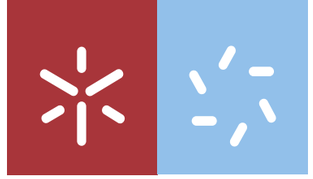


Deteção de anomalias em séries temporais de fluxo de água usando métodos de *machine learning*

Elisa Maria Maio de Araújo

UMinho | 2021

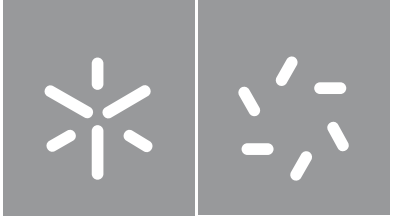


Universidade do Minho
Escola de Ciências

Elisa Maria Maio de Araújo

Deteção de anomalias em séries temporais de fluxo de água usando métodos de *machine learning*

Dezembro de 2021



Universidade do Minho
Escola de Ciências

Elisa Maria Maio de Araújo

Deteção de anomalias em séries temporais de fluxo de água usando métodos de *machine learning*

Dissertação de Mestrado
em Matemática e Computação

Trabalho efetuado sob a orientação do(a)
Professora Doutora
Flora José da Rocha Ferreira

Dezembro de 2021

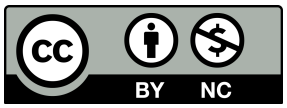
DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Atribuição-NãoComercial

CC BY-NC

`https://creativecommons.org/licenses/by-nc/4.0/`

AGRADECIMENTOS

A elaboração desta dissertação enfatiza o culminar desta jornada de 2 anos que foi o mestrado e, tal não seria possível sem o apoio das pessoas incríveis que me acompanharam durante este percurso, a quem devo um agradecimento especial.

Ao meu namorado e companheiro de uma vida, que me apoiou desde o primeiro momento nesta aventura de voltar a estudar, um enorme obrigada por toda a ajuda e compreensão, pelo apoio incondicional, pelas palavras constantes de incentivo e ânimo e, pela paciência infinita. Serei eternamente grata por tudo o que fez.

Aos meus pais e irmãos, um obrigada especial, pelo amor, carinho e apoio constante, por todos os sacrifícios que já fizeram por mim que permitiram que eu chegasse até aqui e, por estarem sempre presentes quando eu mais preciso, nos bons e maus momentos.

À minha orientadora, Doutora Flora Ferreira, o meu profundo agradecimento, por toda a ajuda e tempo dispensado ao longo deste projeto. Pela partilha de conhecimento, pela disponibilidade a todas as horas, pela compreensão e apoio. Um especial agradecimento pela revisão cuidada da minha dissertação, bem como, pelos comentários e sugestões que permitiram uma grande melhoria da presente dissertação.

Um sincero obrigado à empresa Águas do Norte pela oportunidade em trabalhar neste tema com dados reais. Agradeço especialmente a todos os elementos da equipa que colaborou comigo durante este projeto, muito obrigada por toda a ajuda prestada e disponibilidade.

E, por fim, um obrigada a todos os professores e colegas que, durante estes dois anos me ajudaram e contribuíram para o meu crescimento pessoal e profissional.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

RESUMO

Deteção de anomalias em séries temporais de fluxo de água usando métodos de machine learning

A deteção de anomalias é uma tarefa importante para as empresas de distribuição de água, uma vez que os vazamentos nos sistemas de distribuição de água aumentam significativamente os custos económicos e criam escassez de recursos hídricos. A empresa portuguesa Águas do Norte S.A. com o objetivo de melhorar o processo de monitorização e controlo de perdas na rede de distribuição em 2020 começaram a instalar contadores de água inteligentes nos pontos de consumo. O principal objetivo desta dissertação é a identificação de anomalias (ocorrências que se desviam do padrão considerado normal) presentes no primeiro conjunto de dados fornecido por estes contadores inteligentes, com recurso a técnicas de estatísticas e *machine learning*. Numa primeira fase, procedeu-se ao pré-processamento dos dados onde foram identificadas anomalias ou discrepâncias (valores negativos, muito elevados, duplicados, etc.). De seguida, com o objetivo de identificar os diferentes perfis de consumo dos clientes (483 clientes domésticos e 53 não domésticos), realizou-se uma análise de clusters (com base no algoritmo de *K-Means*) separadamente para cada escala temporal: consumo por hora, por dia, e mês. A análise de componentes principais (PCA) foi aplicada para reduzir a dimensão dos dados antes de aplicar o algoritmo de *K-Means*. Os clientes com comportamentos de consumo semelhante nas três escalas temporais foram agrupados e as curvas de consumo de água foram calculadas para cada grupo. Os resultados obtidos da análise de clusters fornecem para a empresa uma visão resumida do consumo temporal de água dos seus clientes, que podem ser usados para obter estimativas do consumo de água, bem como para detetar consumos fora do padrão normal (que podem ser possíveis anomalias). Tendo em mente a identificação de consumos fora do padrão normal determinou-se os intervalos de consumos habituais, para cada hora de cada perfil de consumo. Os intervalos de consumos foram calculados com base na distribuição de probabilidades que melhor descreve o consumo médio de cada hora em cada perfil. Para cada caso, diferentes funções de distribuição de probabilidades (incluindo Weibull, Gamma, Log-normal e Exponencial) foram investigadas na seleção da que melhor se ajusta aos dados. Por fim, é descrito e exemplificado como estes intervalos de consumos poderão ser usados na deteção de registos de consumos diferentes do esperado.

Palavras-chave: Análise de clusters, Deteção de Anomalias, Consumo de Água, Pré-processamento de Dados

ABSTRACT

Detection of anomalies in water flow time series using machine learning methods

Detecting anomalies is an important task for water distribution companies since leaks in water distribution systems significantly increase economic costs and create scarcity of water resources. The Portuguese company Águas do Norte S.A., with the objective of improving the process of monitoring and controlling losses in the distribution network, in 2020 began installing smart water meters at consumption points. The main objective of this dissertation is to identify anomalies (occurrences that deviate from the normal pattern) present in the first dataset provided by these smart water meters, using statistical and machine learning techniques. In the first phase, data was pre-processed, where anomalies or discrepancies were identified (negative values, very high, duplicates, etc.). Then, in order to identify the different consumption profiles of the customers (483 domestic and 53 non-domestic), a cluster analysis was performed (based on the K-Means algorithm) separately for each scale temporal: hourly, monthly, and monthly consumption. Principal component analysis (PCA) was applied to reduce the data dimension before applying K-means clustering. Customers with similar consumption behaviors across the three temporal scales were grouped and water consumption curves were calculated for each cluster. The results obtained from the cluster analysis provide to the company a summarized view of the temporal water consumption of their customers, which can be used to obtain estimates of water consumption, as well as to detect out of normal consumptions (which may be possible anomalies). Keeping in mind the identification of out-of-normal pattern behavior of customers' water consumption, usual consumption intervals were determined for each hour of each consumption profile. Consumption intervals were calculated based on the probability distribution that best describes the average consumption of each hour in each profile. For each case, different probability distribution functions (including Weibull, Gamma, Log-normal, and Exponential) were investigated in selecting the one that best fits the data. Finally, it is described and exemplified how these consumption intervals can be used to detect consumption records that are different from what was expected.

Keywords: Anomaly Detection, Clusters Analysis, Data Pre-processing, Water Consumption

ÍNDICE

1 Introdução 2
1.1 Enquadramento e motivação 2
1.2 Contextualização e objetivos 2
1.3 Estado da Arte 3
1.4 Estrutura da dissertação 6
2 Base de dados 9
2.1 Conjunto de dados 9
2.2 Pré-processamento dos dados 12
3 Métodos e ferramentas 20
3.1 *Machine learning*: Uma breve introdução 20
3.2 Linguagem de Programação 21
3.3 Técnicas de *machine learning* implementadas 22
3.4 Métodos estatísticos implementados 32
4 Resultados 38
4.1 Anomalias detetadas no pré-processamento dos dados 38
4.2 Identificação de diferentes padrões de consumo 41
4.3 Identificação de ocorrências fora do padrão de consumo habitual 51
5 Conclusões e trabalho futuro 55

LISTA DE FIGURAS

1 Estrutura do *dataset* leituras 10
2 Estrutura do conjunto de dados Cadastro 11
3 Informação utilizada do conjunto de dados Cadastro 12
4 Exemplo de registos do consumo de água em intervalos de 15 minutos de um determinado consumidor 13
5 Exemplo de registos de consumo de água por hora de um determinado consumidor nas primeiras horas de um dia 15
6 Exemplo de registos do consumo médio por hora (em 7 meses) referentes às primeiras horas do dia de um determinado consumidor 16
7 Exemplo de registos do consumo médio em cada dia da semana (em 7 meses) de um determinado consumidor 16
8 Exemplo de registos do consumo médio em cada um dos 7 meses de um determinado consumidor 17
9 Conjunto de dados: consumo médio por hora em consumidores domésticos (em 7 meses) . 18

10 Parte do conjunto de dados para a aplicação do PCA (consumo médio por hora) 24
11 Determinação do número de componentes principais 25
12 Representação de três etapas do algoritmo *k-means* 29
13 Método do Cotovelo: soma de quadrados intra-clusters 31
14 Método do Cotovelo: número ótimo de clusters 31
15 Distribuições utilizadas na analise do comportamento do consumo por hora. 33
16 Ilustração de uma função de distribuição de probabilidades e a posição dos quantis de ordem $\frac{\alpha}{2}$ (L_{lower}) e ordem $1 - \frac{\alpha}{2}$ (L_{upper}). 35
17 Comparação do consumo médio por hora de um local com os intervalos de consumos de cada hora 36

18 Número de ocorrências de *reverse flow* por local 39
19 Exemplo de um pico de consumo provocado por um acerto 40
20 Exemplo de um pico de consumo provocado por uma quebra de registos 40
21 Determinação do número de componentes principais para o consumo médio por dia dos locais domésticos 42
22 Determinação do número de componentes principais para o consumo médio por mês dos locais domésticos 42

23 Consumos médios por hora, dia e mês dos 53 locais não domésticos 43
24 Determinação do número de componentes principais para o consumo médio por hora dos locais não domésticos 44
25 Determinação do número de componentes principais para o consumo médio por dia dos locais não domésticos 45
26 Determinação do número de componentes principais para o consumo médio por mês dos locais não domésticos 45
27 *Clusters* de clientes tendo em conta o consumo médio por hora, dia e mês 46
28 Locais em comum com os *clusters* por consumo médio por dia 47
29 Curvas de consumo dos 6 perfis encontrados 48
30 *Clusters* de clientes tendo em conta o consumo médio por hora, dia e mês 49
31 Locais em comum com os *clusters* por consumo médio por dia 50
32 Curvas de consumo do perfil encontrado 51
33 Ocorrências fora do padrão esperado de um cliente (nos 7 meses em estudo) 52

LISTA DE TABELAS

1 Listagem dos calibres existente 11

1. INTRODUÇÃO

1.1 ENQUADRAMENTO E MOTIVAÇÃO

Nos últimos tempos, a população mundial tem vindo a aumentar exponencialmente, levando a uma enorme preocupação na gestão dos recursos da terra, principalmente, na gestão de um dos recursos mais preciosos: a água. Garantir o acesso à água potável tem sido um dos maiores desafios da humanidade, uma vez que, a água potável desempenha um papel crucial quer para promover a saúde pública, quer para a produção de alimentos ou quer para praticamente todos os aspetos da vida quotidiana. Assim, a preocupação com as perdas nos abastecimentos urbanos de água, das empresas responsáveis pelas redes de abastecimento de água, tem vindo a aumentar. As empresas neste setor, procuram cada vez mais, implementar sistemas de controle e avaliação de anomalias, com o objetivo principal de detetar perdas de água ou fraudes dos consumidores ou erros em todos estes processos (por exemplo, imprecisões do medidor de leitura do cliente).

A deteção de anomalias, nos sistemas de abastecimento de água, tem sido, nos últimos tempos, um campo de pesquisa muito importante. Uma vez que, a perda de água, é um problema não só ambiental e humanitário, mas também, é um problema económico que leva a uma ineficiência de produção, de transporte e de distribuição de um bem essencial como a água. Atualmente, existem cada vez mais , programas de medição inteligente de água para monitorizar e reportar o consumo de água de cada consumidor. Portanto, a quantidade de dados em séries temporais, recolhidos nos sistemas de distribuição de água, tem vindo a aumentar significativamente e, consequentemente, a análise automatizada destes dados torna-se uma necessidade. Assim, com o custo computacional cada vez maior nas últimas décadas, o uso de técnicas de *machine learning* tem sido potencialmente requerido na monitorização destes sistemas Mounce et al. (2011); Braei e Wagner (2020). Mais concretamente, o uso de técnicas de machine learning baseadas na detecção de anomalias, como classificação e deteção de padrões (Chandola et al. (2009); Braei e Wagner (2020)), uma vez que, estas poderão melhorar os sistemas de detecção de fluxo anormal de água (como por exemplo, vazamentos, roturas, roubos, vandalismo, imprecisões) Rahim et al. (2020).

1.2 CONTEXTUALIZAÇÃO E OBJETIVOS

O presente trabalho foi desenvolvido em parceria com a empresa Águas do Norte, S.A. e, tem como principal objetivo, detetar anomalias em séries temporais de fluxo de água usando métodos de estatística e *machine learning*.

A empresa Águas do Norte é a entidade gestora responsável pela captação, tratamento e abastecimento de água para consumo público da região Norte de Portugal, bem como, pela recolha, tratamento e rejeição

de efluentes domésticos, urbanos e industriais e, ainda, de efluentes provenientes de fossas séticas do Norte (2015).

Em 2018, a Águas do Norte S.A., teve perdas que atingiram o volume total de $2.700.000$ metros cúbicos de água não faturada, correspondendo a $3,67\%$ do valor da água ingressada no sistema. Como consequência disto e, decorrente de uma maior conscientização sobre a racionalização dos recursos hídricos, teve a necessidade de implementar novas metodologias de controle de perdas de água, a fim de melhorar a eficiência de seus sistemas Silva et al. (2019). Para tal, desenvolveu uma ferramenta que permitiu obter dados sobre a eficiência de um subsistema de abastecimento de água em tempo real, bem como, a monitorização *online* e preditiva do estado dos equipamentos de medição da Águas do Norte S.A. Silva et al. (2019). Nos meados de 2020, a empresa iniciou o processo de instalação de cerca de 950 contadores de água inteligentes nos diferentes pontos de consumo. Os dados obtidos em cada contador são integrados em tempo real num software da empresa. Um dos objetivos da empresa é converter estes dados em informação útil tanto para os consumidores, que poderão ter um melhor entendimento sobre os seus consumos, assim como para a empresa que através da informação retirada dos dados, poderá melhorar a monitorização e gestão da rede de abastecimento de águas, incluindo a deteção de ocorrências de consumos de água se desviam do padrão de consumo considerado normal.

Pretende-se assim, com este projeto, analisar e tratar estes dados provenientes dos contadores de água inteligentes instalados e, detetar situações anômalas (que poderão ser possíveis fugas de água, fraudes, ou ainda erros de medição) e, também, detetar padrões de consumo.

1.3 ESTADO DA ARTE

Obter um fornecimento de água com o mínimo de perdas de água nos sistemas de distribuição possível, é um dos maiores desafios enfrentados pelas concessionárias de água em todo o mundo. Estima-se que, as perdas de água a nível mundial, atinguem valores na ordem de 50% Sardinha et al. (2017). Liemberger e Wyatt Liemberger e Wyatt (2019) relataram perdas de, aproximadamente, 346 milhões de m^3 por dia, o equivalente a 126 bilhões de m^3 em perdas em um ano, mundialmente. Só na Europa, exitem perdas de, aproximadamente, $26,8$ milhões de m^3 por dia, cerca de 9,8 milhões de m^3 ao ano. Assim, a medição exata e precisa do consumo de água dos clientes é uma componente crucial na sustentabilidade de uma concessionária de água. Por esta razão, durante a última década, surgiram novas tecnologias de medição Arregui et al. (2020) com o intuito de monitorizar o consumo de água dos respetivos clientes para detetar anomalias na distribuição e, consequentemente, detetar perdas de água.

A deteção de anomalias é um tópico crítico em vários setores da indústria, abordando questões como deteção de fraude e manutenção preditiva. Uma quantidade significativa de pesquisas sobre anomalias

baseadas em dados tem sido realizada usando modelagem estatística e, mais recentemente, algoritmos de *Machine Learning*. Hodge e Austin [2004], por exemplo, forneceram um vasto levantamento de técnicas de detecção de anomalias desenvolvidas em *Machine Learning* (tais como: *Classification Based*, *Clustering Based*, *Nearest Neighbor Based*) e em domínios estatísticos. Várias técnicas de detecção de anomalias baseadas em *clustering* foram desenvolvidas ao longo dos tempos. *Clustering* é usado para agrupar instâncias de dados semelhantes em *clusters* Chandola et al. (2009). As técnicas de detecção de anomalias baseadas em *clustering* podem ser agrupadas em diferentes categorias Chandola et al. (2009). A primeira categoria baseia-se no facto de as instâncias de dados normais pertencem a um certo *cluster*, enquanto que, as anomalias não pertencem a nenhum. Estas técnicas aplicam um algoritmo baseado em *cluster* conhecido ao conjunto de dados e, declaram qualquer instância de dados que não pertença a nenhum cluster como anômala. Existem vários algoritmos de *clustering* que não forçam todas as instâncias de dados a pertencerem a um *cluster*, como DBSCAN Ester et al. (1996), ROCK Guha et al. (2000) e o agrupamento SNN Ertöz et al. (2003). O algoritmo FindOut é uma extensão do algoritmo Wave-Cluster em que os *clusters* detectados são removidos dos dados e as instâncias residuais são declaradas como anomalias Chandola et al. (2009). Uma desvantagem de tais técnicas é o facto de que elas não são otimizadas para encontrar anomalias, uma vez que, o objetivo principal do algoritmo de agrupamento é encontrar *clusters*. A segunda categoria baseia-se no facto de que instâncias de dados normais ficam perto de seu centróide de *cluster* mais próximo, enquanto que, as anomalias estão longe do seu centróide de cluster mais próximo. E, estas técnicas consistem em duas etapas diferentes. Na primeira etapa, os dados são agrupados através de um algoritmo de agrupamento. Na segunda etapa, para cada instância de dados, a distância até ao seu centróide mais próximo é calculada como uma pontuação de anomalia. Uma série de técnicas de detecção de anomalias, que seguem esta abordagem de duas etapas, foram propostas usando diferentes algoritmos de agrupamento. Smith estudou *Self-Organizing Maps* (SOM), *K-means Clustering* e *Expectation-Maximization* (EM) para agrupar os dados de treino e, em seguida, usar os *clusters* para classificar os dados de teste Chandola et al. (2009). No entanto, observa-se que, se os dados anomalos formarem *clusters*, as técnicas mencionadas anteriormente não serão capazes de detetar anomalias. Para resolver este problema, uma terceira categoria de técnicas baseadas em *clustering* foi proposta e, baseia-se no facto de instâncias de dados normais pertencem a *clusters* grandes e densos, enquanto que as anomalias pertencem a *clusters* pequenos ou esparsos Chandola et al. (2009). Foram várias as variações propostas a esta terceira categoria (Pires e Santos-Pereira (2005); Eskin et al. (2002); He et al. (2003)). A técnica proposta em He et al. (2003), por exemplo, denominado de *FindCBLOF*, atribui uma pontuação de anomalia conhecida como *Cluster-Based Local Outlier Factor* (CBLOF) para cada instância de dados. A pontuação CBLOF captura o tamanho do *cluster* ao qual a instância de dados pertence, bem como a distância da instância de dados ao seu centróide.

Devido à complexidade dos sistemas de distribuição de água, várias anomalias podem ocorrer devido a diferentes causas subjacentes, por exemplo, vazamento de água, bloqueio da tubagem, condição do

dispositivo ou eficiência energética Duan e Lee (2016); Duan et al. (2017); Duan e Lee (2016). A deteção de uma anomalia, num sistema de distribuição de água, requer muitas vezes de uma combinação apropriada do conhecimento do domínio humano, de dados observacionais e de técnicas de modelagem precisas. Um dos desafios no desenvolvimento de modelos inteligentes de monitorização, planeamento do consumo de água e deteção de anomalias é a heterogeneidade dos consumidores atendidos pelo sistema regional de água Avni et al. (2015a). Embora seja típico para empresas de energia e telecomunicações classificar os seus clientes em grupos que têm padrões de consumo semelhantes, tendo em consideração o seu consumo anual, tais práticas não são comuns para concessionárias de água Yang et al. (2018); Laspidou et al. (2015); Ioannou et al. (2017).

Os estudos sobre o reconhecimento do padrão de consumo de água classificando os clientes em diferentes segmentos é limitado Yang et al. (2018); Laspidou et al. (2015); Avni et al. (2015b); Ioannou et al. (2021).

Recentemente em Ioannou et al. (2021) uma metodologia baseada em *Self Organizing Maps* (SOM) combinada com métodos de agrupamento (*K-means and Hierarchical Agglomerative Clustering*), foi aplicada a um conjunto de dados de consumo de água doméstico, consumo este registado de meia hora em meia hora, para 21 clientes durante um período de 7 meses. Clientes individuais com um comportamento de consumo semelhante foram assim agrupados e as curvas de consumo de água foram calculadas para cada agrupamento. Os resultados deste estudo, revelaram melhores estimativas do uso de água pelo cliente quando o *clustering* foi empregado, em comparação com as previsões quando o *clustering* não foi aplicado.

A curva de um perfil de consumo de água mostra a quantidade de água que um cliente gasta ao longo do tempo. Por analogia com as curvas de carga de eletricidade, os perfis de consumo de água podem fornecer estimativas da distribuição temporal e espacial do consumo, dando assim uma visão dos picos de consumo em diferentes locais Räsänen et al. (2010).

Uma deteção de uma anomalia é uma identificação de pontos dos dados que se comportam de maneira diferente dos padrões normais ou expectáveis, ou seja, é uma identificação de *outliers*. A deteção de anomalias (ou identificação de *outliers*) com base num intervalo de predição construído a partir de desvios padrão que abrangem a média é uma prática comum. Em vez de usar a média como o indicador central para deteção de *outliers*, vários métodos baseados em mediana foram alternativamente propostos devido à sua robustez Leys et al. (2013). Particularmente, métricas estatísticas robustas, mediana e desvio absoluto da mediana (MAD), foram introduzidos para detetar anomalias com precisão em Hochenbaum et al. (2017). A estimativa da MAD é comumente usada para a análise de anomalias devido ao seu desempenho desejável. É insensível a uma grande dispersão de anomalias. Além disso, é fácil de implementar e apenas um pequeno esforço computacional é necessário Mehrang et al. (2015). Outro importante método baseado em estatística é a predição do intervalo de confiança que se baseia

em dados vizinhos para identificar anomalias de uma série temporal. Um valor observado é classificado como anomalo se ficar fora da faixa dos limites superior e inferior do intervalo Hill e Minsker (2010). A sua principal vantagem é a capacidade de fornecer um limite automático para diferenciar entre pontos normais (expectáveis) e anormais (não expectáveis). Devido às distribuições de dados contrastantes entre várias fontes, um ajuste de parâmetros e uma estimativa dos limites do intervalo são os principais desafios. Estudos anteriores aplicaram diferentes técnicas baseadas na predição do intervalo de valores expectáveis para detetar padrões de anomalias em séries temporais hidrológicas Kulanuwat et al. (2021); Zhao et al. (2018). Por exemplo, em Zhao et al. (2018), foi proposto um método de controlo da qualidade dos dados hidrológicos que incluiu dois modelos de previsão e um modelo estatístico para determinarem três intervalos de confiança. Especificamente, os intervalos de confiança para detetar potenciais *outliers*. Os resultados mostraram que o método proposto fornece um desempenho desejável na deteção de dados suspeitos e erróneos. No entanto, para determinar os intervalos de confiança assim intervalos de consumos habituais é importante saber qual a distribuição de probabilidade que melhor se aproxima dos dados. Em estudos anteriores Mehrang et al. (2015); Kofinas et al. (2018); Kossieris e Makropoulos (2018); Surendran e Tota-Maharaj (2018); Gargano et al. (2017); Surendran e Tota-Maharaj (2020); Santos et al. (2021) as distribuições Gama, Weibull, Log-Normal, Log-Logística e Exponencial são identificadas como as mais adequadas para modelar variáveis hidrológicas, incluindo o consumo de água.

1.4 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação encontra-se dividida em cinco capítulos: Introdução, Base de dados, Metodologia, Resultados, conclusões e trabalho futuro. O conteúdo de cada capítulo é brevemente descrito nos seguintes tópicos:

- Capítulo 1: Introdução

  Este capítulo inicia com uma introdução ao problema abordado nesta dissertação e motivação, seguido da contexttualização e objetivos, onde é feita uma breve apresentação da empresa que colaborou com este projeto. Neste capítulo também é apresentado o estado da arte, onde é apresentado uma descrição de estudos anteriores relacionados com o tema desta dissertação, nomeadamente na deteção de anomalias em séries temporais de fluxo de água utilizando métodos estatísticos e de machine learning.

- Capítulo 2: Base de dados

  Neste capítulo é feita uma descrição da base de dados fornecida pela empresa para a realização deste estudo, seguido do pré-processamento dos dados efetuado.

- Capítulo 3: Métodos e ferramentas

  Neste capítulo é feita inicialmente uma breve introdução sobre *machine learning*, seguidamente é apresentada a linguagem de programação (*Python* e suas bibliotecas) usado na implementação dos métodos. Posteriormente, é feito uma contextualização teórica dos métodos de *machine learning* e estatísticos usados neste estudo, assim como é feito uma descrição de como estes foram implementados.

- Capítulo 4: Resultados

  Neste capítulo são apresentados todos os resultados obtidos nas três fases do estudo realizado: 1) Os resultados sobre as anomalias detetadas no pré-processamento dos dados; 2) Os resultados da análise de clusters realizada para identificação de diferentes padrões de consumos; 3) Alguns resultados obtidos na identificação de ocorrências fora do padrão de consumo habitual com base em intervalos de consumos. Parte dos resultados apresentados na fase 2 foram publicados em Araújo et al. (2021).

- Capítulo 5: Conclusões e trabalho futuro

  Neste capítulo as conclusões, limitações e trabalho futuro são apresentadas.

2. BASE DE DADOS

2.1 CONJUNTO DE DADOS

Os dados utilizados para o desenvolvimento deste projeto foram fornecidos pela empresa Águas do Norte S.A.. São ficheiros do tipo csv (*Comma-separated values*) onde, cada ficheiro, contém informações sobre os consumidores da empresa.

Os ficheiros em questão dividem-se em três categorias denominadas (pela empresa) por:

- **Leituras**: ficheiros que contêm informações sobre as leituras de cada consumidor;
- **Cadastro**: ficheiros que contêm informações sobre os consumidores;
- **Alarmes**: ficheiros que contêm informações sobre alarmes destacados pela empresa.

Esta última categoria, alarmes, não foi utilizada no presente estudo e, portanto, não será apresentada.

2.1.1 Conjunto de dados: Leituras

Os ficheiros que estão inseridos na categoria Leituras, estão separados por dias e, contêm leituras (em litros) desde o dia 24/09/2020 até ao dia 06/07/2021. O conjunto de dados Leituras resulta assim do agrupamento de todos estes ficheiros num só. Cada leitura é realizada em um intervalo de 15 minutos e, portanto, cada linha deste conjunto de dados contém uma leitura num determinado instante juntamente com os valores das restantes variáveis deste conjunto de dados, uma vez que, para além do registo da leitura, este conjunto de dados contém as seguintes variáveis:

- **Local**: representa o local do contador do consumidor em estudo (identificação do consumidor);
- **Data/Hora**: representa a data e a hora do registo da leitura;
- **Módulo**: representa uma outra forma, idêntica ao local, de identificar o consumidor em estudo.

Na Figura 1, pode ser visualizada uma pequena amostra do conjunto de dados Leituras. Este conjunto de dados contém 19109099 registos e 929 locais.

Index	Local	Data/Hora	Módulo	Leitura
0	819450	2021-03-30 01:15:00	12810	640708
1	819450	2021-03-30 01:30:00	12810	640708
2	819450	2021-03-30 01:45:00	12810	640708
3	819450	2021-03-30 02:00:00	12810	640708
4	819450	2021-03-30 02:15:00	12810	640708
5	819450	2021-03-30 02:30:00	12810	640708
6	819450	2021-03-30 02:45:00	12810	640708
7	819450	2021-03-30 03:00:00	12810	640708
8	819450	2021-03-30 03:15:00	12810	640708
9	819450	2021-03-30 03:30:00	12810	640708

(...)

Figura 1. Estrutura do *dataset* leituras

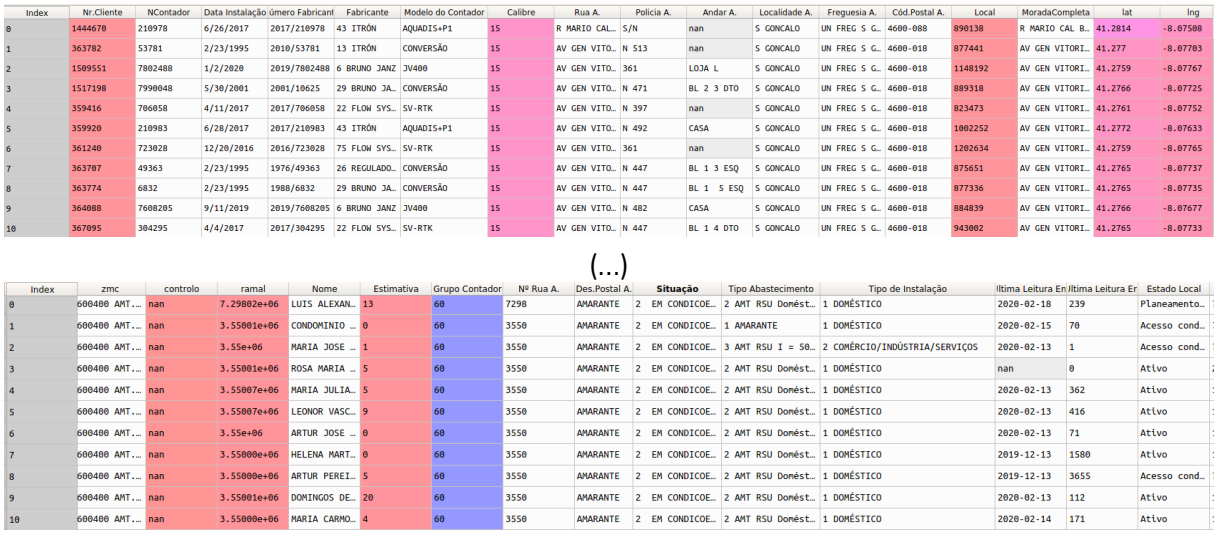
2.1.2 Conjunto de dados: Cadastro

Os ficheiros que estão inseridos na categoria Cadastro foram agrupados num só conjunto de dados dando origem ao conjunto de dados Cadastro. Este conjunto de dados contém informações sobre cada um dos 929 consumidores, tais como o número de cliente, informações sobre o contador (número de contador, data de instalação, fabricante, modelo e calibre), informações geográficas, tipo de instalação, tipo de abastecimento, entre outros. Isto é, contém informações que permitem fazer a correspondência entre o cliente e a leitura através da chave: local ou módulo. A Figura 2 contém uma pequena parte deste conjunto de dados.

De todas estas informações fornecidas por este conjunto de dados, para o estudo em causa, foram utilizadas: o calibre do contador e o tipo de instalação.

O **calibre** representa o valor do diâmetro da entrada do contador. Este valor depende das necessidades de cada consumidor. Isto é, o tipo de calibre varia caso seja um cliente que necessite de muita ou de pouca quantidade de água.

Existem 8 tipos diferentes de calibres e, para cada um, existe um limite máximo de volume de água por hora (em m^3/h) possível a passar pelo contador. A Tabela 1 contém as informações dos 8 tipos de calibres existentes.



Index	Nr.Cliente	NContador	Data Instalação	úmero Fabricant	Fabricante	Modelo do Contador	Calibre	Rua A.	Polícia A.	Andar A.	Localidade A.	Freguesia A.	Cód.Postal A.	Local	MoradaCompleta	lat	lng
0	1444670	210978	6/26/2017	2017/210978	43 ITRÓN	AQUADIS+P1	15	R MARIO CAL…	S/N	nan	S GONCALO	UN FREG S G…	4600-088	890138	R MARIO CAL B…	41.2814	-8.07508
1	363782	53781	2/23/1995	2010/53781	13 ITRÓN	CONVERSÃO	15	AV GEN VITO…	N 513	nan	S GONCALO	UN FREG S G…	4600-018	877441	AV GEN VITORI…	41.277	-8.07703
2	1509551	7802488	1/2/2020	2019/7802488	6 BRUNO JANZ	JV400	15	AV GEN VITO…	361	LOJA L	S GONCALO	UN FREG S G…	4600-018	1148192	AV GEN VITORI…	41.2759	-8.07767
3	1517198	7990048	5/30/2001	2001/10625	29 BRUNO JA…	CONVERSÃO	15	AV GEN VITO…	N 471	BL 2 3 DTO	S GONCALO	UN FREG S G…	4600-018	889318	AV GEN VITORI…	41.2766	-8.07725
4	359416	706058	4/11/2017	2017/706058	22 FLOW SYS…	SV-RTK	15	AV GEN VITO…	N 397	nan	S GONCALO	UN FREG S G…	4600-018	823473	AV GEN VITORI…	41.2761	-8.07752
5	359920	210983	6/28/2017	2017/210983	43 ITRÓN	AQUADIS+P1	15	AV GEN VITO…	N 492	CASA	S GONCALO	UN FREG S G…	4600-018	1002252	AV GEN VITORI…	41.2772	-8.07633
6	361240	723028	12/20/2016	2016/723028	75 FLOW SYS…	SV-RTK	15	AV GEN VITO…	361	nan	S GONCALO	UN FREG S G…	4600-018	1202634	AV GEN VITORI…	41.2759	-8.07765
7	363707	49363	2/23/1995	1976/49363	26 REGULADO…	CONVERSÃO	15	AV GEN VITO…	N 447	BL 1 3 ESQ	S GONCALO	UN FREG S G…	4600-018	875651	AV GEN VITORI…	41.2765	-8.07737
8	363774	6832	2/23/1995	1988/6832	29 BRUNO JA…	CONVERSÃO	15	AV GEN VITO…	N 447	BL 1 5 ESQ	S GONCALO	UN FREG S G…	4600-018	877336	AV GEN VITORI…	41.2765	-8.07735
9	364088	7608205	9/11/2019	2019/7608205	6 BRUNO JANZ	JV400	15	AV GEN VITO…	N 482	CASA	S GONCALO	UN FREG S G…	4600-018	884839	AV GEN VITORI…	41.2766	-8.07677
10	367095	304295	4/4/2017	2017/304295	22 FLOW SYS…	SV-RTK	15	AV GEN VITO…	N 447	BL 1 4 DTO	S GONCALO	UN FREG S G…	4600-018	943002	AV GEN VITORI…	41.2765	-8.07733

(...)

Index	zmc	controlo	ramal	Nome	Estimativa	Grupo Contador	Nº Rua A.	Des.Postal A.	Situação	Tipo Abastecimento	Tipo de Instalação	Ultima Leitura En	Ultima Leitura Er	Estado Local
0	600400 AMT…	nan	7.29802e+06	LUIS ALEXAN…	13	60	7298	AMARANTE	2 EM CONDICOE…	2 AMT RSU Domést…	1 DOMÉSTICO	2020-02-18	239	Planeamento…
1	600400 AMT…	nan	3.55001e+06	CONDOMINIO …	0	60	3550	AMARANTE	2 EM CONDICOE…	1 AMARANTE	1 DOMÉSTICO	2020-02-15	70	Acesso cond…
2	600400 AMT…	nan	3.55e+06	MARIA JOSE …	1	60	3550	AMARANTE	2 EM CONDICOE…	3 AMT RSU I = 50…	2 COMÉRCIO/INDÚSTRIA/SERVIÇOS	2020-02-13	1	Acesso cond…
3	600400 AMT…	nan	3.55001e+06	ROSA MARIA …	5	60	3550	AMARANTE	2 EM CONDICOE…	2 AMT RSU Domést…	1 DOMÉSTICO	nan	0	Ativo
4	600400 AMT…	nan	3.55007e+06	MARIA JULIA…	5	60	3550	AMARANTE	2 EM CONDICOE…	2 AMT RSU Domést…	1 DOMÉSTICO	2020-02-13	362	Ativo
5	600400 AMT…	nan	3.55007e+06	LEONOR VASC…	9	60	3550	AMARANTE	2 EM CONDICOE…	2 AMT RSU Domést…	1 DOMÉSTICO	2020-02-13	416	Ativo
6	600400 AMT…	nan	3.55e+06	ARTUR JOSE …	0	60	3550	AMARANTE	2 EM CONDICOE…	2 AMT RSU Domést…	1 DOMÉSTICO	2020-02-13	71	Ativo
7	600400 AMT…	nan	3.55000e+06	HELENA MART…	0	60	3550	AMARANTE	2 EM CONDICOE…	2 AMT RSU Domést…	1 DOMÉSTICO	2019-12-13	1580	Ativo
8	600400 AMT…	nan	3.55000e+06	ARTUR PEREI…	5	60	3550	AMARANTE	2 EM CONDICOE…	2 AMT RSU Domést…	1 DOMÉSTICO	2019-12-13	3655	Acesso cond…
9	600400 AMT…	nan	3.55001e+06	DOMINGOS DE…	20	60	3550	AMARANTE	2 EM CONDICOE…	2 AMT RSU Domést…	1 DOMÉSTICO	2020-02-13	112	Ativo
10	600400 AMT…	nan	3.55000e+06	MARIA CARMO…	4	60	3550	AMARANTE	2 EM CONDICOE…	2 AMT RSU Domést…	1 DOMÉSTICO	2020-02-14	171	Ativo

(...)

Figura 2. Estrutura do conjunto de dados Cadastro

Tabela 1. Listagem dos calibres existente

Calibre do contador	Caudal máximo (m^3/h)
DN 15	3.125
DN 20	5
DN 25	7.5875
DN 30	12.5
DN 32	12.5
DN 40	20
DN 50	32.25
DN 65	50

O **tipo de instalação** consiste na classificação do tipo de consumidor, isto é, diferencia um consumidor doméstico de um consumidor de não doméstico (comércio, escolas, etc.). Portanto, cada local está associado ao seu correspondente tipo de instalação. Esta informação é importante porque agrupa os consumidores em categorias e, por essa razão, foi tida em conta no presente estudo.

Os vários tipos de instalação dividem-se em 10 tipologias e, estão apresentados no conjunto de dados Cadastro da seguinte forma:

- "1 DOMÉSTICO";
- "2 COMÉRCIO/INDÚSTRIA/SERVIÇOS";
- "3 IPSS/ISF";
- "4 AUTARQUIAS";

- “5 ESTADO”;
- “7 TARIFA FAMÍLIA – 5 PESSOAS”;
- “10 TARIFA SOCIAL – TS1”
- “11 TARIFA SOCIAL – TS2”
- “14 JUNTAS DE FREGUESIA”
- “15 OUTROS”

Assim, do conjunto de dados Cadastro, foi retirada a informação do tipo de instalação e do tipo de calibre. A Figura 3 contém alguns exemplos do que foi utilizado.

Local	Tipo de Instalação	Calibre
1202634	1 DOMÉSTICO	15
875651	1 DOMÉSTICO	15
877336	1 DOMÉSTICO	15
884839	1 DOMÉSTICO	15
943002	1 DOMÉSTICO	15
951404	5 ESTADO	30
966380	1 DOMÉSTICO	15
992429	1 DOMÉSTICO	15
1201018	2 COMÉRCIO/INDÚSTRIA/SERVIÇOS	15
1028707	2 COMÉRCIO/INDÚSTRIA/SERVIÇOS	15
1023047	5 ESTADO	30
967394	1 DOMÉSTICO	15
1028715	1 DOMÉSTICO	15

(...)

Figura 3. Informação utilizada do conjunto de dados Cadastro

2.2 PRÉ-PROCESSAMENTO DOS DADOS

Antes da aplicação de qualquer técnica para a procura de anomalias nos dados em estudo, é necessário fazer um pré-processamento e análise dos mesmos.

Numa primeira análise, alguns dos problemas encontrados foram:

- dados separados por muitos ficheiros;
- registos duplicados;
- falhas de registos;
- consumos com valores negativos;
- muitos consumos a zero.

Começou-se por analisar e tratar todos os dados que continham os registos das leituras de cada consumidor. Todos os ficheiros com estes registos foram agrupados num só conjunto de dados, como já mencionado na Secção 2.1.

Numa primeira fase, observou-se que existiam, para o mesmo local, leituras diferentes para instantes iguais. Constatou-se que tal acontece porque quando ocorre um acerto, de uma leitura já existente, este é acrescentado ao conjunto de dados (num novo ficheiro) mas não é eliminado o registo anterior da leitura em questão a corrigir. Ficando assim, dois registos com a mesma data e hora, mas leituras diferentes. Posto isto, foi eliminado o registo mais antigo destes instantes repetidos.

Numa segunda fase, com o intuito de estudar o consumo por local e, posteriormente, procurar anomalias, foi criado uma nova variável no conjunto de dados Leituras: consumo (como se pode observar na Figura 4). O **consumo** (em litros) em intervalos de 15 minutos é dado pela diferença entre leituras consecutivas (diferença entre a leitura atual e a leitura anterior).

Data/Hora	Local	Leitura	Consumo
2020-11-10 19:45:00	1219065	24670403	1
2020-11-10 20:00:00	1219065	24670405	2
2020-11-10 20:15:00	1219065	24670409	4
2020-11-10 20:30:00	1219065	24670411	2
2020-11-10 20:45:00	1219065	24670414	3
2020-11-10 21:00:00	1219065	24670416	2

Figura 4. Exemplo de registos do consumo de água em intervalos de 15 minutos de um determinado consumidor

Nesta fase, foram observados os seguintes problemas:

1. consumos nulos;

2. consumos com valores negativos.

Estes casos foram denominados de:

- **Consumidor zero**: representa um consumidor com registo de consumo sempre igual a zero;
- **Consumidor pontualmente zero**: representa um consumidor que tem registo de pelo menos uma ocorrência de consumo diário zero durante 7 ou mais dias consecutivos;
- ***Reverse flow***: representa um consumo no sentido inverso e é detetado quando o consumo num dado instante é inferior ao consumo no instante anterior, ou seja, se o resultado da diferença entre a leitura atual e a leitura anterior for um valor negativo.

Os consumidores zero foram identificados e retirados do conjunto de dados inicial. Assim como os consumidores com ocorrências de reverse flow mas, no entanto, para estes casos, foi criado um novo conjunto de dados com todos os registos dos consumidores com ocorrências de reverse flow para ser estudado à parte.

De seguida, verificou-se que, os dados em estudo não representavam séries temporais contínuas. Uma vez que, existiam locais com falhas de registos e, também existiam locais onde os registos de leituras não iniciavam nem terminavam todos no mesmo instante. Isto é, o primeiro registo dos dados em estudo foi a 24/09/2020 mas, existiam locais cujo o primeiro registo de leitura foi em abril de 2021, por exemplo. Posto isto, para o estudo seguinte da procura de outras anomalias, fez-se uma seleção de locais que continham registos de leituras durante 7 meses, a iniciar em 01/12/2020 e terminar em 30/06/2021, inclusive, com locais que continham menos de 500 falhas de 15 minutos. Assim, nestas circunstâncias, dos 929 locais existentes inicialmente ficaram 635. No entanto, a estes 635 locais ainda foram excluídos os locais com *reverse flow* e os consumidores zero, ficando assim um total de 536 locais para o estudo em 7 meses.

Numa terceira fase, determinou-se o consumo de água em diferentes escalas temporais: consumo por hora, consumo por dia e consumo por mês, para cada consumidor. E, posteriormente, calculou-se o consumo médio por hora, dia e mês.

O **consumo por hora** é calculado através da soma do consumo referente a intervalos de 15 minutos. Ou seja, somando o consumo dos quatro intervalos de 15 minutos obtém-se o consumo por hora. Por exemplo: o consumo às 19 horas, num certo local, é dado pela soma do consumo de 15 minutos nos instantes: 19:15; 19:30; 19:45; 20:00. A Figura 5 apresenta um exemplo de registos de consumo de água por hora de um determinado consumidor nas primeiras horas de dia (uma segunda-feira de dezembro de 2020).

Local	Consumo_hora	hora	mes	ano	dia_da_semana
819450	0	0	12	2020	1
819450	0	1	12	2020	1
819450	0	2	12	2020	1
819450	0	3	12	2020	1
819450	0	4	12	2020	1
819450	0	5	12	2020	1
819450	0	6	12	2020	1
819450	0	7	12	2020	1
819450	30	8	12	2020	1
819450	69	9	12	2020	1
819450	53	10	12	2020	1
819450	14	11	12	2020	1

Figura 5. Exemplo de registos de consumo de água por hora de um determinado consumidor nas primeiras horas de um dia

O **consumo médio por hora** é dado pela média de consumo em cada hora (por consumidor). Por exemplo, num certo local, o consumo médio às 3 horas é dado pelo quociente entre a soma de todas as ocorrências de consumo a essa hora nesse local e a quantidade de ocorrências. Daqui resultou um conjunto de dados que contém, para cada local (consumidor), o consumo médio de cada hora nos 7 meses estudados (ver exemplo na Figura 6).

O cálculo do consumo por dia e mês, bem como, o cálculo do consumo médio por dia e por mês seguem a mesma linha de raciocínio da apresentada em cima. As Figuras 7 e 8, apresentam exemplos de registos do consumo médio de um determinado consumidor por dia e por mês, respetivamente.

Local	hora	Consumo_medio_hora
819450	0	0.0377358
819450	1	0.00471698
819450	2	0.0188679
819450	3	0.0141509
819450	4	0.0660377
819450	5	0.0566038
819450	6	1.56604
819450	7	14.8726
819450	8	17.3679

Figura 6. Exemplo de registos do consumo médio por hora (em 7 meses) referentes às primeiras horas do dia de um determinado consumidor

Local	dia_da_semana	Consumo_medio_dia
819450	0	110.467
819450	1	154.806
819450	2	159.387
819450	3	166
819450	4	171.7
819450	5	180.233
819450	6	105.467

Figura 7. Exemplo de registos do consumo médio em cada dia da semana (em 7 meses) de um determinado consumidor. O dia da semana 0 representa o domingo, o dia da semana 1 representa a segunda-feira e assim sucessivamente.

Local	mes	Consumo_medio_mes	ano
819450	1	4867	2021
819450	2	4571	2021
819450	3	4483	2021
819450	4	5509	2021
819450	5	4374	2021
819450	6	3420	2021
819450	12	4532	2020

Figura 8. Exemplo de registos do consumo médio em cada um dos 7 meses de um determinado consumidor

Após o tratamento de dados nos registos de leituras, fez-se a análise e o tratamento dos registos com informações dos consumidores (Conjunto de dados Cadastro). E foram detetados os seguintes problemas:

1. Muitos ficheiros separados;
2. Registos duplicados;
3. Valores em falta;
4. Informação não necessária para o estudo em causa.

Inicialmente agruparam-se todos os ficheiros num só, dando origem ao conjunto de dados Cadastro (como descrito na Secção 2.1). De seguida, observou-se que para o mesmo consumidor existiam várias linhas repetidas e, posto isto, eliminaram-se todas as linhas repetidas ficando apenas com uma linha de informação para cada local. Posteriormente, verificou-se a existência de vários valores em falta, no entanto, estes não foram tratados porque ocorriam em variáveis não utilizadas no estudo em causa. Uma vez que, para o estudo em causa, foram apenas usadas deste conjunto de dados as variaveis: tipo de instalação e calibre.

Numa segunda fase, inseriu-se, em cada um dos conjuntos de dados descritos anteriormente, o respetivo tipo de calibre de cada local bem como o valor do caudal máximo estipulado para cada caso (quer o máximo por hora, ou por dia ou por mês), com o objetivo de analisar, posteriormente, a possível existência de consumidores com consumos superiores aos previstos. E, inseriu-se também, o tipo de instalação de

cada local. Nesta fase, alteraram-se 3 tipos de instalação, "10 TARIFA SOCIAL - TS1", "11 TARIFA SOCIAL - TS2", "7 TARIFA FAMILIAR - 5 PESSOAS" (ver na Secção 2.1), para o tipo "1 DOMÉSTICO", uma vez que, todos representam consumos domésticos.

Por fim, com o intuito de obter uma melhor análise, uma melhor procura de perfis de consumidores e de anomalias, fez-se uma separação dos locais que eram classificados como consumidores domésticos dos não domésticos. Portanto, de todo este processo resultaram 6 conjuntos de dados, 3 com locais domésticos e 3 com locais não domésticos, com os consumos médios nas diferentes escalas temporais (consumo médio por hora, por dia e por mês) durante 7 meses. Para, posteriormente, lhes ser aplicado algumas técnicas de *machine learning* para a deteção de perfis de consumidores (domésticos e não domésticos) bem como, para a deteção de anomalias. Segue, na Figura 9, uma amostra de um dos 6 conjuntos de dados descritos anteriormente.

Local	hora	Consumo_medio_hora	Tipo de Instalação	calibre	caudal
819450	0	0.0377358	1 DOMÉSTICO	15	3125
819450	1	0.00471698	1 DOMÉSTICO	15	3125
819450	2	0.0188679	1 DOMÉSTICO	15	3125
819450	3	0.0141509	1 DOMÉSTICO	15	3125
819450	4	0.0660377	1 DOMÉSTICO	15	3125
819450	5	0.0566038	1 DOMÉSTICO	15	3125
819450	6	1.56604	1 DOMÉSTICO	15	3125
819450	7	14.8726	1 DOMÉSTICO	15	3125
819450	8	17.3679	1 DOMÉSTICO	15	3125
819450	9	16.1934	1 DOMÉSTICO	15	3125

Figura 9. Conjunto de dados: consumo médio por hora em consumidores domésticos (em 7 meses)

3. MÉTODOS E FERRAMENTAS

3.1 *MACHINE LEARNING*: UMA BREVE INTRODUÇÃO

Machine learning é um ramo da inteligência artificial e da ciência da computação que fornece aos sistemas a capacidade de aprender e melhorar automaticamente com a experiência. O processo de aprendizagem, começa com observações ou dados e, baseia-se na ideia de que os sistemas podem aprender com os dados fornecidos, a fim de identificar padrões e tomar decisões com o mínimo de intervenção humana. Nesta secção será descrito brevemente os principais métodos de machine learning bem como os algoritmos mais utilizados.

Os métodos de machine learning dividem-se em duas principais categorias:

Aprendizagem com supervisão (*Supervised learning*): é um método de aprendizagem que trabalha em cima de dados e respostas conhecidos e pré-estabelecidos. Ou seja, a partir da análise de um conjunto de dados de treino conhecido, o algoritmo de aprendizagem produz uma função inferida para fazer previsões sobre os valores de saída. Assim, os algoritmos de *Supervised learning* usam, para aprender, casos que já contêm informação acerca dos resultados pretendidos, sendo possível estabelecer uma relação entre os valores pretendidos e os valores produzidos pelo sistema e, encontrar erros para modificar o modelo. Os algoritmos de *Supervised learning* podem ser divididos em duas categorias: Classificação e Regressão.

As técnicas de classificação, como o próprio nome diz, são algoritmos que atuam para classificar os dados em grupos. Estas técnicas reconhecem entidades especificas no conjunto de dados e, a partir deste, retiram conclusões de como essas entidades devem ser classificadas. Os algoritmos de classificação mais comuns são: *support vector machines* (SVM), *decision trees,K-nearest neighbor, random forest* e *Naive Bayes*.

As técnicas de regressão, são algoritmos que atuam na previsão de dados e respostas continuas, como ,por exemplo, nos casos de previsão de temperatura. A regressão linear e a regressão logística são algoritmos de regressão populares.

Aprendizagem sem supervisão (*Unsupervised learning*): é um método de aprendizagem que trabalha em cima de dados cujo as respostas não são conhecidas previamente. Assim, os algoritmos de aprendizagem sem supervisão são usados quando a informação usada para treinar não está classificada nem rotulada. Ou seja, quando existem dados de entrada mas não existem os correspondentes resultados.

Este método de aprendizagem é aplicado em problemas que se enquadram dentro destas três categorias: Segmentação, Associação e Redução de Dimensão.

As técnicas de segmentação (*clustering*) aplicam-se quando se pretende organizar os dados não rotula-

dos em grupos coerentes, com base nas suas semelhanças ou diferenças (por exemplo, agrupar clientes com o mesmo tipo de consumo de água).

As técnicas de associação aplicam-se quando se pretende conhecer regras que associam o comportamento demonstrado pelos dados.

As técnicas de redução de dimensão aplicam-se quando é necessário reduzir o número de variáveis num conjunto de dados.

3.2 LINGUAGEM DE PROGRAMAÇÃO

Para a realização deste projeto foi utilizada uma linguagem de programação *Python* (na versão 3.7). Python é uma linguagem de programação de alto nível, simples, mas potente e flexível. Tem uma grande quantidade de recursos e documentação de alta qualidade, que facilitam o seu uso. Além disso, Python tem uma ampla seleção de ferramentas e bibliotecas construídas especificamente para análise de dados e para implementação de técnicas de *Machine Learning*, que permitem economizar tempo no desenvolvimento de novos scripts. Tem bibliotecas para carregamento de dados, visualização, estatísticas, processamento de linguagem natural, processamento de imagens e muito mais Müller e Guido (2016a). Para o desenvolvimento deste trabalho foram utilizadas as seguintes bibliotecas Müller e Guido (2016b); Foundation (2021):

- *Matplotlib*: biblioteca para geração de gráficos e visualização de dados em geral.
- *Seaborn*: biblioteca usada muitas vezes para geração de gráficos estatísticos.
- *Pandas*: biblioteca usada para a manipulação e análise de conjunto de dados.
- *Mglearn*: biblioteca usada muitas vezes na geração de gráficos.
- *NumPy*: biblioteca fundamental para a programação científica, uma vez que suporta objetos de array multidimensionais de alto desempenho e fornece as ferramentas e funções para manipulá-los.
- *SciPy*: biblioteca usada para resolver um grande número de problemas matemáticos.
- *Fitter*: biblioteca usualmente utilizada na identificação da distribuição de probabilidades que melhor se ajusta a um conjuntos de dados.
- *Scikit-learn*: biblioteca desenvolvida especificamente para aplicação prática de algoritmos de *machine learning*.

3.3 TÉCNICAS DE *MACHINE LEARNING* IMPLEMENTADAS

3.3.1 Redução de dimensão

A análise de dados complexos e de grandes dimensões é cada vez mais frequente e pode originar quer um enorme custo computacional quer um enorme custo de tempo, tornando a análise impraticável ou inviável. Assim, as técnicas de redução de dimensão surgem para combater este problema, quando necessário, reduzindo conjuntos de dados multivariados sem que exista uma grande perda de informação. Estas técnicas consistem em obter uma representação reduzida do conjunto de dados que, apesar de conter muito menos volume de dados, mantém a integridade dos dados originais. Ou seja, estas técnicas identificam conjuntos de variáveis não correlacionadas entre si que explicam a maior parte da variabilidade dos dados.

Análise de Componentes Principais (PCA)

Existem muitas técnicas para reduzir a dimensão de um conjunto de dados, mas umas das técnicas mais conhecidas e mais utilizadas é a análise de componentes principais (mais conhecida como PCA, do inglês*Principal Component Analysis*).

O PCA consiste em transformar variáveis de um conjunto de dados multivariado em novas variáveis não correlacionadas entre si, de modo a reduzir a dimensão do conjunto de dados sem que haja perda significativa de informação, uma vez que, a análise recai sobre as componentes principais que caracterizam o conjunto de dados.

O procedimento básico desta técnica é dado por Li (2019):

- Centrar os dados na média;
- Cálculo da matriz de covariância;
- Cálculo dos valores e vetores próprios da matriz de covariância;
- Seleção do valor de redução;
- Projeção dos dados.

O primeiro passo, para o desenvolvimento desta técnica, consiste numa **normalização** dos dados, isto é, subtrair a média do conjunto de dados a todos os elementos do mesmo, de modo a obter um novo conjunto que tem de média 0.

De seguida, determina-se a **matriz de covariância**. Esta relaciona a covariância entre todas as variáveis de um conjunto. Onde a fórmula é dada pela equação (3.1).

$$B = \frac{1}{N} A^T \cdot A \tag{3.1}$$

Note-se que, A é a matriz com os dados de treino centrados na média, A^T é a sua correspondente matriz transposta e N representa o número de elementos de A. A matriz de covariância é uma matriz quadrada, simétrica e definida positiva e, portanto, todos os seus valores próprios são positivos.

Numa fase seguinte, calculam-se os **valores e vetores próprios** da matriz de covariância, uma vez que, os vetores próprios associados aos valores próprios de maior valor representam as componentes principais de uma matriz.

Os **valores próprios** de uma matriz $B_{n\times n}$ são as raízes do polinómio característico de B. E, o polinómio característico de B é dado pela equação (3.2).

$$\Delta_B(\lambda) = det(\lambda I - B) \tag{3.2}$$

Ou seja, λ é um valor próprio de B se $det(\lambda I - B)$.

Daqui resulta que, se λ é valor próprio de B, então $(\lambda I - B)\cdot v = 0$ é um sistema possível indeterminado. Isto é, existe um vetor $v \neq 0$ tal que:

$$(\lambda I - B) \cdot v = 0 \tag{3.3}$$

Ou seja, para um valor próprio λ, existe $v \neq 0$ tal que:

$$Bv = \lambda v \tag{3.4}$$

e, quando tal acontece, diz-se que v é **vetor próprio** de B associado a λ.

Posteriormente, é feita a **seleção do valor de redução**, ou seja, é feita a escolha do número de componentes principais necessários para descrever todos os dados, com o objetivo de reduzir a dimensão inicial. Esta seleção é feita através valores próprios. Isto é, após determinar os valores próprios da matriz de covariância, calcula-se a variância explicada para cada um e, o número de componentes principais é dado por aqueles valores próprios que têm maior variância explicada.

A variância explicada resulta da razão entre cada um dos valores próprios com a soma de todos os valores próprios. Isto é, seja λ_i um valor próprio da matriz covariância, a variância explicada de cada valor próprio é dada por:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_n} \tag{3.5}$$

De seguida, forma-se uma matriz que contém os vetores próprios associados aos valores próprios selecionados.

Por fim, é feita a **projeção dos dados**. E, esta projeção é dada através da multiplicação da matriz transposta que contém os vetores próprios resultantes da redução de dimensão, isto é, os mais relevantes, com a matriz transposta que contém os dados de treino centrados na média, ou seja, $V^T \cdot A^T$, sendo V a matriz dos vetores próprios resultantes da redução de dimensão e A a matriz com os dados de treino centrados na média.

Implementação

Esta técnica de redução de dimensão foi aplicada, neste estudo, com recurso à biblioteca *Scikit-Learn* Müller e Guido (2016a).

A técnica PCA foi aplicada aos 6 conjuntos de dados descritos no final da Secção 2.2, com o objetivo de reduzir as variáveis: hora, dia e mês. Apenas será explicado a aplicação desta técnica a um destes conjunto de dados, uma vez que, o processo repete-se em todos os restantes casos. No entanto, os resultados da aplicação aos restantes conjuntos de dados serão apresentados a seguir, no Capítulo 4. Pegando, por exemplo, no conjunto de dados que contém o consumo médio por hora em cada local, em 7 meses e nos locais domésticos, a técnica PCA foi aplicada com o intuito de reduzir as 24 horas existentes. Segue, na Figura 10, parte do conjunto de dados utilizado na aplicação do PCA do exemplo que está a ser explicado.

Index	0	1	2	3	4	5	6	7	8	9	10
819450	0.0377358	0.00471698	0.0188679	0.0141509	0.0660377	0.0566038	1.56604	14.8726	17.3679	16.1934	13.1887
820342	0.00943396	0	0	0	0	0	0	0	0.382075	1.00943	0.367925
821454	1.20283	0.59434	0.212264	0.301887	0.0990566	0.207547	0.551887	6.23585	39.5849	22.7358	28.0802
823473	16.3443	13.7075	10.1085	10.0094	10.0708	11.4623	17.3915	25.6604	29.25	29.6132	33.4953
824089	0	0	0	0	0	0	0.820755	4.51887	8.87264	1.99528	1.10849
826782	0.693396	0.660377	1.43396	1.28774	1.09906	1.23113	1.99528	1.78302	5.19811	21.7406	15.2783

Figura 10. Parte do conjunto de dados para a aplicação do PCA (consumo médio por hora)

Numa primeira fase, centralizou-se os dados na média, transformando os dados que se pretende reduzir num conjunto de dados com média=0 e variancia=1, usando *StandardScaler*. Depois, determinou-se o número de componentes principais necessários, obtendo assim o valor da redução de dimensão a efetuar. E, para obter este valor, recorreu-se ao cálculo da variância explicada. Uma vez que, ao visualizar o gráfico que relaciona a variância explicada acumulada com diversos números de componentes principais, Figura 11, é possível verificar que componentes são mais importante.

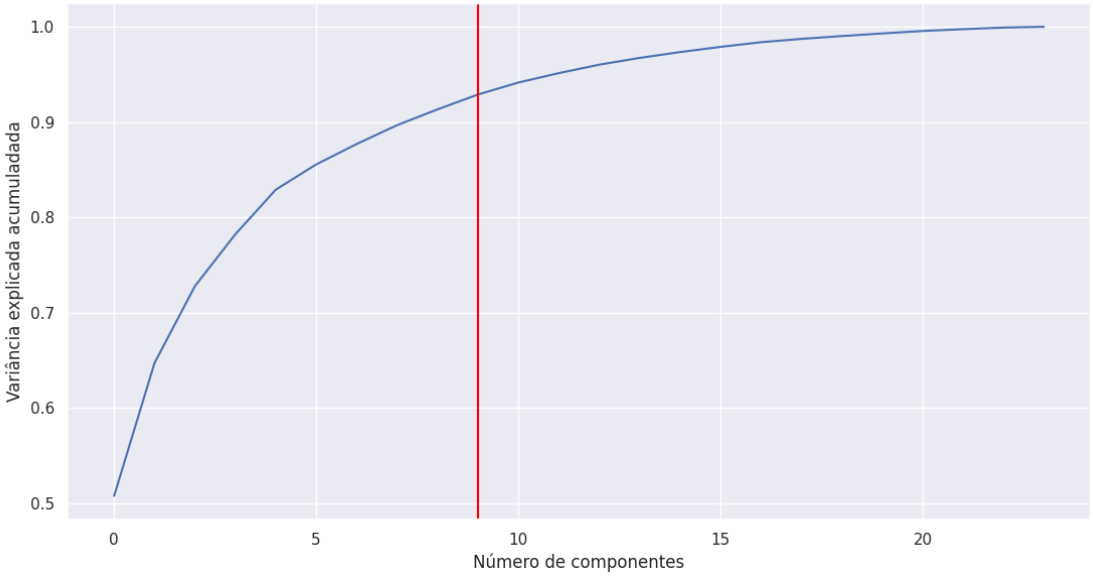


Figura 11. Determinação do número de componentes principais

Pela observação do gráfico, verifica-se que, as nonas primeiras componentes contêm mais de 90% da variação, ou seja, com 9 componentes é possível reter mais de 90% da informação inicial dos dados estudados. Assim, pode-se reduzir as 24 variáveis iniciais, para este caso, a 9 variáveis. Por fim, com recurso à classe PCA, obteve-se a redução de dimensão. Este procedimento origina um *dataset* que contém os valores resultantes da redução de dimensão para, posteriormente, ser utilizado na aplicação do algoritmo *kmeans*, ver em 3.3.3, com o objetivo de criar grupos de clientes com consumos idênticos.

3.3.2 *Clustering*

Clustering é uma técnica de *machine learning* não-supervisionada que é frequentemente utilizada em análise de dados e tem como objetivo agrupar elementos de um conjunto de dados em subconjuntos diferentes, de forma a que, elementos semelhantes sejam agrupados num mesmo subconjunto, enquanto que, elementos diferentes são agrupados em subconjuntos diferentes. Isto é, esta técnica consiste em particionar um conjunto de dados , D, em diferentes subconjuntos (chamados de *clusters*), C^k , com $k = 1, ..., K$, tal que: $D = \bigcup_{k=1}^{K} C^k$ e $C_i \cap C_j = \varnothing$, para $i \neq j$. Ou seja, a união de todos os clusters formados resulta no conjunto incial de dados mas, cada elemento de D, pertence apenas a um e um só cluster Boulicaut e Jeudy (2005) Rao e Govindaraju (2013).

Métricas ponto-a-ponto

Dado que, esta técnica consiste no agrupamento de elementos semelhantes de um conjunto de dados, é então necessário algum tipo de medida que possa determinar se dois elementos são semelhante ou diferentes. Medir a diferença entre os dados é então uma tarefa fundamental e, devido ao grande espectro de tipos de dados (categóricos, ordinais, binários, numéricos), a introdução de várias métricas é necessária para lidar com tal diversidade de informações. Existem dois tipos principais de métricas para medir a diferença entre elementos: dissimilaridade e similaridade.

Dissimilaridade

Seja D um conjunto de dados, $D = D_1 \times D_2 \times ... \times D_I$ e, $x = (x_1, x_2, ..., x_I)^T$ um elemento genérico de D, onde $x_i \in D_i$ e, seja a função $d : D \times D \to \mathbb{R}$, tal que $d(x, y) \in \mathbb{R}$ para qualquer $x, y \in D$, uma função que quantifica a distância entre dois elementos de D. A definição da métrica dissimilaridade é dada por Rao e Govindaraju (2013):

Definition 3.3.1. *A função d é uma dissimilaridade se satisfaz as seguintes condições:*

1. $\forall x, y \in D, d(x, y) \geq 0$;
2. $\forall x \in D, d(x, x) = 0$.

A dissimilaridade quantifica então a diferença entre dois elementos de um certo conjunto de dados e, esta diferença é pequena quando os elementos estão próximos e é grande quando estes estão afastados.

O conceito de distância surge da definição de dissimilaridade.

Definition 3.3.2. *A função de dissimilaridade $D \times D \to [0, +\infty[$ é uma distância se satisfaz as seguintes propriedades* (Rao e Govindaraju (2013)):

1. $\forall x, y \in D, d(x, y) = d(y, x)$;
2. $\forall x, y \in D, d(x, y) = 0 \Leftrightarrow x = y$;
3. $\forall x, y, z \in D, d(x, z) \leq d(x, y) + d(y, z)$.

Existem várias medidas de distância para calcular a dissimilaridade entre elementos descritos por atributos numéricos. Mas, as que são frequentemente utilizadas em *machine learning*, mais precisamente na técnica *clustering* são: a distância Euclidiana, de *Manhattan* e de *Chebyshev*.

Para $D = \mathbb{R}^I$, as distâncias são dadas por:

- **Distância *Manhattan***: $d(x,y) = \sum_{i=1}^{I} |x_i - y_i|$
- **Distância Euclidiana**: $d(x,y) = \sqrt{\sum_{i=1}^{I}(x_i - y_i)^2}$
- **Distância de *Chebyshev***: $d(x,y) = \max_{i=i}^{I} |x_i - y_i|$

Similaridade

Seja agora a função $s : D \times D \to \mathbb{R}$, tal que $s(x,y) \in \mathbb{R}$, para qualquer $x, y \in D$, a função de similaridade. É um conceito diferente ao de distância, uma vez que, compara dois elementos, x e y , de um dado conjunto de dados. E, ao contrário da dissimilaridade, a similaridade tem um valor alto quando x e y são próximos (semelhantes) e tem um valor pequeno quando estão bastante afastados.

Definition 3.3.3. *A função s é uma similaridade se satisfaz as seguintes condições* Rao e Govindaraju (2013):

1. $\forall x, y \in D, s(x,x) \geq s(x,y)$ e $s(x,x) \geq s(y,x)$.

Representante de um subconjunto

Quando se trabalha com técnicas de *Clustering*, consequentemente, trabalha-se com subconjuntos de dados. E, determinar um elemento que seja representativo de todo o subconjunto é útil quer para simplificar as operações quer para reduzir o custo computacional Rao e Govindaraju (2013). Este representante tanto pode ser um elemento do conjunto de dados em estudo (medóide) como pode ser um elemento que não pertence à base de dados (centróide). A escolha do representante depende de dois fatores essenciais: da métrica utilizada na quantificação ponto-a- ponto e das operações algébricas disponíveis no espaço dos dados.

Existem vários tipos de representantes de um subconjunto Rao e Govindaraju (2013), a média e a mediana são um exemplo. Mas, apenas a média será abordada como um tipo de representante, uma vez que, os restantes não foram utilizados no estudo realizado.

Média

Seja A o espaço de atributos e $D = (x^n)_{n=1}^{N}$ o conjunto de dados, onde $x_n \in A$ e C é um subconjunto de D.

O valor médio, m, de cada subconjunto de D é um tipo de representate e é denominado de centróide. E é dado por:

$$m = \frac{1}{|C|} \sum_{x \in C} x \tag{3.6}$$

Assim, m é um centróide se não pertencer ao conjunto de dados e se é o valor médio de um subconjunto de D.

Métodos de *Clustering*

Existem, na literatura Boulicaut e Jeudy (2005) e Aggarwal e Reddy , muitos métodos de *clustering* e muitos algoritmos de agrupamento inseridos nestes métodos. A noção de agrupamento não está definida com precisão e, portanto, é difícil fornecer uma categorização precisa dos métodos de *clustering*. No entanto, é útil apresentar uma ideia relativamente organizada acerca dos métodos de *clustering*. Assim, os principais métodos de *clustering* podem ser classificados nas seguintes categorias:

- **Métodos de Particionamento**: dividem os elementos de um conjunto de dados em grupos (*clusters*) não sobrepostos. Ou seja, nenhum elemento pode ser membro de mais de um *cluster* e cada *cluster* deve ter pelo menos um objeto. Os algoritmos inseridos nesta classificação, requerem que seja especificado o número de clusters, normalmente indicado pela variável k e, funcionam por meio de um processo iterativo para atribuir subconjuntos de pontos de dados em k *clusters*. São algoritmos não determinísticos, o que significa que, podem produzir resultados diferentes em duas execuções separadas, mesmo se as execuções forem baseadas na mesma entrada. Um exemplo de algoritmo inserido nesta classificação é o algoritmo K-Means.

- **Métodos de Hierarquização**: agrupam os elementos de um conjunto de dados seguindo uma sequência hierárquica de partições. Existem dois tipos de sequencias possíveis: as de aglomeração e as de divisão. As sequencias de aglomeração começam por formar *clusters* com um elemento apenas e , efetuam sucessivas fusões dos *clusters*, até se formarem *clusters* de maiores dimensões. Construindo assim uma hierarquia ascendente (estratégia *bottom-up*). Enquanto que , as sequencias de divisão, começam com um *cluster* que contém todos os elementos do conjunto de dados e, num processo iterativo, divide-o em clusters mais pequenos, até se obter a partição fina ou até se atingir uma determinada condição (por exemplo, o número pretendido de *clusters*). Construindo uma hierarquia descendente (estratégia *top-down)*.

3.3.3 *K-Means*

O algorimo *K-Means* é um dos algoritmos de *clustering* mais simples e mais frequentemente utilizado. Muito sucintamente, este algoritmo consiste em selecionar, numa primeira fase, k elementos de um conjunto de dados (com o objetivo de o dividir em k *clusters*). Cada um destes elementos representa inicialmente a média ou o centro do *cluster* e , é denominado de centroíde. Esta divisão dos dados tem que ser de forma a que os diferentes *clusters* fiquem o mais separados possível entre eles, enquanto que, os elementos dentro de cada *cluster* fiquem o mais próximos possível. Assim, cada um dos elementos do conjunto de dados, não selecionado inicialmente, é atribuído ao cluster que lhe é mais semelhante, com recurso à distância euclideana entre o elemento e o centroíde. Depois, para cada *cluster*, o algoritmo calcula uma nova média com os novos elementos do *cluster* em questão e encontra novos centroídes. As iterações continuam até que, os *clusters* formados na iteração atual são os mesmo que os formados na iteração anterior Han et al. (2011).

A Figura 12, retirada a biblioteca *Mglearn* Müller e Guido (2016a), ilustra um exemplo de três iterações do algoritmo *K-Means*.

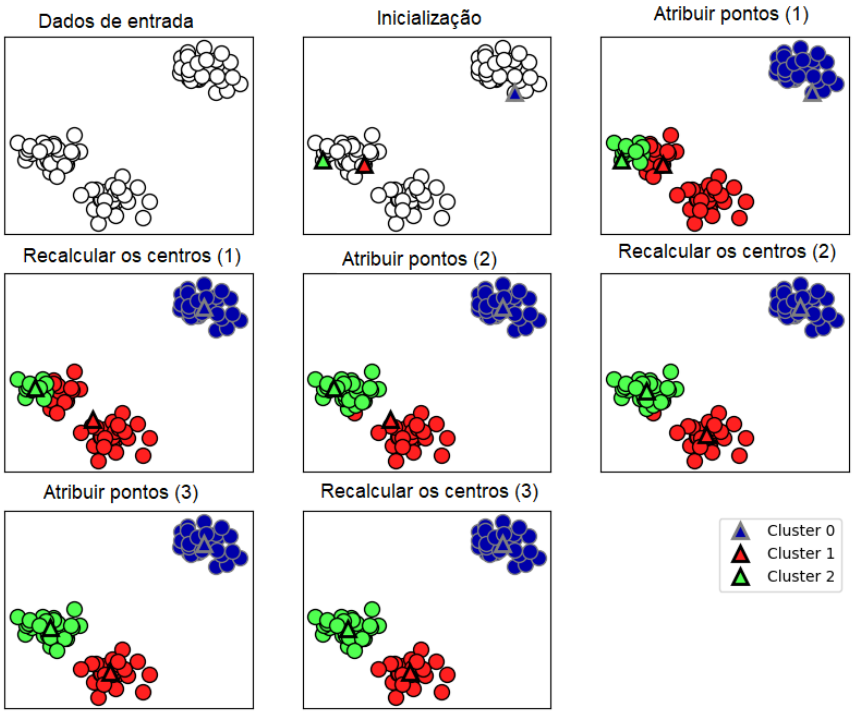


Figura 12. Representação de três etapas do algoritmo *k-means* (Adaptada de Müller e Guido (2016a))

Este algoritmo, apesar de ser de fácil compreensão, tem o problema da escolha inicial da quantidade de *clusters* necessária. E, como na maioria dos casos, os dados em estudo têm várias dimensões, fica difícil visualiza-los e , consequentemente, o número ótimo de clusters a utilizar não é óbvio.

Existem, na literatura Aggarwal e Reddy, vários métodos para determinar o melhor valor para o número de *clusters* a utilizar num conjunto de dados. Um dos métodos mais frequentes e mais antigo é o **Método do Cotovelo**. Trata-se de um método visual , uma vez que, consiste na observação de um gráfico que representa a relação entre o aumento do número de clusters com uma determinada métrica. E, chama-se Método do Cotovelo porque à medida que o número de *clusters* aumenta a métrica utilizada como função diminui e, forma um ponto onde a curva da função representada dobra, ficando com um formato de um 'cotovelo'. É nesse ponto que é determinado o número ideal de *clusters*, uma vez que, o valor ideal para k é aquele que tem o menor número de *clusters* possível e ao mesmo tempo um menor valor da métrica utilizada (diferente de zero).

No estudo em questão, para a visualização do Método do Cotovelo e para a determinação do número ótimo de *clusters* , utilizou-se como métrica a soma dos quadrados dentro do *cluster* (mais conhecida como WCSS, *within clusters sum of squares*). WCSS é defenida como a soma dos quadrados das distâncias entre cada membro do cluster e o respectivo centróide, cujo a fórmula é dada por learn developers (2021):

$$\sum_{i=1}^{k} \sum_{x_i \in C_i} \| x_i - m_j \|^2 \tag{3.7}$$

onde x_i representa um elemento do conjunto de dados, m_j representa a média (ou o centróide) dos elementos do cluster C_i.

O número de *clusters* é aquele que minimiza esta soma. Assim, pode-se encontrar o valor ótimo de *clusters* através do Método do Cotovelo, procurando o ponto que minimiza WCSS.

Observe-se graficamente, na Figura 13, a relação entre o número de *clusters* e a soma dos quadrados dentro do *cluster* para o *dataset* resultante da redução de dimensão descrita na Secção 3.3.1.

Cada ponto assinalado na curva, $a_0, a_1, ..., a_{18}$, representa uma quantidade de *clusters*, que, para este caso, pode variar entre 2 a 20 *clusters*. E, pela observação do gráfico, verifica-se que, o número ótimo de *clusters* não é obvio só pela visualização do mesmo. Uma vez que, pela aplicação do método do cotovelo, o número ótimo de *clusters* pode ser em qualquer dos pontos: a_5, a_6 ou a_7.

Assim, para descobrir o número ótimo de *clusters* a utilizar, aplicou-se uma outra forma apresentada em Kaufman e Rousseeuw (2009).

Observa-se na Figura 14 que o número ótimo de clusters é dado pelo ponto da curva cuja a distância à reta que une os pontos a_0 e a_{18} seja máxima.

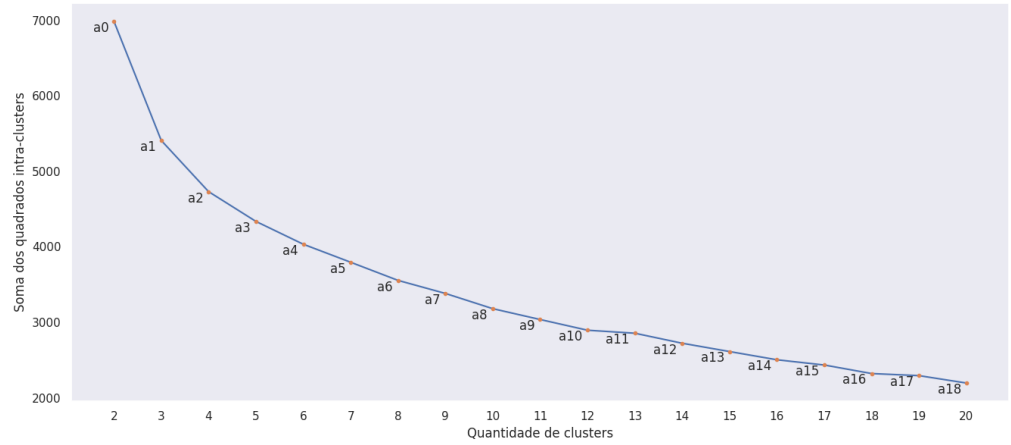


Figura 13. Método do Cotovelo: soma de quadrados intra-clusters (Adaptado de Kaufman e Rousseeuw (2009))

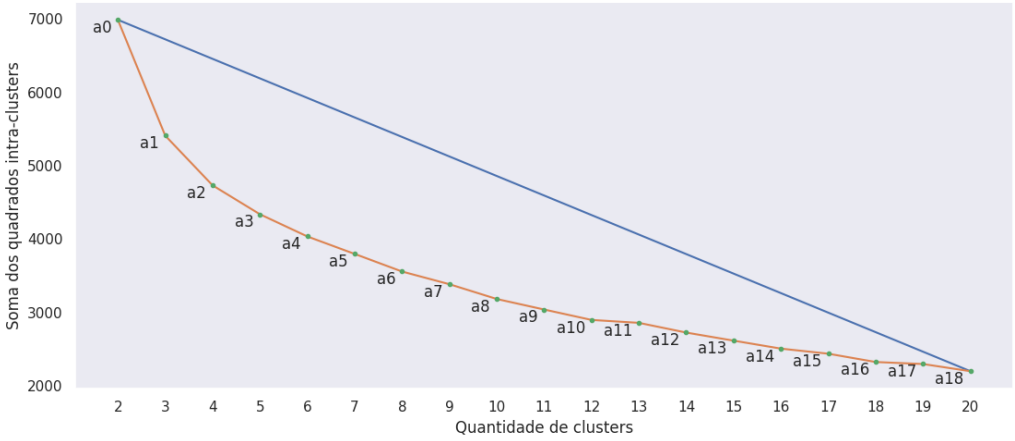


Figura 14. Método do Cotovelo: número ótimo de clusters (Adaptado de Kaufman e Rousseeuw (2009))

A distância entre um ponto qualquer da curva , (x, y) e a reta que passa por a_0 e a_{18}, é dada pela fórmula seguinte:

$$d(P_0, P_1, (x, y)) = \frac{|(y_1 - y_0)x - (x_1 - x_0)y + x_1 y_0 - y_1 x_0|}{\sqrt{(y_1 - y_0)^2 + (x_1 - x_0)^2}} \tag{3.8}$$

onde, neste caso, $P_0 = (x_0, y_0) = a_0$ e $P_1 = (x_1, y_1) = a_{18}$.

Todo este processo explicado do algoritmo *K-Means*, foi aplicado a cada um dos 6 conjuntos de dados estudados, depois da aplicação da técnica de redução de dimensão, PCA.

Implementação

Este algoritmo foi aplicado, neste projeto, com recurso à biblioteca *Scikit-Learn* Müller e Guido (2016a); Kaufman e Rousseeuw (2009).

Primeiro calculou-se a WCSS através do atributo *inertia* da classe *KMeans*, com o intuito de determinar o número de *clusters* a utilizar. De seguida, fez-se a partição do conjunto de dados para o números de *clusters* achado, recorrendo à classe *K-Means* e utilizando o método *fit_predict*.

3.4 MÉTODOS ESTATÍSTICOS IMPLEMENTADOS

A estatística é um dos ramos mais importantes da matemática. É essencialmente utilizada no desenvolvimento de técnicas que envolvam a análise de dados, como a sua interpretação e representação, mas também, as técnicas estatísticas, podem ser utilizadas como base no desenvolvimento de métodos de detecção de anomalias. Assim, para além das técnicas de *machine learning* utilizadas e, vistas anteriormente, também se recorreu a técnicas estatísticas para a determinação de anomalias. Mais precisamente, recorreu-se ao uso de funções de distribuição de probabilidade para analisar o comportamento dos resultados obtidos, provenientes da aplicação das técnicas de *machine learning*. Isto é, após a aplicação do algoritmo *K-Means*, cujo os resultados serão apresentados mais à frente no Capítulo 4, foram encontrados padrões de consumo para os clientes estudados, domésticos e não domésticos e, com o objetivo de detetar consumos anómalos, nestes locais estudados, procedeu-se à analise do comportamento do consumo por hora, de cada um dos perfis, através da procura do tipo de distribuição de probabilidade que melhor descrevia cada curva de consumo, para cada hora, em cada perfil.

Implementação

Para determinar as distribuições de probabilidade que melhor descreviam os dados, utilizaram-se as bibliotecas: *Fitter* e *Scipy*. A biblioteca *Fitter* determina a distribuição que mais se adequa aos dados em estudo, recorrendo às 80 distribuições que a biblioteca *Scipy* dispõe Cokelaer (2019).

Distribuições de probabilidade

Muito sucintamente, uma distribuição de probabilidade é uma descrição matemática da probabilidade de ocorrência de determinados eventos. Assim, a distribuição de probabilidade corresponde a uma função estatística que descreve todos os valores possíveis e probabilidades que uma variável aleatória pode tomar dentro de um dado intervalo.

Com o intuito de encontrar as distribuições que melhor se ajustavam ao dados em estudo, foram utilizadas todas as 80 distribuições fornecidas pelas bibliotecas utilizadas na implementação do algoritmo. E, no final, verificou-se que, o comportamento do consumo de cada hora, dos perfis encontrados, pode ser descrito pelas distribuições apresentadas na Figura 15.

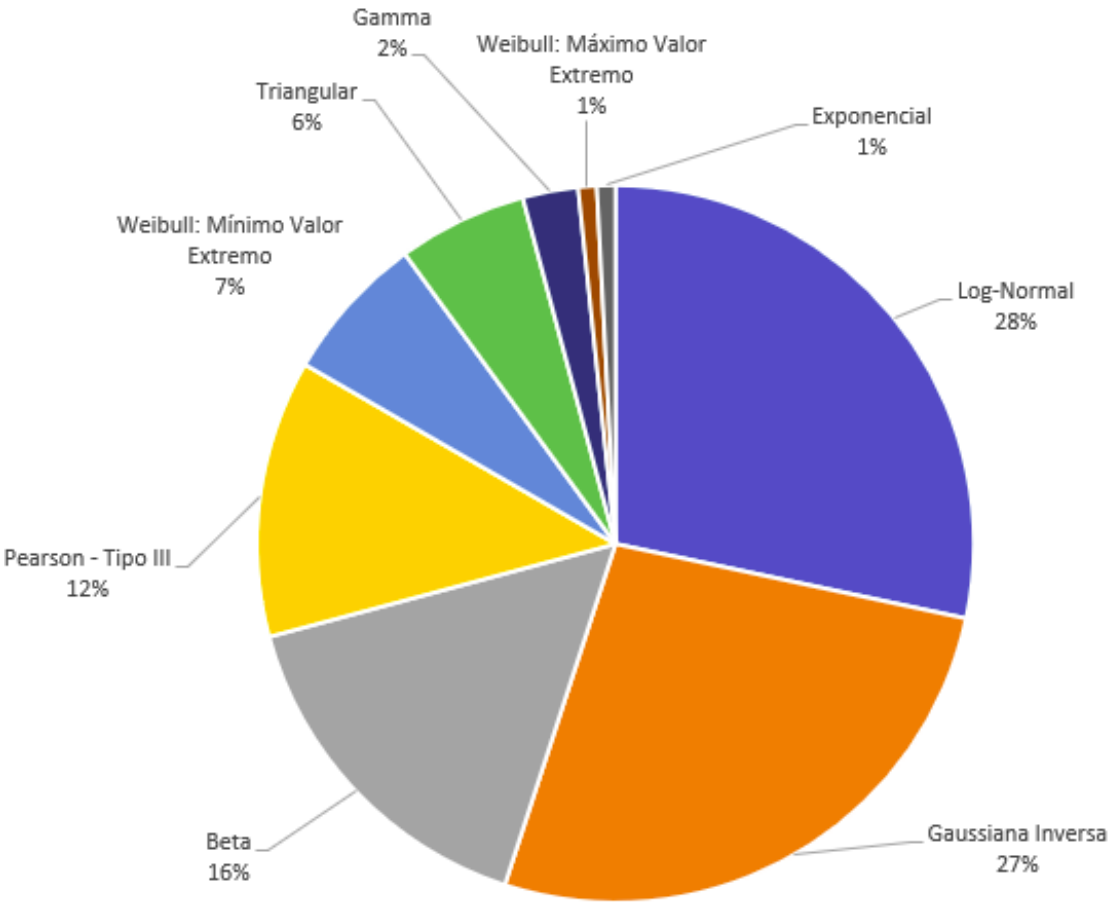


Figura 15. Distribuições utilizadas na analise do comportamento do consumo por hora.

Todas as distribuições determinadas são distribuições de probabilidade contínuas, o que faz sentido, uma vez que, a variável a ser modelada (caudal volúmico medido pelos contadores) é descrito por uma escala contínua, podendo assumir valores pertencentes ao intervalo: $[0, +\infty[$.

De seguida, serão apresentadas com mais detalhe, apenas as 3 distribuições que tiveram maior expressão. A frequência relativa das distribuições, apresentada na Figura 15, foi calculada com base no número de vezes que tal distribuição foi atribuída ao subconjunto de dados: perfil e hora.

Log-Normal

Seja x uma variável aleatória, μ a média e σ, o desvio padrão, a função densidade de probabilidade de uma distribuição log-normal é dada por Weisstein (2021) Johnson et al. (1995a):

$$f(x, \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right] \tag{3.9}$$

Uma particularidade relevante das distribuições log-normais é o facto de o limite inferior para x ser 0, a função logarítmica impõe essa condição. Esse aspeto da distribuição alinha-se perfeitamente com a variável a modelar (caudal volúmico) o que pode ajudar a explicar a sua predominância como distribuição que melhor modela a frequência relativa dos caudais volúmicos.

Gaussiana Inversa

Seja x uma variável aleatória, μ a média e λ, o fator de forma, a função densidade de probabilidade de uma distribuição log-normal é dada por Chhikara (1988):

$$f(x, \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x - \mu)^2}{2\mu^2 x}\right) \tag{3.10}$$

A distribuição gaussiana inversa tem como limite inferior para x, 0 e como tal a média segue esse limite inferior. O fator de forma é limitado inferior também por 0.

Beta

Seja x uma variável aleatória compreendida entre 0 e 1, a função densidade de probabilidade de uma distribuição beta é dada por Johnson et al. (1995b):

$$f(x, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \tag{3.11}$$

Onde:

- $\Gamma(z)$ é a função gama aplicada à variável z. A função gama é, em termos simples, uma extensão da função fatorial para o conjunto dos números reais e complexos.
- α, β são dois parâmetros de ajuste que controlam a forma da distribuição. São forçosamente positivos.

A função beta é a terceira mais usada para descrever o caudal volúmico e corresponde a horas de pouco consumo de água.

Intervalo de consumos (IC)

Após ser determinado a função de distribuição de probabilidades (Figura 15) que melhor se ajusta aos valores de consumo de água na hora i para o perfil j, o objetivo é definir um intervalo que com uma dada probabilidade, $1-\alpha$ (com $\alpha \in [0,1]$), inclua um possível valor do consumo de água na hora i para o perfil j. Para tal, determinou-se os intervalos de consumo do tipo $[L_{lower}, L_{upper}]$, de modo que se verifique a seguinte equação:

$$P(L_{lower} < x < L_{upper}) = 1-\alpha, \alpha \in [0,1], \tag{3.12}$$

onde L_{lower} representa o quantil de ordem $\frac{\alpha}{2}$ e L_{upper} o quantil de ordem $1-\frac{\alpha}{2}$ (ver Figura 16).

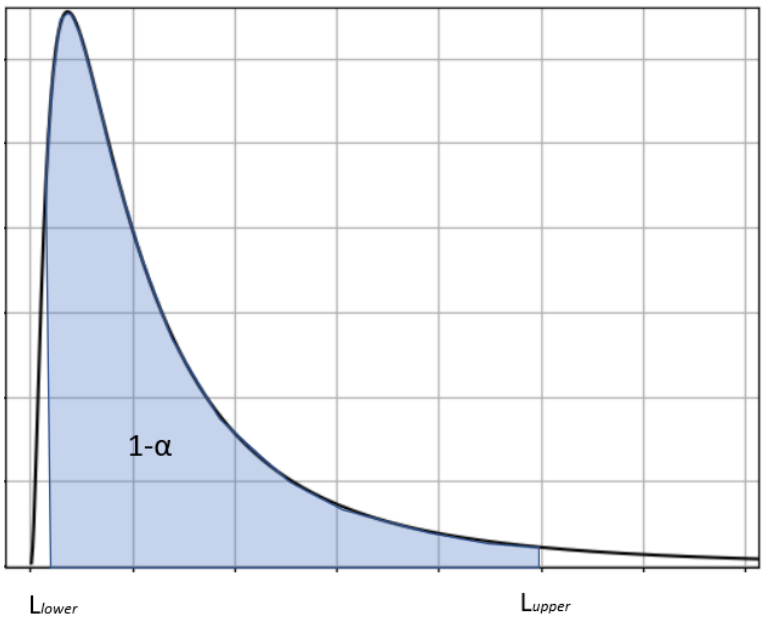


Figura 16. Ilustração de uma função de distribuição de probabilidades e a posição dos quantis de ordem $\frac{\alpha}{2}$ (L_{lower}) e ordem $1-\frac{\alpha}{2}$ (L_{upper}).

Desta forma, espera-se com uma probabilidade de $(1-\alpha)\times 100$ % o valor do consumo de água na hora i para o perfil j esteja contido no intervalo $[L_{lower}, L_{upper}]$.

A Figura 17 contém um exemplo do consumo médio por hora de um local com os seus repetivos intervalos de consumos com probabilidade igual a 99%.

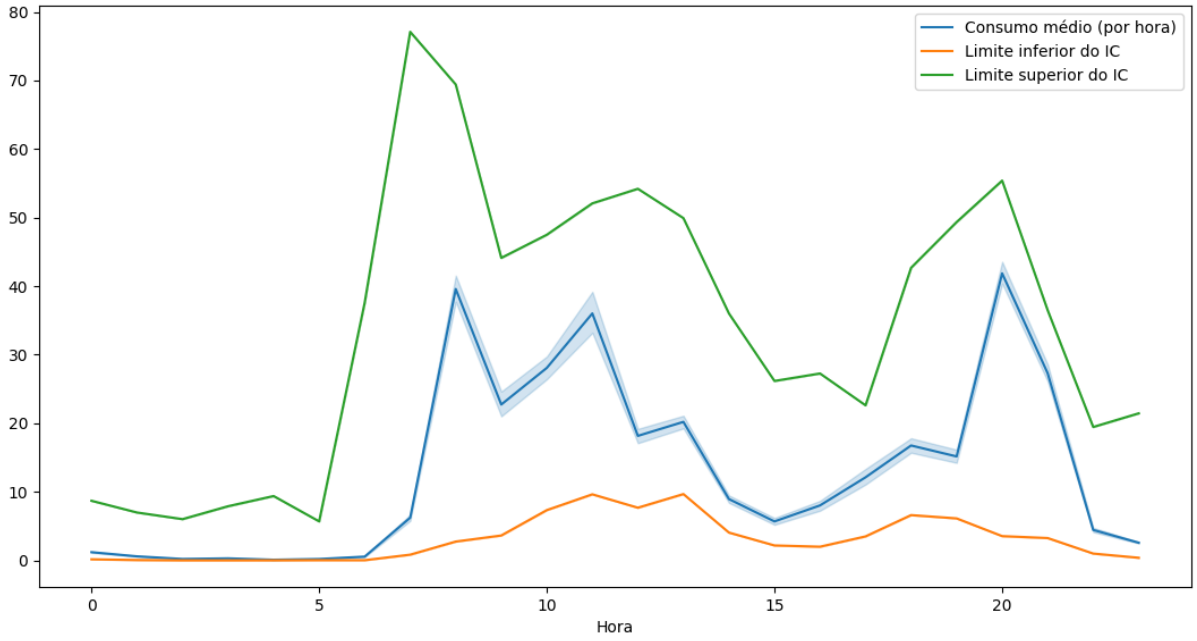


Figura 17. Comparação do consumo médio por hora de um local com os intervalos de consumos de cada hora

4. RESULTADOS

O presente estudo centrou-se nas seguintes fases:

1. Pré-processamento dos dados;
2. Análise de clusters para identificação de diferentes padrões de consumo de água;
3. Determinação de intervalos de consumos para identificação de ocorrências fora do padrão de consumo habitual.

Neste capítulo, será apresentado os resultados obtidos em cada uma das fases.

4.1 ANOMALIAS DETETADAS NO PRÉ-PROCESSAMENTO DOS DADOS

No fase de pré-processamento dos dados foram detetadas as seguinte anomalias:

1. Consumidor zero (ausência de registos);
2. *Reverse flow* (valores de consumo negativo);
3. *Overflow* (valores superiores ao caudal máximo);

4.1.1 Consumidor zero

Como já foi descrito na Secção 2.2, um consumidor zero representa um cliente com registo de consumo sempre igual a zero. Dos 929 locais iniciais, cerca de 3% são consumidores zero. Estes locais foram extraídos do estudo e reportados à empresa para que, esta os possa avaliar e perceber o porquê da inexistência de consumo. Até ao momento, ainda não foi possível verificar todos os locais denominados de consumidor zero mas, aqueles que foram verificados, têm como motivo mais frequente a danificação do contador ou a extração deste pelo cliente.

4.1.2 *Reverse flow*

Este tipo de anomalia, refere-se à existência de pelo menos um valor negativo nos registos de consumo de água (ver Secção 2.2). Os locais que tinham ocorrências com *reverse flow* foram excluídos da conjunto de dados usadas nas fases seguintes e analisados à parte. Portanto, foi criado um conjunto de dados com os registos de consumo negativo (*reverse flow*) de todos estes locais.

De forma a sumariar as ocorrências de *reverse fow* em termos de número de locais e o número de ocorrências em cada local construiu-se o gráfico de barras apresentado na Figura 18. Podemos observar

que a maioria dos locais apenas tem um registo de *reverse fow* mas, no entanto, existem locais com vários casos de consumo no sentido inverso.

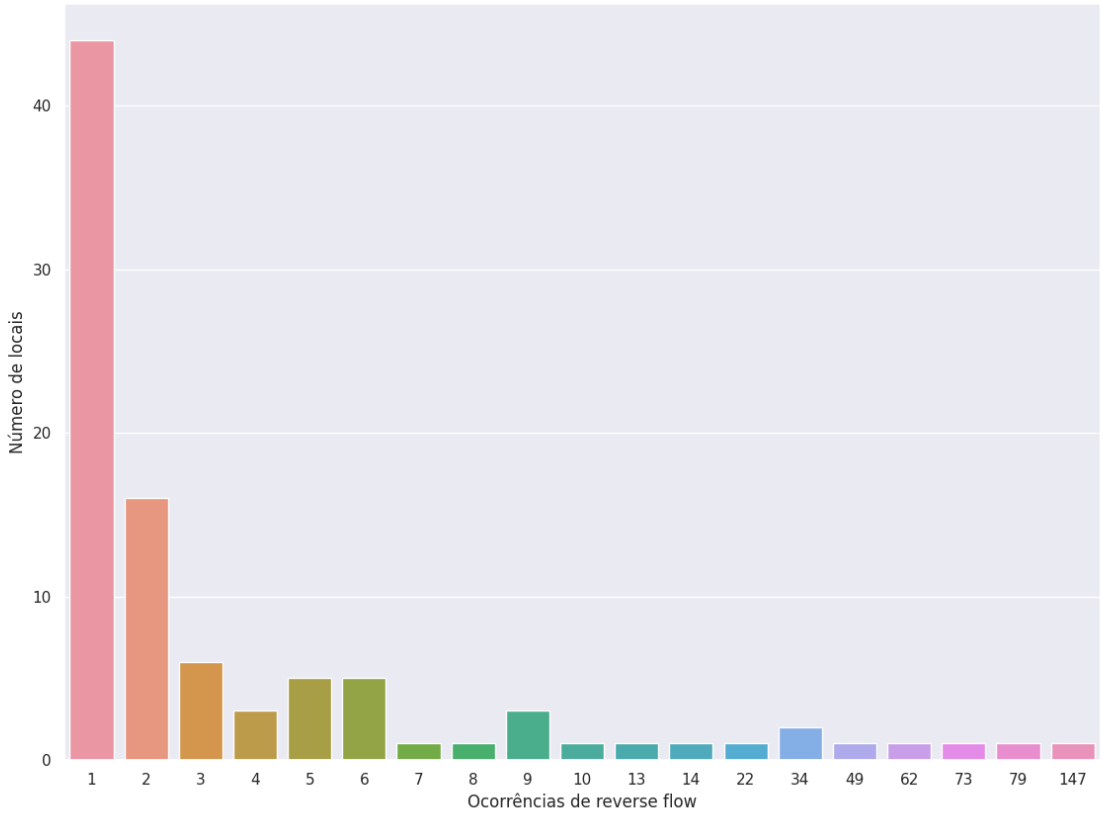


Figura 18. Número de ocorrências de *reverse flow* por local

Estes resultados foram reportados à empresa mas ainda não foram detetadas todas as causas que deram origem aos casos de *reverse flow* encontrados. Mas, um dos principais motivos identificados no passado, foi o facto de, existirem consumidores com captações de água próprias com ligação (ilegal) à rede pública. A entrada de água própria na rede pública leva a que haja um fluxo invertido e, consequentemente, um consumo com valores negativos. A identificação destas situações é de grande importância, uma vez que água proveniente de captações próprias não é muitas vezes tratada e poderá contaminar a água na rede pública, o que constitui um risco para a saúde pública.

4.1.3 *Overflow*

Outro tipo de anomalia detetado no pré-processamento dos dados foi a existência de consumidores que, em determinadas horas, excediam o caudal máximo por hora que lhes está estipulado. Isto é, cada local tem um tipo de calibre associado (como o descrito na Secção 2.1.2) e, cada calibre, tem um caudal máximo possível (em m^3/h), no entanto, foram detetadas ocorrências que excediam este caudal máximo.

Estes casos foram originados, principalmente, por estas duas situações:

- **Existência de acertos**: no instante em que um consumidor sofre um acerto de uma leitura, se este for um valor muito discrepante do registo anterior, a diferença entre a leitura atual com a anterior será um valor elevado e, por esta razão, existe um pico de consumo nestes instantes. A Figura 19 mostra um exemplo de um caso de *overflow* detetado que resultou de um acerto de uma leitura.

Data/Hora	Local	Leitura	Consumo
2021-04-07 17:00:00	1135651	713895	0
2021-04-07 17:15:00	1135651	713895	0
2021-04-07 17:30:00	1135651	1713884	999989
2021-04-07 17:45:00	1135651	1713884	0
2021-04-07 18:00:00	1135651	1713890	6

Figura 19. Exemplo de um pico de consumo provocado por um acerto

- **Registos de leituras não consecutivos**: quando existe uma quebra nos registos de leituras, levando a uma inexistência de leituras num intervalo de tempo, e se a primeira leitura após a paragem for muito mais elevada que a que estava antes da quebra, o consumo neste instante sofrerá um pico. No entanto, não é um verdadeiro pico de consumo, uma vez que é originado pela quebra de registos. A Figura 20 mostra um exemplo de um caso de *overflow* detetado que resultou de uma quebra nos registos dos consumos.

Data/Hora	Local	Leitura	Consumo
2021-04-07 08:45:00	1219065	25124010	1
2021-04-07 09:00:00	1219065	25124010	0
2021-04-07 13:30:00	1219065	29206973	4082963
2021-04-07 14:00:00	1219065	29207033	60
2021-04-07 14:15:00	1219065	29207043	10

Figura 20. Exemplo de um pico de consumo provocado por uma quebra de registos

Como estes casos de consumos não correspondem à realidade do consumidor em questão, então foram identificados e excluídos do conjunto de dados usado nas fases seguintes.

4.2 IDENTIFICAÇÃO DE DIFERENTES PADRÕES DE CONSUMO

4.2.1 Redução de dimensionalidade

A técnica de redução de dimensão, PCA, foi aplicada aos 6 casos de estudo, como já foi mencionado na Secção 3.3.1. Nesta secção será reportado, mais detalhadamente, os resultados obtidos para cada caso, à exceção do caso já descrito anteriormente, como exemplo da aplicação do PCA (na Secção 3.3.1).

Locais com consumo doméstico

Dos 536 locais em estudo, durante 7 meses (de dezembro de 2020 a junho de 2021), 483 são locais com consumo de água doméstico. Portanto, foi aplicada esta técnica de redução aos três casos de consumo em estudo dos registos de leituras destes 483 locais.

Consumo médio por dia (em 7 meses)

A técnica de análise de PCA foi aplicada ao conjunto que contém os consumos médios por dia em cada local, durante 7 meses, com o intuito de reduzir os 7 dias da semana existentes.

Pela observação do gráfico da Figura 21, verifica-se que, as duas primeiras componentes contêm mais de 94% da variação, ou seja, com 2 componentes é possível reter mais de 94% da informação inicial dos dados. Posto isto, reduziu-se as 7 variáveis iniciais a 2 variáveis. Por fim, com recurso à análise de PCA (ver na Secção 3.3.1), obteve-se a redução de dimensão deste caso. Posteriormente, será aplicado ao conjunto de dados resultante desta redução, o algoritmo *K-Means*, com o objetivo de identificar grupos de clientes com o mesmo padrão de consumo.

Consumo médio por mês (em 7 meses)

Neste caso, a técnica de análise de PCA foi aplicada ao conjunto de dados que contém os consumos médios por mês em cada local, durante 7 meses, com o intuito de reduzir as 7 variáveis existentes (os 7 meses).

Novamente, através da visualização do gráfico da Figura 22, verifica-se que, as duas primeiras componentes contêm mais de 90% da variação, ou seja, com 2 componentes é possível reter mais de 90% da informação inicial dos dados. Posto isto, reduziu-se as 7 variáveis iniciais a 2 variáveis.

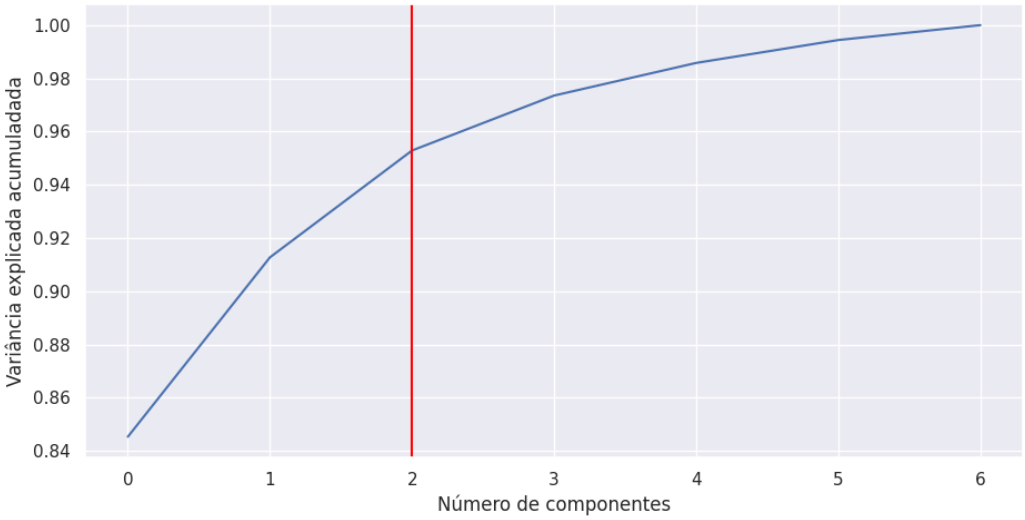


Figura 21. Determinação do número de componentes principais para o consumo médio por dia dos locais domésticos

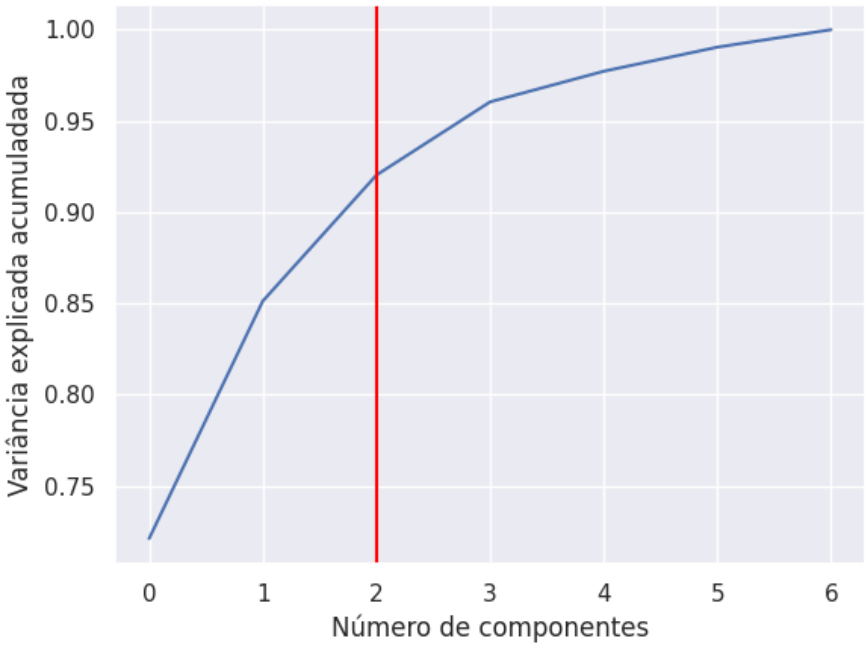


Figura 22. Determinação do número de componentes principais para o consumo médio por mês dos locais domésticos

Locais com consumo não doméstico

Há 53 locais, dos 536 em estudo, que têm um consumo de água não doméstico. Numa primeira fase, idêntico ao descrito na secção anterior, visualizaram-se as curvas de consumo de cada um desses

locais, nos diferentes espaços temporais, consumo médio por hora, por dia e por mês (Figura 23). E, concluiu-se também que, existiam dois locais com consumos muito elevados. Mais uma vez, estes locais foram reportados e removidos do estudo, uma vez que, posteriormente, iriam prejudicar o agrupamento de locais e a procura de perfis de consumidores não domésticos. Ficando assim em estudo, 51 locais com consumos de água não domésticos.

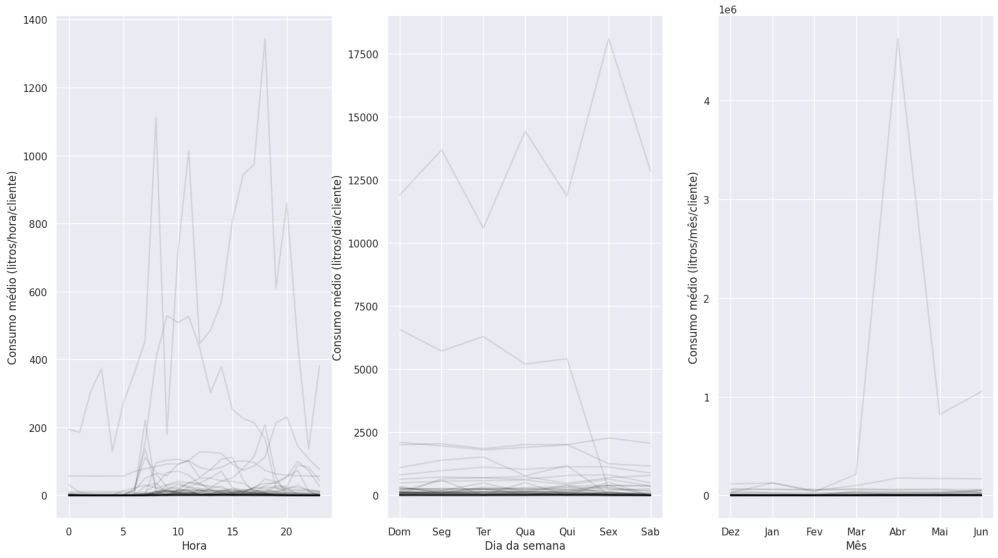


Figura 23. Consumos médios por hora, dia e mês dos 53 locais não domésticos

Consumos médios por hora (em 7 meses)

A técnica de análise PCA foi aplicada ao conjunto de dados que contém os consumos médios por hora em cada local não doméstico, à semelhança do caso dos locais com consumo doméstico (ver Secção 3.3.1), com intuito de reduzir as 24 variáveis existentes.

Neste caso, verifica-se que, as quatro primeiras componentes contêm mais de 95% da variação, ou seja, com 4 componentes é possível reter mais de 94% da informação inicial dos dados (ver Figura 24). Posto isto, reduziu-se as 24 variáveis iniciais a 4 variáveis.

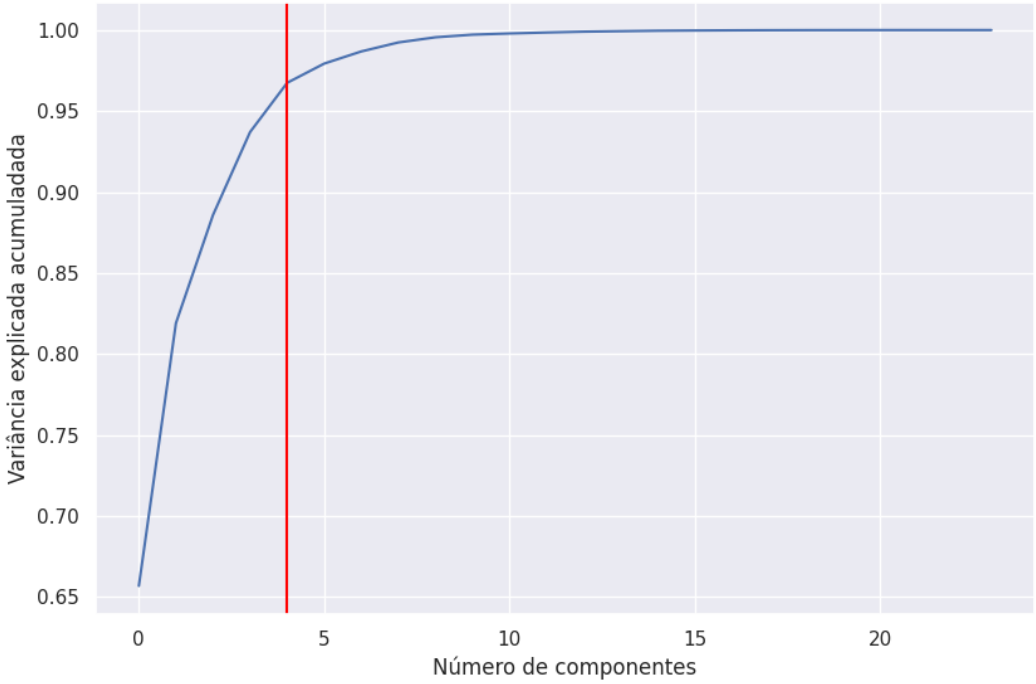


Figura 24. Determinação do número de componentes principais para o consumo médio por hora dos locais não domésticos

Consumos médios por dia (em 7 meses)

A técnica de análise de PCA foi aplicada ao conjunto de dados que contém os consumos médios por dia em cada local não doméstico com o intuito de reduzir as 7 variáveis existentes.

Para este caso, observa-se que, a primeira componente contém mais de 97% da variação (ver Figura 25). Portanto, reduziu-se as 7 variáveis a 1 variável.

Consumos médios por mês (em 7 meses)

A técnica PCA foi aplicada a este conjunto de dados que contém os consumos médios por mês em cada local não doméstico em estudo, com o objetivo de reduzir as 7 variáveis existentes, dos 7 meses existentes.

Pela análise da Figura 26, verifica-se que, 2 componentes principais contêm mais de 96% da informação inicial. Posto isto, para este caso, as 7 variáveis inicais foram reduzidas a 2 variáveis.

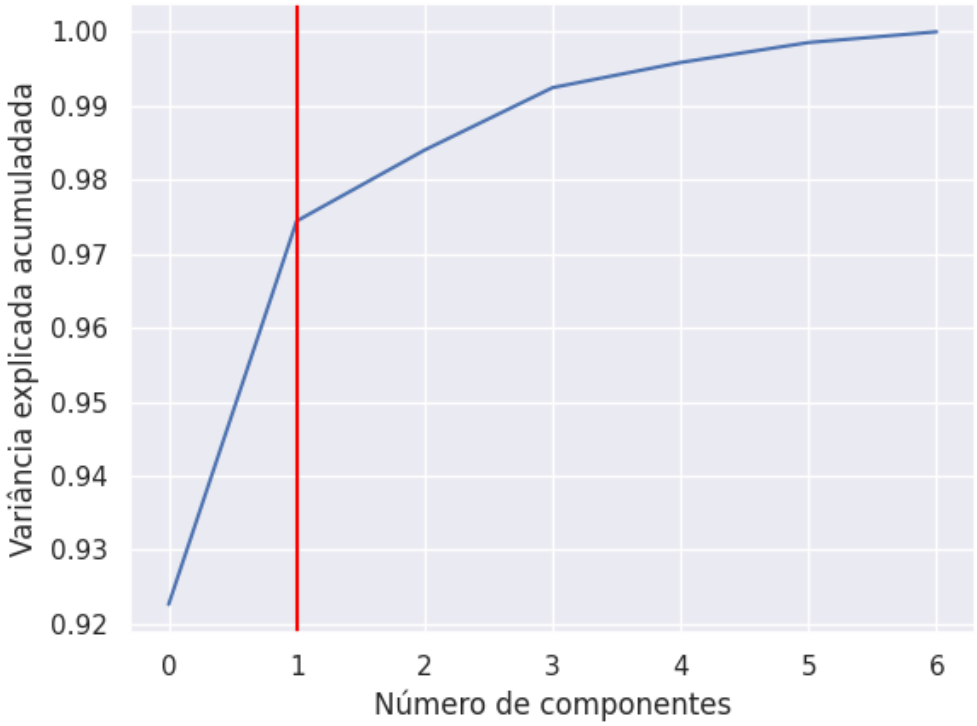


Figura 25. Determinação do número de componentes principais para o consumo médio por dia dos locais não domésticos

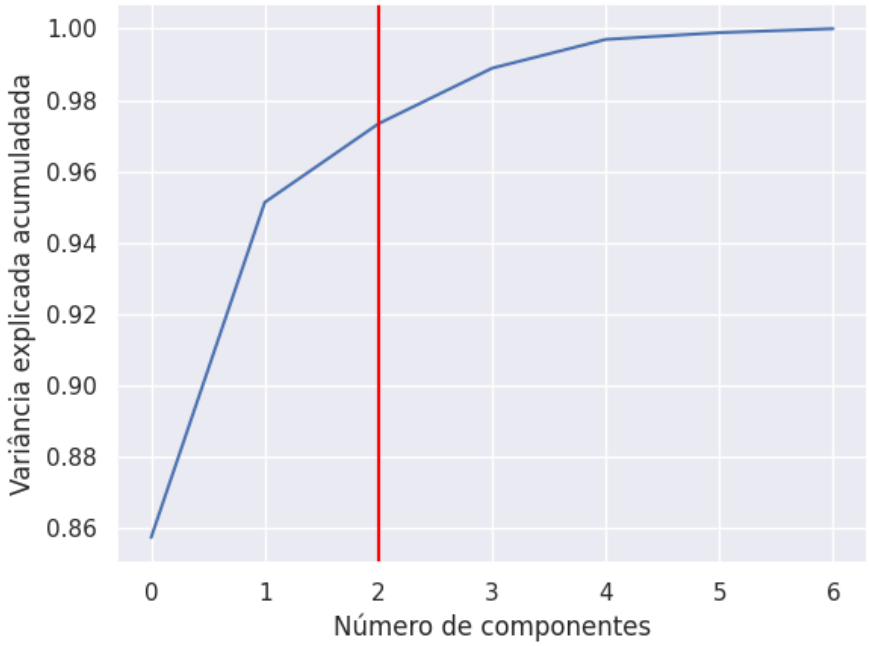


Figura 26. Determinação do número de componentes principais para o consumo médio por mês dos locais não domésticos

4.2.2 *Clustering*

Após a aplicação da técnica de análise PCA a todos os casos em estudo, aplicou-se o algoritmo *K-Means* a todos os resultados provenientes da redução de dimensão, tendo em conta a separação de locais efetuada, com o principal objetivo de encontrar diferentes padrões de consumo de água.

Clusters de locais com consumo doméstico

Numa primeira etapa, como já foi mencionado na Secção 3.3.3, começou-se por determinar o número ótimo de *clusters* para cada um dos casos em estudo, consumo por hora, dia e mês. E, concluiu-se que, para o estudo do consumo médio por hora e para o consumo médio por dia o número ótimo de *clusters* era 6, mas para o consumo médio por mês, o número ótimo de *clusters* já era 7. De seguida, aplicou-se o algoritmo *K-Means* e obtiveram-se os *clusters* para cada situação. A Figura 27, ilustra os *clusters* encontrados para o consumo nas diferentes escalas temporais.



Figura 27. *Clusters* de clientes tendo em conta o consumo médio por hora, dia e mês

Depois de determinar os *clusters* de clientes, tendo em conta o consumo médio por hora, dia e mês, fez-se uma contabilização dos locais que estavam em comum nos diferentes *clusters* obtidos. Com o intuito de achar os consumidores mais semelhantes. E, dessa contabilização (Figura 28) resultaram 6

agrupamentos principais de consumidores, contendo cerca de 85% dos domicílios em estudo. Uma vez que, foram selecionados os 6 grupos com mais locais em comum, agrupando assim, os clientes com comportamentos de consumo semelhantes em todas as escalas temporais estudadas.

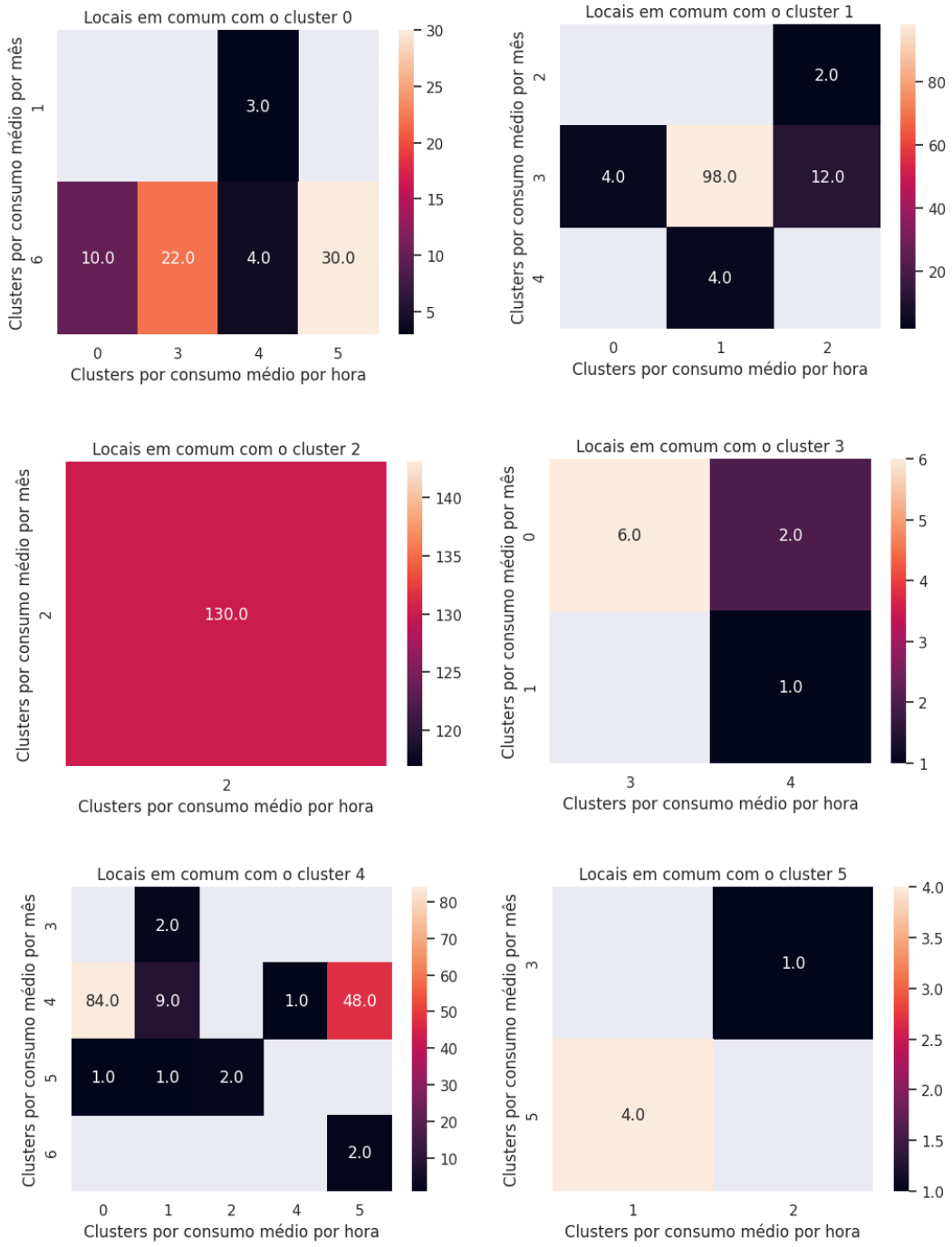


Figura 28. Locais em comum com os *clusters* por consumo médio por dia

Através da observação da Figura 28, os novos *clusters* foram classificados tendo em conta o número de locais em comum, isto é, o *cluster* 1 contém o maior número de locais em comum nas três escalas temporais (130 locais), o *cluster* 2 contém a segunda maior quantidade de locais em comum (98 locais) e, assim sucessivamente, até ao *cluster* 6 que contém 22 locais em comum. Os restantes locais não foram tidos em conta, uma vez que, tinham comportamentos de consumo pouco semelhantes.

De seguida, com o objetivo de perceber melhor os comportamentos de consumo dos perfis encontrados, foram calculadas as curvas de consumo para cada um destes 6 *clusters*, em todas as escalas temporais estudadas (consumo médio horário, diário e mensal), como se pode visualizar na Figura 29.

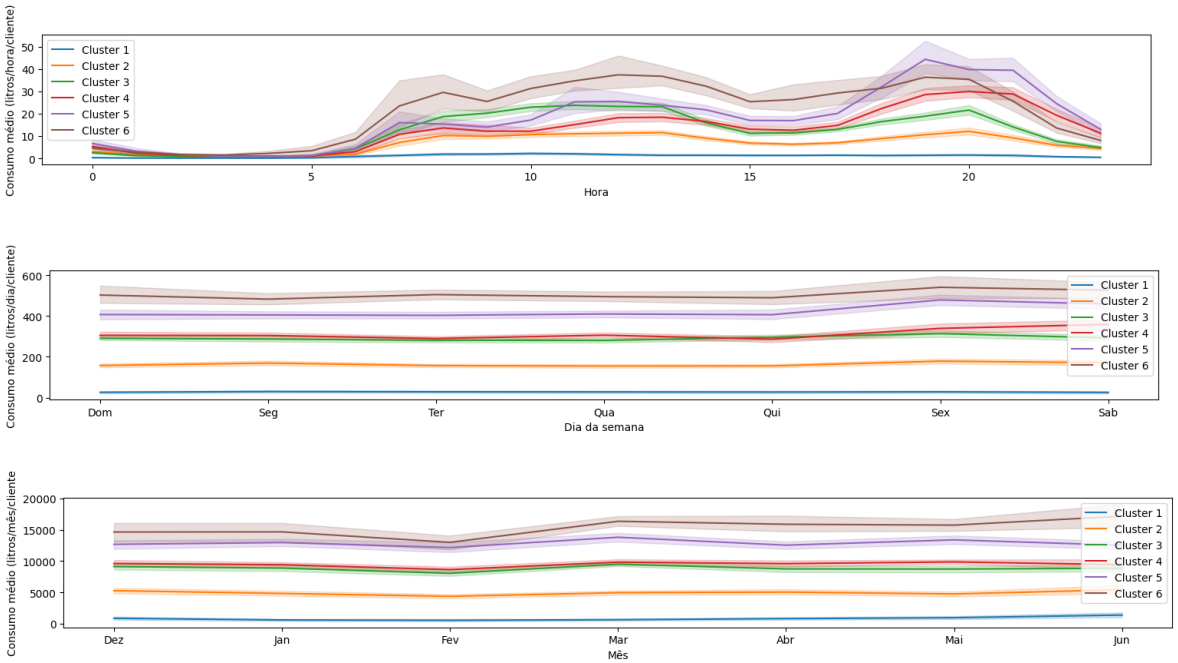


Figura 29. Curvas de consumo dos 6 perfis encontrados

Após a análise das curvas de consumo, verifica-se que, existe uma pequena variabilidade no consumo médio mensal. E, que a principal diferença entre os 6 *clusters* é a quantidade de água consumida, uma vez que, o *cluster* 1, apesar de ser o cluster com o maior número de clientes, é o que regista o menor consumo médio, enquanto que, o *cluster* 6 é o que regista o maior consumo. Portanto, concluíu-se que, os 130 agregados familiares que constituem o *cluster* 1, apresentam um baixo consumo de água (cerca de 28 litros diários) e, uma pequena variabilidade diária e horária. Os *clusters* 2, 3 e 6, com 98, 84 e 22 domicílios, respetivamente, mostram um perfil de consumo diário semelhante, todos têm um pico de consumo pela manhã, um declínio à tarde e, voltam novamente a ter um pico de consumo à noite. Observou-se ainda que, o consumo destes 3 *clusters* era ligeiramente superior na sexta-feira em relação aos restantes dias da semana. Os *clusters* 4 e 5, com 48 e 30 agregados familiares, respetivamente, apresentam um perfil de consumo diário semelhante, ambos têm valores mais elevados de consumo à

noite e, têm também um pico de consumo, com valores inferiores aos da noite, no final da manhã. Por fim, observou-se ainda que, ambos apresentavam um maior consumo às sextas-feiras e sábados.

***Clusters* de locais com consumo não doméstico**

Inicialmente, à semelhança dos locais com consumo doméstico, determinou-se o número ótimo de *clusters* para cada um dos casos em estudo, para o consumo médio por hora, por dia e por mês. Concluiu-se que, para o consumo médio por hora e para o consumo médio por mês, o número ótimo de *clusters* era 6, mas para o consumo médio por dia, era 5. De seguida, aplicou-se o algoritmo *K-Means* e obtiveram-se os clusters para cada situação (Figura 30).

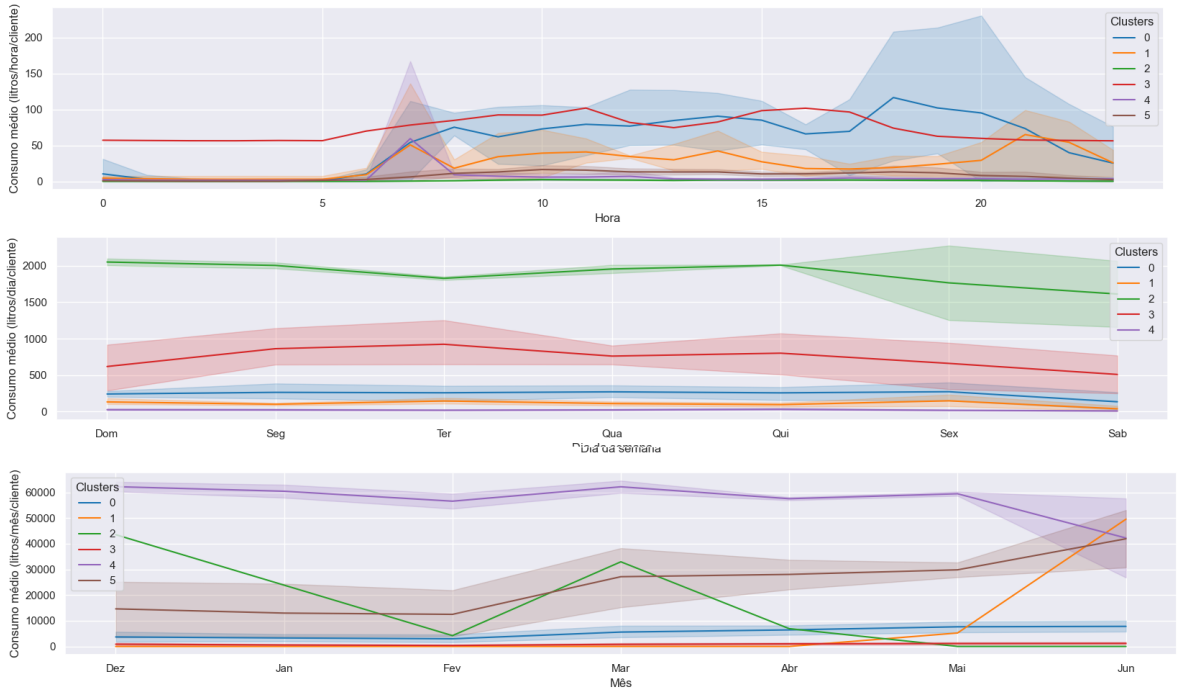


Figura 30. *Clusters* de clientes tendo em conta o consumo médio por hora, dia e mês

Numa segunda fase, contabilizou-se quantos locais estavam em comum nos diferentes *clusters* encontrados, com o objetivo de achar consumidores não domésticos semelhantes. No entanto, devido ao facto de existirem, em estudo, poucos locais com consumos não domésticos e também, devido ao facto de serem consumidores mais inconstantes, mais heterogéneos e ainda, locais que foram afetados pelo período pandémico (como por exemplo fecho de escolas ou empresas), apenas foi encontrado um grupo de consumidores semelhantes em todas as escalas temporais estudadas.

Como se pode observar na Figura 31, o único grupo de locais semelhantes encontrados é composto por 27 locais.

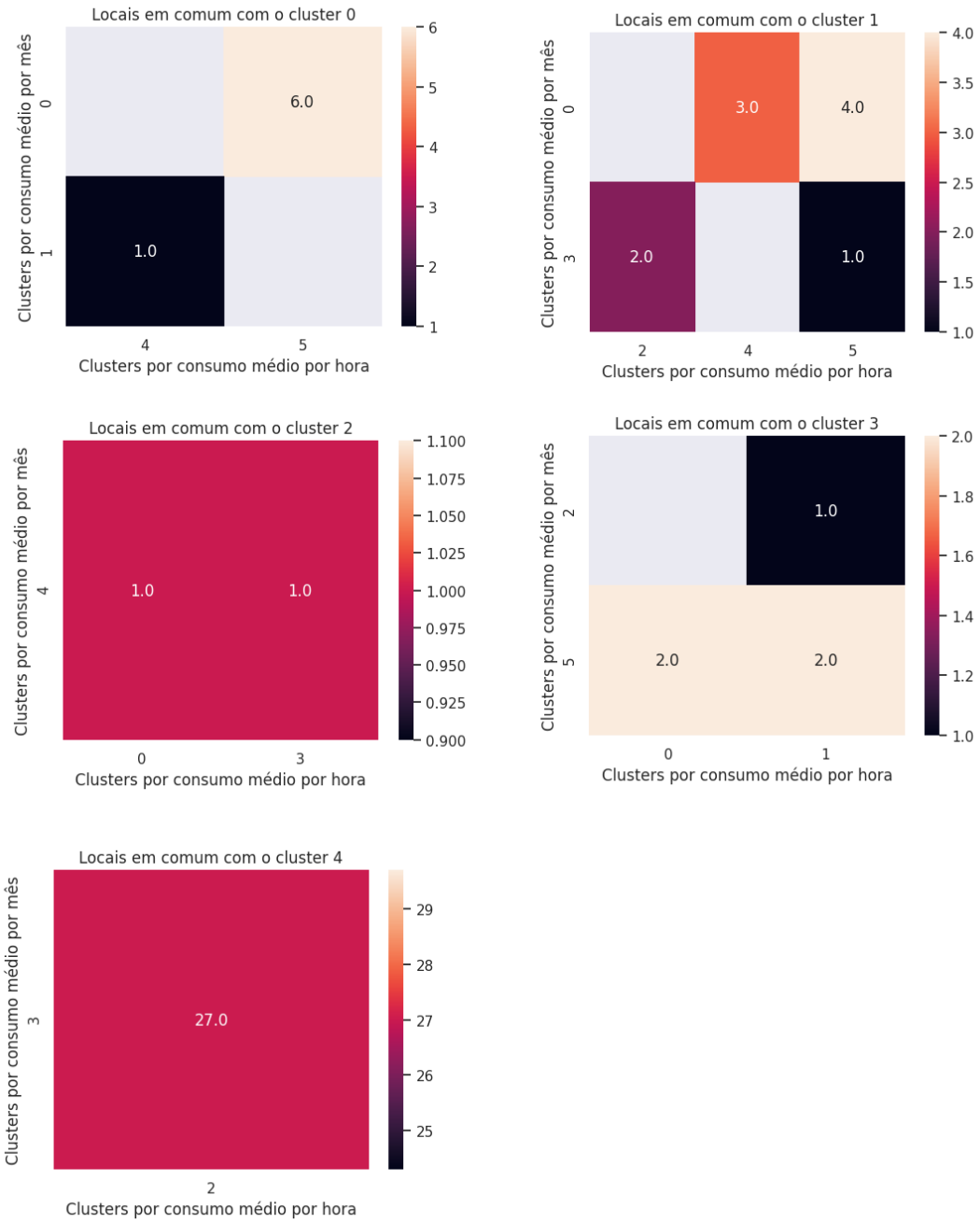


Figura 31. Locais em comum com os *clusters* por consumo médio por dia

De seguida, com o objetivo de perceber o comportamento deste perfil, foram calculadas as curvas de consumo nas diferentes escalas temporais, representadas na Figura 32.

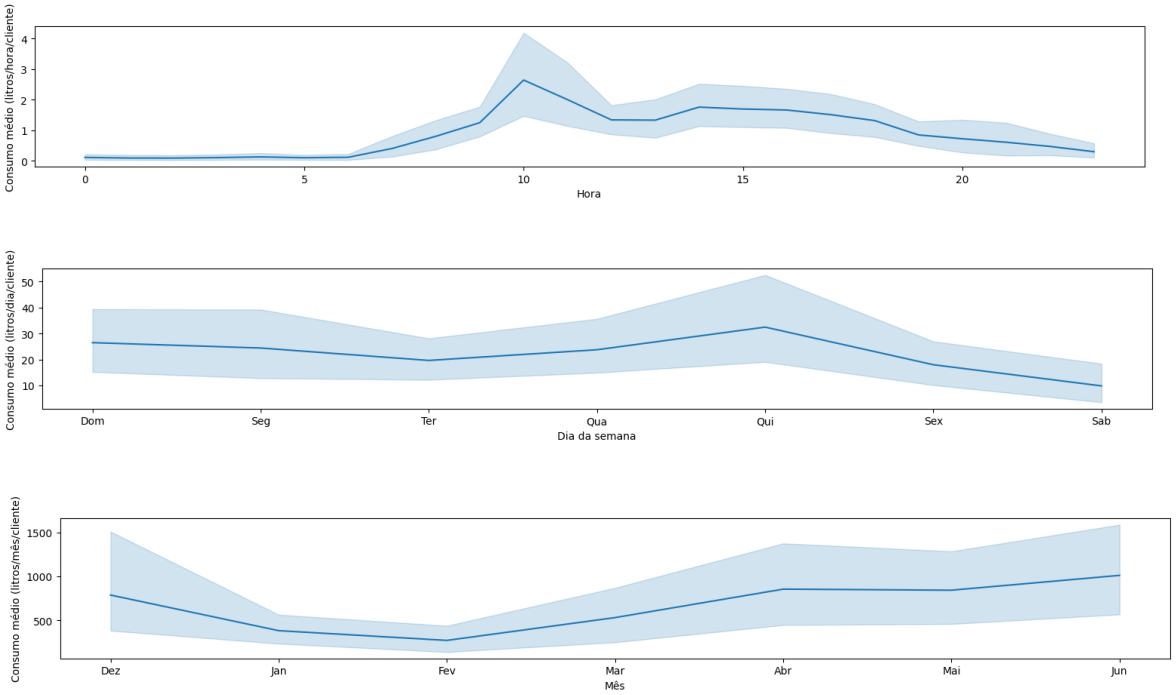


Figura 32. Curvas de consumo do perfil encontrado

Pela observação da Figura 32, concluiu-se que, apesar de serem consumidores não domésticos registam um consumo quer diário, quer semanal quer mensal baixo. Observa-se que têm um pico de consumo pela manhã e, um consumo baixo, sendo até nulo, durante a noite. Verificou-se também que têm um ligeiro aumento de consumo às quintas-feiras e, verificou-se ainda que, de janeiro a fevereiro não têm praticamente nenhum consumo, o que coincidiu com o período de confinamento imposto pela pandemia provocada pelo COVID-19.

4.3 IDENTIFICAÇÃO DE OCORRÊNCIAS FORA DO PADRÃO DE CONSUMO HABITUAL

Registos de consumo de água fora do padrão de consumo habitual, ou seja, desviados de um intervalo de valores de consumo considerados normais para um determinado consumidor poderão ser anomalias. Com o objetivo de identificar este tipo de registos foi determinado os intervalos de consumos com probabilidade de 99% do consumo médio para cada hora, em cada perfil (cluster) com base na função de distribuição de probabilidade que melhor se ajusta aos dados em cada caso. Por outras palavras, o objetivo é identificar registos de consumos de locais que, apesar de pertencerem a um determinado perfil, possam conter consumos fora dos respetivos intervalos de consumos de cada padrão, em determinados

momentos. Uma vez que, usar como *threshold* o intervalo de consumos (IC) poderá ser limitativo, o que pode resultar na identificação de grande número de valores fora desse *threshold*, também se implementou como *threshold* o dobro e o triplo do intervalo de consumos obtido (2IC e 3IC). A Figura 33 apresenta um exemplo da procura de ocorrências fora do padrão de consumo habitual (pontos anómalos), tendo em conta os três *thresholds* (IC, 2IC e 3IC), durante os 7 meses em estudo, para um cliente com consumo doméstico.

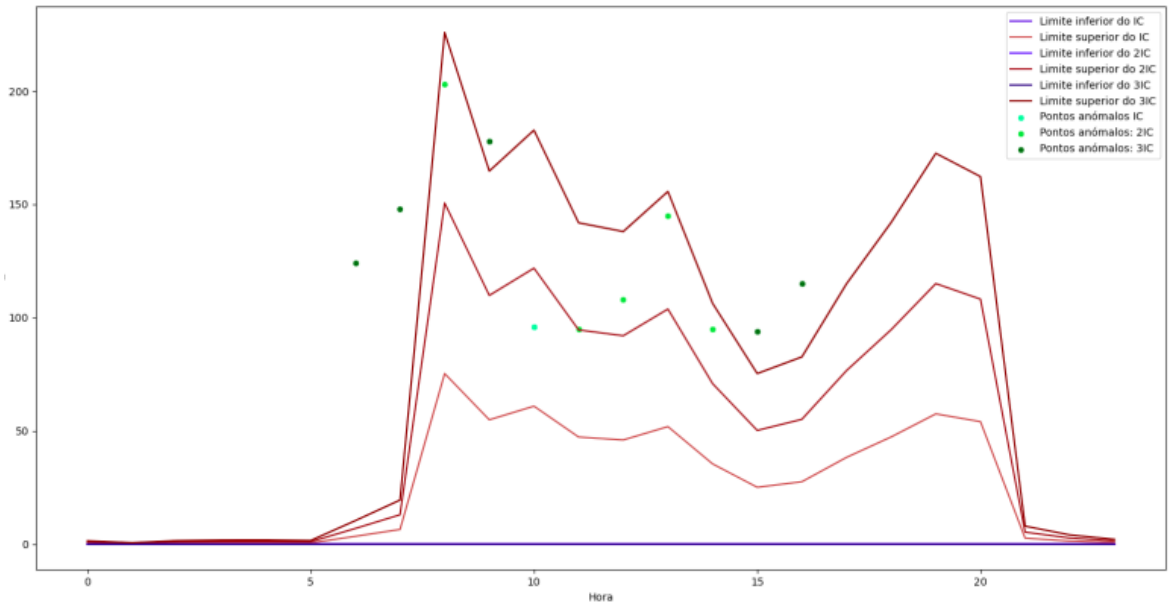


Figura 33. Ocorrências fora do padrão esperado de um cliente (nos 7 meses em estudo)

Como se pode observar na Figura 33, existem 11 ocorrências fora do intervalo de consumos (IC), 9 fora do dobro do intervalo (2IC) e 5 fora do triplo (3IC). Estas ocorrências detetadas, que estão fora do expectável, representam registos de consumos (l/h) fora do padrão estabelecido, para cada hora do perfil onde este local do exemplo está inserido.

Estes casos, denominados nesta análise como pontos anómalos, podem ser classificados como:

- Desvios pontuais do padrão de consumo: caso estes pontos surjam pontualmente, então representam situações fora do padrão considerado normal num determinado momento (numa hora especifica). Estes pontos anómalos poderão representar reais anomalias pontuais.

- Alteração de padrão de consumo: caso estes pontos surjam com mais frequência, podem indicar que este consumidor sofreu uma alteração na sua rotina de consumo de água e isto leva a que a associação deste consumidor a um determinado cluster tenha que ser revista. Uma alteração de

padrão de consumo pode ocorrer para um dado consumidor. Assim, um consumidor associado a um determinado cluster poderá em algum momento mudar para outro cluster.

5. CONCLUSÕES E TRABALHO FUTURO

A gestão do consumo de água doméstico ou não doméstico é, obviamente, a principal preocupação das empresas de distribuição de água. Evitar a existência de perdas de água ou de consumos elevados desnecessários são, certamente, os seus maiores problemas. Assim, e com objetivo de auxiliar a empresa Águas do Norte S.A. com estes problemas, este projeto apresentou uma abordagem para a deteção de anomalias e de padrões de consumo, utilizando técnicas de estatísticas e de *machine learning*.

Numa primeira fase, durante a análise, tratamento e pré-processamento dos dados fornecidos, obtiveram-se as seguintes anomalias e conclusões:

- Existência de consumidores zero, isto é, locais sem registo de consumo durante o período analisado. Concluiu-se que, algumas destas situações, tiveram origem em casos de vandalismo ao contador ou locais desabitados e não declarados;

- Existência de consumos negativos (*reverse flow*). Estas ocorrências, foram indicadas pela empresa, como possível entrada anormal de água na rede pública. Um dos exemplos, são os clientes com captações de água própria com ligação (ilegal) à rede pública. Para garantir a qualidade da água é importante que não exista qualquer interligação entre as captações próprias e a rede pública. Embora, nenhuma das ocorrências encontradas neste estudo foi identificado pela empresa como sendo causa de uma interligação particular com a rede publica, é de grande importância a identificação destes valores negativos e posterior averiguação da causa e/ou razão.

- Existência de consumos que excediam o caudal máximo por hora previsto (*overflow*). Concluiu-se que estas anomalias tiveram origem em erros de leitura referentes a acertos de leitura ou registos de leituras não consecutivos (ausência de registos).

Numa segunda fase, efetuou-se uma divisão dos clientes estudados em *clusters*, tendo em conta os seus consumos médios por hora, dia e por mês. E, daqui concluiu-se que estes locais poderiam ser agrupados em padrões de consumo com base nas suas semelhanças de comportamento no consumo da água por hora, dia e mês. Este processo de segmentação de clientes e de padronização tornar-se-á útil para a empresa, uma vez que lhe fornece uma melhor compreensão da carga da rede, ou seja, permite à empresa obter uma melhor previsão e distribuição do consumo de água por hora, por dia e por mês. E, permite também que a empresa, futuramente, obtenha uma faturação mais precisa e consiga prestar serviços mais personalizados e mais específicos a cada cliente.

Num terceira e última fase, procedeu-se à determinação de intervalos de consumos para identificação de ocorrências fora do padrão de consumo habitual, para cada padrão de consumo encontrado na fase anterior. A ideia geral, é apresentada e exemplificada com um exemplo de um consumidor em particular. Foram propostos três tipos de *thresholds* para detetar registos fora do padrão considerado normal. No entanto, é necessário aprofundar este estudo e analisar com mais detalhe os pontos anómalos encontra-

dos (que pontos correspondem a uma real anomalia, quando é que houve uma alteração de padrão de consumo por parte do cliente, etc.), para que futuramente, possa existir um critério que classifique em que situações, estes registos fora do expectável, podem ser considerados de possíveis reais anomalias ou quando são situações onde o local em questão está a sofrer uma alteração de padrão de consumo e, nestes casos, o local terá que ser reavaliado e, possivelmente, terá que ser atribuído a outro perfil de consumo.

No geral, pode-se afirmar que, as metodologias apresentadas ao longo desta dissertação foram bem sucedidas e permitirão à empresa melhorar o conhecimento em relação aos seus consumidores e aos seus dados. No entanto, existem algumas críticas e limitações a apontar ao trabalho feito, tais como:

- Foi apenas aplicada uma técnica de *clustering*, a técnica de agrupamento *K-Means*. E, apesar de ser uma das técnicas mais utilizadas e conhecidas, tem algumas limitações, como por exemplo, a determinação prévia do valor de k *clusters*, ou então, o possível impacto do valor do centroide escolhido inicialmente no resultado final;
- Somente foi tido em conta, para toda a análise, o consumo médio por hora, dia e mês de cada local. E, teria sido vantajoso ter tido em conta parâmetros como a área de cada local em estudo, ou o número de pessoas que habitam ou frequentam o local, ter em conta se o local é uma casa, um apartamento ou um edifício, se tem ou não jardim/quintal. Todos estes atributos teriam sido úteis para obter uma melhor segmentação dos clientes.

Neste sentido, existem outras perpectivas de trabalho futuro. Tais como:

- Aplicar e avaliar outras técnicas de *clustering* (por exemplo: k-medoids, Self- Organizing Maps (SOM)) com base na série temporal (em vez de tomar o valor médio do consumo de água em cada hora, dia e mês);
- Incluir outros parâmetros na análise, para além do consumo de cada cliente, como por exemplo, a área de cada local, o número de pessoas, se é uma casa ou apartamento, se tem ou não jardim;
- Expandir a análise feita nesta dissertação a novos dados, com um intervalo de analise maior que 7 meses;
- Consolidar e fixar os padrões de consumo e, identificar os perfis de consumo dos clientes que não foram incluídos na análise ou, até mesmo, identificar os perfis de novos clientes, através da correlação entre a curva de consumo do local a identificar e cada uma das curvas dos padrões encontrados. Assim, o cliente a classificar ficará atribuído ao perfil de consumo cujo o valor da correlação, entre a série temporal dada pelo consumo em relação ao tempo do cliente e do perfil,

seja o menor. A correlação entre duas séries temporais a e b pode ser dada por Rodrigues e Gama (2009):

$$corr(a,b) = \frac{P - \frac{AB}{n}}{\sqrt{A_2 - \frac{A^2}{n}}\sqrt{B_2 - \frac{B^2}{n}}} \tag{5.1}$$

onde, $A = \sum(a_i)$, $B = \sum(b_i)$, $A_2 = \sum(a_i)^2$, $B_2 = \sum(b_i)^2$ e $P = \sum(a_i)(b_i)$.

BIBLIOGRAPHY

Aggarwal, C. e Reddy, C. Data clustering: algorithms and applications. 2013.

Araújo, E., Ferreira, F., Silva, D., Bicho, E., e Erlhagen, W. (2021). Clustering domestic water consumption profiles: a portuguese case study. In *XXVIII Meeting of the Portuguese Association for Classification and Data Analysis (JOCLAD 2021)*, pages 77–78.

Arregui, F. J., Pastor-Jabaloyes, L., Mercedes, A. V., e Gavara, F. J. (2020). Accuracy of solid-state residential water meters under intermittent flow conditions. *Sensors*, 20(18):5339.

Avni, N., Fishbain, B., e Shamir, U. (2015a). Water consumption patterns as a basis for water demand modeling. *Water Resources Research*, 51(10):8165–8181.

Avni, N., Fishbain, B., e Shamir, U. (2015b). Water consumption patterns as a basis for water demand modeling. *Water Resources Research*, 51(10):8165–8181.

Boulicaut, J.-F. e Jeudy, B. (2005). The data mining and knowledge discovery handbook, chapter constraint-based data mining.

Braei, M. e Wagner, S. (2020). Anomaly detection in univariate time-series: A survey on the state-of-the-art. *arXiv preprint arXiv:2004.00433*.

Chandola, V., Banerjee, A., e Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Chhikara, R. (1988). *The Inverse Gaussian Distribution: Theory: Methodology, and Applications*, volume 95. CRC Press.

Cokelaer, T. (2019). Fitter documentation. Último acesso: 05 Dez 2021.

do Norte, A. (2015). Quem somos. Último acesso: 16 Set 2021.

Duan, H. e Lee, P. (2016). Transient-based frequency domain method for dead-end side branch detection in reservoir pipeline-valve systems. *Journal of Hydraulic Engineering*, 142(2):04015042.

Duan, H., Lee, P., Che, T., Ghidaoui, M., Karney, B., e Kolyshkin, A. (2017). The influence of non-uniform blockages on transient wave behavior and blockage detection in pressurized water pipelines. *Journal of Hydro-Environment Research*, 17:1–7.

Ertöz, L., Steinbach, M., e Kumar, V. (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of the 2003 SIAM international conference on data mining*, pages 47–58. SIAM.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., e Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*, pages 77–101. Springer.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Foundation, P. S. (2021). Python standard library. Último acesso: 12 Set 2021.

Gargano, R., Tricarico, C., Granata, F., Santopietro, S., e De Marinis, G. (2017). Probabilistic models for the peak residential water demand. *Water*, 9(6):417.

Guha, S., Rastogi, R., e Shim, K. (2000). Rock: A robust clustering algorithm for categorical attributes. *Information systems*, 25(5):345–366.

Han, J., Pei, J., e Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

He, Z., Xu, X., e Deng, S. (2003). Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24(9-10):1641–1650.

Hill, D. J. e Minsker, B. S. (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software*, 25(9):1014–1022.

Hochenbaum, J., Vallis, O. S., e Kejariwal, A. (2017). Automatic anomaly detection in the cloud via statistical learning. *arXiv preprint arXiv:1704.07706*.

Ioannou, A., Kofinas, D., Spyropoulou, A., Laspidou, C., et al. (2017). Data mining for household water consumption analysis using self-organizing maps. *European Water*, 58:443–448.

Ioannou, A. E., Creaco, E. F., e Laspidou, C. S. (2021). Exploring the effectiveness of clustering algorithms for capturing water consumption behavior at household level. *Sustainability*, 13(5):2603.

Johnson, N. L., Kotz, S., e Balakrishnan, N. (1995a). Continuous univariate distributions, volume 2. volume 289, chapter 14: Lognormal Distributions. John wiley & sons.

Johnson, N. L., Kotz, S., e Balakrishnan, N. (1995b). Continuous univariate distributions, volume 2. volume 289, chapter 25:Beta Distributions. John wiley & sons.

Kaufman, L. e Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Kofinas, D. T., Spyropoulou, A., e Laspidou, C. S. (2018). A methodology for synthetic household water consumption data generation. *Environmental modelling & software*, 100:48–66.

Kossieris, P. e Makropoulos, C. (2018). Exploring the statistical and distributional properties of residential water demand at fine time scales. *Water*, 10(10):1481.

Kulanuwat, L., Chantrapornchai, C., Maleewong, M., Wongchaisuwat, P., Wimala, S., Sarinnapakorn, K., e Boonya-aroonnet, S. (2021). Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series. *Water*, 13(13):1862.

Laspidou, C., Papageorgiou, E., Kokkinos, K., Sahu, S., Gupta, A., e Tassiulas, L. (2015). Exploring patterns in water consumption by clustering. *Procedia Engineering*, 119:1439–1446.

learn developers, S. (2021). 2.3. clustering. Último acesso: 30 Out 2021.

Leys, C., Ley, C., Klein, O., Bernard, P., e Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766.

Li, L. (2019). Principal component analysis for dimensionality reduction. *Towards Data Science*.

Liemberger, R. e Wyatt, A. (2019). Quantifying the global non-revenue water problem. *Water Supply*, 19(3):831–837.

Mehrang, S., Helander, E., Pavel, M., Chieh, A., e Korhonen, I. (2015). Outlier detection in weight time series of connected scales. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1489–1496. IEEE.

Mounce, S. R., Mounce, R. B., e Boxall, J. B. (2011). Novelty detection for time series data analysis in water distribution systems using support vector machines. *Journal of hydroinformatics*, 13(4):672–686.

Müller, A. C. e Guido, S. (2016a). *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media, Inc.

Müller, A. C. e Guido, S. (2016b). *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media, Inc.

Pires, A. e Santos-Pereira, C. (2005). Using clustering and robust estimators to detect outliers in multivariate data.

Rahim, M. S., Nguyen, K. A., Stewart, R. A., Giurco, D., e Blumenstein, M. (2020). Machine learning and data analytic techniques in digital water metering: a review. *Water*, 12(1):294.

Rao, C. R. e Govindaraju, V. (2013). *Machine learning: theory and applications*. Newnes.

Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K., e Kolehmainen, M. (2010). Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Applied Energy*, 87(11):3538–3545.

Rodrigues, P. P. e Gama, J. (2009). A system for analysis and prediction of electricity-load streams. *Intelligent Data Analysis*, 13(3):477–496.

Santos, M. C., Borges, A. I., Carneiro, D. R., e Ferreira, F. J. (2021). Synthetic dataset to study breaks in the consumer's water consumption patterns. In *2021 4th International Conference on Mathematics and Statistics*, pages 59–65.

Sardinha, J., Serranito, F., Donnelly, A., Marmelo, V., Saraiva, P., Dias, N., e Rocha, V. (2017). Controlo ativo de perdas de água. Último acesso: 20 Set 2021.

Silva, D., Sampaio, M., Milagres, C., Alves, V., e Ferreira, F. (2019). Flow4link - the flow in the hand. In *18th International Flow Measurement Conference (Flomeko2019)*.

Surendran, S. e Tota-Maharaj, K. (2018). Effectiveness of log-logistic distribution to model water-consumption data. *Journal of Water Supply: Research and Technology—AQUA*, 67(4):375–383.

Surendran, S. e Tota-Maharaj, K. (2020). 3-parameter log-logistic distribution modelling and scenario development to evaluate the united kingdom's water demand. *Institute of Water Journal*, pages 4–9.

Weisstein, E. W. (2021). Log normal distribution. Último acesso: 05 Dez 2021.

Yang, A., Zhang, H., Stewart, R. A., e Nguyen, K. (2018). Enhancing residential water end use pattern recognition accuracy using self-organizing maps and k-means clustering techniques: Autoflow v3. 1. *Water*, 10(9):1221.

Zhao, Q., Zhu, Y., Wan, D., Yu, Y., e Cheng, X. (2018). Research on the data-driven quality control method of hydrological time series data. *Water*, 10(12):1712.