimestre de 2017, aplicando o método de Ward.

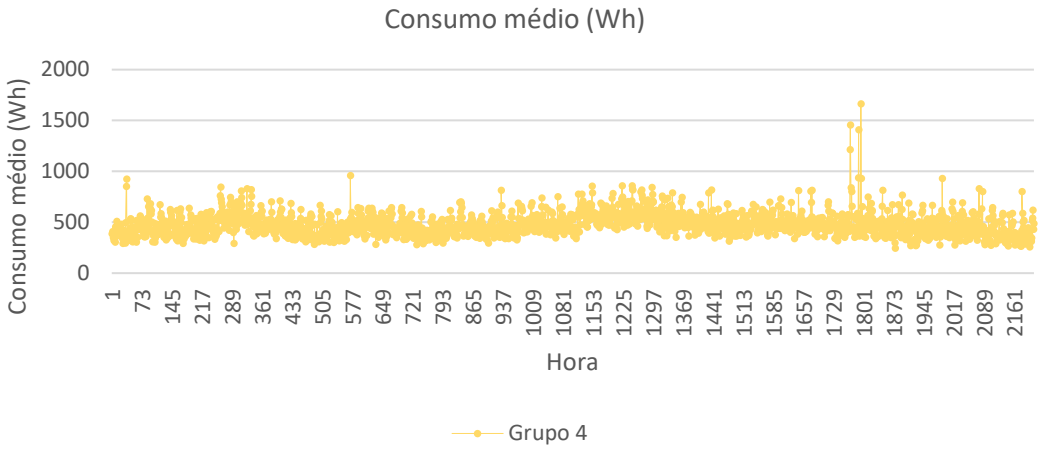

Figura A.1.11: Consumo médio (Wh) para o grupo 4, no terceiro trimestre de 2017, aplicando o método de Ward.

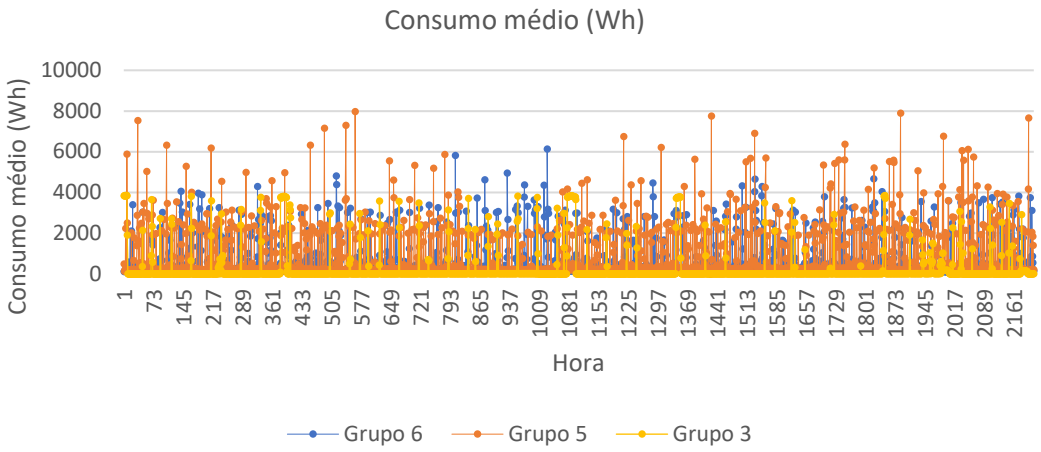

Figura A.1.12: Consumo médio (Wh) para os grupos 3, 5 e 6, no terceiro trimestre de 2017, pelo método de Ward.

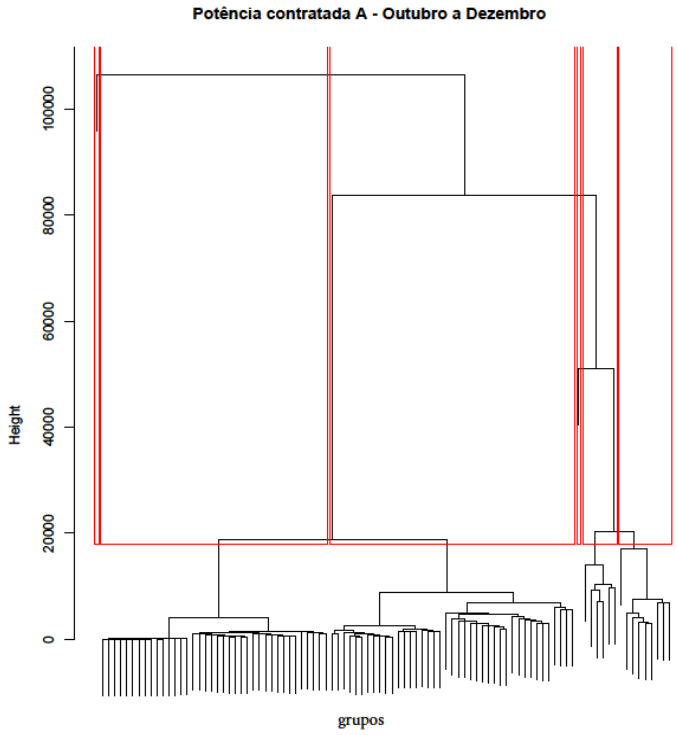

Figura A.1.13: Dendrograma associado à PC A, para o quarto trimestre de 2017, pelo método de Ward.

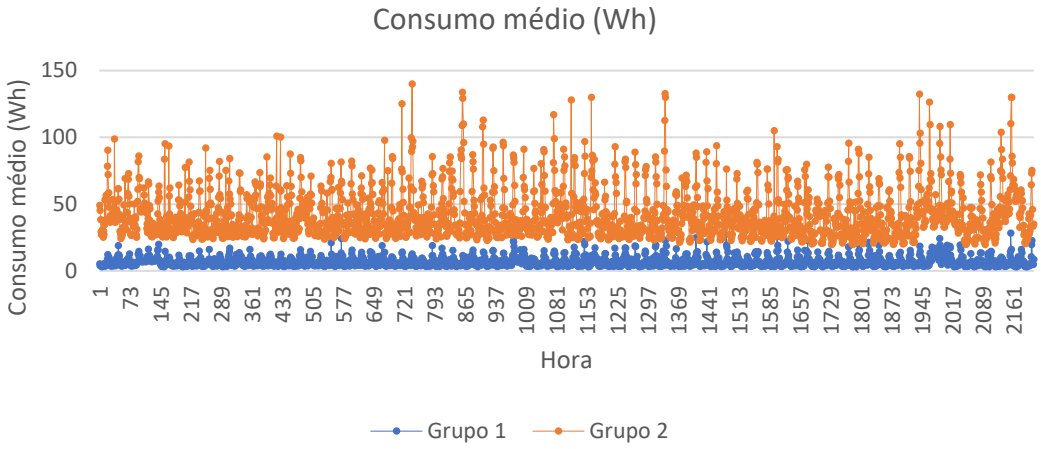

Figura A.1.14: Consumo médio (Wh) para os grupos 1 e 2, no quarto trimestre de 2017, aplicando o método de Ward.

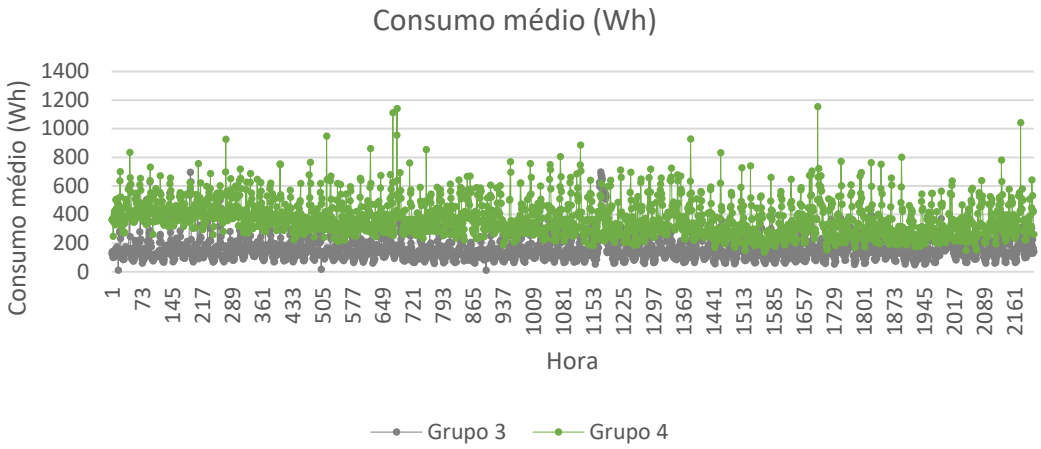

Figura A.1.15: Consumo médio (Wh) para os grupos 3 e 4, no quarto trimestre de 2017, aplicando o método de Ward.

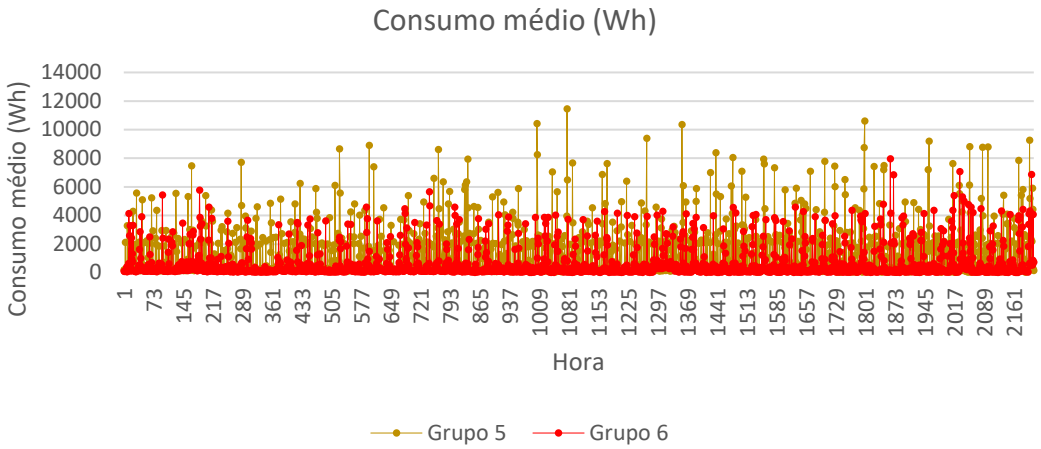

Figura A.1.16: Consumo médio (Wh) para os grupos 5 e 6, no quarto trimestre de 2017, aplicando o método de Ward.

Anexo A.2. Potência Contratada B

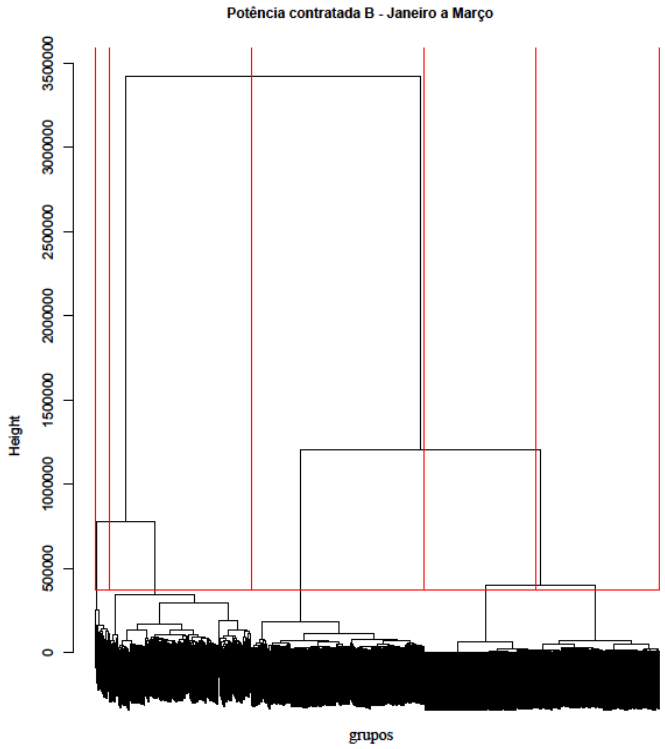

Figura A.2.1: Dendrograma associado à PC B, para o primeiro trimestre de 2017, aplicando o método de Ward.

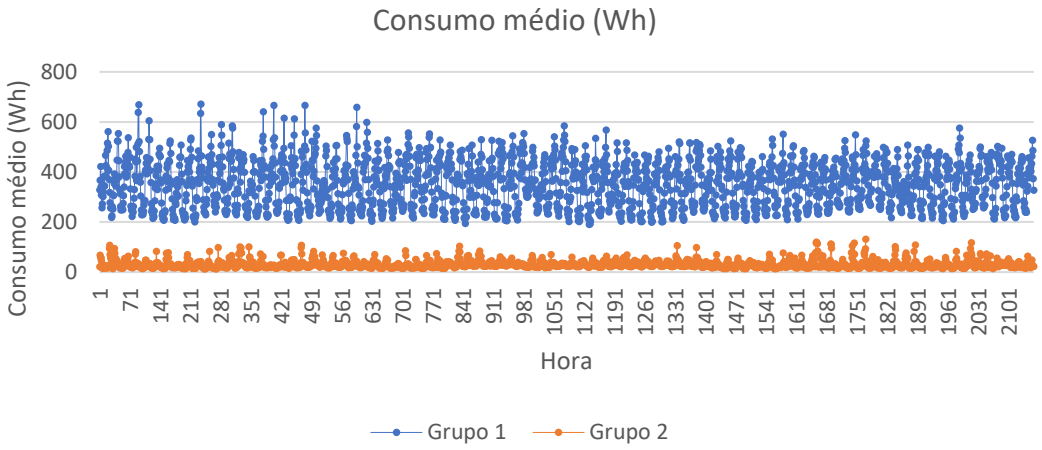

Figura A.2.2: Consumo médio (Wh) para os grupos 1 e 2, no primeiro trimestre de 2017, aplicando o método de Ward.

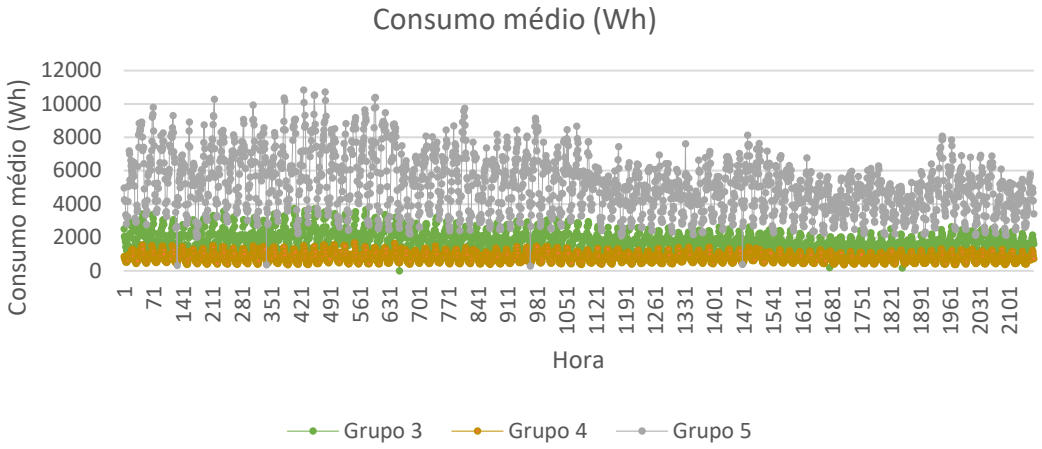

Figura A.2.3: Consumo médio (Wh) para os grupos 3, 4 e 5 no primeiro trimestre de 2017, aplicando o método de Ward.

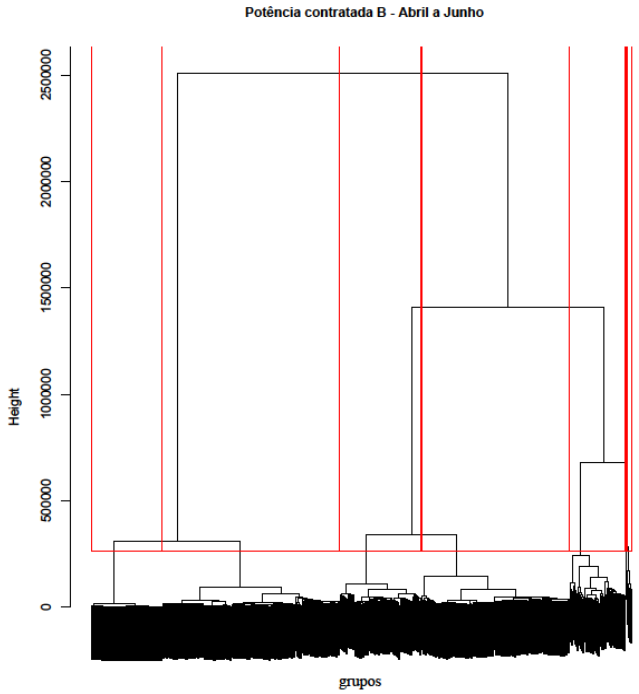

Figura A.2.4: Dendrograma associado à PC B, para o segundo trimestre de 2017, aplicando o método de Ward.

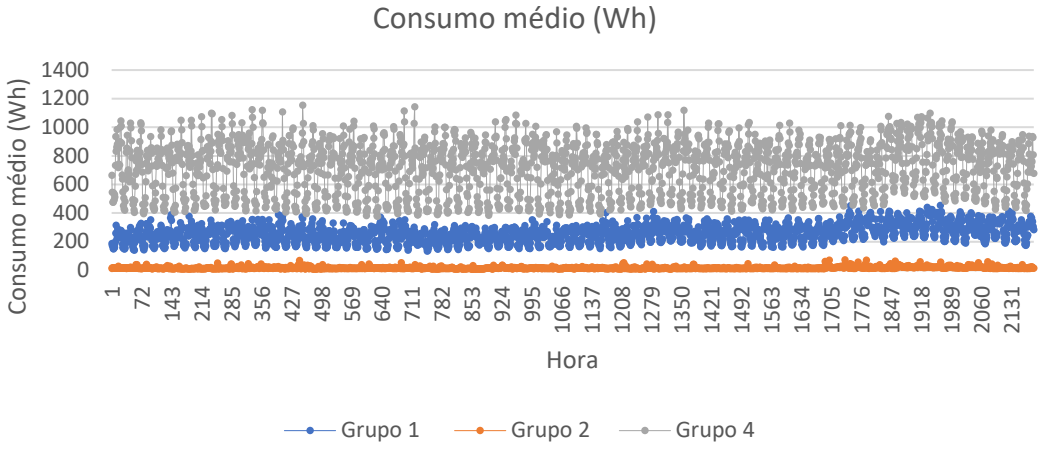

Figura A.2.5: Consumo médio (Wh) para os grupos 1, 2 e 4 no segundo trimestre de 2017, aplicando o método de Ward.

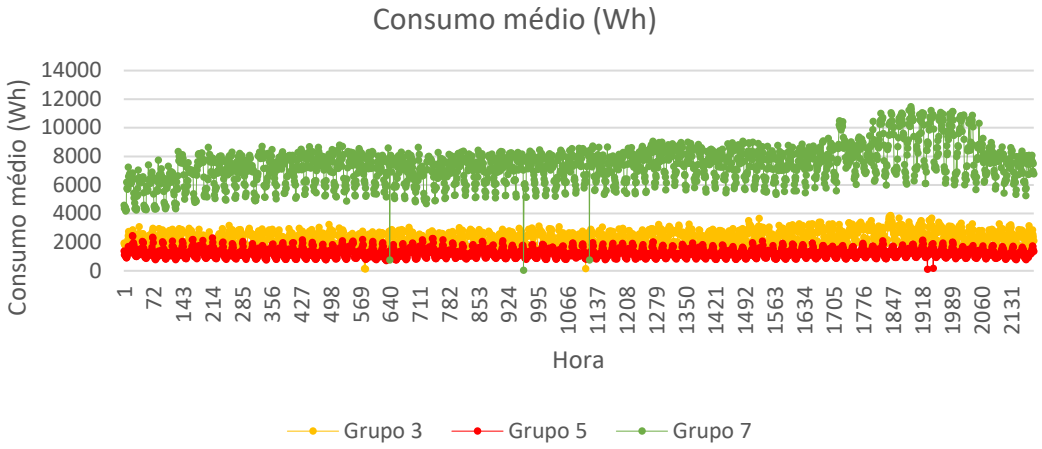

Figura A.2.6: Consumo médio (Wh) para os grupos 3, 5 e 7 no segundo trimestre de 2017, aplicando o método de Ward.

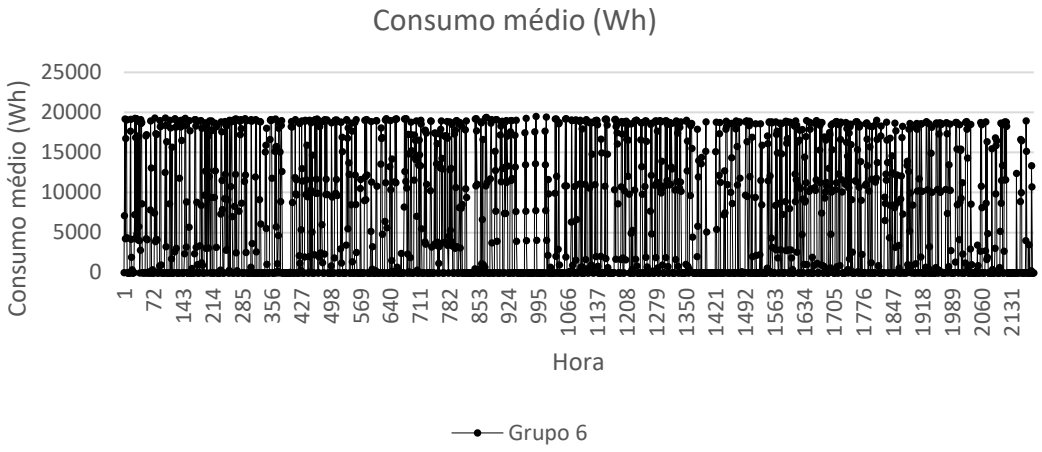

Figura A.2.7: Consumo médio (Wh) para o grupo 6 no segundo trimestre de 2017, aplicando o método de Ward.

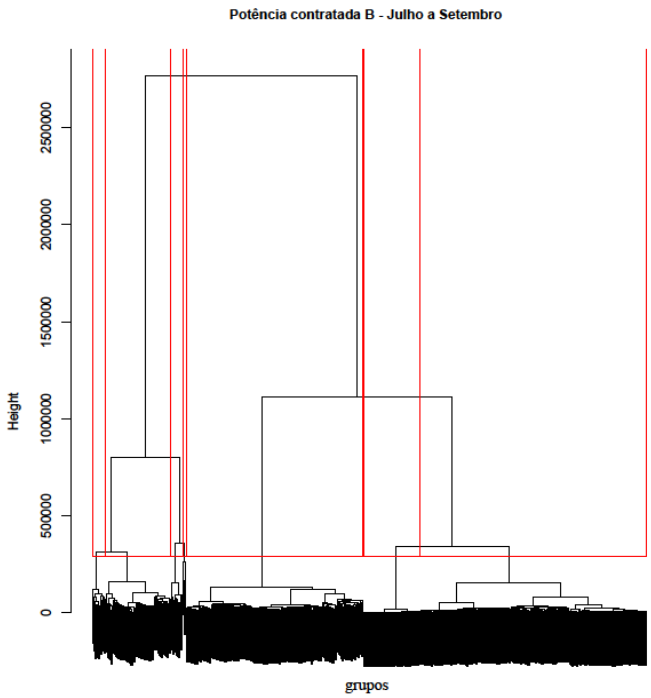

Figura A.2.8: Dendrograma associado à PC B, para o terceiro trimestre de 2017, aplicando o método de Ward.

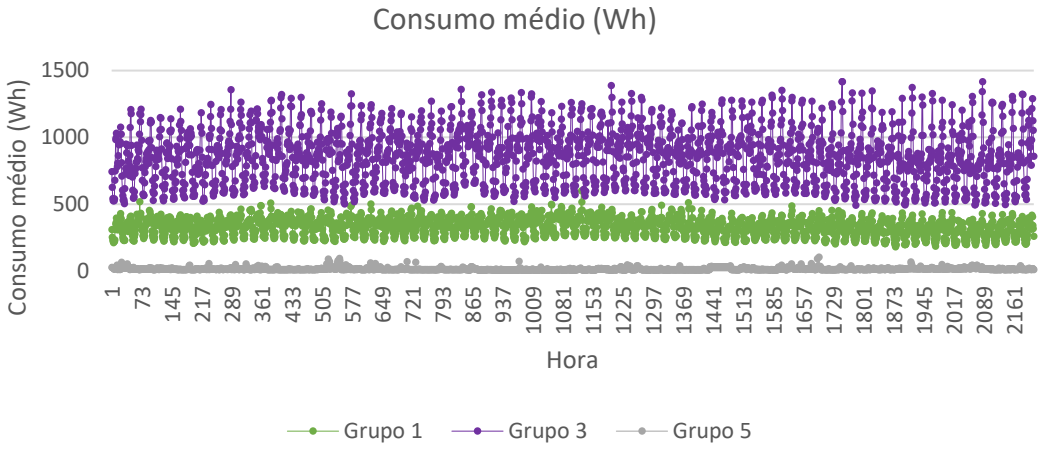

Figura A.2.9: Consumo médio (Wh) para os grupos 1, 3 e 5 no terceiro trimestre de 2017, aplicando o método de Ward.

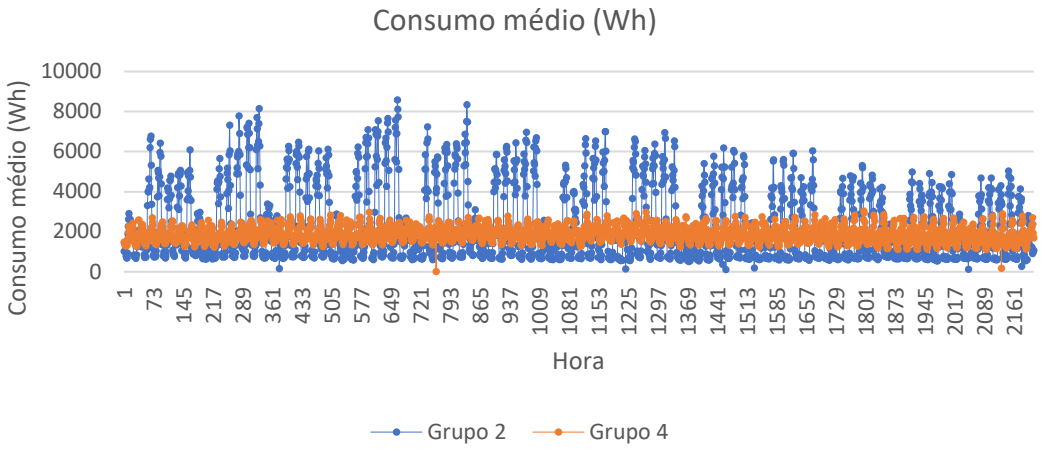

Figura A.2.10: Consumo médio (Wh) para os grupos 2 e 4 no terceiro trimestre de 2017, aplicando o método de Ward.

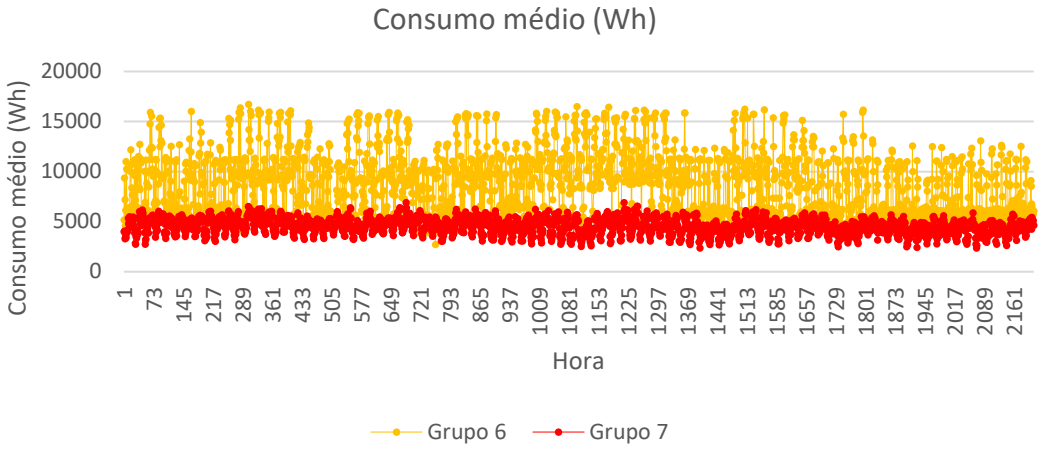

Figura A.2.11: Consumo médio (Wh) para os grupos 6 e 7 no terceiro trimestre de 2017, aplicando o método de Ward.

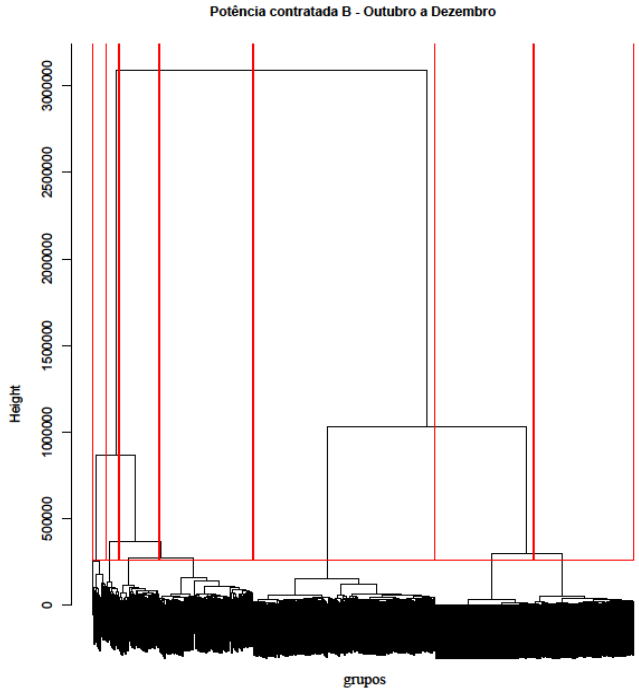

Figura A.2.12: Dendrograma associado à PC B, para o quarto trimestre de 2017, aplicando o método de Ward.

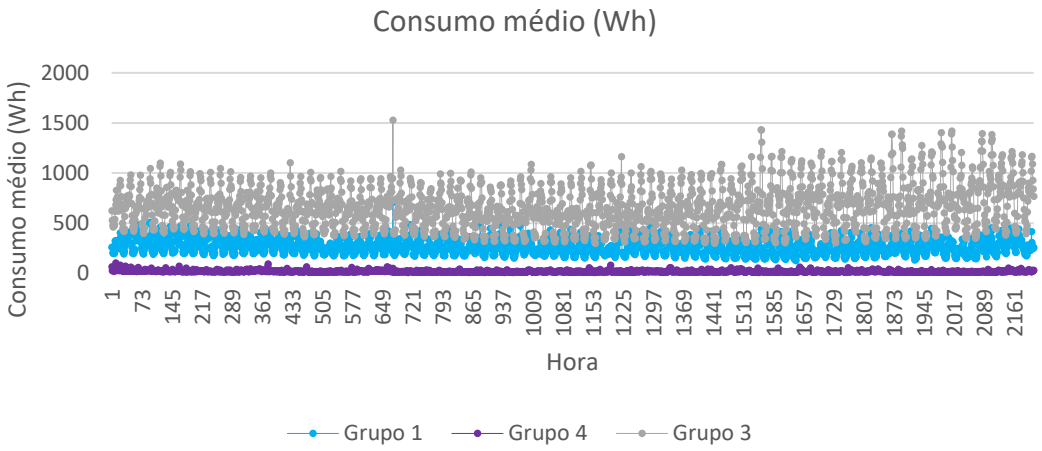

Figura A.2.13: Consumo médio (Wh) para os grupos 1, 3 e 4 no quarto trimestre de 2017, aplicando o método de Ward.

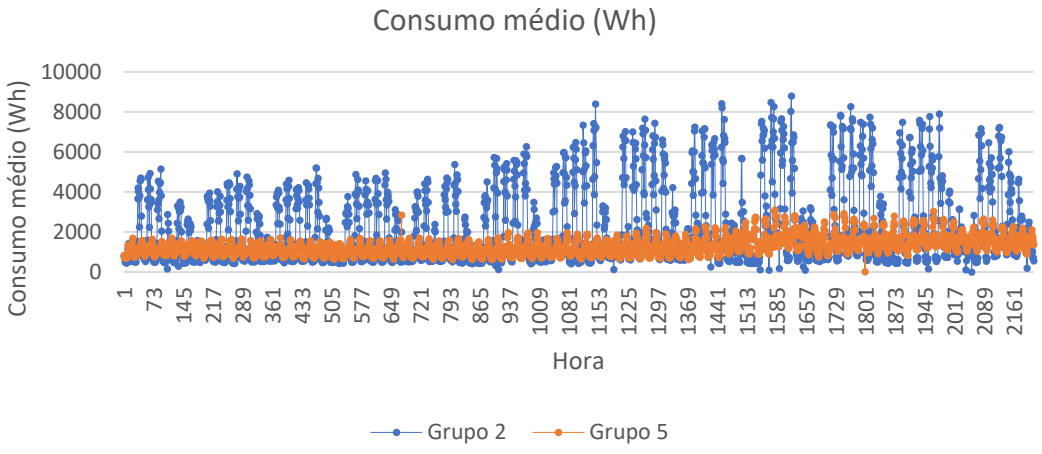

Figura A.2.14: Consumo médio (Wh) para os grupos 2 e 5 no quarto trimestre de 2017, aplicando o método de Ward.

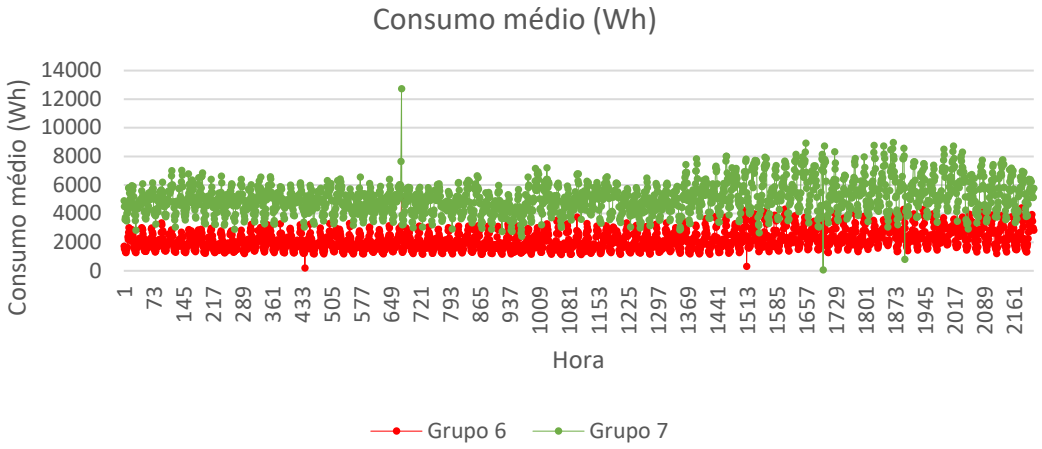

Figura A.2.15: Consumo médio (Wh) para os grupos 6 e 7 no quarto trimestre de 2017, aplicando o método de Ward.

Anexo A.3. Potência Contratada J

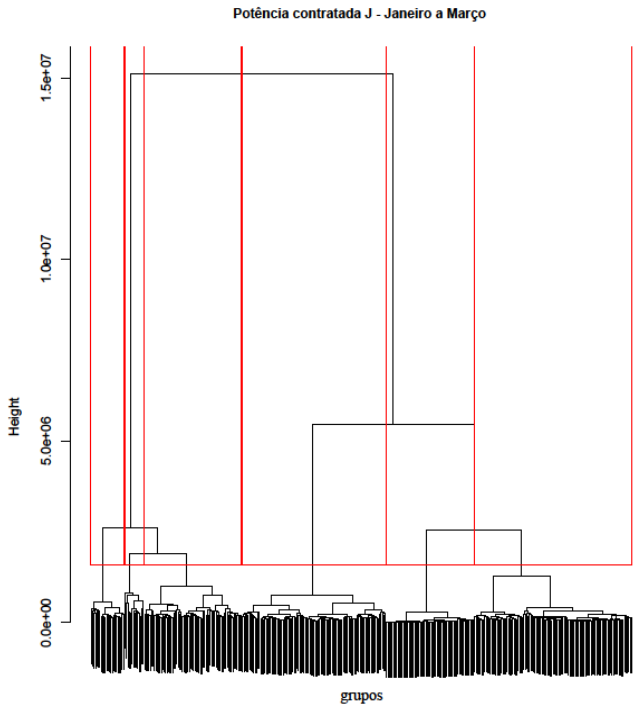

Figura A.3.1: Dendrograma associado à PC J, para o primeiro trimestre de 2017, aplicando o método de Ward.

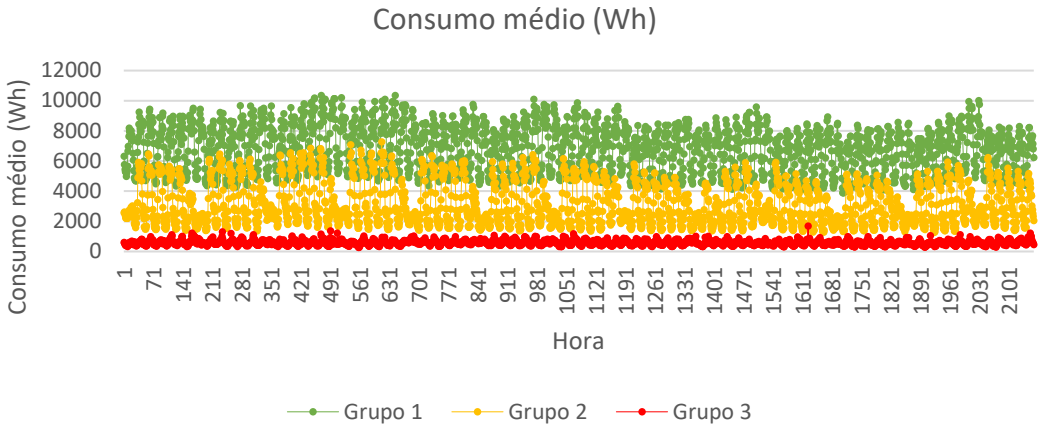

Figura A.3.2: Consumo médio (Wh) para os grupos 1, 2 e 3 no primeiro trimestre de 2017, aplicando o método de Ward.

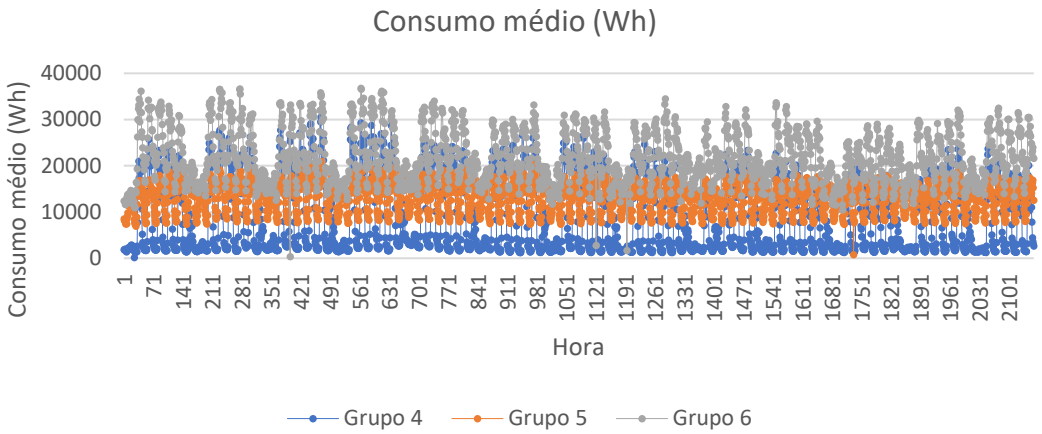

Figura A.3.3: Consumo médio (Wh) para os grupos 4, 5 e 6 no primeiro trimestre de 2017, aplicando o método de Ward.

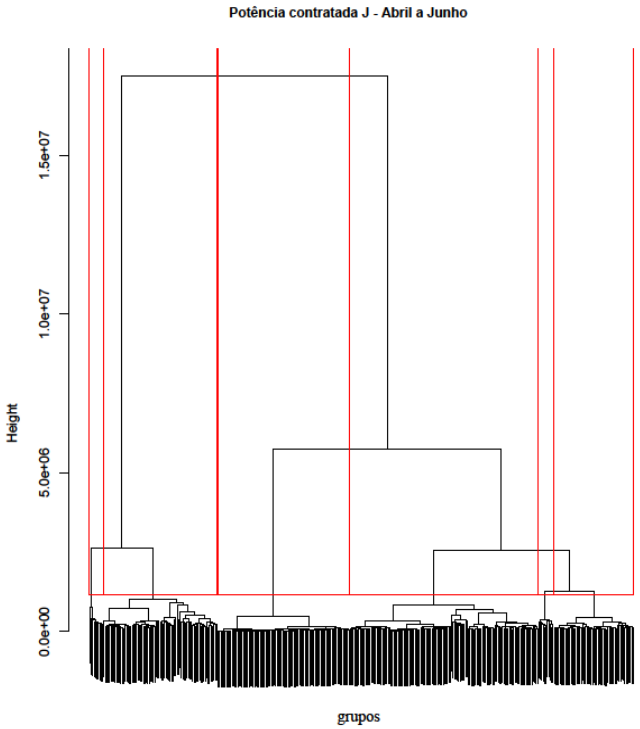

Figura A.3.4: Dendrograma associado à PC J, para o segundo trimestre de 2017, aplicando o método de Ward.

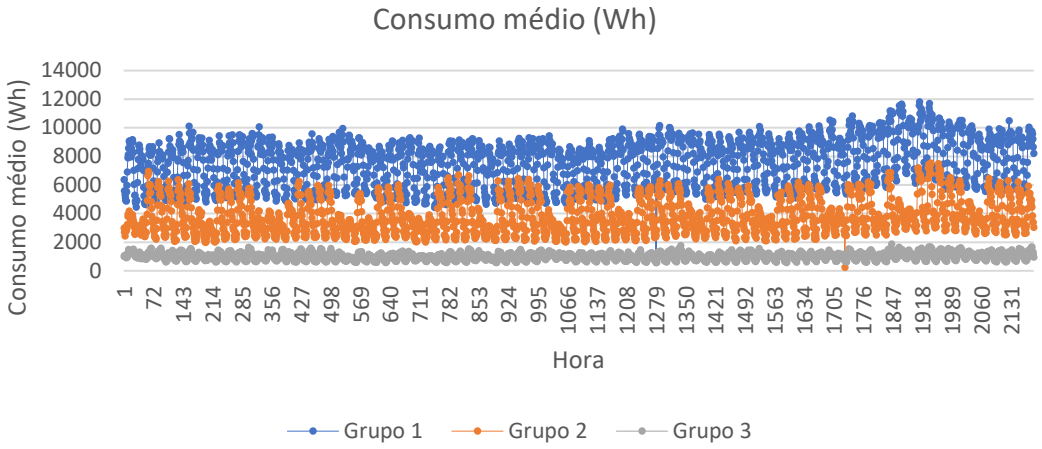

Figura A.3.5: Consumo médio (Wh) para os grupos 1, 2 e 3 no segundo trimestre de 2017, aplicando o método de Ward.

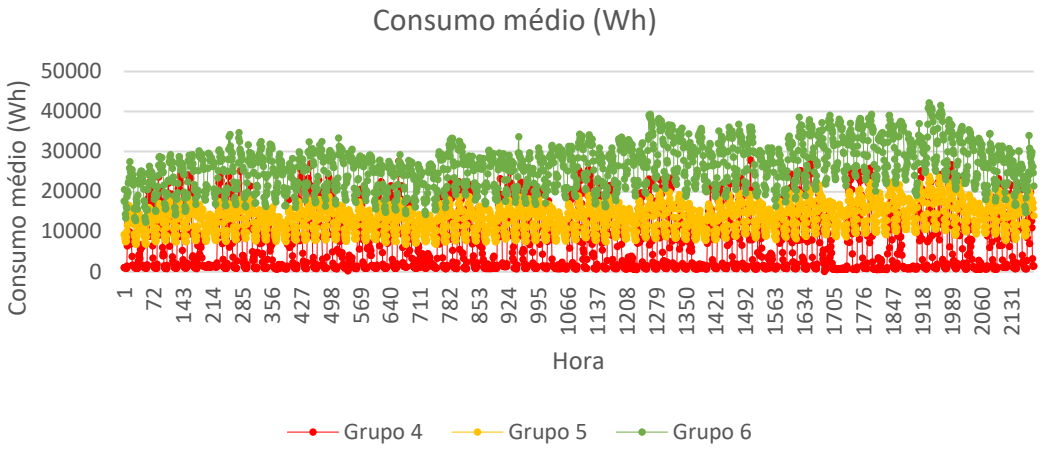

Figura A.3.6: Consumo médio (Wh) para os grupos 4, 5 e 6 no segundo trimestre de 2017, aplicando o método de Ward.

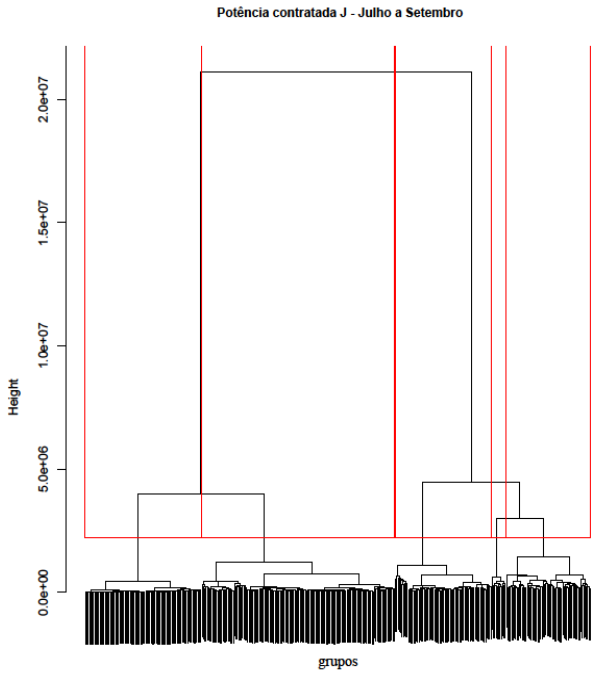

Figura A.3.7: Dendrograma associado à PC J, para o terceiro trimestre de 2017, aplicando o método de Ward.

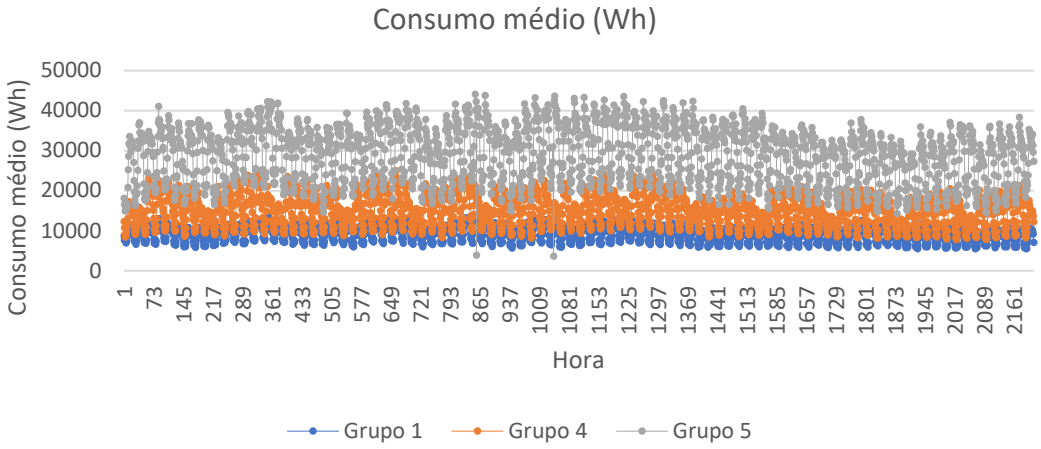

Figura A.3.8: Consumo médio (Wh) para os grupos 1, 4 e 5 no terceiro trimestre de 2017, aplicando o método de Ward.

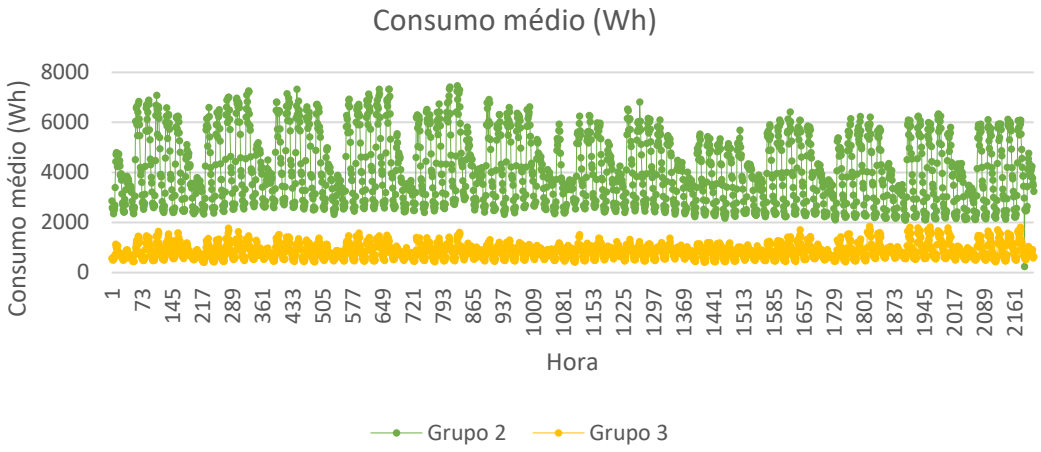

Figura A.3.9: Consumo médio (Wh) para os grupos 2 e 3 no terceiro trimestre de 2017, aplicando o método de Ward.

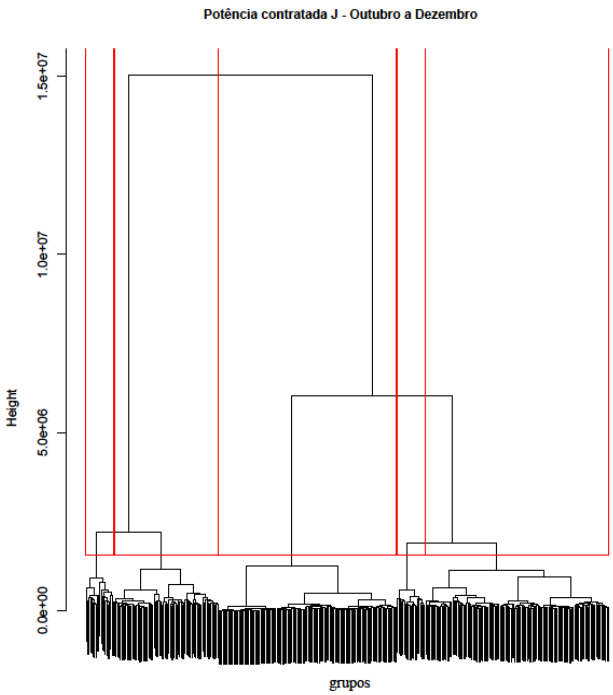

Figura A.3.10: Dendrograma associado à PC J, para o quarto trimestre de 2017, aplicando o método de Ward.

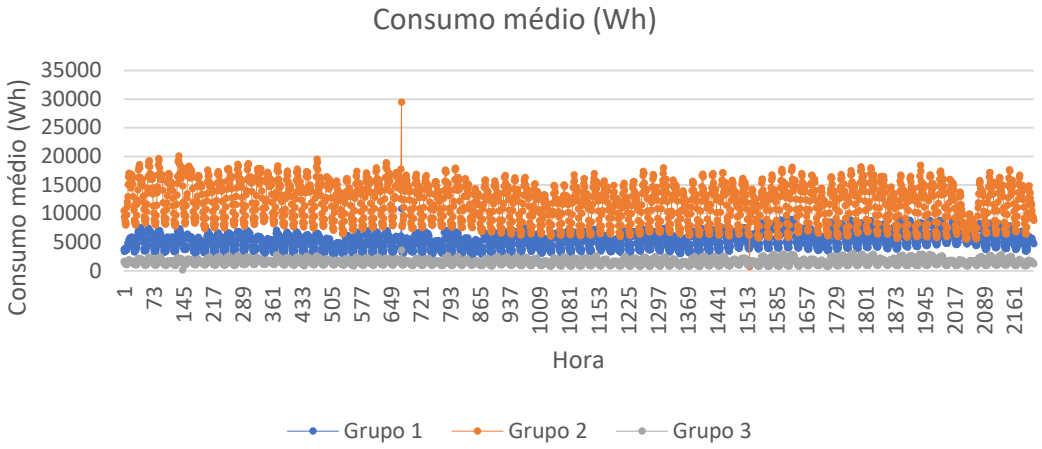

Figura A.3.11: Consumo médio (Wh) para os grupos 1, 2 e 3 no quarto trimestre de 2017, aplicando o método de Ward.

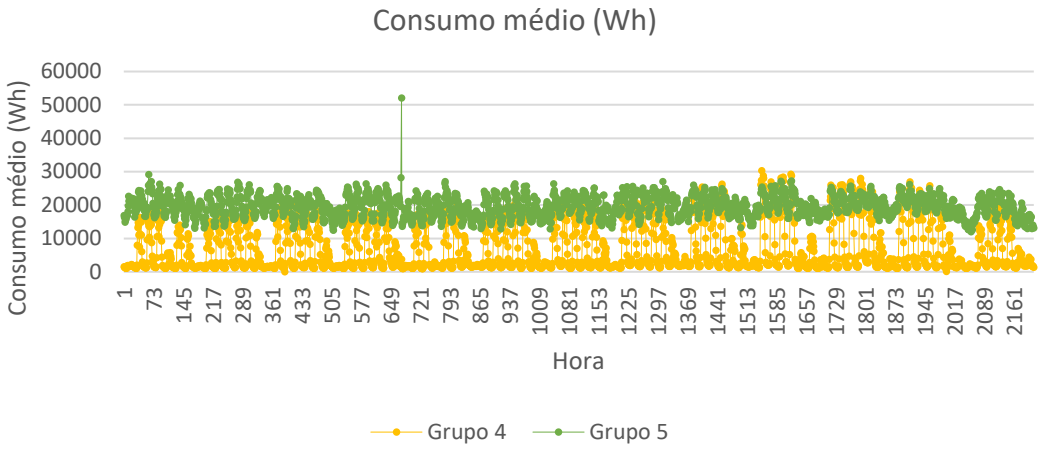

Figura A.3.12: Consumo médio (Wh) para os grupos 4 e 5 no quarto trimestre de 2017, aplicando o método de Ward.

Anexo B.1. Potência Contratada A

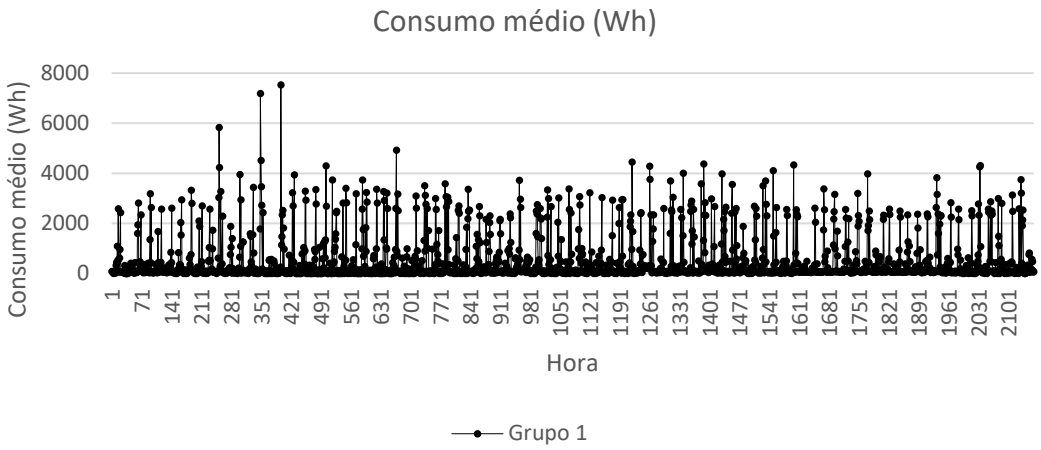

Figura B.1.1: Consumo médio (Wh) para o grupo 1 no primeiro trimestre de 2017, aplicando o método de *k*-means.

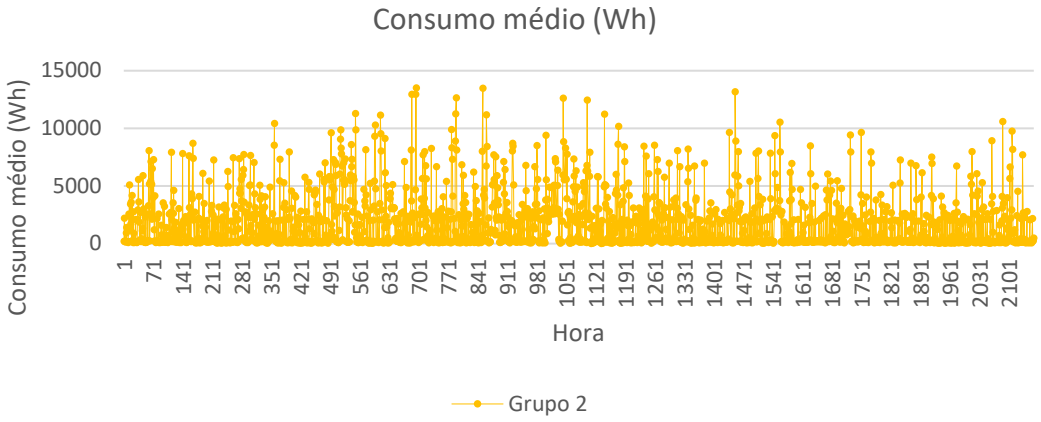

Figura B.1.2: Consumo médio (Wh) para o grupo 2 no primeiro trimestre de 2017, aplicando o método de *k*-means.

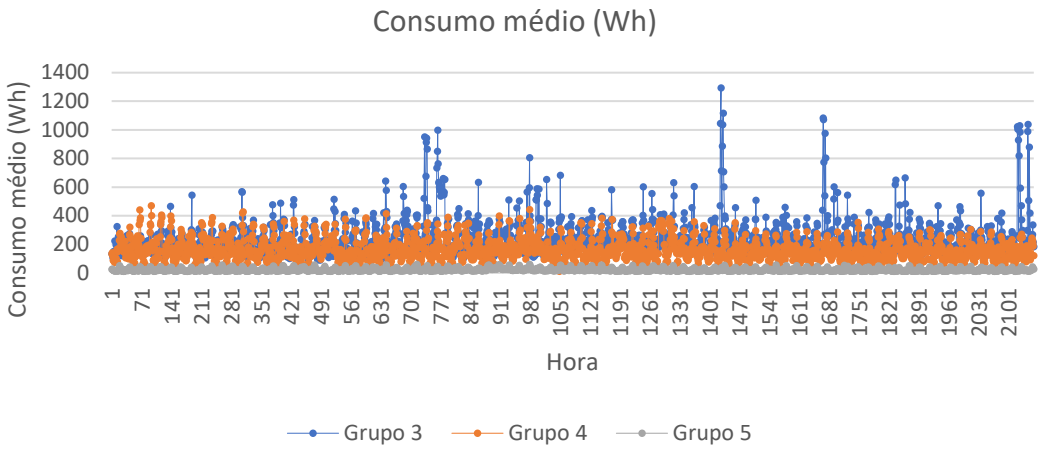

Figura B.1.3: Consumo médio (Wh) para os grupos 3, 4 e 5 no primeiro trimestre de 2017, pelo método de *k*-means.

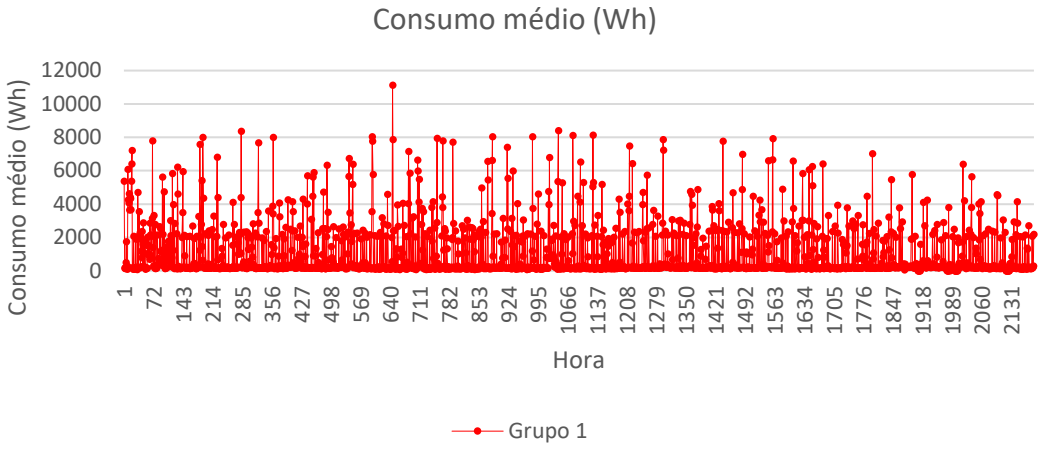

Figura B.1.4: Consumo médio (Wh) para o grupo 1 no segundo trimestre de 2017, aplicando o método de *k*-means.

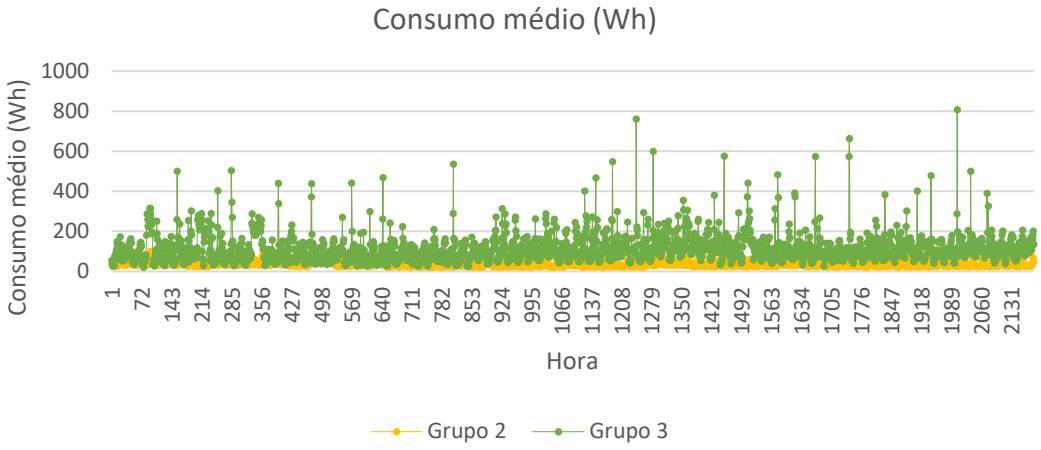

Figura B.1.5: Consumo médio (Wh) para os grupos 2 e 3 no segundo trimestre de 2017, pelo método de *k*-means.

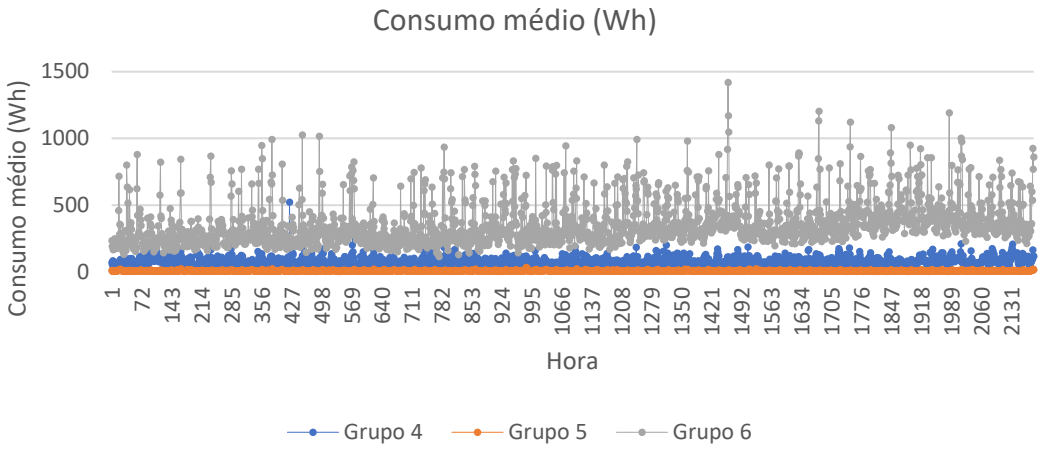

Figura B.1.6: Consumo médio (Wh) para os grupos 4, 5 e 6 no segundo trimestre de 2017, pelo método de *k*-means.

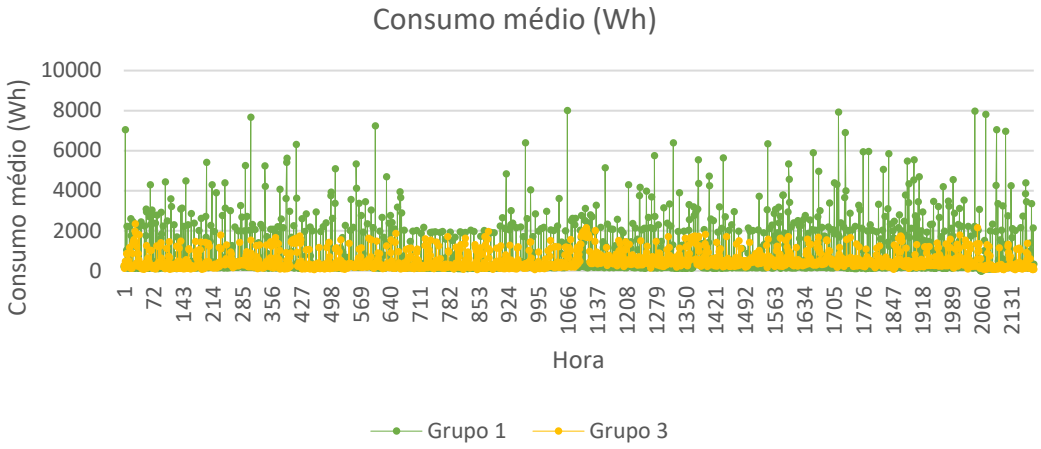

Figura B.1.7: Consumo médio (Wh) para os grupos 1 e 3 no terceiro trimestre de 2017, pelo método de *k*-means.

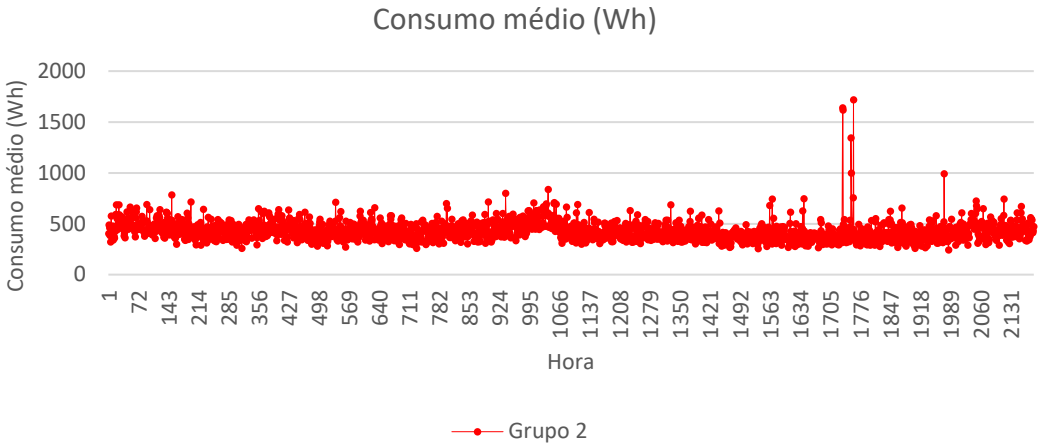

Figura B.1.8: Consumo médio (Wh) para o grupo 2 no terceiro trimestre de 2017, aplicando o método de *k*-means.

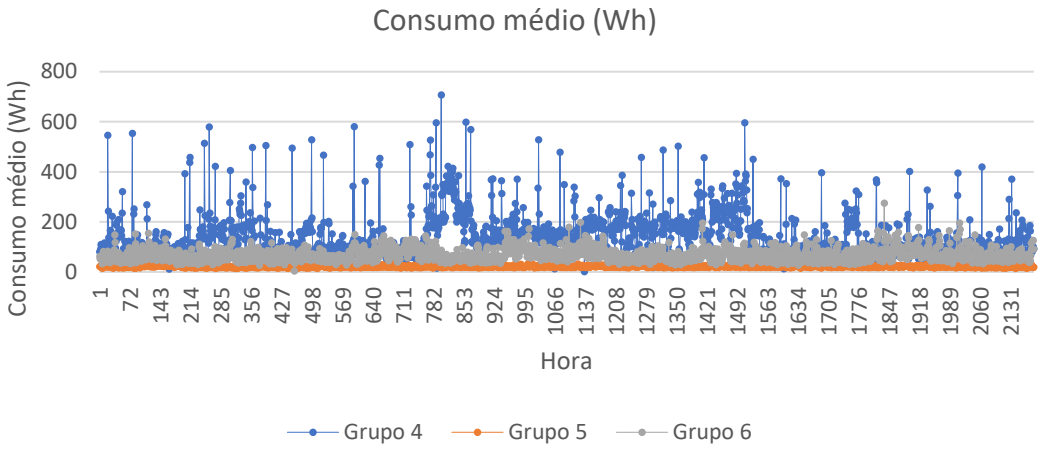

Figura B.1.9: Consumo médio (Wh) para os grupos 4, 5 e 6 no terceiro trimestre de 2017, pelo método de *k*-means.

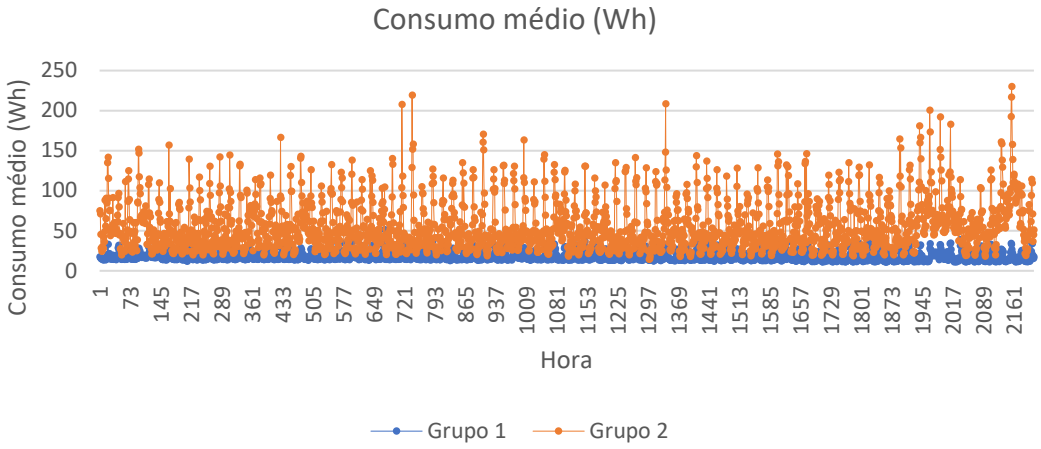

Figura B.1.10: Consumo médio (Wh) para os grupos 1 e 2 no quarto trimestre de 2017, pelo método de k-means.

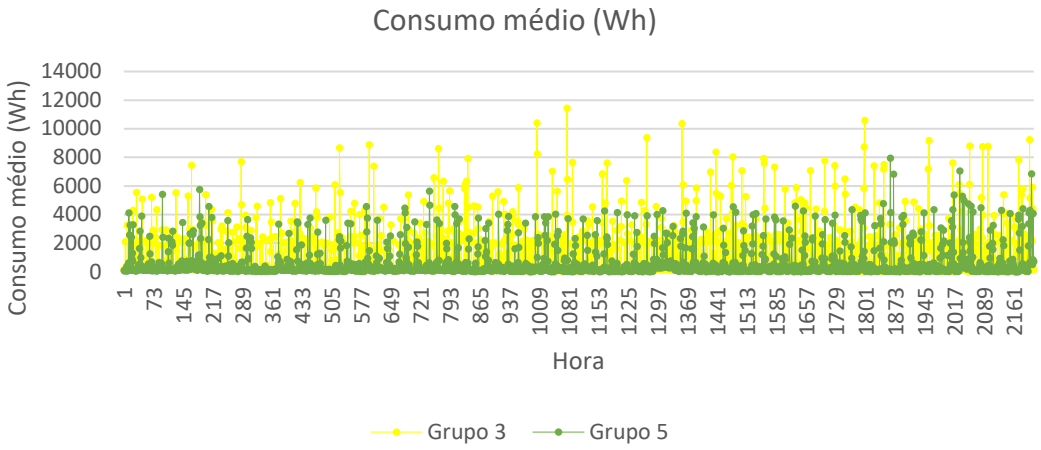

Figura B.1.11: Consumo médio (Wh) para os grupos 3 e 5 no quarto trimestre de 2017, pelo método de *k*-means.

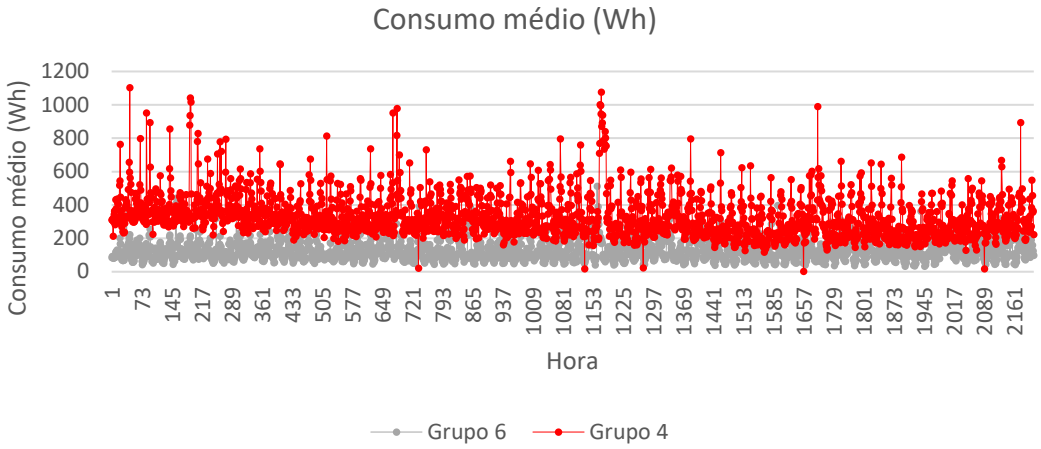

Figura B.1.12: Consumo médio (Wh) para os grupos 4 e 6 no quarto trimestre de 2017, pelo método de *k*-means.

Anexo B.2. Potência Contratada B

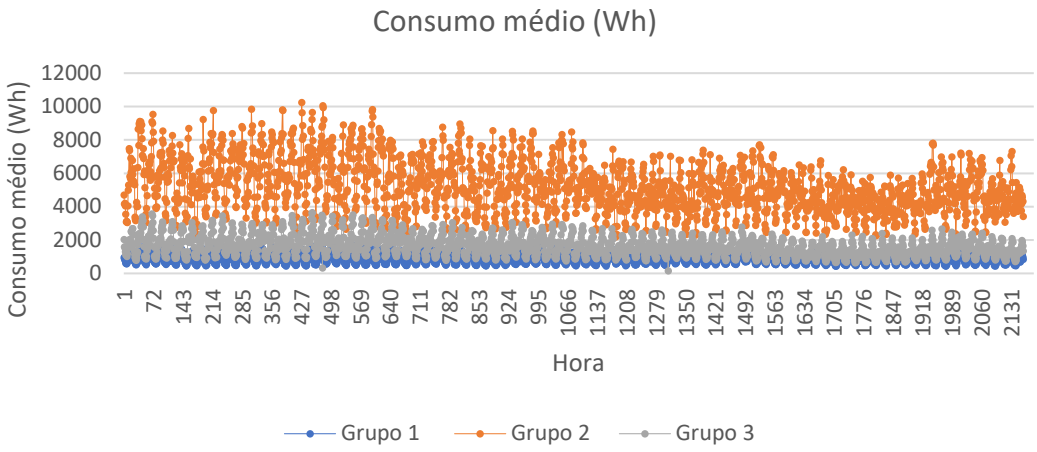

Figura B.2.1: Consumo médio (Wh) para os grupos 1, 2 e 5 no primeiro trimestre de 2017, aplicando o método de *k*-means.

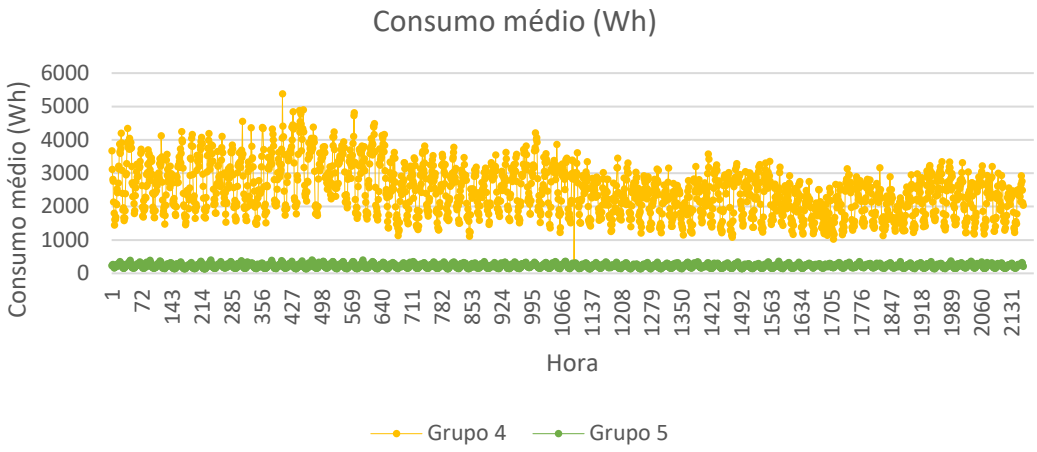

Figura B.2.2: Consumo médio (Wh) para os grupos 4 e 5 no primeiro trimestre de 2017, aplicando o método de *k*-means.

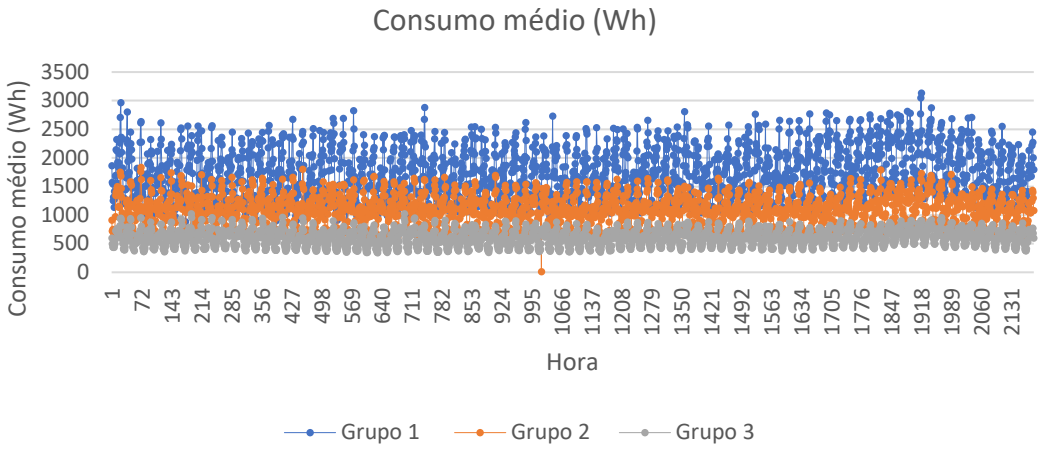

Figura B.2.3: Consumo médio (Wh) para os grupos 1, 2 e 3 no segundo trimestre de 2017, aplicando o método de *k*-means.

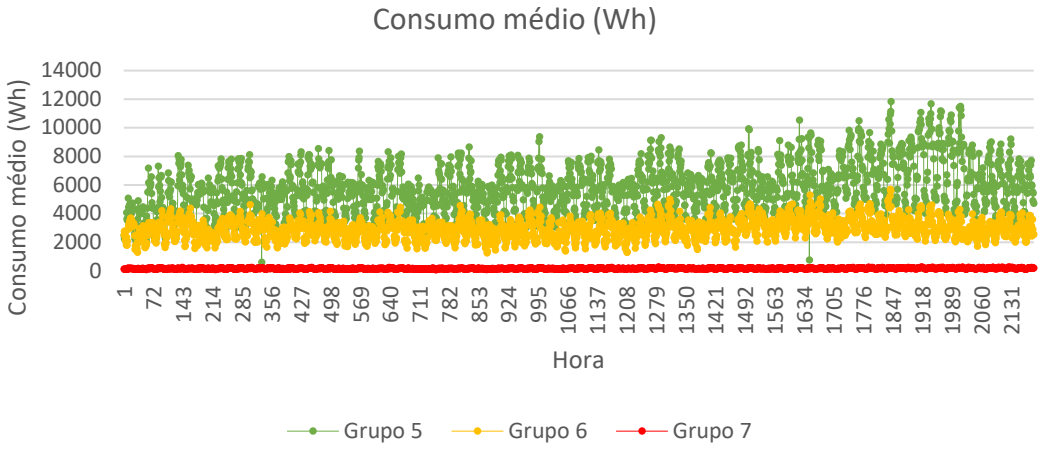

Figura B.2.4: Consumo médio (Wh) para os grupos 5, 6 e 7 no segundo trimestre de 2017, aplicando o método de *k*-means.

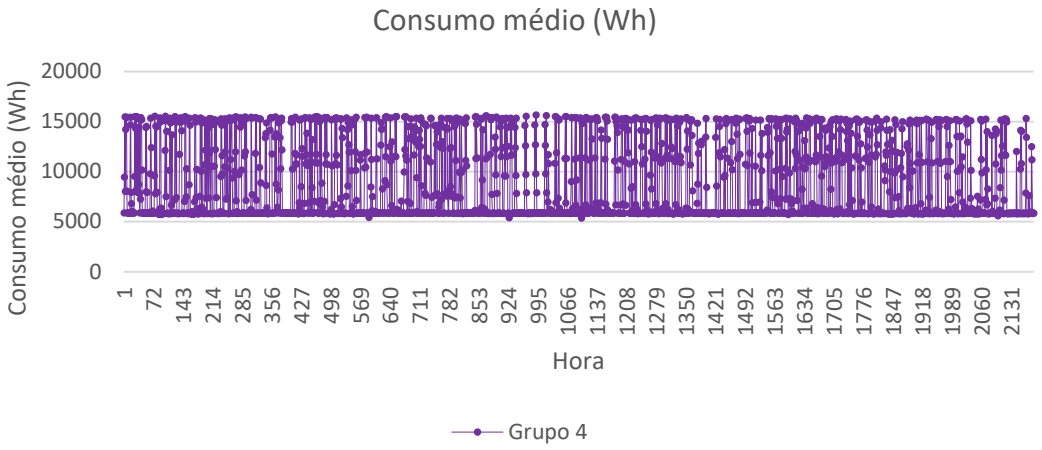

Figura B.2.5: Consumo médio (Wh) para o grupo 4 no segundo trimestre de 2017, aplicando o método de *k*-means.

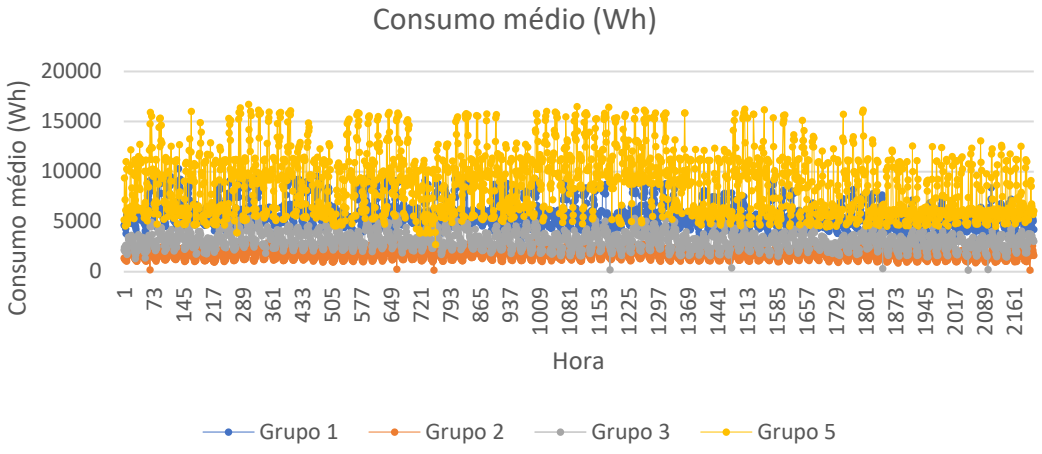

Figura B.2.6: Consumo médio (Wh) para os grupos 1, 2, 3 e 5 no terceiro trimestre de 2017, aplicando o método de *k*-means.

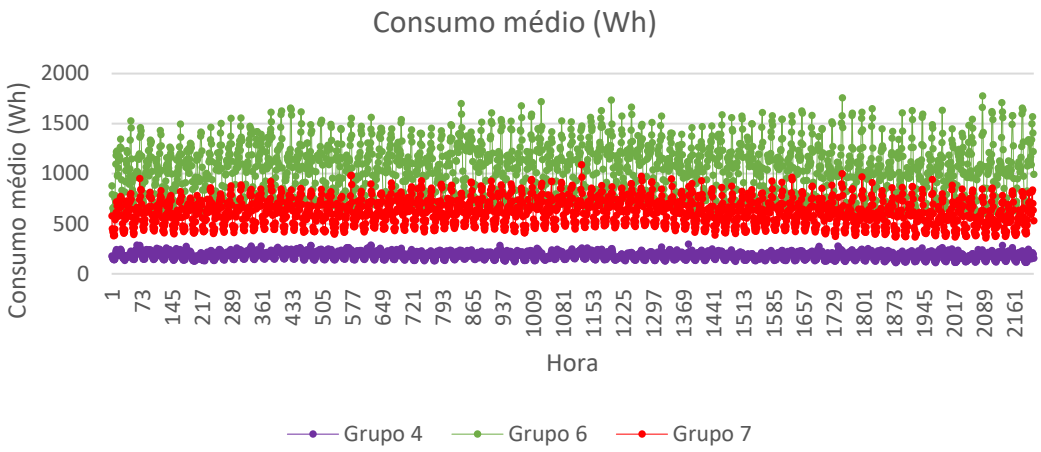

Figura B.2.7: Consumo médio (Wh) para os grupos 4, 6 e 7 no terceiro trimestre de 2017, aplicando o método de *k*-means.

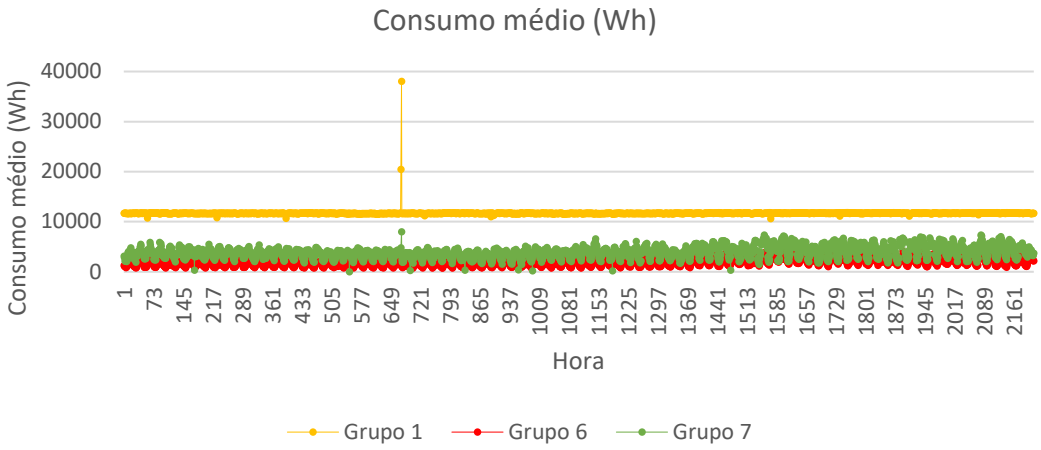

Figura B.2.8: Consumo médio (Wh) para os grupos 1, 6 e 7 no quarto trimestre de 2017, aplicando o método de k-means.

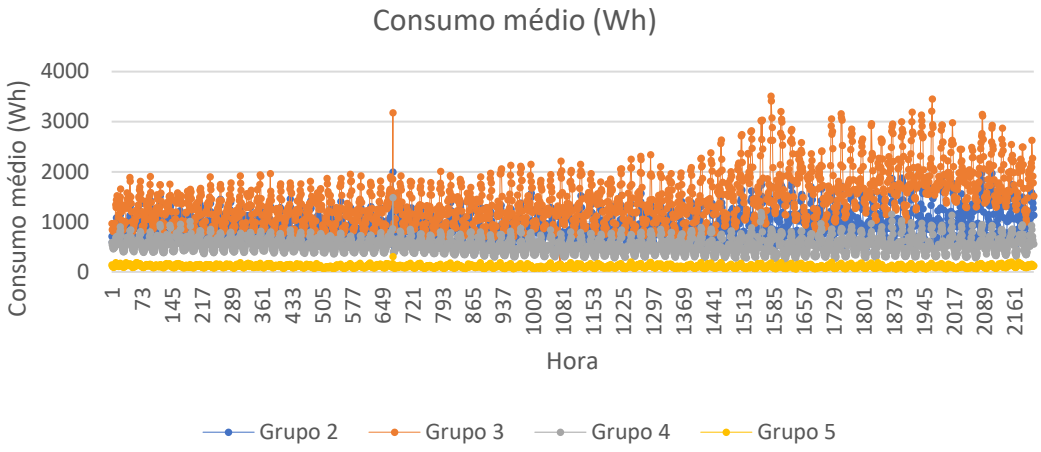

Figura B.2.9: Consumo médio (Wh) para os grupos 2, 3, 4 e 5 no quarto trimestre de 2017, aplicando o método de *k*-means.

Anexo B.3. Potência Contratada J

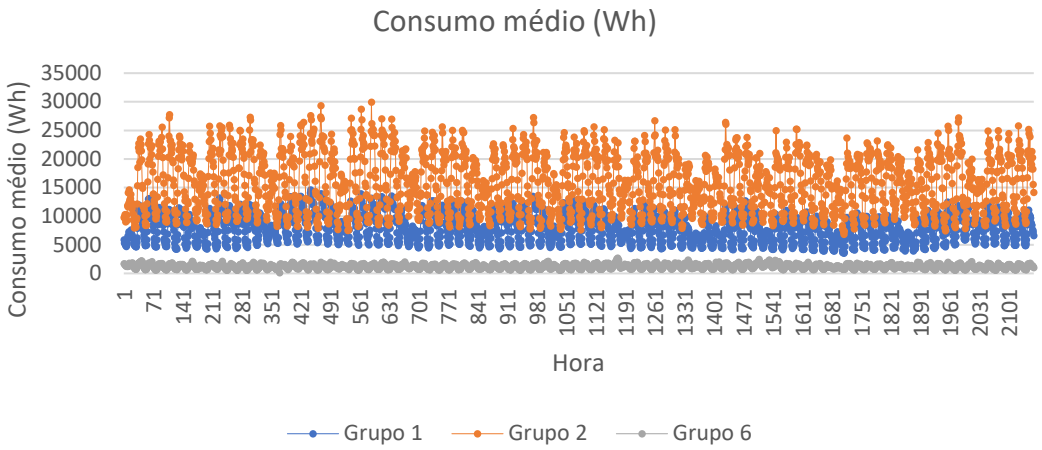

Figura B.3.1: Consumo médio (Wh) para os grupos 1, 2 e 6 no primeiro trimestre de 2017, aplicando o método de *k*-means.

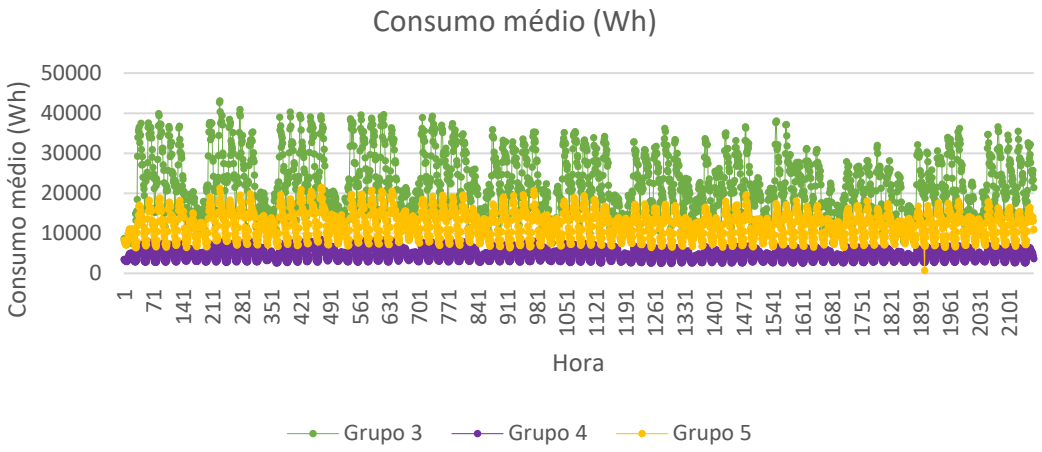

Figura B.3.2: Consumo médio (Wh) para os grupos 3, 4 e 5 no primeiro trimestre de 2017, aplicando o método de *k*-means.

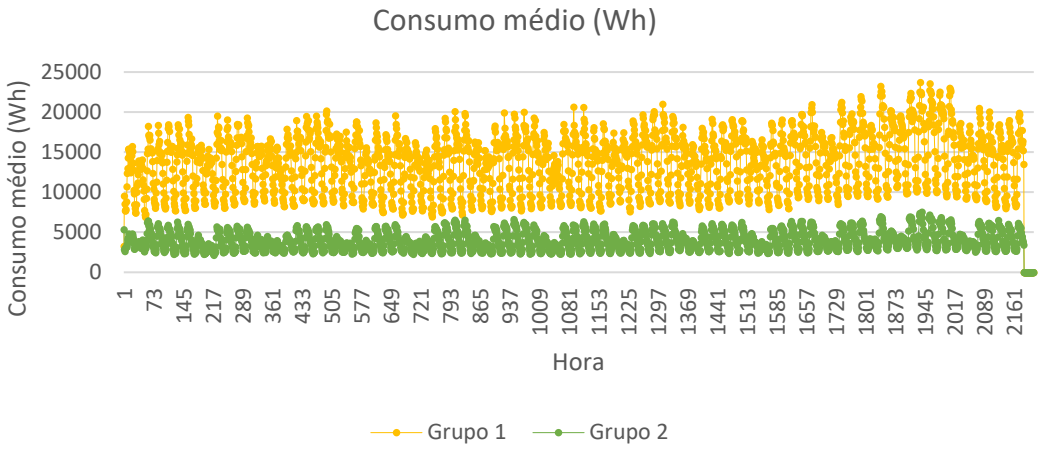

Figura B.3.3: Consumo médio (Wh) para os grupos 1 e 2 no segundo trimestre de 2017, aplicando o método de *k*-means.

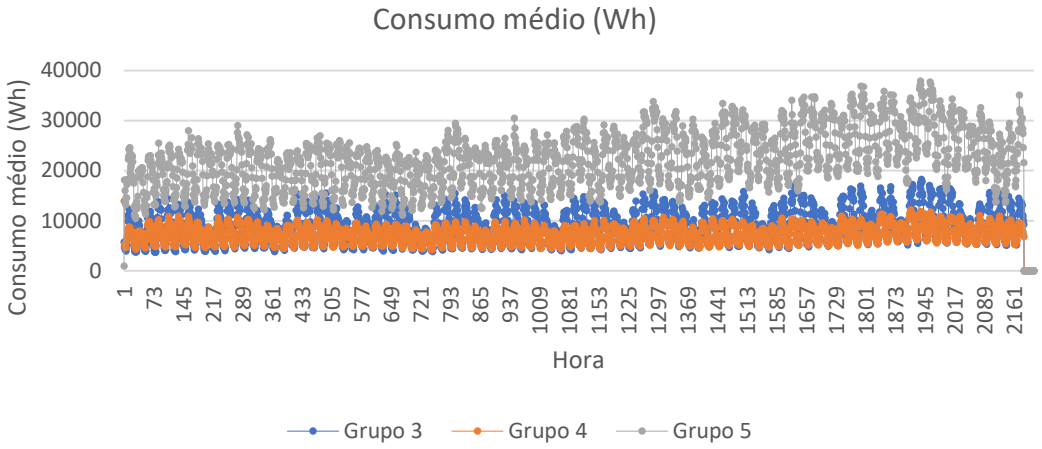

Figura B.3.4: Consumo médio (Wh) para os grupos 3, 4 e 5 no segundo trimestre de 2017, aplicando o método de *k*-means.

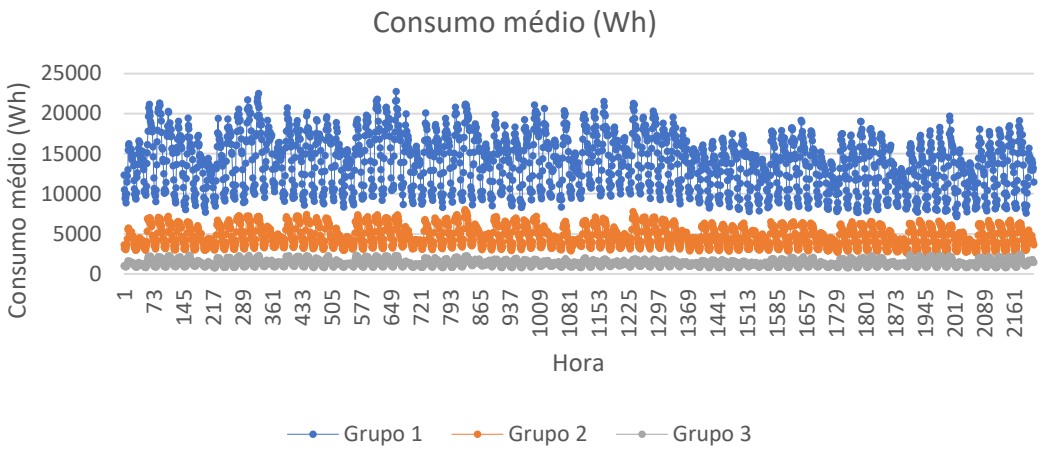

Figura B.3.5: Consumo médio (Wh) para os grupos 1, 2 e 3 no terceiro trimestre de 2017, aplicando o método de *k*-means.

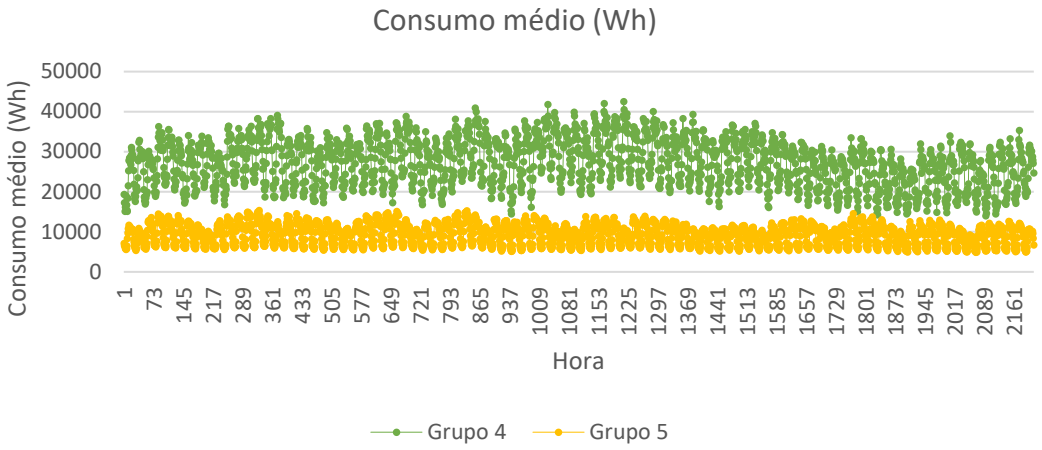

Figura B.3.6: Consumo médio (Wh) para os grupos 4 e 5 no terceiro trimestre de 2017, aplicando o método de *k*-means.

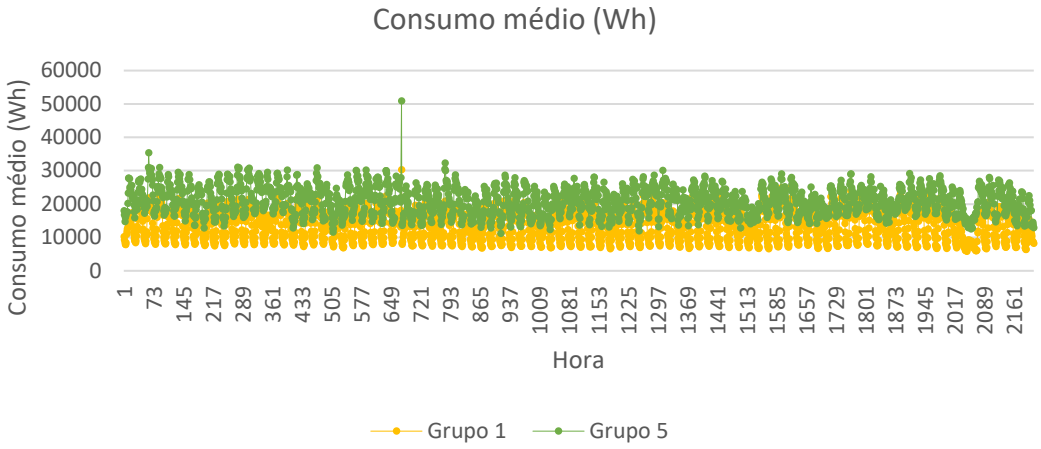

Figura B.3.7: Consumo médio (Wh) para os grupos 1 e 5 no quarto trimestre de 2017, aplicando o método de *k*-means.

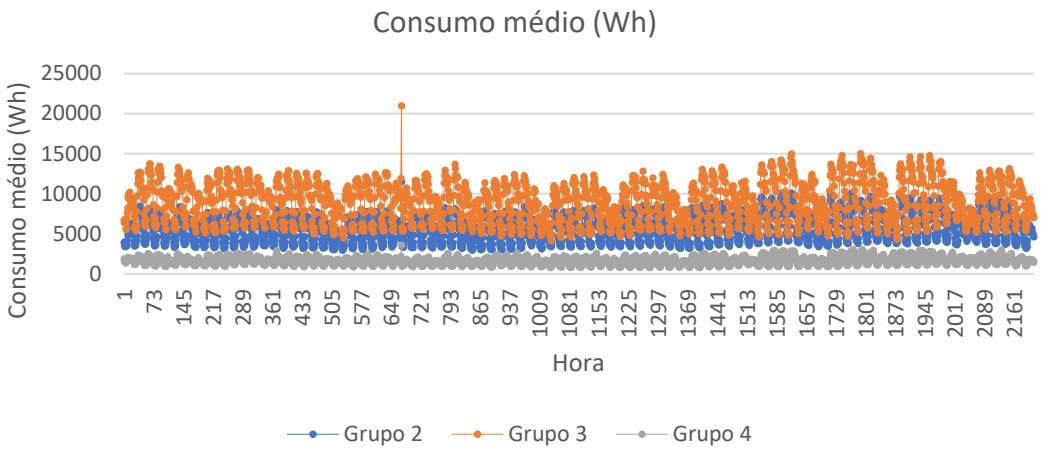

Figura B.3.8: Consumo médio (Wh) para os grupos 2, 3 e 4 no quarto trimestre de 2017, aplicando o método de *k*-means.