UAlg ISE
UNIVERSIDADE DO ALGARVE
INSTITUTO SUPERIOR DE ENGENHARIA

IPS
Instituto Politécnico de Setúbal
Escola Superior de Tecnologia do Barreiro

ESTUDO DE MODELOS DE RECONSTRUÇÃO DE SÉRIES TEMPORAIS DE CAUDAL EM SISTEMAS DE ABASTECIMENTO DE ÁGUA

Carlos Ferreira Pinto Ascensão

Mestrado em Engenharia Civil
Área de Especialização: Estruturas
Dissertação

ORIENTADOR(ES): Professor Doutor Nelson Jorge Gaudêncio Carriço
Professora Doutora Maria Raquel Feliciano Barreira

outubro de 2023

Dissertação submetida no Instituto Politécnico de Setúbal

ESTUDO DE MODELOS DE RECONSTRUÇÃO DE SÉRIES TEMPORAIS DE CAUDAL EM SISTEMAS DE ABASTECIMENTO DE ÁGUA

Mestrado em Engenharia Civil

DECLARAÇÃO DE AUTORIA DO TRABALHO

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

Carlos Ferreira Pinto Ascensão

Assinado por: **Carlos Ferreira Pinto Ascensão**
Num. de Identificação: 14598138
Data: 2023.10.09 21:04:09+01'00'

(assinatura)

DIREITOS DE COPIA OU COPYRIGHT

AGRADECIMENTOS

Gostaria de expressar meu profundo agradecimento aos meus pais, pelo apoio incondicional que me deram durante todo o processo de realização da minha dissertação. Agradeço pelo amor, pelo incentivo e pelo apoio emocional que me deram.

Quero agradecer à Joana, pelo apoio e pelo incentivo que me deu durante a realização da minha dissertação. Agradeço pelo carinho, pelo apoio emocional e pelo ombro amigo que me ofereceu, mesmo em momentos difíceis.

Quero agradecer à Professora Doutora Raquel Barreira e ao Professor Doutor Nelson Carriço, pelo apoio, orientação e disponibilidade durante a realização da minha dissertação. Agradeço pelo tempo e pelo esforço que dedicaram ao meu projeto, pelo conhecimento e pela experiência que compartilharam comigo. Sem esta orientação e o apoio, não teria sido possível concluir este trabalho com sucesso.

Quero agradecer ao meu amigo e colega Bruno Ferreira, pelo apoio e companheirismo que me deu durante a realização da minha dissertação. Agradeço pelo tempo que dedicou a discutir e a ajudar a desenvolver o meu projeto, pelos conhecimentos e pela experiência que compartilhou comigo.

RESUMO

O presente trabalho de mestrado, apresenta um estudo comparativo de técnicas de reconstrução de séries temporais de caudal dos sistemas de abastecimento de água, recorrendo a modelos de previsão. Normalmente, nas séries temporais de caudal dos sistemas de abastecimento de água, são encontrados dados erróneos que devem ser tratados e validados. Estas falhas nos dados, podem ter origem durante o processo de aquisição e/ou serem resultantes de problemas nos sensores que recolhem a informação. A presença destes dados erróneos, nas séries temporais de caudal dos sistemas de abastecimento de água, restringe o seu uso em tarefas de gestão, de operacionalização e monitorização dos sistemas. O processo de validação, identifica os dados anómalos e remove-os da série temporal, originando dados omissos. Estes dados podem ser estimados, recorrendo a modelos de previsão. Com o intuito de reconstruir as séries temporais de caudal de sistemas de abastecimento de água, comparou-se o desempenho e o tempo de computação entre um modelo autorregressivo (ARIMA sazonal), dois modelos de suavização exponencial (Holt Winters e Holt Winters de dupla sazonalidade), um modelo de aprendizagem automática (SVR), um modelo híbrido (abordagem Quevedo) e uma melhoria ao modelo híbrido. O desempenho e o tempo de computação dos modelos foram avaliados considerando três casos de estudo reais, representativos de uma grande percentagem das entidades gestoras portuguesas. Foi considerado, no máximo, um mês e cinco dias de registos históricos com intervalos de 1 hora e 10 minutos, para a previsão de um dia da semana e de um feriado, respetivamente. Na previsão de um dia da semana, com intervalos de 10 minutos entre cada medição, o modelo SVR obteve o melhor desempenho e foi dos modelos mais rápidos a realizar a previsão, à semelhança da abordagem preconizada por Quevedo. Na previsão de um feriado com intervalos de 10 minutos entre cada medição, nenhum modelo conseguiu prever o feriado, apenas abordagem de Quevedo modificada conseguiu aproximar-se dos valores reais de caudal, sendo o mais rápido a obter uma previsão.

PALAVRAS-CHAVE: Caudal, previsão, métodos de reconstrução, séries temporais, sistemas de abastecimento de água

ABSTRACT

The present master's thesis presents a comparative study of techniques for the reconstruction of flow time series of water supply system using forecasting models. Erroneous data is often found in water supply system flow rate time series and must be treated and validated. These data errors may occur during the acquisition process and/or be the result of sensor problems. The presence of these data errors in water supply system flow rate time series restrict their use in management tasks, operationalization, and monitoring of the systems. The validation process identifies anomalous data and removes it from the time series, resulting in missing data. These data can be estimated using forecasting models. To reconstruct the flow rate time series of water supply systems, the performance and computational time of an autoregressive model (seasonal ARIMA), two exponential smoothing models (Holt Winters and double-seasonality Holt Winters), a machine learning model (SVR), a hybrid model (Quevedo approach), and an improvement to the hybrid model were compared. The performance and computational time of the models were evaluated based on three real-life case studies, representative of a large percentage of Portuguese management entities. A maximum of one month and five days of historical records with intervals of 1 hour and 10 minutes were considered for the prediction of a weekday and a holiday, respectively. In the prediction of a weekday with 10-minute intervals between each measurement, the SVR model achieved the best performance and was the fastest to perform the prediction, similar to the approach proposed by Quevedo. In the prediction of a holiday with 10-minute intervals between each measurement, no model was able to predict the holiday, only the modified Quevedo approach was able to approximate the actual flow rate values, being the fastest to obtain a prediction.

KEYWORDS: Flow rate, forecasting, reconstruction methods, time series, water supply systems.

ÍNDICE GERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

SÍMBOLOS E ABREVIATURAS

ADF - Augmented Dickey - Fuller
AIC - Critério de Informação de Akaike
AR - Auto Regressive
ARFIMA -Auto Regressive Fractionally Integrated Moving Average
ARIMA -Auto Regressive Integrated Moving Average
ARMA -Auto Regressive Moving Average
BIC -Critério de Informação Bayesiano
CE1 -Caso de Estudo 1
CE2 -Caso de Estudo 2
CE3 -Caso de Estudo 3
CSV - Comma Separated Values
EG - Entidade Gestora
FAC - Função de Autocorrelação
FACP - Função de Autoorrelação Parcial
GARCH -Generalized Auto Regressive Conditional Heteroskedasticity
IA - Inteligência Artificial
MA - Médias Móveis
ML - Machine Learning
RMSE - Root Mean Square Error
RNA - Redes Neuronais Artificiais
RNCC - Redes Neuronais de Correlação em Cascata
RNFF - Redes Neuronais Feed
RNRG - Redes Neuronais de Regressão Generalizada
SAA - Sistemas de Abastecimento de Água
SVM - Support Vector Machine
SVR -Support Vector Regression
WISDom -Water Intelligence System Data
ZMC - Zona de Medição e Controlo

1. INTRODUÇÃO

1.1. ENQUADRAMENTO

As entidades gestoras (EG) de sistemas de abastecimento de água (SAA), em áreas urbanas, enfrentam novos desafios em tarefas como a gestão, a operação e a monitorização dos seus sistemas, devido à escassez dos recursos hídricos, ao aumento das necessidades energéticas, ao envelhecimento das redes, à regulamentação mais rigorosa e à constante preocupação pelo impacto ambiental provocado pelo uso da água (Puig *et al.*, 2017).

O objetivo de qualquer SAA é o de fornecer água aos utilizadores (i.e., população em geral, comércio, serviços e indústria) em quantidade, qualidade e pressão satisfatórias, sem qualquer tipo de interrupção, de maneira a conseguir-se um sistema fiável. A operação e a monitorização de SAA requerem ferramentas de previsão de caudais e/ou consumos, a fim de manter um equilíbrio entre a oferta e a procura (Groppo *et al* 2019).

A investigação na área da previsão de caudais nos SAA tem sido bastante intensa nas últimas duas décadas, o número de artigos publicados tem aumentado exponencialmente, como se pode facilmente constatar por uma rápida pesquisa nas plataformas *Web of Science* (Clarivate) ou *Scopus* (Elsevier). Segundo Ghalehkhondabi *et al.* (2017), entre 2010 e 2015 os artigos publicados com títulos incluindo as palavras água, previsão e consumo, passaram de 60 publicações para 160 publicações, respetivamente. Vários investigadores e especialistas demonstram nas suas publicações as diferentes aplicações possíveis e metodologias adotadas, como resumem Ghalehkhondabi *et al.* (2017) e Groppo *et al.* (2019).

A previsão de caudais é um procedimento recorrente na fase de projeto (e.g., planeamento de novas expansões da rede) e na fase de exploração dos SAA (e.g., previsão de consumos). No entanto, a previsão poderá ter um horizonte diferente mediante cada objetivo: a previsão com uma longa duração é usada para o planeamento e dimensionamento da rede, enquanto que a previsão com uma curta duração é utilizada para a operação dos sistemas (Donkor *et al.*, 2014). A operação dos sistemas deve ser realizada de forma preditiva, sendo que as ações de controlo devem ser programadas com antecedência e com um horizonte adequado. Nas redes de distribuição de água esse horizonte é, geralmente, de 24 horas (Puig *et al.*, 2017).

A implementação de ferramentas que permitam a previsão de curta duração de caudais apresenta bastantes benefícios, nomeadamente, do ponto de vista de operação, permite que as EG determinem a regulação de válvulas e otimização das bombagens para corresponder aos caudais previstos, melhorando a eficiência energética através da redução dos consumos de energia. Do ponto de vista de tratamento da água, permite o tratamento de água necessária para responder à previsão de caudais, e do ponto de vista da vulnerabilidade, é possível estabelecer uma operação em tempo real, comparando os valores de caudal previsto com os valores de caudal reais, conseguindo-se assim identificar possíveis falhas no sistema e criar alertas no sistema de telegestão (Herrera *et al.*, 2010).

A operação de SAA em tempo real passa, numa primeira fase, pela aplicação de sistemas de telemedição, que recolhem e armazenam os dados das medições de caudal com uma dada frequência de aquisição, de forma sistemática e contínua. Os dados armazenados formam séries temporais que podem ser utilizadas para a previsão de caudais. No entanto, estas séries apresentam algumas falhas resultantes do processo de aquisição de dados dos sensores de telemedição e do sistema de comunicação, nomeadamente, valores extremos, valores nulos, e descontinuidades nas medições. Estas falhas põem em causa a validade dos registos históricos, indispensáveis para processos que necessitam de séries de dados completas para obter análises ou conclusões significativas. Como tal, para utilização das séries temporais de caudal, estas têm de ser validadas, no caso de valores corretos, e reconstruídas, no caso de valores anormais e omissos. Após este processamento da série (i.e., validação e reconstrução) é possível aplicar técnicas avançadas (e.g., aprendizagem automática) para a identificação, em tempo real, de eventos anómalos (e.g., roturas) (Puig *et al.*, 2017).

1.2. OBJETIVOS DA DISSERTAÇÃO

A presente dissertação de mestrado tem como principal objetivo o estudo de modelos de reconstrução de séries temporais para previsão de caudais, em sistemas de abastecimento de água (SAA), aplicado a três casos de estudo reais. Para esse efeito, numa primeira instância, será imprescindível proceder a uma revisão bibliográfica relacionada com modelos de previsão de séries temporais de caudais, para posterior análise e identificação das principais falhas, conseguindo-se, assim, realizar uma escolha dos modelos de previsão mais adequados ao problema e aplicação ao caso de estudo real.

A presente dissertação tem os seguintes objetivos específicos:

- Revisão da literatura relacionada com a validação e reconstrução de séries temporais de caudal a partir de modelos de previsão;
- Seleção dos modelos mais adequados para a resolução das falhas identificadas;
- Aplicação dos modelos de previsão selecionados aos casos de estudo;
- Análise das séries temporais dos casos de estudo e identificação das suas principais falhas;
- Compilação e análise dos resultados obtidos;
- Apresentação das principais conclusões e recomendações para trabalhos futuros.

1.3. ESTRUTURA DA DISSERTAÇÃO

A presente dissertação de mestrado é constituída por cinco capítulos. No capítulo introdutório (Capítulo 1), realiza-se o enquadramento da dissertação e apresentam-se os objetivos gerais e específicos a serem atingidos. No segundo capítulo, realiza-se a síntese de conhecimentos (Capítulo 2), onde se apresenta a revisão de literatura realizada e algumas considerações importantes a serem tomadas. No Capítulo 3, apresenta-se a metodologia utilizada para a

reconstrução das séries temporais que, posteriormente, será aplicada aos casos de estudo. No Capítulo 4, analisam-se os resultados obtidos de cada modelo aplicado, aos casos de estudo, a partir da metodologia apresentada no capítulo anterior. No Capítulo 5, apresentam-se as principais conclusões e recomendações para trabalhos futuros.

2. SÍNTESE DE CONHECIMENTOS

No presente capítulo apresenta-se a síntese de conhecimentos sobre validação e reconstrução de séries temporais de caudal, em sistemas de abastecimento de água (SAA), através de técnicas que recorrem a métodos de previsão. As séries temporais só serão reconstruídas caso não sejam consideradas como válidas, pelo que é essencial realizar-se uma revisão geral sobre os processos de validação e reconstrução de séries temporais de abastecimento de água.

2.1. SÉRIES TEMPORAIS DE CAUDAL DOS SISTEMAS DE ABASTECIMENTO DE ÁGUA

2.1.1. MEDIÇÃO DE CAUDAL

A medição do caudal nos SAA é essencial para a gestão e operação desses sistemas, e, devido ao menor custo de aquisição e maior disponibilidade de equipamentos de medição, a quantidade destes equipamentos nas redes tem aumentado consideravelmente. Com o objetivo de aumentar o conhecimento acerca dos caudais nos SAA, as EG usualmente dividem a sua rede de distribuição por zonas devidamente delimitadas e identificadas, facilitando o controlo de entradas e saídas da água através do balanço de caudais e estimando o comportamento dos consumos. Normalmente, estas zonas são designadas por Zona de Medição e Controlo (ZMC). A delimitação e dimensão da zona é definida tendo em conta diversos fatores, nomeadamente, a topologia da rede, a densidade populacional e de ramais na rede. O número de pontos de entrada de água, deve ser o menor possível (Henriques *et al.*, 2006).

Geralmente, os valores de caudal de uma ZMC são obtidos recorrendo a um medidor de caudal (caudalímetro), localizado à entrada da rede. Estes equipamentos podem ser divididos em três grupos, caudalímetros ultrassónicos, deprimogênios e eletromagnéticos, sendo estes últimos o tipo de caudalímetro mais usual (Henriques *et al.*, 2006).

Os equipamentos de telemetria são constituídos por sensores, existindo diversos tipos de sensores com diferentes características. Normalmente, nos SAA são instalados sensores que registam as medições de caudal (Henriques *et al.*, 2006). Habitualmente, os sensores de caudal realizam os registos em intervalos de tempo iguais, tipicamente entre 5 e 15 minutos (Barrela *et al.*, 2017; Romano e Kapelan, 2014). No entanto, existem sensores que devido às suas diferentes características (e.g., sensores de caudal por impulso) realizam registos com intervalos de tempo irregulares. Os sensores de impulso não registam os dados em intervalos de tempo fixo, mas sim por um volume fixo de caudal que passa pelo sensor (e.g., um impulso por metro cúbico) (Boyle *et al.*, 2013).

Os dados medidos pelos caudalímetros são, tipicamente, armazenados num *data logger*, que se encontra associado a uma rede de comunicações que os transmite ao sistema de

telegestão (Quevedo *et al.*, 2010a). Quando a EG não tem capacidade financeira ou não existem condições para haver uma rede de comunicações, os dados são armazenados no *data logger* até à sua recolha no local. O *data logger* tem uma capacidade de armazenamento limitada, pelo que é desejável que a recolha dos dados seja efetuada antes de atingir esse limite sob pena de haver perda dos dados mais antigos. Os sistemas de telegestão são compostos por equipamentos elétricos e eletromecânicos, equipamentos de instrumentação, redes de comunicação e alarmes (Henriques *et al.*, 2006).

Os conjuntos de dados de caudal são normalmente organizados em ficheiros em formato *CSV (Comma Separated Values),* em que os dados se encontram separados por vírgulas, ou outro tipo de separador, e não por colunas, tornando, assim, menos complexo trabalhar os dados entre diferentes *softwares*. Na Figura 1, apresenta-se um excerto de um ficheiro *CSV* com registo da data e do valor de caudal, sendo estes separados por uma vírgula. Estes ficheiros constituem séries temporais que podem conter dados erróneos, podendo conduzir a análises e conclusões incorretas. Como tal, é imperioso ter dados fiáveis sendo necessário proceder à validação dos mesmos (Kirstein *et al.*, 2019). A realização da validação dos dados garante a sua precisão e fiabilidade, o que permite às EG construir a base para a aplicação de ferramentas avançadas que possibilitem a deteção de roturas, a otimização do tratamento da água e a operação do sistema (Puig *et al.*, 2017).

```
1  date,value
2  2018-01-01 00:00:52,12.32
3  2018-01-01 00:01:16,10.47
4  2018-01-01 00:01:49,12.44
5  2018-01-01 00:02:20,10.6
6  2018-01-01 00:02:38,8.8
7  2018-01-01 00:03:19,10.68
8  2018-01-01 00:03:30,11.51
9  2018-01-01 00:03:38,12.72
10 2018-01-01 00:03:49,14.54
11 2018-01-01 00:04:39,12.63
12 2018-01-01 00:05:24,14.44
13 2018-01-01 00:06:03,12.49
14 2018-01-01 00:06:17,14.31
15 2018-01-01 00:06:23,16.31
16 2018-01-01 00:06:57,18.28
17 2018-01-01 00:07:20,16.37
18 2018-01-01 00:07:35,14.5
19 2018-01-01 00:08:05,16.38
20 2018-01-01 00:08:13,18.39
21 2018-01-01 00:08:21,20.26
22 2018-01-01 00:08:39,22.32
23 2018-01-01 00:08:52,20.28
24 2018-01-01 00:08:59,18.32
25 2018-01-01 00:09:09,16.36
26 2018-01-01 00:09:52,18.25
27 2018-01-01 00:11:27,20.14
28 2018-01-01 00:12:35,18.31
```

Figura 1 – Excerto de um ficheiro CSV com registo de dados de caudal de um SAA.

Independentemente do tipo de sensor utilizado na aquisição dos dados, é frequente encontrar problemas nos registos (i.e., valores extremos isolados, longos períodos sem medição ou com medição constante) (Mounce *et al.*, 2010). Normalmente, estes problemas são gerados devido a falhas no sensor ou no sistema de comunicação entre o sensor e o *data logger* ou no próprio sistema de telegestão, devido a falhas de energia, originando assim dados erróneos ou

perdidos (Loureiro *et al.*, 2016; Xu *et al.*, 2020). Outro problema comum, prende-se com a falta de fiabilidade dos sensores, equipamentos mecânicos que normalmente se encontram enterrados sob pavimentos de estradas ou arruamentos e que sofrem perturbações devido às vibrações de ações externas (e.g., tráfego).

Frequentemente, os sensores não são alvo de manutenções adequadas devido à dificuldade de acesso, o que não permite verificar se estão a funcionar corretamente. A conjugação destes fatores resulta em medições incorretas (i.e., duplicação de leituras, valores em falta). Estes dados, falsos ou em falta, devem ser considerados como não válidos e substituídos por valores estimados (Cugueró-Escofet *et al.*, 2016). Muitas vezes, esses dados são preenchidos manualmente por técnicos das EG que recorrem ao conhecimento empírico adquirido ao longo dos anos de análise de dados. No entanto, este processo manual é moroso e reduz a fiabilidade dos dados para a aplicação de ferramentas de aprendizagem automática (por exemplo, para deteção de roturas). Para além disso, o aumento do número de equipamentos de medição resulta, por consequência, no aumento do número de dados a processar, tornando a sua análise humanamente inviável. Na Figura 2 apresenta-se o registo histórico de um dia completo de medições de caudal, representado a azul, sendo os problemas mais comummente encontrados, em séries temporais de caudal dos SAA, representados por simbologia a vermelho. Primeiramente, verificam-se leituras duplicadas, ou seja, para a mesma hora existe mais que uma medição. Em seguida, podem-se observar leituras com valores anormalmente altos, seguidos de longos períodos sem medições e de patamares estáticos. Por último, observam-se valores anormalmente baixos.

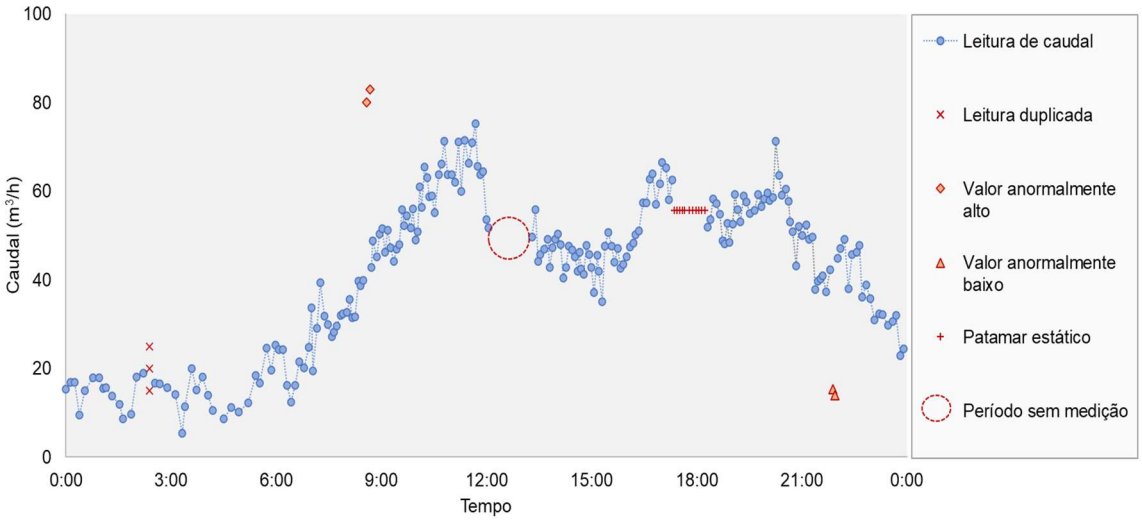

Figura 2 - Exemplo de problemas identificados numa série bruta de dados de caudal

Os dados recolhidos pelas EG são utilizados na gestão e exploração dos sistemas de abastecimento de água. Um dos objetivos principais e comum a todas as EG portuguesas é a minimização das perdas de água por razões ambientais, sociais, de saúde pública e económicas que se traduzem na diminuição do custo de produção e de manutenção, nomeadamente, na redução dos custos com encargos energéticos associados ao processo de bombagem de água, ao processo de tratamento e reparação das roturas na rede,

conseguindo-se assim reduzir o custo da água para as EG e da mesma forma da tarifa para os consumidores (Henriques *et al.*, 2006).

Posteriormente à recolha dos dados, estes são transmitidos aos sistemas de telegestão que permitem a visualização dos mesmos para análise das eventuais causas dos problemas, de modo a possibilitar a prevenção de situações futuras. A telegestão dos SAA baseia-se numa ferramenta de suporte para a exploração do sistema, permitindo centralizar, comandar e monitorizar as operações do sistema, neste caso, de captação, tratamento, transporte e distribuição de água. Torna-se, assim, possível, compreender em tempo real o estado funcional da rede de modo a ser possível intervir de acordo com as necessidades, garantindo assim a qualidade, a eficiência do sistema e o melhoramento global da qualidade do serviço prestado, permitindo uma melhor segurança na exploração dos sistemas, através da implementação de automatismos que mantem os valores para o funcionamento do sistema corretos (Val, 2013).

A possibilidade de controlar as diferentes variáveis (e.g., de qualidade e quantidade) na exploração do sistema também é um dos aspetos que tornam os sistemas de telegestão imprescindíveis, nos dias de hoje, na gestão dos SAA. O armazenamento do conjunto de dados em registos históricos visa uma melhoria na gestão estatística possibilitando assim análises das condições técnicas e económicas da exploração, bem como informações relevantes que permitam um adequado planeamento e apoiem a tomada de decisão para a prevenção de situações futuras que possam comprometer o serviço (Henriques *et al.*, 2006).

2.1.2. SAZONALIDADES DAS SÉRIES TEMPORAIS DE CAUDAL

Diz-se que uma série temporal é sazonal, toda a série que seja afetada por fatores sazonais, como a época do ano ou o dia da semana (Hyndman e Athanasopoulos, 2012), enquanto as séries temporais não sazonais são aquelas que não apresentam um padrão percetível e repetitivo ao longo do tempo. Os modelos de previsão devem ser selecionados tendo em conta as características das séries temporais (e.g., sazonalidade) visto que as séries temporais dos SAA demonstram grande evidência de ciclos diários e semanais (Marinis *et al.*, 2008; Mamade, 2013).

A previsão de caudal nos SAA com modelos que recorrem a padrões de consumo para as suas previsões tem sido uma prática recorrente (Bakker *et al.*, 2013). No setor urbano da água, que inclui o consumo doméstico, o consumo comercial e o consumo público, é possível definir facilmente padrões de consumo devido aos hábitos da população. Numa cidade, independentemente da sua localização, são percetíveis picos de consumo de água no início da manhã e no início da noite, ou seja, quando a maioria da população se encontra nas suas habitações, a preparar as suas refeições, a tomar duches, a fazer limpezas, entre outras tarefas. Estes picos de consumo também podem ser associados às horas de refeição, pois também é usual verificar-se um pico, embora de menor dimensão, nas horas de almoço. Na Figura 3, apresenta-se um consumo de caudal numa zona urbana, para um dia útil (i.e., segunda-feira), para um dia de sábado e para um dia de domingo.

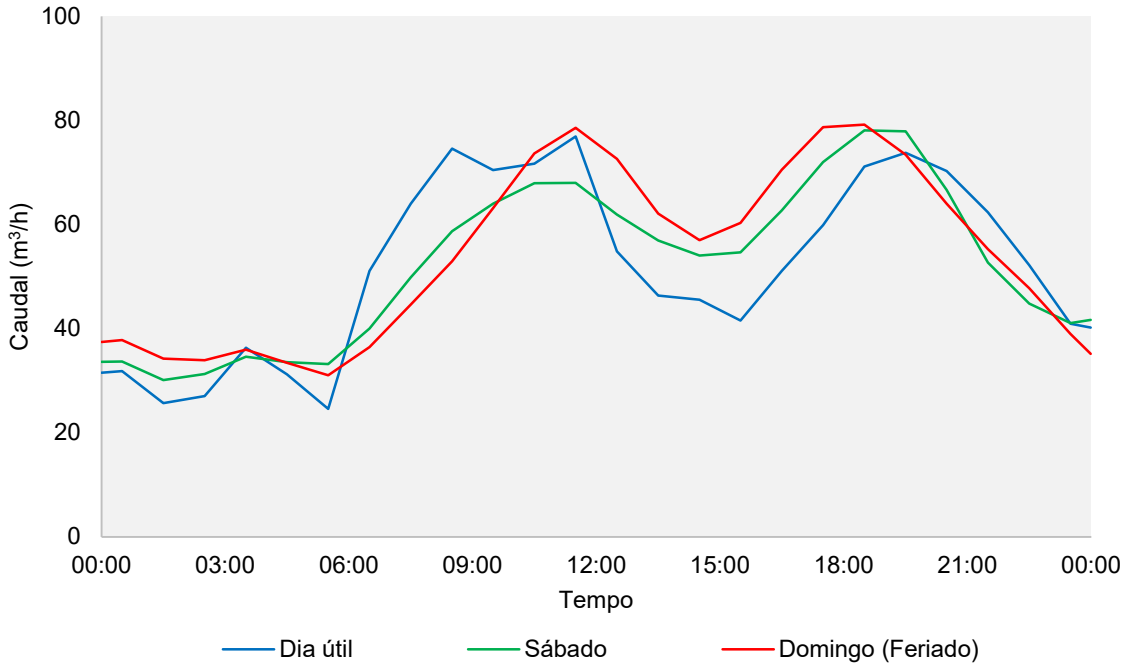

Figura 3 - Consumo de caudal numa zona urbana, num dia útil, num sábado e num domingo

Na Figura 3, para um dia útil, observa-se um aumento acentuado do caudal no período da manhã (i.e., das 6 da manhã às 9 da manhã), um ligeiro aumento do caudal no período de almoço e por último, um pico de consumo no período da noite. No entanto, estes picos de consumo podem alterar-se consoante o dia da semana (i.e., dias úteis, sábados, domingos e feriados).

Analisando o caudal de um sábado, observa-se que os picos anteriormente mencionados para um dia útil, alteraram-se um pouco. Num sábado, o pico de consumo no período da manhã verifica-se mais tarde e não existe o pico de consumo da hora de almoço, apenas se observa um pico de consumo no período da noite.

O caudal de um domingo, demonstra um comportamento semelhante ao consumo de caudal diário num sábado, existindo apenas dois picos de consumo, um pico de consumo no período da manhã e outro pico de consumo no período da noite.

Os padrões de consumo de água são normalmente gerados com base no valor médio das medições passadas sendo, por isso, necessário ter dados históricos tratados e validados. Na literatura existente, nota-se que é inevitável ter em conta o dia da semana, ou seja, se estamos perante um dia útil (i.e., de segunda a sexta), ou dia de fim-de-semana (i.e., sábado ou domingo), sendo que, um feriado tem um comportamento mais próximo do domingo. Jowitt e Xu (1992) desenvolveram um estudo em que aplicaram três padrões de consumo de água diferentes, um para os dias úteis, um para os sábados e outro para os domingos, enquanto que Zhou *et al.* (2002), no seu modelo de previsão, decidiu usar apenas dois padrões de consumo, um para os dias úteis e outro para os fins-de-semana, incluindo feriados.

Na análise realizada por Alvisi *et al.* (2007), com uma série temporal de caudais/consumos diários, observou-se a existência de padrões nos quais é possível identificar sazonalidades ao nível da semana. Com uma série temporal de caudais/consumos horários observou-se sazonalidades diárias, mostrando um comportamento variável ao longo do dia, com padrões diferentes dependendo da hora do dia e do dia da semana. Alvisi *et al.* (2007) definiu para o seu modelo de previsão um padrão para cada dia da semana e, como a análise dos feriados mostrou um comportamento idêntico aos fins de semana, decidiu, assim, tratar os feriados como sendo dias semelhantes aos do fim de semana.

2.2. TÉCNICAS DE VALIDAÇÃO DE SÉRIES TEMPORAIS DE CAUDAL

O processo de validação tem sido abordado nas últimas décadas por diferentes autores, devido ao aumento da procura, por parte das EG, de políticas de gestão mais eficientes. Normalmente, os processos de validação baseiam-se em heurísticas simples e em alguns dados estatísticos das séries temporais dos SAA. Quevedo *et al.* (2010) desenvolveu um modelo que permite detetar valores anómalos numa série temporal de caudal real, numa zona da rede de distribuição de água da cidade de Barcelona (Espanha), substituindo os valores considerados como não válidos por previsões, seguindo uma abordagem similar à de Alvisi *et al.* (2007). O processo de validação apresentado tem em vista comparar as medições com um intervalo de valores, definindo um máximo e um mínimo. Quando a medição excede este intervalo, o valor dessa medição é considerado como "não válido", bem como o valor da medição anterior e posterior. Adicionalmente, cada medição é comparada com a previsão e se ultrapassar um determinado valor de erro residual, essa medição também é considerada como não válida, sendo substituída pelo valor da estimativa.

O processo de validação apresentado por Cugueró-Escofet *et al.* (2016) é baseado num conjunto de testes de "baixo nível", constituído por quatro testes que verificam a consistência do sinal produzido pelo sensor, e num conjunto de testes de "alto nível" que verificam a consistência das medições do sensor. O primeiro teste pretende avaliar o sistema de comunicação e detetar os seus problemas frequentes (e.g., longos períodos sem medição). O segundo teste tem em vista detetar as medições com valores negativos, verificando se os valores se encontram dentro da gama de leitura do sensor. O terceiro teste pretende sinalizar as medições com picos altos ou baixos, analisando a magnitude dos dados. O quarto teste procura verificar o estado do sensor através da relação entre os valores medidos de caudal e o estado das válvulas do sistema. O quinto teste verifica a consistência espacial das medições, tirando partido de mais do que uma variável do sistema, realizando uma relação entre as medições realizadas à entrada do sistema, as medições da altura de água do reservatório e o registo histórico. O sexto e último teste faz a verificação da consistência temporal, designado de modelo da série temporal. Este modelo recorre ao registo histórico de medições de caudal para realizar a estimativa através de uma alternativa, amplamente usada na modelação de séries temporais devido à sua simplicidade, baixos requisitos computacionais e facilidade de automação – método de suavização exponencial Holt-Winters de dupla sazonalidade.

Kirstein *et al.* (2019) desenvolveram uma metodologia com o objetivo de identificar e categorizar os dados anómalos nas séries temporais dos SAA, sugerindo dividir os dados pela

relevância da informação presente nos mesmos. O processo de identificação dos dados anómalos baseia-se num conjunto de 7 testes, com o intuito de identificar os dados que não contêm informação relevante, devido à ausência de valores, horas ou data duplicadas e medições constantes, consequência das falhas internas dos sensores ou falhas no sistema de comunicação e de armazenamento. Dados anómalos com alguma relevância na sua informação também são identificados. Estes afetam negativamente a qualidade dos dados, por possuírem mudanças discrepantes ou algum tipo de inconsistência temporal. Os dados correspondentes a eventos anómalos na rede (i.e., roturas na rede, consumos irregulares, abertura de válvulas) são identificados pela informação valiosa que contêm, imprescindível na aplicação de ferramentas avançadas (e.g., deteção de roturas). Depois de identificados todos os dados anómalos, estes são armazenados numa base de dados com um indicador de mau funcionamento, para análise e visualização futura, referindo que estes podem ser, posteriormente, validados através de técnicas de reconstrução, com modelos de análise de séries temporais e modelos de aprendizagem automática.

Ferreira *et al.* (2022) apresenta uma metodologia para o tratamento e validação de séries temporais de caudal dos SAA, para o uso em técnicas avançadas (e.g., aprendizagem automática) na deteção e localização de eventos anómalos (e.g., deteção e localização de roturas). Estas técnicas avançadas requerem séries temporais tratadas, validadas e que a sua frequência temporal esteja normalizada pelo que é imprescindível uma ferramenta com estas características antes de se realizar qualquer tipo de estudo em séries temporais de caudal dos SAA. A metodologia desenvolvida por Ferreira *et al.* (2022), foi implementada numa ferramenta computacional em código aberto e de livre distribuição. A documentação de apoio e o código fonte estão disponíveis no repositório GitHub, em https://github.com/Ferreira-B/Flowrate-time-series-processing, para livre utilização dos técnicos de entidade gestoras e por membros da comunidade académica.

A ferramenta computacional desenvolvida por Ferreira *et al.* (2022), tem implementados quatro passos para a validação e tratamento de séries temporais de caudal, nomeadamente, para: 1) identificação e remoção de valores anómalos, 2) reconstrução de falhas de curta duração, 3) normalização da frequência temporal e 4) reconstrução de falhas de longa duração. O primeiro passo identifica e remove as anomalias típicas em séries temporais de caudal dos SAA, com recurso a um conjunto de testes automáticos. A tipificação dos valores anómalos tem como base os problemas mais comuns encontrados nas séries temporais de caudal, conforme supracitado (i.e., duplicação de leituras, valores negativos, valores anormalmente baixos ou altos, períodos de variação anormalmente baixa e longos períodos sem medições). O segundo passo, realiza a imputação de valores pontuais recorrendo a técnicas de interpolação. Esta imputação é realizada quando a duração da falha resultante do primeiro método é inferior a um determinado valor (e.g., 1 minuto). O terceiro passo, realiza a normalização temporal da série para um intervalo pré-definido, através da integração numérica, recorrendo à regra dos trapézios entre as falhas de longa duração. O último passo efetua a reconstrução de falhas de longa duração, recorrendo a um modelo de previsão híbrido, resultado do estudo realizado na presente dissertação sendo o mesmo apresentado no seguinte Capítulo (i.e., Capítulo 3). A ferramenta foi utilizada para o tratamento das séries temporais dos casos de estudo, a apresentar no Capítulo 4. No entanto, visto que o quarto passo é resultado do estudo apresentado na presente dissertação, apenas foi possível recorrer à ferramenta computacional para identificar e remover os valores anómalos (i.e.,

passo 1), reconstruir as falhas de curta duração (i.e., passo 2) e normalizar a frequência temporal (i.e., passo 3) das séries temporais dos casos de estudo.

2.3. TÉCNICAS DE RECONSTRUÇÃO DE SÉRIES TEMPORAIS DE CAUDAL

A reconstrução de séries temporais tem como principal tarefa a estimativa de valores de dados não validados, sendo usual ser realizada por previsões geradas através de modelos. Um dos principais requisitos para a aplicação de um modelo de previsão é o registo histórico disponível. O volume de dados a utilizar irá depender do horizonte de previsão pretendido. De maneira a conseguir-se trabalhar com as diferentes sazonalidades dos SAA, a reconstrução requer modelos de previsão de curto prazo, com um horizonte de previsão de 24 horas (Gagliardi *et al.*, 2017; Antunes *et al.*, 2018).

Foram desenvolvidos vários modelos que permitem acomodar as diferentes sazonalidades das séries temporais dos SAA. O modelo preconizado por Quevedo *et al.* (2010), para trabalhar com séries temporais com duas sazonalidades, diária e semanal, é composta por dois módulos: um primeiro, encarregue de realizar a estimativa do caudal diário através de um modelo de previsão *Auto Regressive Integrated Moving Average* (ARIMA) e um segundo módulo, encarregue de realizar padrões de consumo com intervalos de 10 minutos. A previsão é obtida quando é realizada a distribuição do caudal diário pelo padrão de consumo com intervalos de 10 minutos.

O processo de reconstrução adotado por Cugueró-Escofet *et al.* (2016) baseia-se nos dois testes de "alto nível" utilizados no processo de validação. Os testes que verificam a consistência espacial e temporal do sensor dependem de modelos de previsão que também são usados na reconstrução da série temporal. Na fase de validação desenvolvida por Cugueró-Escofet *et al.* (2016), se algum dos testes apresentados detetar alguma falha, é iniciado o processo de reconstrução dos dados através de um dos testes apresentados. O erro residual associado às observações e às previsões é calculado através do erro quadrático médio. Consoante o erro mais baixo de cada um dos testes, é escolhido o valor previsto e substituído até a série estar validada. No último teste do processo de reconstrução é possível adotar um modelo ARIMA em alternativa ao modelo de suavização exponencial (i.e., Holt-Winters) (Puig *et al.*, 2017).

Donkor *et al.* (2014) fez uma revisão dos modelos de previsão publicados entre os anos de 2000 a 2010, sintetizando a literatura e classificando-os por métodos de previsão, a fim de identificar os modelos úteis para problemas relacionados com a tomada de decisão das EG, sendo o objetivo principal o de chamar a atenção para as fraquezas, problemas e lacunas na previsão de caudal nos SAA e sugerir como estes problemas podem ser tratados. Nas suas considerações finais, mencionou algumas fragilidades na literatura referente à previsão de séries temporais dos SAA, sugerindo que o desenvolvimento de modelos de previsão deve levar a modelos em que as variáveis de entrada sejam possíveis de adquirir e de fácil acesso. Modelos que necessitem de muitas variáveis representam um maior desafio na sua aplicação e na aquisição de dados, tornando assim a operacionalização dos modelos por parte das EG uma tarefa difícil. A dificuldade de aquisição de variáveis e a complexidade de tais modelos faz com que as entidades procurem modelos de previsão simples e de aplicação acessível. É

recomendado que os investigadores tenham em consideração a capacidade das EG em adquirir dados e monitorizar os modelos de previsão, perfazendo que os modelos devem ser económicos e simples o quanto possível, sem comprometerem a integridade estrutural e a qualidade da previsão.

As técnicas de reconstrução de séries temporais de caudal dos SAA requerem modelos de previsão que possibilitem explorar as sazonalidades deste tipo de séries. Diversos autores referem-se ao problema e às necessidades de prever os consumos dos SAA (Donkor *et al.*, 2014; Groppo *et al.*, 2019). No entanto, a temática da validação e reconstrução das séries temporais é pouco aprofundada e dessa forma é essencial compreender a literatura relacionada com modelos de previsão que tenham em conta as múltiplas sazonalidades das séries temporais dos SAA. Ainda assim, as séries temporais de consumo de eletricidade têm características idênticas às séries temporais dos SAA, devido aos hábitos de consumo da população, existindo assim também muita literatura relativa a previsões do consumo de eletricidade que pode contribuir de forma enriquecedora para o desenvolvimento de técnicas de reconstrução das séries temporais dos SAA.

Taylor *et al.* (2006) compararam as previsões obtidas através de vários modelos para séries temporais univariadas, um modelo ARIMA de sazonalidade dupla, um modelo de suavização exponencial de sazonalidade dupla, um modelo de redes neurais artificiais, um modelo de regressão e dois modelos tradicionais, aplicados à previsão de consumo energético, concluindo que o modelo de suavização exponencial tinha um melhor desempenho. Mohamed *et al.* (2010) desenvolveram um modelo ARIMA de sazonalidade dupla para a previsão do consumo de eletricidade. Os dados foram adquiridos com intervalos de meia hora e o modelo inclui sazonalidades diárias e semanais. Hassan *et al.* (2012) comparara os modelos ARIMA de sazonalidade dupla e ARFIMA (*Auto Regressive Fractionally Integrated Moving Average*) também de sazonalidade dupla, com o intuito de prever o consumo energético em intervalos de meia hora, sendo que o modelo ARFIMA produziu resultados sensivelmente melhores. Dudek (2013) apresentou um modelo com redes neurais simples para prever séries temporais com múltiplas sazonalidades, aplicando-o à previsão do consumo energético e comparando os resultados com modelos ARIMA e abordagens de suavização exponencial.

Relativamente a previsões de séries temporais de caudal dos SAA, foram explorados vários modelos e métodos com horizontes de previsão de curto prazo e que permitem, também, acomodar as diferentes sazonalidades. Caiado (2010) analisou o desempenho da previsão de caudal dos SAA, utilizando de vários modelos univariados, Holt-Winters, ARIMA e GARCH *(Generalized Auto Regressive Conditional Heteroskedasticity*), tendo em consideração sazonalidades semanais e anuais, e diferentes combinações de previsões com a finalidade de melhorar a precisão do modelo.

Nas últimas duas décadas, a inteligência artificial (IA) na forma de *Machine Learning* (ML) ou em português, modelos de Aprendizagem Automática, tem sido dos temas mais abordados nas publicações relacionadas com a previsão de caudal de água nos SAA (Brentan *et al.*, 2017). Segundo Donkor *et al.* (2014) e Ghalehkhondabi *et al.* (2017), os modelos de IA têm um desempenho razoável quando o horizonte da previsão é de curto prazo.

Na previsão mensal de caudal de água de um SAA, LeFirat *et al.* (2010) compararam várias técnicas de ML, tais como, Redes Neuronais Artificiais (RNA), como Redes Neuronais de

Regressão Generalizada (RNRG), Redes Neuronais de Correlação em Cascata (RNCC) e Redes Neuronais *Feed-Forward* (RNFF). Foram construídos seis modelos para cada uma das técnicas e comparados os seus desempenhos. O modelo M5-RNCC, que se baseia nos 5 valores mensais anteriores de caudal de água, teve um melhor desempenho, comparando com as restantes técnicas.

Li e Huicheng *et al.* (2010) aplicaram um modelo com dois módulos: um pretende modelar a tendência da série por meio de uma regressão linear múltipla e o segundo módulo, através de uma rede neuronal difusa (*fuzzy neural network)*, pretende modelar os componentes cíclicos anuais do caudal dos SAA. Herrera *et al.* (2010) compararam vários métodos de inteligência artificial, como RNA, regressão adaptativa multivariada por *splines*, florestas aleatórias (*Random Forests*) e regressões por vetor suporte (*Support Vector Regression*), para a previsão de caudal de água com um horizonte de previsão de 1 hora.

Mounce *et al.* (2011) desenvolveu uma abordagem para a detetar eventos anómalos nos SAA, recorrendo ao modelo de regressões por vetor de suporte (SVR) para obter previsões de caudal. Os resultados apresentados, demonstraram que o modelo SVR tem um bom desempenho na previsão e apresenta potencial para aplicação na operação em tempo real dos SAA.

Brentan *et al.* (2017) comparam SVR com um modelo híbrido de SVR e séries de Fourier adaptativas, para a previsão horária de caudal de água. Os resultados obtidos demonstraram que o modelo SVR teve problemas com os valores extremos, enquanto o modelo híbrido conseguiu ter um desempenho razoável na previsão.

Antunes *et al.* (2018) recorreu a vários modelos de aprendizagem automática para a previsão de caudal de água de duas EG portuguesas. As técnicas de aprendizagem automática utilizadas, tais como, modelos de RNA, *Random Forests,* SVR *e* modelos de k-vizinhos mais próximos *(k-nearest neighbours)*, foram comparadas com modelos tradicionais (i.e., ARIMA). Os modelos de aprendizagem automática demonstraram um desempenho global superior, embora os modelos ARIMA por vezes obtenham melhores resultados quando comparados com certos modelos de aprendizagem automática (i.e., RNA e SVR).

3. TÉCNICAS DE RECONSTRUÇÃO DE SÉRIES TEMPORAIS DE CAUDAL

No presente capítulo apresentam-se as técnicas implementadas na reconstrução das séries temporais de caudal de sistemas de abastecimento de água (SAA). Nos subcapítulos seguintes, apresenta-se a formulação de cinco métodos que permitem a reconstrução das séries temporais de caudal dos SAA e a implementação de uma melhoria a um desses métodos. Na primeira secção é apresentado o modelo autorregressivo (i.e., ARIMA sazonal), na segunda a técnica de reconstrução desenvolvida por Quevedo *et al.* (2010) e a melhoria à sua abordagem, na terceira apresentam-se os métodos de suavização exponencial (i.e., Holt-Winters simples e Holt-Winters de dupla sazonalidade). Na quarta secção é apresentado um modelo de aprendizagem automática (i.e., SVR). Por último, apresenta-se a métrica adotada na avaliação do desempenho das técnicas de reconstrução.

3.1. MODELO AUTORREGRESSIVO

Um dos modelos mais utilizados para a reconstrução de séries de caudal é o Autorregressivo integrado de médias móveis (ARIMA) que derivam da família dos modelos *Auto Regressive Moving Average* (ARMA). Estes modelos foram desenvolvidos por Box e Jenkins (1976), também conhecidos como modelos estocásticos, que descrevem a probabilidade de uma sequência de observações. Ou seja, fazer uma previsão entende-se como a distribuição da probabilidade de uma observação futura em uma determinada série temporal, dada uma amostra de observações anteriores.

Os modelos ARMA são modelos lineares para análise de séries temporais e são compostos por duas componentes principais: uma componente que trabalha os processos autorregressivos e outra que trabalha os processos de médias móveis. Os modelos ARMA requerem séries temporais estacionárias, ou seja, séries temporais em que as suas propriedades estatísticas (i.e., média, variância, autocorrelação, etc.) sejam constantes ao longo do tempo. Logo, as séries temporais com evidências de sazonalidade são séries temporais não estacionárias (Box *et al.*, 2018). A diferença entre um modelo ARMA e um modelo ARIMA é a componente de integração que existe neste último (daí o I no acrónimo) e que permite diferenciar a série de maneira a convertê-la numa série estacionária. A diferenciação de uma série temporal, geralmente é usada para remover os efeitos de tendência, estabilizar a média e variância de uma série temporal. A diferenciação de uma série temporal é a diferença do valor no período (i.e., t), com o valor do período anterior (i.e., t-1).

Para fazer previsões, os modelos ARIMA usam um polinómio dos valores anteriores junto com os erros de previsão anteriores. Os modelos ARIMA sazonais consideram um polinómio adicional para a componente da sazonalidade.

A função dos modelos ARIMA pode ser representada pelos graus do modelo (p, d, q), onde p representa o número de termos autorregressivos, d representa o número de diferenciações e q o número de erros de previsão desfasados na equação de previsão. A função polinomial dedicada à componente sazonal funciona apenas com uma periodicidade. Semelhante ao polinómio da componente regular, também pode ser representado por (P, D, Q)s onde P, D e Q representam os graus do modelo e s representa o número de períodos sazonais. Os modelos ARIMA sazonais podem ser representados pela seguinte expressão:

$$\emptyset_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (1)$$

Onde $\emptyset_p(B)$ é o polinómio autorregressivo de ordem p, $\theta_q(B)$ é o polinómio de médias móveis de ordem q, d é a ordem da diferenciação simples, $\Phi_P(B^s)$ é o polinómio autorregressivo sazonal de grau P em B^s, $\Theta_Q(B^s)$ é o polinómio de médias móveis sazonal de grau P em B^s e D é a ordem da diferenciação sazonal.

Na seleção dos graus do modelo ARIMA sazonal (p, d, q)(P, D, Q)s é essencial compreender se a série é estacionária ou não, com o objetivo de definir quantas ordens terão as componentes de diferenciação do modelo (i.e., diferenciação não sazonal e a diferenciação sazonal) podendo fazer-se a análise através do teste *Augmented Dickey-Fuller* (ADF). A diferenciação das séries deve ser usada apenas quando se verificar ser necessário estabilizar a média e a variância das séries, pois ao efetuar-se cada diferenciação está-se a perder informação sobre a série original.

Recorrendo aos gráficos da função de autocorrelação (FAC) e da função de autocorrelação parcial (FACP) é possível identificar os termos das componentes autorregressivas (AR) e das componentes de médias móveis (MA). Estas funções desempenham um papel fundamental na seleção dos graus do modelo visto que as duas funções possibilitam a identificação dos modelos a serem ajustados ao conjunto de dados. Nas Figura 4 e Figura 5, apresenta-se a descrição gráfica da FAC e FACP, respetivamente. A série temporal de caudal, representada nos exemplos, está normalizada temporalmente em intervalos de 1 hora entre cada medição.

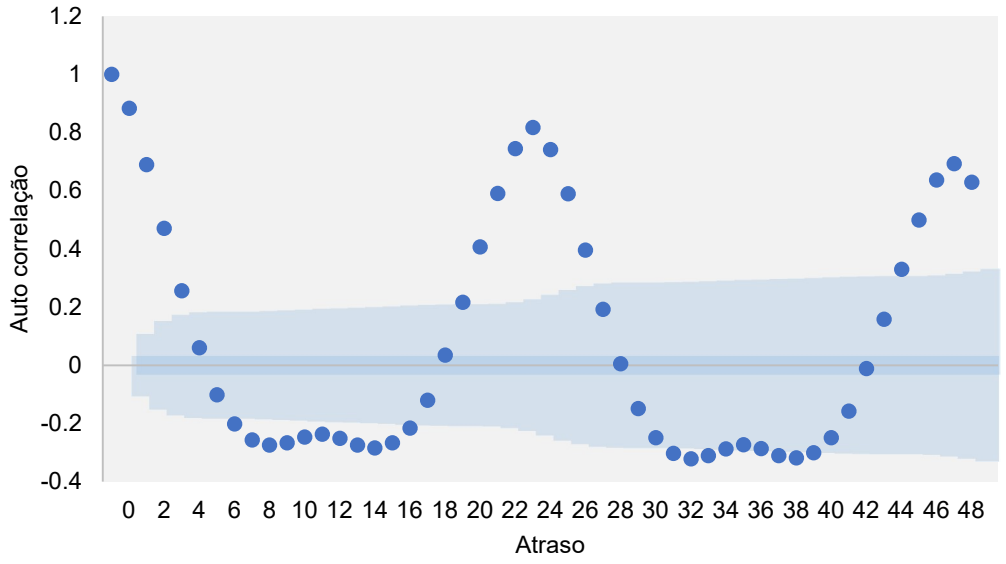

Figura 4 – Descrição gráfica da função de auto correlação

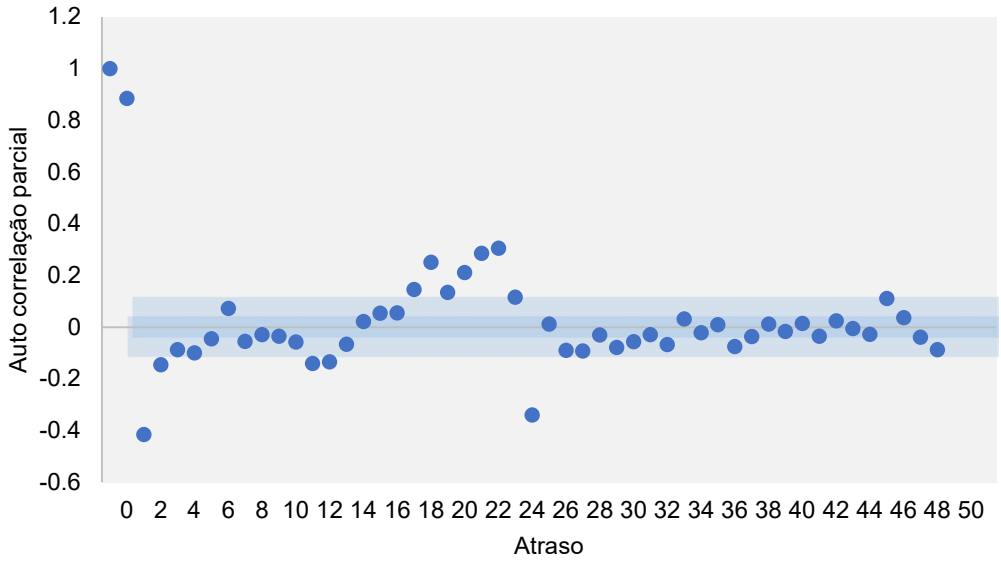

Figura 5 - Descrição gráfica da função de auto correlação parcial

Os gráficos de FAC demonstram a autocorrelação entre os diferentes valores. Quanto mais próximo estiver de zero, mais correlacionado aquele valor está com as medições anteriores. As sombras representadas nos gráficos das Figura 4 e Figura 5, demonstram os intervalos de confiança da autocorrelação: quando os pontos azuis estão dentro dos intervalos de confiança podemos concluir que as observações estão diretamente correlacionadas com os valores anteriores. No entanto, se uma observação (i.e., y_t) está correlacionado com a observação anterior (i.e., y_{t-1}), e y_{t-1} está correlacionado com y_{t-2}, então poderá existir uma correlação entre

y_t e y_{t-2}. Para analisar este tipo de correlação, recorre-se aos gráficos da FACP (Hyndman e Athanasopoulos, 2012).

Segundo Hyndman e Athanasopoulos (2012), podemos estimar os termos das componentes AR e MA do modelo ARIMA, como (p,d,0) se o gráfico FAC da série temporal diferenciada descrever um comportamento exponencial decrescente e existir um pico significativo no gráfico FACP mas mais nenhum para além desse, ou então, como (0,d,q) se o gráfico FACP da série temporal diferenciada descrever um comportamento exponencial decrescente e existir um pico significativo no gráfico FAC, mas mais nenhum para além desse. Os picos azuis iniciais fora dos intervalos de confiança, podem ser utilizados para determinar a ordem dos termos das componentes AR e MA.

O processo de seleção dos graus do modelo através dos gráficos FAC e FACP não consegue identificar um modelo único que melhor se ajuste à série temporal, mas sim um conjunto de modelos que se conseguem ajustar à série. De maneira a ser possível medir a qualidade de ajustamento entre os modelos selecionados, recorreu-se aos critérios de informação, baseados na comparação dos modelos contruídos com base na maximização do logaritmo da função de verosimilhança, penalizando os modelos com mais parâmetros. Os critérios de informação mais usais são o Critério de Informação de Akaike (AIC) proposto por Akaike (1974) e o Critério de Informação Bayesiano (BIC), desenvolvido por Schwarz (1978).

3.2. MODELO HÍBRIDO

3.2.1. ABORDAGEM QUEVEDO

Nesta secção será apresentada, essencialmente, uma abordagem desenvolvida por Quevedo *et al.*, (2010) para a reconstrução dos dados das séries temporais de SAA. Adicionalmente, é apresentada a implementação de uma melhoria dessa abordagem. O procedimento proposto para reconstruir os dados em falta consiste em dois níveis. O primeiro nível fornece a previsão do caudal diário agregado com base nos modelos ARIMA sazonais e o segundo determina um conjunto de padrões de distribuição em intervalos de 10 minutos, consistindo em 144 valores médios de caudal para cada padrão.

As séries temporais de caudal dos SAA consistem em vários valores de caudal intra-diários sendo assim possível construir uma série temporal de caudal diário agregado recorrendo ao volume diário. Quevedo *et al.*, (2010) analisou a sazonalidade das séries temporais de caudal diário agregado e concluiu que todas mostram grande evidência de sazonalidade semanal e de componentes periódicos determinísticos. Dessa forma, formulou a expressão para a previsão do caudal diário agregado com base nas três componentes principais a seguir apresentadas, onde $y(k)$ é o caudal diário agregado para o dia $k \in \mathbb{N}$:

1. Um polinómio trigonométrico com um período de uma semana para ter em conta o comportamento determinístico das séries temporais:

$$y(k) = 2\cos\left(\frac{2\pi}{7}\right) y(k-1) - y(k-2). \quad (2)$$

2. Um integrador para se conseguir considerar possíveis tendências:

$$y(k) = y(k-1). \quad (3)$$

3. Uma componente autorregressiva para considerar a influência de valores de caudal em períodos de uma semana:

$$y(k) = -a_1 y(k-1) - a_2 y(k-2) - a_3 y(k-3) - a_4 y(k-4). \quad (4)$$

Combinando as três componentes principais apresentadas, obtém-se a expressão geral para a previsão do caudal diário agregado, dado por $y_p(k)$ que é o caudal diário agregado previsto para o dia:

$$y_p(k) = -b_1 y(k-1) - b_2 y(k-2) - b_3 y(k-3) - b_4 y(k-4) - b_5 y(k-5) - b_6 y(k-6) - b_7 y(k-7). \quad (5)$$

Sendo que:

- $b_1 = a_1 - \left(2\cos\left(\frac{2\pi}{7}\right) + 1\right)$,
- $b_2 = a_2 - \left(2\cos\left(\frac{2\pi}{7}\right) + 1\right) a_1 + \left(2\cos\left(\frac{2\pi}{7}\right) + 1\right)$,
- $b_3 = a_3 - \left(2\cos\left(\frac{2\pi}{7}\right) + 1\right) a_2 + \left(2\cos\left(\frac{2\pi}{7}\right) + 1\right) a_1 - 1$,
- $b_4 = a_4 - \left(2\cos\left(\frac{2\pi}{7}\right) + 1\right) a_3 + \left(2\cos\left(\frac{2\pi}{7}\right) + 1\right) a_2 - a_1$,
- $b_5 = -\left(2\cos\left(\frac{2\pi}{7}\right) + 1\right) a_4 + \left(2\cos\left(\frac{2\pi}{7}\right) + 1\right) a_3 - a_2$,
- $b_6 = \left(2\cos\left(\frac{2\pi}{7}\right) + 1\right) a_4 - a_3$,
- $b_7 = -a_4$.

Os parâmetros do modelo (i.e., a_i) devem ser ajustados usando um método de estimativa de parâmetros, como o método dos mínimos quadrados, minimizando o erro quadrático médio – *RMSE* (resulta da terminologia anglo-saxónica *Root Mean Square Error*). A previsão do caudal diário agregado requer um registo histórico do volume diário sem dados em falta.

O segundo modelo determina um conjunto de padrões de distribuição do caudal em intervalos de 10 minutos, consistindo em 144 valores médios de caudal para cada padrão. Os padrões de distribuição consideram a variação nas medidas entre os dias úteis e os finais de semana. Por este motivo, os padrões devem ser determinados para os dias úteis (i.e., segunda a sexta-feira) e para os fins de semana (i.e., sábados e domingos). Os feriados também devem ser considerados pelo impacto que têm na análise. Quevedo *et al.,* (2010) determinaram que os hábitos de consumo nos feriados são semelhantes aos dos domingos. No entanto, ao considerar-se o modelo ARIMA sazonal como o modelo de previsão do caudal diário agregado, não é possível considerar o efeito de um feriado durante um dia da semana.

Para a determinação do número de padrões de distribuição de 10 minutos, Quevedo *et al.* (2010) recorreu a duas abordagens diferentes: uma usando um estudo de correlação entre diferentes agrupamentos de dias, como os dias úteis e dias de fins de semana e outra recorrendo a um classificador não supervisionado baseado em lógica difusa (*fuzzy logic*). Ambas as abordagens resultam em duas classes de padrões, ou seja, um padrão de distribuição para os dias úteis e um para os dias de fins de semana. No entanto, no âmbito do

presente documento considera-se um padrão para cada dia da semana, por forma a obter uma maior precisão na previsão.

O modelo de caudal de 10 minutos determina a previsão para o dia k a partir da distribuição do caudal diário agregado previsto ($y_p(k)$) pelo padrão associado ao dia que se pretende prever ($y_{pat}(k,i)$). A expressão geral para a previsão é dada por:

$$y_{p10}(k+i) = \frac{y_{pat}(k,i)}{\sum_{j=1}^{144} y_{pat}(k,j)} y_p(k), \quad i = 1, \dots, 144. \tag{6}$$

3.2.2. ABORDAGEM QUEVEDO MODIFICADA

A abordagem apresentada em Quevedo *et al. (*2010), nem sempre permite obter um desempenho razoável na previsão, podendo surgir problemas ao tentar utilizar o modelo ARIMA sazonal para prever o caudal diário agregado de um feriado, quando este ocorra durante um dia útil. Dessa forma, na presente secção é apresentada uma melhoria à abordagem original proposta por Quevedo *et al. (*2010), para estimar o caudal diário agregado de um feriado.

O objetivo é estimar o caudal diário agregado do dia a prever (i.e., feriado) com base apenas em domingos e feriados passados. Por esse motivo, como entrada do modelo é necessário um subconjunto de dados com datas de feriados. Ao inicializar o modelo, é verificado se a data a ser prevista está dentro do subconjunto com datas de feriados. Se não estiver, o modelo é executado de acordo com a abordagem original Quevedo *et al. (*2010) e estima o caudal diário agregado com o modelo ARIMA sazonal. Se estiver, a estimativa do caudal diário agregado começa com um modelo de suavização exponencial simples, para o qual a entrada é um subconjunto com o caudal diário agregado dos domingos passados (i.e., *s*), sendo o valor dos dois domingos passados representado por, s_{k-2}. A estimativa para o caudal diário agregado de um feriado é realizada como um novo domingo e pode ser definida pela seguinte expressão:

$$\hat{s}_k = \alpha s_{k-1} + \alpha(1-\alpha)s_{k-2} + \alpha(1-\alpha)^2 s_{k-2} + \dots \tag{7}$$

O parâmetro de suavização (i.e., α), controla a importância relativa das observações passadas em comparação com as observações mais recentes. Este parâmetro pode ser estimado usando o método de mínimos quadrados, minimizando o erro quadrático médio – *RMSE*.

3.3. MODELOS DE SUAVIZAÇÃO EXPONENCIAL

O modelo de suavização exponencial mais bem-sucedido em diversas áreas de aplicação foi inicialmente desenvolvido por Holt (1957) e considerava apenas séries temporais que apresentavam tendência. Winters (1960) expandiu o modelo para conseguir modelar séries temporais que apresentavam tendência e componentes sazonais, criando-se, assim, o método de Holt-Winters (Spyros *et al.*, 1997). No setor da água urbana, os métodos de

suavização exponencial são bem conhecidos e têm sido usados em modelos de previsão automática (Puig *et al.*, 2017). A principal característica é a sua simplicidade, visto que pode ser otimizado usando apenas um método de estimativa de parâmetros, como o método dos mínimos quadrados.

3.3.1. HOLT-WINTERS SIMPLES

O método de Holt-Winters é baseado em três componentes principais, nível, tendência e sazonalidade. O método pode ser dividido em duas versões com base em padrões de sazonalidade e tendência, ou seja, sazonalidade aditiva ou multiplicativa e tendência aditiva ou multiplicativa (Hyndman *et al.*, 2008). Dependendo do tipo de padrão sazonal e de tendência apresentados nos dados, uma das versões mencionadas pode ser escolhida (Galvas, 2016). Na sazonalidade aditiva, a diferença na flutuação sazonal entre medições sucessivas é constante, enquanto que, na sazonalidade multiplicativa a variação é uma percentagem (Galvas, 2016). Similarmente, na tendência aditiva a diferença na flutuação da tendência entre medições sucessivas também é constante e na tendência multiplicativa a variação é de igual forma uma percentagem. Neste documento, apenas é considerado o método Holt-Winters com sazonalidade multiplicativa e tendência aditiva, uma vez que a variação na flutuação sazonal entre medições sucessivas, em séries temporais de caudal, é uma percentagem, e a variação na flutuação da tendência é constante. Deste modo, o modelo pode ser apresentado pela seguinte expressão:

$$\hat{y}_t = (l_{t-1} + b_{t-1})s_{t-m} \tag{8}$$

As três componentes principais do método Holt-Winters são designadas por nível (l_t), tendência (b_t) e sazonalidade (s_t), onde *t* é o intervalo de tempo em análise e *m* a periodicidade do ciclo sazonal. As equações das três componentes principais do método de suavização exponencial com sazonalidade multiplicativa foram obtidas através da formulação de Taylor (2003):

$$l_t = \alpha\left(\frac{y_t}{s_{t-m}}\right) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{9}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \tag{10}$$

$$s_t = \gamma\left(\frac{y_t}{l_t}\right) + (1-\gamma)s_{t-m} \tag{11}$$

As componentes principais do modelo são baseadas em três parâmetros de suavização: α, β e γ. Estes parâmetros representam o parâmetro de nível, parâmetro de tendência e o parâmetro sazonal para a periodicidade do ciclo sazonal (diário no presente trabalho), respetivamente. Estes parâmetros podem ser estimados minimizando o erro quadrático médio – *RMSE* (resulta da terminologia anglo-saxónica *Root Mean Square Error*) – e geralmente ficam restritos entre 0 e 1.

A expressão para a previsão desde o instante do último valor observado, *T*, até *h* passos de tempo à frente, é dada por:

$$\hat{y}_{T+h} = (l_T + h \cdot b_T) s_{T+h-m}, \qquad h = 1,2,3, \ldots \tag{12}$$

Para inicializar o modelo de sazonalidade multiplicativa de Holt-Winters são necessários valores iniciais das componentes principais, ou seja, nível, tendência e um índice sazonal. De acordo com Spyros *et al.*, (1997) o nível inicial (l_0) é obtido pela média das observações do primeiro período sazonal. A tendência inicial (b_0) estimada usa uma média móvel do primeiro período sazonal e os índices sazonais (s_i) são estimados usando a média do primeiro período de sazonalidade:

$$l_0 = \frac{y_1 + y_2 + \cdots + y_m}{m}, \tag{13}$$

$$b_0 = \frac{\sum_{t=m+1}^{2m} y_t - \sum_{t=m}^{m} y_t}{m^2}, \tag{14}$$

$$s_i = \frac{y_i}{l_0}, \; i = 1,2, \ldots, m. \tag{15}$$

3.3.2. HOLT-WINTERS DE DUPLA SAZONALIDADE

O Holt-Winters de dupla sazonalidade acomoda dois períodos sazonais, permitindo modelar as séries temporais com variações de sazonalidade, como é o caso das séries temporais de caudal que evidenciem mais do que um padrão sazonal (i.e., diário e semanal). Desta forma, a formulação matemática do modelo Holt-Winters de dupla sazonalidade contém duas componentes de sazonalidade distintas, designadas por D_t e W_t, sendo que a primeira representa a sazonalidade diária e a segunda a sazonalidade semanal, com a periodicidade associada a cada ciclo sazonal, m_1 e m_2, respetivamente.

$$\hat{y}_t = (l_{t-1} + b_{t-1}) D_{t-m_1} W_{t-m_2} \tag{16}$$

As quatro componentes principais do método Holt-Winters de dupla sazonalidade são designadas por nível (l_t), tendência (b_t) e pelas duas componentes de sazonalidade já mencionadas. As equações das quatro componentes principais do método de suavização exponencial com dupla sazonalidade multiplicativa foram obtidas através da formulação de Taylor (2003):

$$l_t = \alpha\left(y_t - D_{t-m_1} - W_{t-m_2}\right) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{17}$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \quad (18)$$

$$D_t = \gamma\left(y_t - l_t - W_{t-m_2}\right) + (1-\gamma)D_{t-m_1} \quad (19)$$

$$W_t = \delta\left(y_t - l_t - D_{t-m_1}\right) + (1-\delta)W_{t-m_2} \quad (20)$$

Os componentes do modelo são baseados em quatro parâmetros de suavização α, β, ϒ e δ. Os primeiros três parâmetros são semelhantes aos apresentados anteriormente para o modelo Holt-Winters simples. Além disso, um parâmetro sazonal para o maior ciclo sazonal (semanal no nosso caso) é considerado. Da mesma forma, esses parâmetros podem ser estimados minimizando o RMSE e geralmente são restritos a estar entre 0 e 1.

A expressão para a previsão desde o instante, T, do último valor observado até h passos temporais à frente, é dada por:

$$\hat{y}_{T+h} = (l_T + h \cdot b_T)D_{T+h-m_1}W_{T+h-m_2}, h = 1,2,3,\ldots \quad (21)$$

O modelo duplo requer valores iniciais para tendência, nível, sazonalidade diária e índice de sazonalidade semanal. A formulação de Taylor (2003) foi usada para representar os valores iniciais, usando um ciclo de período m_1 para a sazonalidade diária e um ciclo de período m_2 para a sazonalidade semanal. Similar ao modelo Holt Winters simples, as quatro componentes iniciais são designadas por tendência inicial (b_0), nível inicial (l_0), índice inicial sazonal diário (D_0) e o índice inicial sazonal semanal (W_0), seguindo assim a seguinte formulação:

$$b_0 = \frac{1}{2m_2}\left[\left(\sum_{t=1}^{m_2} y_t - \sum_{t=m_2}^{2m_2} y_t\right) + \left(\sum_{t=1}^{m_2} y_t - y_{t-1}\right)\right] \quad (22)$$

$$l_0 = \frac{1}{2m_2}\sum_{t=1}^{2m_2} y_t - (m_2 + 0.5)b_0 \quad (23)$$

$$D_0 = \frac{1}{7}\sum_{k=1}^{7} \frac{y_{n+(k-1)m_1}}{\overline{d_k}} \quad (24)$$

$$W_0 = \frac{1}{2}\frac{1}{D_{(n\ mod\ m_1)}}\sum_{k=1}^{2} \frac{y_{n+(k-1)m_2}}{\overline{w_k}} \quad (25)$$

A Expressão (24) do índice inicial sazonal diário, requer uma componente definida como $\overline{d_k}$, podendo ser obtida através da média de caudal diário para cada *k* (i.e., dia). Similarmente, a Expressão (25) do índice inicial sazonal semanal, requer uma componente definida como $\overline{w_k}$, que pode ser obtida através da média de caudal semanal para cada *k* (i.e., semana).

3.4. MODELO DE APRENDIZAGEM AUTOMÁTICA

A Aprendizagem Automática, ou *Machine Learning* (ML) em língua inglesa, é um dos campos da Inteligência Artificial, que se tem focado no desenvolvimento de algoritmos que vão usar os dados para aprender com a sua experiência e ajustarem-se por forma a maximizar o seu desempenho (Han *et al.*, 2012). De um modo geral, os algoritmos de aprendizagem são classificados por: Aprendizagem Supervisionada, Aprendizagem Semi-Supervisionada, Aprendizagem Não Supervisionada e Aprendizagem por Reforço.

Neste subcapítulo, apresenta-se um algoritmo de aprendizagem supervisionada. Os algoritmos de aprendizagem supervisionada podem ser separados em dois tipos de problemas: classificação e regressão. Em problemas de classificação pretende-se prever uma variável categórica ou qualitativa enquanto nos problemas de regressão pretende prever-se uma varável numérica ou quantitativa (Hastie *et al.*, 2009).

O que distingue um problema de aprendizagem supervisionada de um problema de aprendizagem não supervisionada é que, em relação ao primeiro tipo, temos acesso à classificação ou ao valor numérico da variável de interesse a prever para um conjunto de dados que utilizamos para treinar e avaliar o nosso modelo. No que se segue, temos como pressuposto que uma parte da série temporal que nos propomos reconstruir se apresenta completamente validada, permitindo assim a utilização de uma técnica supervisionada.

O método Máquina de Vetores de Suporte, ou *Support Vector Machine* (SVM) em língua inglesa, é um modelo de aprendizagem automática supervisionada. O modelo SVM tem como propósito, definir um hiperplano ótimo que classifique os dados no espaço n-dimensional. Inicialmente, o SVM foi usado para problemas de classificação, tendo sido estendido para problemas de regressão e previsão, dando origem ao modelo de Regressão Vetorial de Suporte (SVR) (Hastie *et al.*, 2009).

O SVR, segue a mesma ideia base do SVM, ou seja, pretende definir um hiperplano que irá ajudar a prever o valor alvo. O problema de regressão do SVR requer um conjunto de dados de treino a partir do qual tenta aprender um correto mapeamento da função, *f*, do seguinte modo (Vapnik, 1995):

O conjunto de dados de treino consiste em *n* pares,

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \quad (26)$$

Sendo que, x_n representa o valor de entrada e y_n representa o valor real. O objetivo principal do SVR é encontrar uma função de perda que seja insensível a ε (i.e., *épsilon*), também conhecido como tamanho do tubo ou margens dentro do qual nenhuma penalidade está

associada, e apenas conte como erro as previsões que estão ε distantes dos dados medidos. Desse modo, a função *f* satisfaz:

$$-\varepsilon - \xi^{-} \leq y_i - f(x_i) \leq \varepsilon + \xi^{+} \quad (27)$$

Sendo, todo o *i = 1, …, n*, onde ξ^{-}, ξ^{+} são normalmente designadas por variáveis de folga que representam o desvio que o tamanho do tubo pode compreender. Na Figura 6 apresenta-se a descrição gráfica do modelo SVR que ilustra as suas diferentes variáveis.

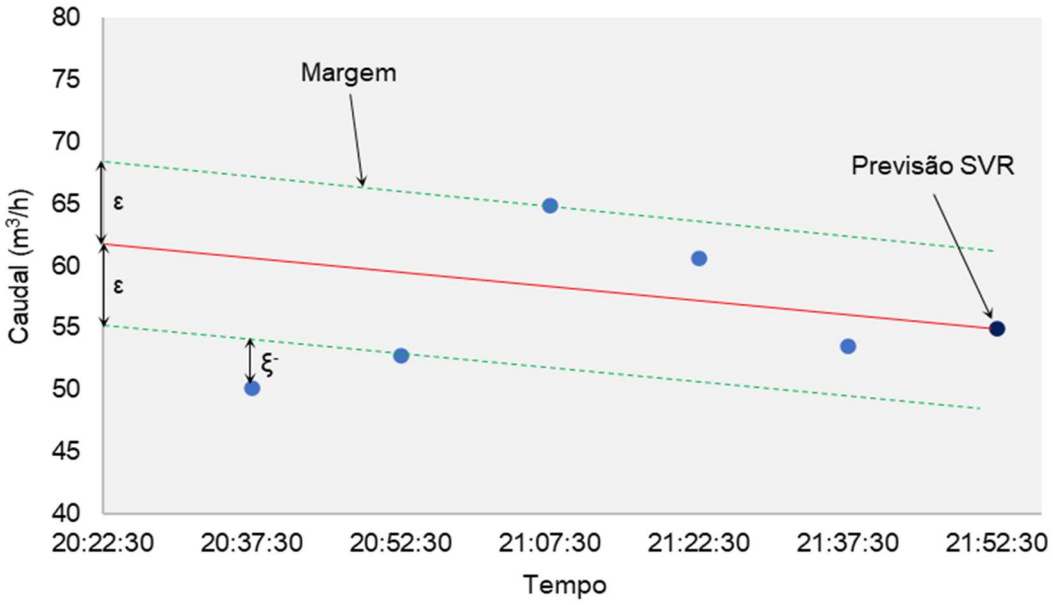

Figura 6 - Descrição gráfica do modelo SVR

A delimitação das margens é definida com o auxílio dos vetores de suporte. Na Figura 6, os pontos azuis são os vetores temporais das últimas observações, sendo que, os pontos próximos das margens são os nossos vetores de suporte e os pontos fora das margens estão a uma distância inferior às nossas variáveis de folga (i.e., ξ). A linha vermelha representada na Figura 6 é definida como o nosso hiperplano, que irá prever o nosso próximo valor de caudal.

O presente método também tem em consideração uma constante de regularização *C*, designada por hiperparâmetro de custo, sendo a sua ideia base, alterar o problema de modo a otimizar tanto o ajuste do hiperplano, como a penalização da quantidade de amostras dentro do tubo, ao mesmo tempo. O hiperparâmetro define a quantidade de amostras dentro do tubo que contribuem para o erro geral, desse modo, podemos ajustar o tamanho da margem para o conjunto de dados. Logo, quando *C* aumenta, a tolerância para os pontos fora do tubo também aumenta (Smola and Scholkopf, 2004). É importante que esta tolerância exista para reduzir a possibilidade do *overfitting*, muito típico em aprendizagem automática

supervisionada. O *overfitting* consiste no modelo ficar demasiado ajustado aos dados utilizados para o construir e ser mais falível perante um conjunto novo de dados.

3.5. AVALIAÇÃO DO DESEMPENHO

Para avaliar o desempenho das diferentes técnicas de reconstrução das séries temporais aplicadas pode-se utilizar uma métrica para medir a precisão e eficácia dos resultados obtidos. As métricas mais usais para avaliar o desempenho de previsões baseiam-se no cálculo do erro médio sendo elas o Erro Quadrático Médio, o Erro Absoluto Médio e a Raiz do Erro Quadrático Médio.

No presente documento os diferentes parâmetros dos modelos de previsão apresentados foram calibrados minimizando o RMSE (i.e., do inglês, root-mean-squared-error) dos modelos. O desempenho das previsões dos modelos foi avaliado usando o RMSE entre a medição real e prevista. Sendo que $e_t(h)$, representa o erro entre a medição real e a medição prevista e n o número de observações, como se apresenta na seguinte expressão:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2(h)} \tag{28}$$

3.6. IMPLEMENTAÇÃO DOS MODELOS DE RECONSTRUÇÃO DE SÉRIES TEMPORAIS DE CAUDAL

Normalmente, para facilitar o processamento de dados e automatização de tarefas repetitivas, os modelos de previsão de séries temporais têm sido implementados recorrendo a linguagem de programação. No entanto, e numa primeira fase, alguns dos modelos apresentados anteriormente (i.e., Modelo Híbrido e Modelos de suavização exponencial) foram implementados em folhas de cálculo que permitem, com maior facilidade, observar e compreender o funcionamento dos modelos, assim como os parâmetros e cada umas das suas componentes. Embora as folhas de cálculo possam não apresentar a eficiência desejada para análise de um elevado volume de dados, foram úteis na introdução dos modelos de reconstrução.

Numa segunda fase, a implementação dos modelos aplicados ao caso de estudo foi realizada recorrendo à linguagem de programação *Python*. Esta linguagem é conhecida por dispor de mecanismos de modulação robustos, sendo um dos mais relevantes, a existência de bibliotecas com diferentes módulos. As bibliotecas, organizadas em módulos, permitem dar suporte a mecanismos de cálculo, apresentando ainda capacidades de visualização gráfica.

Nas subsecções seguintes apresenta-se o processo de implementação e a estruturação dos modelos de reconstrução das séries temporais, as principais bibliotecas utilizadas e quais as suas funcionalidades.

3.6.1. ARIMA SAZONAL

Como exposto no Capítulo 3, o processo de seleção dos graus do modelo ARIMA sazonal realizado através da análise da função de auto correlação (FAC) e da função de auto correlação parcial (FACP) não consegue identificar um modelo único, mas sim um conjunto de modelos que melhor se ajustam à série temporal. Adicionalmente, a análise dos gráficos das funções FAC e FACP pode exigir alguma experiência por parte do utilizador. Neste sentido, considerou-se que a avaliação do ajustamento dos modelos seria realizada mediante os valores dos critérios de informação.

Antes da comparação dos valores dos critérios de informação dos modelos, importa definir quais os modelos a avaliar. Desta forma, e para cada grau do modelo, assumiu-se uma variação no seu valor, podendo ser definidos quais os valores mínimos e máximos para cada grau do modelo. Assim, e se forem definidos para 6 graus do modelo (i.e., p,d,q e P,D,Q) os valores de 0 a 2, obtém-se uma lista com 729 configurações de modelos possíveis.

Para a avaliação das configurações dos modelos, dividiu-se um conjunto de dados, para o qual o caudal é conhecido, em dados de treino e dados de teste, e recorreu-se a uma pesquisa em grelha (*Grid Search*), cujo fluxograma se apresenta na Figura 7. O *Grid Search* tem como entrada um modelo matemático e o conjunto de dados de treino, permitindo assim treinar o modelo para a primeira configuração da lista. Quando terminado o treino do modelo, o modelo é aplicado aos dados de teste e as previsões assim obtidas são comparadas com os valores reais, o que permite a avaliação do desempenho do modelo. Inicia-se posteriormente o processo de treino para a próxima configuração da lista. Todas as configurações da lista serão treinadas e avaliadas, sendo selecionada a configuração com melhor desempenho.

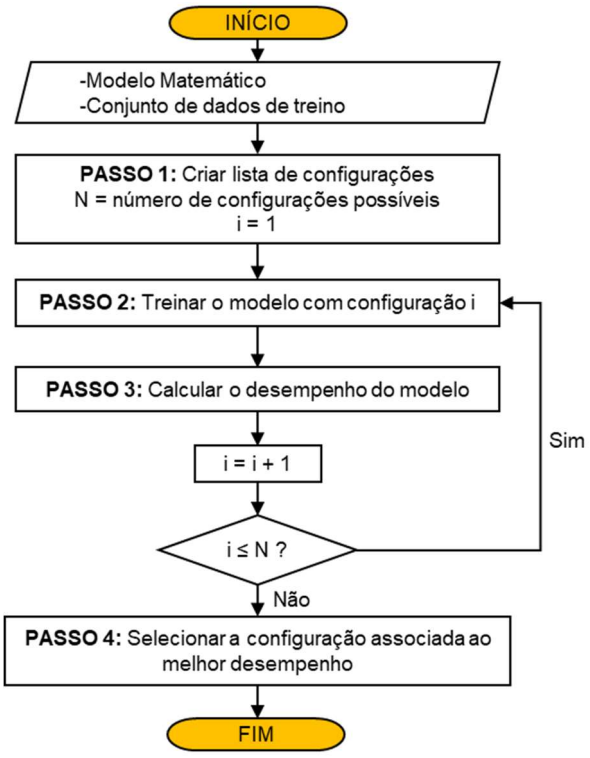

Figura 7 - Fluxograma do funcionamento do Grid Search

O cálculo do desempenho das diversas configurações do modelo ARIMA sazonal é efetuado com base nos critérios de informação, sendo criada uma lista com a configuração e o seu desempenho. Por fim, é selecionada a configuração com o melhor desempenho para ajustar a série temporal do caso de estudo e assim realizar a previsão.

O suporte para a implementação do modelo ARIMA sazonal foi a biblioteca *statsmodel*[1], amplamente utilizada e validada pela comunidade científica. Esta biblioteca fornece funções de diferentes modelos estatísticos, bem como ferramentas e funções para a realização de testes estatísticos e de análise de dados. Os módulos disponíveis na biblioteca permitem implementar modelos de regressão linear, modelos de análise de séries temporais, modelos para análise multivariada e outras funcionalidades que fazem com que seja amplamente usado.

A biblioteca *statsmodel* recorre ainda a duas bibliotecas que permitem a realização de cálculos numéricos, manipulação e análise de dados. A biblioteca *Numpy*[2] é usada principalmente para realizar cálculos em matrizes multidimensionais e a biblioteca *Pandas*[3] oferece estruturas e operações que permitem manipular tabelas numéricas e séries temporais. Para visualização gráfica recorreu-se à biblioteca *Matplotlib*[4], sendo as suas principais funcionalidades a criação de gráficos e visualização de dados de um modo geral.

3.6.2. ABORDAGEM QUEVEDO

Primeiramente, e aquando da importação da série temporal, é verificado se o dia a prever é um feriado. Se estivermos perante um dia de feriado, a previsão do caudal diário agregado é realizada recorrendo a um modelo de suavização exponencial simples, como referido no Capítulo 3. Se o dia a prever não for um feriado, a previsão é obtida através da formulação de Quevedo *et al.*, (2010) ou recorrendo à implementação dos modelos clássicos. A partir do conjunto de dados importados, são estabelecidos os padrões de distribuição para os dias úteis, sábados e domingos, sendo que o caudal diário agregado previsto é distribuído pelo padrão de distribuição consoante o dia da semana. Na Figura 8 apresenta-se um fluxograma onde se representa, passo a passo, a implementação da presente abordagem.

Similar ao modelo ARIMA sazonal, as principais bibliotecas utilizadas foram a *Pandas* e a *Numpy*. Nesta implementação, a biblioteca *statsmodel* também foi utilizada, nomeadamente para criar os modelos de ARIMA e suavização exponencial.

[1] Disponível em https://www.statsmodels.org/stable/index.html
[2] Disponível em https://numpy.org/
[3] Disponível em https://pandas.pydata.org/
[4] Disponível em https://matplotlib.org/

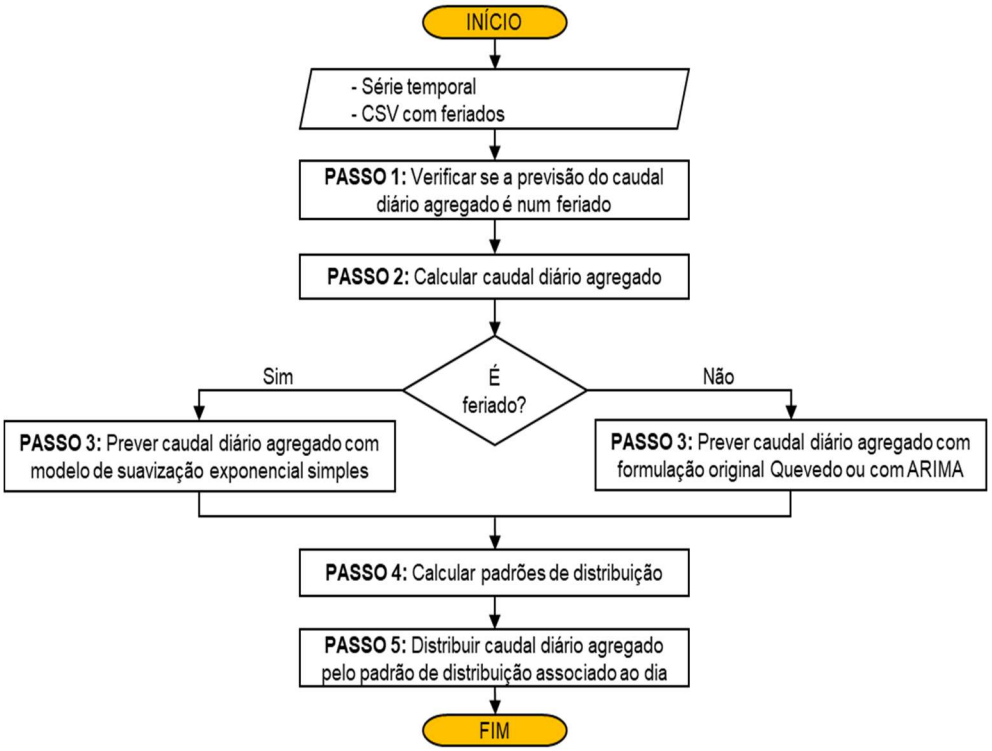

Figura 8 - Fluxograma da implementação da abordagem Quevedo

3.6.3. HOLT-WINTERS

A implementação do modelo de suavização exponencial Holt-Winters é bastante semelhante à do modelo ARIMA sazonal, nomeadamente no facto de ter sido adotado um processo de *grid search* para se encontrarem as configurações do modelo que melhor se ajustam à série temporal do caso de estudo. No entanto, existem outras técnicas que permitem selecionar as configurações que melhor se ajustam à série temporal. A avaliação do desempenho do modelo é realizada através do RMSE, sendo selecionada a configuração que apresentar o menor valor nesta métrica.

Com exposto no Capítulo 3, os modelos de suavização exponencial Holt-Winters, podem ser divididos em duas versões dependendo do tipo de sazonalidade e do tipo de tendência que a série temporal apresenta. Recorrendo à biblioteca *statsmodel*, os modelos de suavização exponencial requerem a definição de três parâmetros adicionais. Um dos parâmetros é chamado de *damped* e permite o amortecimento da tendência, que poderá ser acrescentada quando o modelo faz sobrestimação, algo típico dos modelos de suavização exponencial quando se pretende prever um intervalo de longa duração. O segundo parâmetro realiza transformações *Box-Cox* que normalmente são aplicadas para estabilizar a variância da série temporal. O último parâmetro é designado de *Remove Bias*, sendo utilizado quando existe uma sobre representação de uma dada classe de observações em relação a outra classe de observações. Por esse motivo, o *grid search* foi aplicado de maneira a obter-se a melhor configuração para cada um dos parâmetros, sendo que o tipo de sazonalidade e tendência

poderá ser aditiva, multiplicativa ou nenhuma das duas opções e os restantes parâmetros poderão ser aplicadas ou não, ou seja, apresentar o valor verdadeiro ou falso.

O modelo foi treinado para a configuração selecionada e executando um processo de otimização de modo a minimizar o RMSE ajustando os parâmetros de suavização do modelo (i.e., α, β e Υ). Posto isto, o modelo ficou assim pronto a realizar a previsão.

3.6.4. HOLT-WINTERS DUPLA SAZONALIDADE

O presente modelo de suavização exponencial de dupla sazonalidade não teve suporte de nenhuma biblioteca que permita a sua implementação. Ao contrário dos modelos anteriormente apresentados (i.e., ARIMA e Holt-Winters), que recorriam à biblioteca *statsmodel* para aplicar as funções dos modelos, o presente modelo não se encontra implementado naquela biblioteca. Dessa forma, a implementação baseou-se nas expressões anteriormente apresentadas no Capítulo 3.

Os modelos Holt-Winters de dupla sazonalidade são conhecidos por acomodarem duas componentes de sazonalidade. Por esse motivo, quando a série temporal é importada, os períodos sazonais da série são estimados e atribuídos valores arbitrários aos parâmetros de suavização (i.e., α, β, Υ e δ), sendo executado um processo de otimização que minimiza o RMSE ajustando os mesmos à série temporal do caso de estudo.

Com as componentes principais do modelo e os seus parâmetros definidos, procurou-se assim treinar o modelo e encontrar os parâmetros que melhor se ajustam à série temporal. Desse modo, recorreu-se à expressão (16) para treinar o modelo e à expressão (21) que faz a previsão desde o último valor observado até h passos temporais à frente.

O suporte para a presente implementação foram as bibliotecas já referidas anteriormente, *Pandas*, *Numpy* e *Mathplotlib*. Para além destas também foi utilizada a biblioteca *SciPy*[5] que fornece funções para minimizar funções objetivo, no presente caso para minimizar o RMSE e ajustar os parâmetros de suavização.

3.6.5. SVR

Na implementação do modelo SVR, seguiu-se a abordagem desenvolvida por Mounce *et al.*, (2011) para a previsão de caudal em SAA. Esta abordagem pode ser dividida em três etapas distintas: Numa primeira fase, são selecionados os parâmetros e os dados da série temporal, são formatados e convertidos em vetores temporais das últimas *D* observações anteriores; na segunda fase, com base nos parâmetros selecionados são construídos os regressores (os modelos para previsão de variável numérica, neste caso, o caudal) para cada período do dia e tipo de dia da semana (i.e., dias úteis, sábados, domingos), sendo de seguida treinado o modelo; por último, as previsões de caudal são realizadas de modo a testar o modelo e obter

[5] Disponível em https://scipy.org/

um desempenho razoável. Na Figura 9, apresenta-se o fluxograma que ilustra, passo a passo, a abordagem da presente implementação.

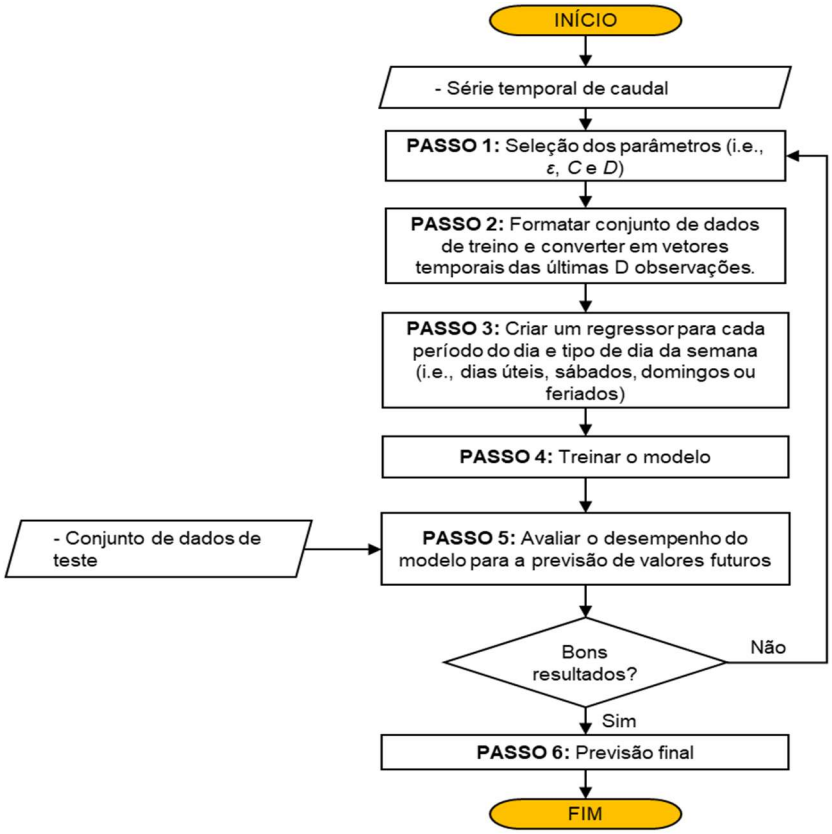

Figura 9 - Fluxograma da implementação do modelo SVR

Na importação da série temporal, os dados são divididos em duas partes, um conjunto de dados de treino e outro conjunto de dados de teste. Com o conjunto de dados de treino, pretende-se determinar os parâmetros ótimos. Com o conjunto de dados de teste e os parâmetros selecionados, o objetivo é avaliar o desempenho do modelo na previsão de caudal, sendo esta obtida com base nas *D* observações anteriores. Na presente implementação, o conjunto de dados de treino é de 4 semanas do conjunto de dados fornecidos pela EG, como sugere Mounce *et al.*, (2011), e o conjunto de dados de teste é definido como 1 dia.

O primeiro passo após a importação da série temporal é a seleção dos parâmetros do modelo (i.e., ε e *C*). Similar aos modelos anteriormente apresentados é possível recorrer a um processo de *grid search* para encontrar as configurações de parâmetros do modelo que melhor se ajustam à série temporal do caso de estudo. No entanto, se sempre que se iniciar o modelo for necessário recorrer a esta abordagem, o processo de seleção dos parâmetros tornar-se-ia muito moroso. Por esse motivo, a seleção dos parâmetros do modelo implementado foi realizada inicialmente com recurso a processos de *grid search* e, por último, teve como base a experimentação empírica, o conhecimento e as recomendações de Mounce *et al.*, (2011), para aplicação do modelo em séries temporais de SAA. Para seleção do ε, foi

definido, metade do desvio padrão do conjunto de dados de teste. O hiperparâmetro de custo foi estimado como *C=10*, e para as observações anteriores foi adotado *D=5*.

A formatação dos dados da série temporal é realizada com o objetivo de criar um conjunto de dados que indique quais as *D* observações anteriores para cada leitura de caudal. Na Tabela 1 apresenta-se um exemplo do conjunto de dados formatado e pronto a ser utilizado no modelo SVR.

Tabela 1 - Conjunto de dados formatado para aplicar no modelo SVR

Data	Hora	Caudal	Índice temporal	Dia da semana	n-1	n-2	n-3	n-4	n-5
02/11/2017	08:52:30	108,85	5	3	114,51	117,20	114,89	110,00	112,50
02/11/2017	09:07:30	103,18	6	3	108,85	114,51	117,20	114,89	110,00
02/11/2017	09:22:30	97,73	7	3	103,18	108,85	114,51	117,20	114,89
02/11/2017	09:37:30	102,57	8	3	97,73	103,18	108,85	114,51	117,20
02/11/2017	09:52:30	117,09	9	3	102,57	97,73	103,18	108,85	114,51
...	...	...	...	...	...	...	...	...	...
02/12/2017	06:22:30	90,22	91	5	91,39	84,06	80,90	81,58	79,94
02/12/2017	06:37:30	85,08	92	5	90,22	91,39	84,06	80,90	81,58
02/12/2017	06:52:30	82,83	93	5	85,08	90,22	91,39	84,06	80,90
02/12/2017	07:07:30	75,72	94	5	82,83	85,08	90,22	91,39	84,06
02/12/2017	07:22:30	63,91	95	5	75,72	82,83	85,08	90,22	91,39

O modelo implementado requer séries temporais validadas e normalizadas temporalmente. Na Tabela 1 apresenta-se uma série temporal com medições a cada 15 minutos, assim sendo, um dia de medições representa 96 leituras, que definimos de índice temporal. Na presente representação, o índice temporal não chega a 96 porque, para manter a coerência com a numeração de índices em Python, que iniciam em 0, optou-se por numerar como 0 a a primeira leitura do dia.

Tendo em conta que as séries temporais evidenciam ciclos de sazonalidade diária e semanal, optou-se por construir um modelo SVR (um regressor) para cada índice temporal, considerando o tipo de dia da semana (i.e., dias úteis, sábados e domingos ou feriados). Por exemplo, para o dia da semana 5 da Tabela 1 (i.e., sábado, pois o Python define como 0 a segunda-feira), o modelo irá construir e treinar 96 regressores, um por cada índice temporal, tendo como base as *D* observações anteriores. O modelo é treinado com o conjunto de dados de treino e é avaliado o desempenho da previsão de um dia completo com o conjunto de dados de teste.

O suporte da abordagem apresentada foram as bibliotecas: *Pandas*, *Numpy*, *Matplotlib* e *Scikit-learn*[6]. Esta última integra especificamente o módulo *sklearn.svm*, que inclui modelo *SVR*.

[6] Disponível em https://scikit-learn.org/stable/

4. CASOS DE ESTUDO

No presente capítulo, pretende-se aplicar os modelos de reconstrução propostos a séries temporais dos sistemas de abastecimento de água e comparar os resultados entre os mesmos. Dessa forma, foram usadas séries temporais de três casos de estudo reais portugueses, cujas características são representativas de uma grande percentagem das EG portuguesas.

A estrutura do presente capítulo inicia-se com a implementação dos modelos de reconstrução de séries temporais de caudal, de seguida apresenta-se a descrição dos diferentes casos de estudo e a análise exploratória dos seus dados. Por último, apresentam-se os resultados obtidos na previsão de caudal dos sistemas de abastecimento de água e a sua discussão.

4.1. CASO DE ESTUDO 1

4.1.1. DESCRIÇÃO DAS SÉRIES TEMPORAIS DO CASO DE ESTUDO 1

O caso de estudo (CE1) é uma zona de medição e controlo (ZMC) localizada na área metropolitana de Lisboa, mais precisamente em Penalva, na freguesia de Santo António da Charneca, do município do Barreiro. A ZMC, tem como objetivo a medição e controlo, do fornecimento de água a sensivelmente 3300 habitantes numa área que é bastante homogénea, composta maioritariamente por moradias unifamiliares e alguns edifícios residenciais até três pisos, bem como algumas lojas e uma escola primária. A EG usa no seu processo de aquisição de dados um medidor de caudal por impulso, na entrada da rede de distribuição de água, que faz o registo das medições a cada 2 m^3 de água, tornando a frequência de medição não igualmente espaçada.

Tendo em conta que os modelos de reconstrução requerem dados validados, tratados e normalizados temporalmente, os dados disponibilizados pela EG são referentes ao ano de 2018 e foram previamente validados e normalizados temporalmente em intervalos de 10 minutos e de 1 hora. Para tal, recorreu-se à ferramenta computacional apresentada anteriormente desenvolvida por Ferreira *et al.* (2022), para a validação de séries temporais de caudal.

Nos dados disponibilizados pela EG observou-se: medições não igualmente espaçadas, leituras duplicadas (i.e., para o mesmo período há mais do que uma medição), valores anormalmente altos ou baixos, períodos sem medição e patamares estáticos, que representaram cerca de 2,7% dos valores anómalos identificados. Deste processo de validação resultou uma série temporal com um espaçamento normalizado e com falhas de longa duração.

4.1.2. ANÁLISE EXPLORATÓRIA DA SÉRIE TEMPORAL CASO DE ESTUDO 1

Para uma melhor compreensão da variável em estudo são apresentadas na Tabela 2 algumas medidas descritivas. Estas medidas são representativas da série temporal de caudal do caso de estudo referente ao ano de 2018, validada e normalizada temporalmente para um período de 10 minutos. Das 52.560 medições contabilizadas, cerca de 3,8% são medições em falta, o que perfaz aproximadamente 14 dias com dados em falta no ano de 2018.

A média de caudal no ano de 2018 foi de 34,2 m^3/h, sendo que o desvio padrão tomou o valor de 18,5 m^3/h. O valor alto do desvio padrão pode ser explicado pela diferença de consumos ao longo do ano, caracterizado por um baixo consumo no primeiro e quarto trimestres, e um consumo elevado no terceiro trimestre. O valor máximo de caudal registado foi de 135,4 m^3/h enquanto o mínimo foi de 3,5 m^3/h.

Tabela 2 - Medidas descritivas da série temporal do CE1

Medida descritiva	**CE1 (2018)**												
	Jan.	**Fev.**	**Mar.**	**Abr.**	**Mai.**	**Jun.**	**Jul.**	**Ago.**	**Set.**	**Out.**	**Nov.**	**Dez.**	**Total**
Medições	4.464	4.032	4.464	4.320	4.464	4.320	4.464	4.464	4.320	4.464	4.320	4.464	52.560
Medições em falta	5	17	195	68	12	4	5	2	0	852	804	16	1 980
Média (m^3/h)	24,1	26,3	21,5	22,7	34,5	40,1	48,1	54,7	50,7	41,1	23,4	22,9	34,2
Desvio Padrão (m^3/h)	8,7	10,7	9,2	10,5	15,7	17,3	17,7	20,2	17,2	15,1	9,1	10,1	18,5
Mínimo (m^3/h)	6,6	6,1	5,3	5,1	6,7	8,3	13,2	16,2	18,0	14,3	6,5	3,5	3,5
P25 (m^3/h)	16,3	16,9	13,2	12,7	21,1	26,4	35,2	39,1	36,7	31,2	15,9	13,8	21,8
P50 (m^3/h)	25,3	26,9	22,9	24,0	34,9	40,1	48,3	52,9	49,7	39,4	24,5	23,7	30,6
P75 (m^3/h)	30,0	33,4	28,0	30,0	45,8	51,2	59,6	67,5	62,5	49,8	29,8	29,7	44,8
Máximo (m^3/h)	54,3	74,7	78,5	77,7	90,8	119,1	112,1	135,4	118,8	110,4	73,4	76,2	135,4

Como supracitado, as séries temporais dos sistemas de abastecimento de água evidenciam ciclos diários e semanais. Dessa forma, apresenta-se na Figura 10 o valor de caudal médio horário para os diferentes dias da semana (i.e., dias úteis, sábados e domingos) da série temporal do CE1.

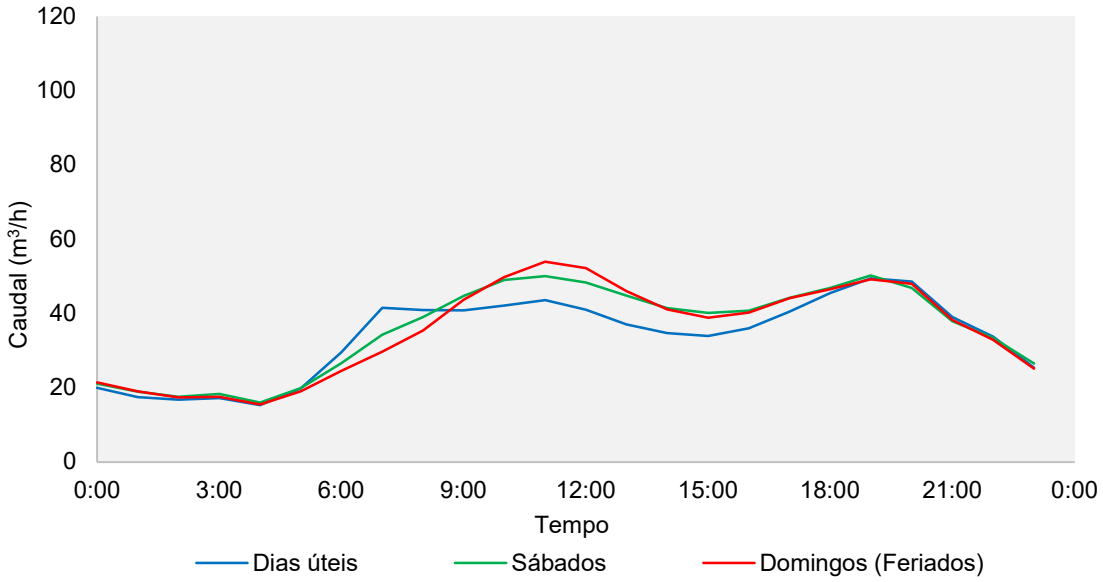

Figura 10 – Caudal horário médio para os diferentes dias da semana da série temporal do CE1

A Figura 10 demonstra claramente que existe um padrão de distribuição específico para os diferentes dias da semana, com principal distinção entre os dias úteis e os dias de fim de semana (i.e., sábados, domingos e feriados). O padrão de distribuição entre sábados, domingos e feriados não sofre alterações significativas apenas um ligeiro aumento do consumo no período da manhã aos domingos. Consegue-se observar no gráfico de linhas que no período noturno (caracterizados pelo baixo consumo) os padrões de distribuição dos diferentes dias da semana são idênticos, enquanto no período da manhã e da tarde (caracterizados pelos altos consumos) surgem alterações significativas.

Sendo o objetivo explorar a sazonalidade diária e semanal da série temporal do CE1, apresenta-se na Figura 11 o valor de caudal médio diário para os dias úteis, sábados e domingos em diferentes meses do ano de 2018. A representação gráfica da totalidade dos meses não será exibida devido ao elevado número de medições que dificulta a leitura da informação. Sendo assim, recorreu-se aos meses de abril, agosto e outubro para exemplificar as sazonalidades da série temporal do caso de estudo, com a segunda-feira como dia inicial e o domingo como o dia final.

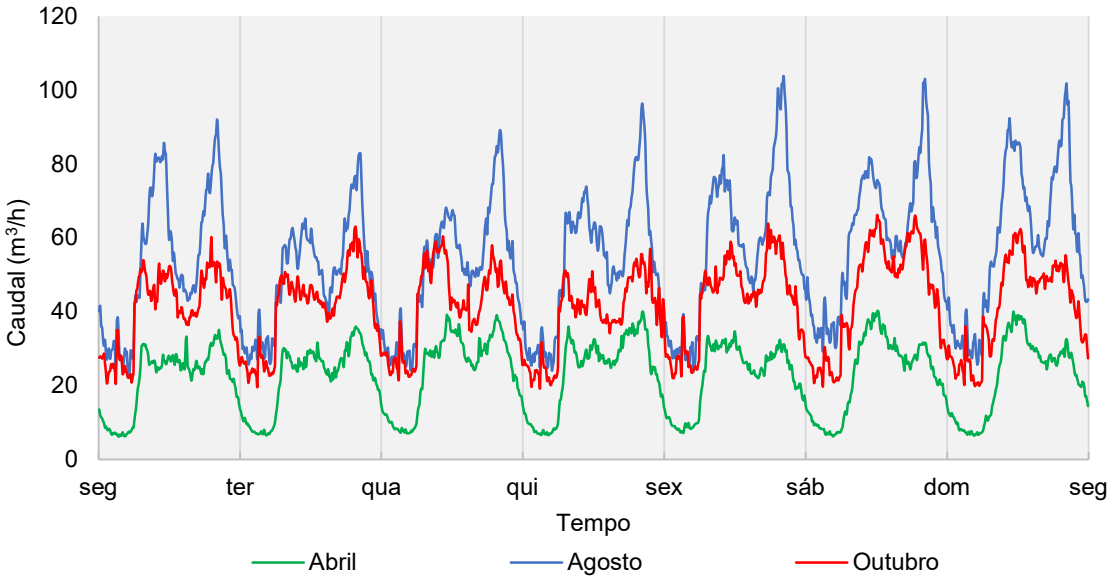

Figura 11 - Caudal diário médio da série temporal do CE1 em m^3/h

Os ciclos que se observam na Figura 11, representam os diversos dias da semana. Como esperado, existe um padrão de distribuição muito idêntico nos dias úteis, com maior enfoque nos meses de abril e outubro. Este padrão altera-se nos fins-de-semanas, onde também existe um padrão de consumo especifico para estes dias. No entanto, no mês de agosto o padrão de consumo entre os diferentes dias úteis não sofre alterações significativas, existindo apenas um aumento no valor médio de caudal no sábado e domingo.

Na Figura 11, a discrepância entre os valores médios de caudal diário é muito expressiva ao longo do ano. Para uma melhor compreensão deste fenómeno, considerou-se que o semestre de inverno ia de novembro até abril, e o semestre de verão ia de maio a outubro, com o objetivo de representar o valor médio caudal mensal do ano de 2018 da série temporal do CE1. Assim sendo, apresenta-se nas Figura 12 e Figura 13 o valor de caudal médio mensal para os meses do semestre de inverno e para os meses do semestre de verão, respetivamente. Em ambas as figuras, também é representado a média diária anual do caudal.

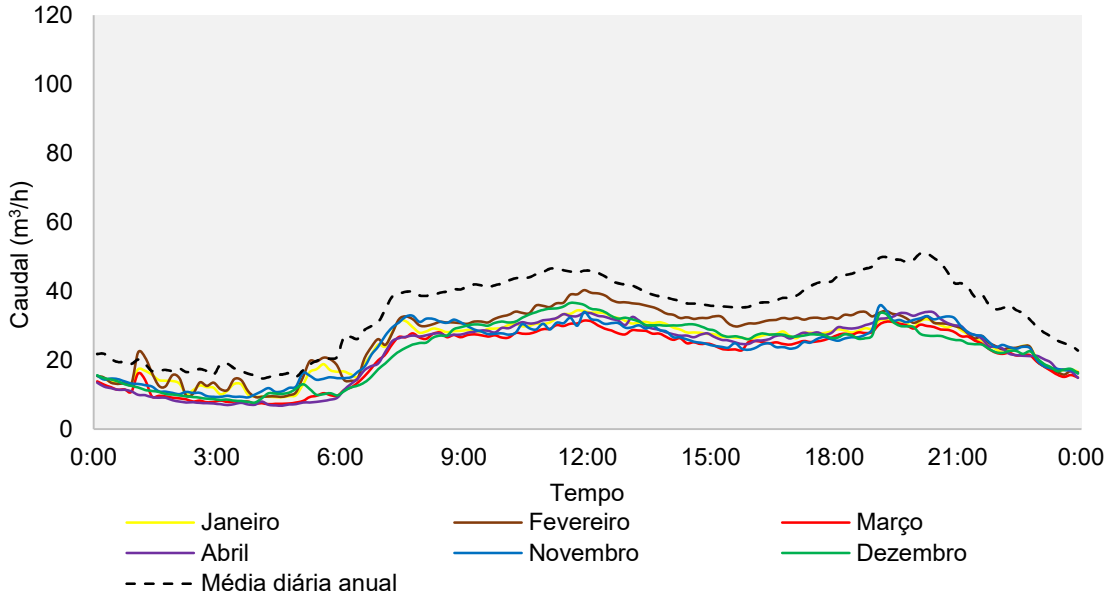

Figura 12 - Caudal mensal médio da série temporal do CE1 em m³/h para os meses do semestre de inverno

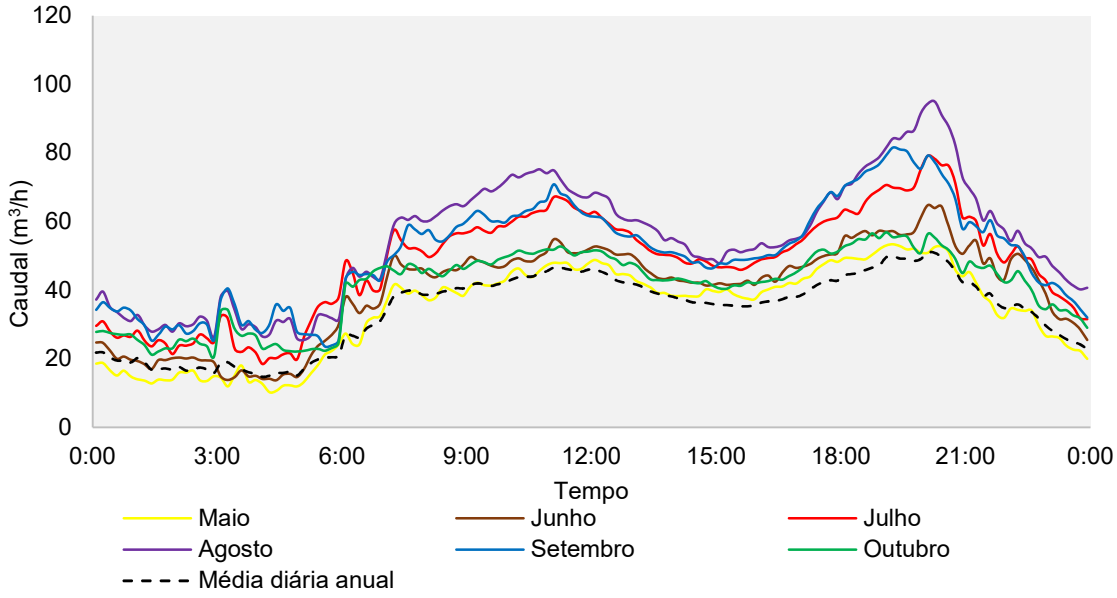

Figura 13 - Caudal mensal médio da série temporal do CE1 em m³/h para os meses do semestre de verão

A partir das Figura 12 e Figura 13 é possível observar o aumento do consumo nos meses de maior calor em Portugal, sendo o mês de agosto aquele que apresenta um valor médio mensal mais elevado e maior variabilidade entre medições (caracterizado pela maior diferença entre o caudal máximo e o caudal mínimo). Esta variabilidade pode ser observada recorrendo a diagramas de extremos e quartis. Por esse motivo, apresenta-se na Figura 14, os diagramas de extremos e quartis para cada mês da série temporal do CE1.

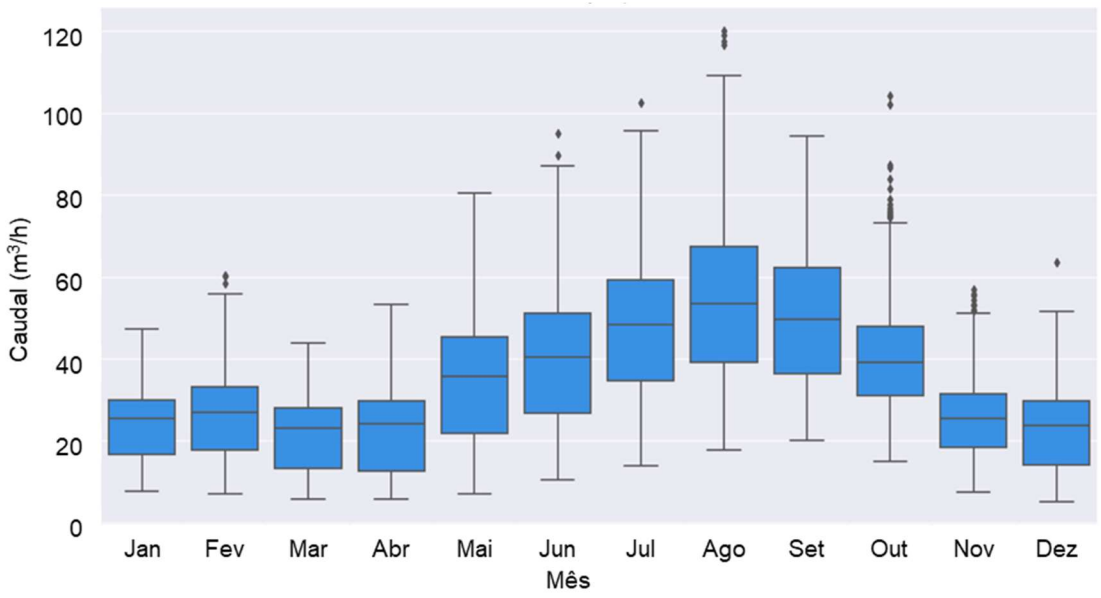

Figura 14 - Diagramas de extremos e quartis para cada mês da série temporal do CE1

Os diagramas de extremos e quartis para cada mês do CE1 são apresentados na Figura 14 evidenciam uma alta dispersão dos dados (i.e., variabilidade entre medições) e uma distribuição simétrica dos dados nos meses de verão. Nos meses de inverno, os diagramas apresentam uma baixa amplitude com uma distribuição assimétrica dos dados.

4.1.3. RESULTADOS E DISCUSSÃO

Na presente subsecção apresenta-se os resultados de três testes comparativos entre os modelos de reconstrução anteriormente apresentados. No primeiro teste considerou-se a série temporal do caso de estudo com espaçamentos entre medições de 1 hora, tendo-se como objetivo avaliar o desempenho dos modelos e o seu tempo de computação na previsão de um dia útil completo. Similarmente, no segundo teste o objetivo também passou por avaliar o desempenho dos modelos e o seu tempo de computação na previsão de um dia útil completo. No entanto, considerou-se a série temporal do caso de estudo com espaçamentos de 10 minutos. No terceiro teste, considerou-se a série temporal com intervalos de 10 minutos para prever um feriado e avaliar o seu desempenho, bem como o seu tempo de computação.

No Teste 1 (avaliação do desempenho e tempo de computação dos modelos de reconstrução na previsão de um dia da semana), foi considerado como histórico o mês de setembro completo, mais os três primeiros dias do mês de outubro, da série temporal do CE1, com intervalos de 1 hora. Assim sendo, os modelos iram prever 24 períodos de 1 hora, completando assim um dia. Na Figura 15, apresentam-se os resultados obtidos para cada modelo bem como as medições reais de caudal.

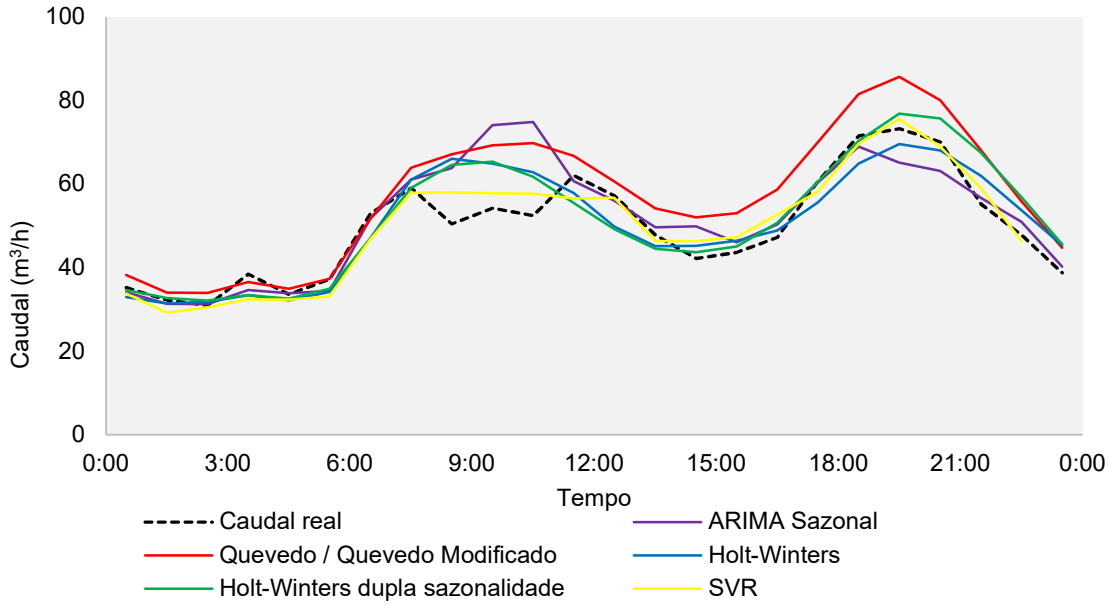

Figura 15 - Comparação dos cinco modelos de reconstrução considerando um dia da semana e intervalos de 1 hora para CE1

Na previsão do dia da semana da série temporal do CE1 com intervalos de 1 hora, os modelos apresentam uma diferença significativa no seu desempenho. Os modelos de suavização exponencial, Holt Winters e Holt Winters de dupla sazonalidade, conseguem o melhor desempenho (RMSE=5,98 e RMSE=6,26, respetivamente), seguidos do modelo de aprendizagem automática SVR (RMSE=6,36), ARIMA sazonal (RMSE=7,00) e, por último, o modelo desenvolvido por Quevedo e o Quevendo modificado (ambos com RMSE=9,00).

Todos os modelos são relativamente rápidos na previsão de um dia completo, para a série temporal do CE1, com intervalos de 1 hora, demorando menos de 30 segundos até obter uma previsão. Na Tabela 3, apresenta-se um resumo do tempo de computação (em segundos) e o valor de RMSE para todos os modelos.

Tabela 3 - Comparação do desempenho dos modelos e o seu tempo de computação considerando um dia da semana e intervalos de 1 hora para CE1

Modelos	**Dia da semana (1h)**	
	RMSE (m^3/h)	Tempo de computação (s)
ARIMA sazonal	7,42	14
Quevedo	9,00	1
Holt-Winters	5,98	30
Holt-Winters dupla sazonalidade	6,26	12
Quevedo modificado	9,00	1
SVR	6,36	1

As séries temporais com intervalos de 1 hora têm um uso muito limitado na operação em tempo real dos sistemas de abastecimento de água. Como referido anteriormente, a aplicação de técnicas avançadas (i.e., modelos aprendizagem automática) requerem séries temporais com intervalos de curta duração para operar sistemas de abastecimento de água. Por esse motivo, no Teste 2 o mesmo dia da semana irá ser previsto com intervalos de 10 minutos.

No Teste 2 (avaliação do desempenho e tempo de computação dos modelos de reconstrução na previsão de um dia da semana) também, foi considerado como histórico o mês de setembro completo, mais os três primeiros dias do mês de outubro, da série temporal do CE1, mas desta vez normalizada em intervalos de 10 minutos. Assim sendo, os modelos iram prever 144 períodos de 10 minutos, completando assim um dia. Na Figura 16, apresentam-se os resultados obtidos para cada modelo bem como as medições reais de caudal.

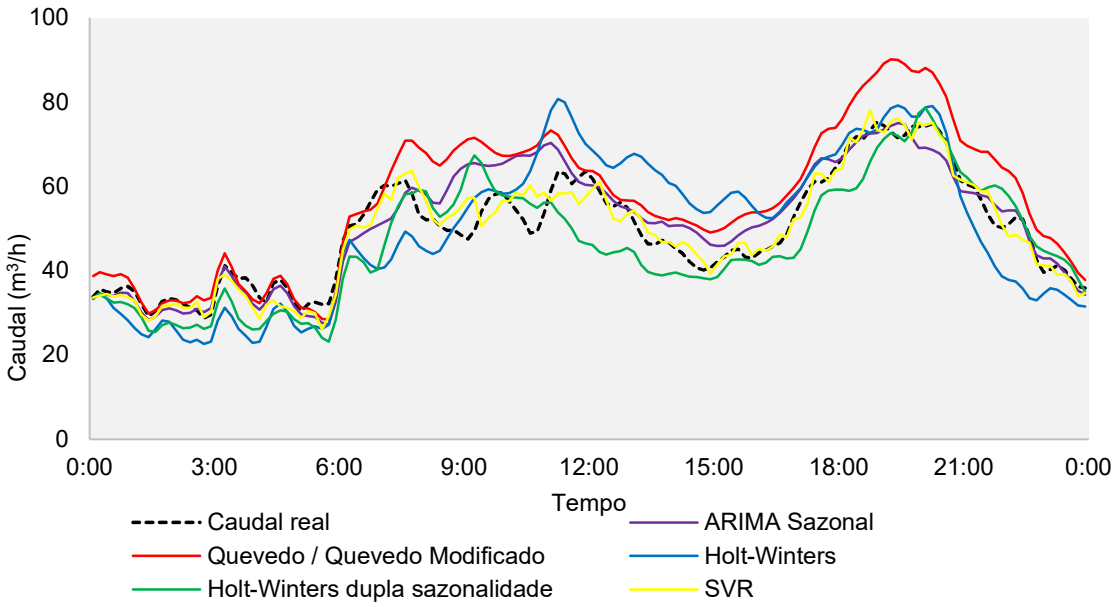

Figura 16 - Comparação dos cinco modelos de reconstrução considerando um dia da semana e intervalos de 10 minutos da série temporal do CE1

Na previsão de um dia completo com intervalos de 10 minutos, o SVR demonstra o melhor desempenho e apresenta o valor mais baixo (RMSE=3,27), seguido do modelo do ARIMA sazonal (RMSE=5,67), Holt-Winters de dupla sazonalidade (RMSE=7,13), abordagem preconizada por Quevedo e Quevedo modificado (ambos com RMSE=9,43) e por último o Holt-Winters (RMSE=9,62).

Contudo a diferença no tempo de computação entre os modelos é bastante significativa, sendo a abordagem preconizada por Quevedo e Quevedo modificado as que apresentam resultados mais rápidos (t=1s), seguido do modelo SVR (t=3s) e dos modelos de suavização exponencial, Holt-Winters e Holt-Winters de dupla sazonalidade (t=71s e t=148s), respetivamente. Em último, o modelo ARIMA sazonal (t=1090s). Na Tabela 4, apresenta-se um resumo do tempo de computação (em segundos) e o valor de RMSE para todos os modelos.

Tabela 4 - Comparação do desempenho dos modelos e o seu tempo de computação considerando um dia da semana e intervalos de 10 minutos da série temporal do CE1

Modelos	Dia da semana (10 minutos)	
	RMSE (m^3/h)	Tempo de computação (s)
ARIMA sazonal	5,76	1090
Quevedo	9,43	1
Holt-Winters	9,62	71
Holt-Winters dupla sazonalidade	7,13	148
Quevedo modificado	9,43	1
SVR	3,27	3

Normalmente, os problemas surgem na previsão de feriados quando estes ocorrem durante os dias úteis, uma vez que o padrão de distribuição de água de um feriado está geralmente relacionado com o padrão de distribuição dos domingos (em oposição aos padrões de distribuição dos dias úteis).

No Teste 3 (avaliação do desempenho e tempo de computação dos modelos de reconstrução na previsão de um feriado), foi, igualmente, considerado como histórico o mês de setembro completo, mais os quatro primeiros dias do mês de outubro, da série temporal do CE1, com intervalos de 10 minutos. Assim, sendo, os modelos iram prever 144 períodos de 10 minutos, completando assim um dia completo. A diferença dos testes anteriores prende-se com o facto de o último dia da série temporal do CE1 ser um feriado (i.e., 5 de outubro). Na Figura 17 apresentam-se os resultados obtidos para cada modelo bem como as medições reais de caudal, para um dia feriado.

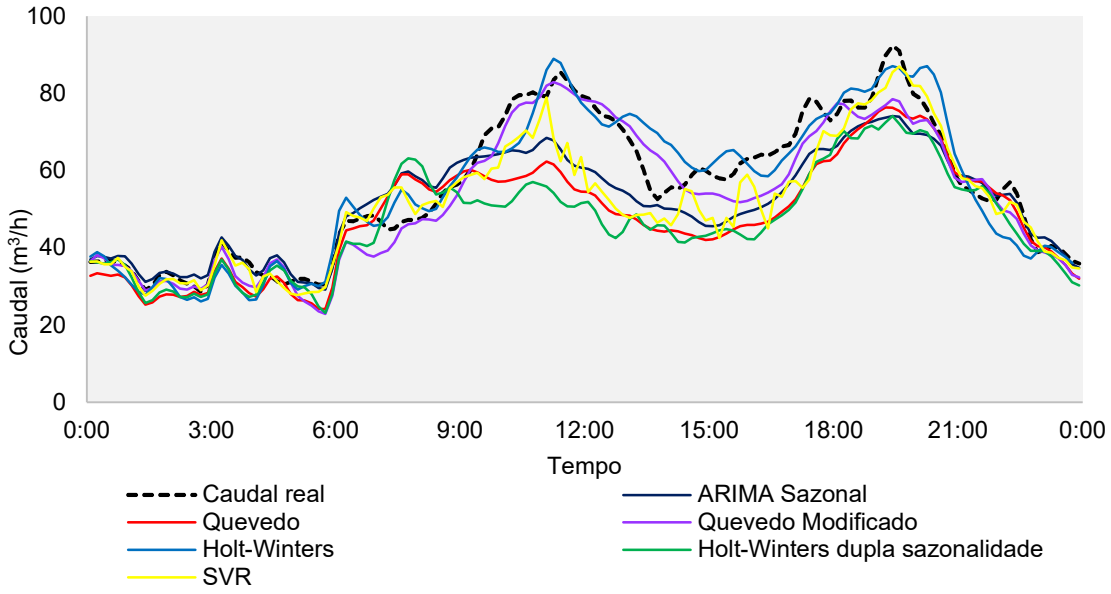

Figura 17 - Comparação dos cinco modelos de reconstrução considerando um feriado e intervalos de 10 minutos da série temporal do CE1

Na previsão do feriado que se apresenta na Figura 17 é possível concluir que os modelos não conseguiram prever a variação do feriado na sua totalidade. Os modelos autorregressivos (i.e., ARIMA sazonal) e os modelos de suavização exponencial (i.e., Holt-Winters e Holt-Winters de dupla sazonalidade) realizam previsões baseados nos períodos sazonais e não permitem "sair" fora destes períodos. Similarmente, a abordagem original de Quevedo realiza as previsões do caudal diário agregado utilizando um modelo autorregressivo e o modelo SVR, efetua as suas previsões com base nas regressões com as observações anteriores. O Quevedo Modificado, consegue prever o pico de consumo matinal associado aos dias de feriados, no entanto, não acerta a totalidade da previsão, falhando a previsão no período da tarde. Na Tabela 5 apresenta-se um resumo do tempo de computação (em segundos) e o valor de RMSE para todos os modelos.

Tabela 5 - Comparação do desempenho dos modelos e o seu tempo de computação considerando um feriado e intervalos de 10 minutos da série temporal do CE1

Modelos	**Feriado (10 minutos)**	
	RMSE (m^3/h)	Tempo de computação (s)
ARIMA sazonal	9,00	598
Quevedo	11,77	1
Holt-Winters	5,76	75
Holt-Winters dupla sazonalidade	13,78	55
Quevedo modificado	5,33	1
SVR	8,63	3

4.2. CASO DE ESTUDO 2

4.2.1. DESCRIÇÃO DAS SÉRIES TEMPORAIS DO CASO DE ESTUDO 2

O caso de estudo 2 (CE2) é uma ZMC localizada na sub-região do Baixo Alentejo, mais precisamente em São Brissos, na freguesia de Trigaches, do município de Beja. A ZMC, tem como objetivo a medição e controlo, do fornecimento de água de sensivelmente 110 habitantes numa área homogénea, composta maioritariamente por moradias unifamiliares, algum comércio e um aeroporto inativo. A EG usa no seu processo de aquisição de dados um medidor de caudal de impulso, na entrada da rede de distribuição de água, que faz o registo das medições em intervalos com cerca de 5 minutos, no entanto, os intervalos nem sempre são exatos, tornanduo o intervalo entre medições não igualmente espaçado.

Tendo em conta que os modelos de reconstrução requerem dados validados, tratados e normalizados temporalmente, os dados disponibilizados pela EG são referentes ao ano de 2018 e foram previamente validados e normalizados temporalmente em intervalos de 10 minutos e 1 hora. Para tal, recorreu-se à ferramenta computacional apresentada anteriormente desenvolvida por Ferreira *et al.* (2022), para a validação de séries temporais de caudal.

Nos dados disponibilizados pela EG observou-se: medições não igualmente espaçadas, valores negativos, valores anormalmente altos ou baixos, períodos sem medição e patamares estáticos, que na totalidade representou cerca de 16% dos valores anómalos identificados. Deste processo de validação resulta a série temporal com um espaçamento normalizado e com falhas de longa duração.

4.2.2. ANÁLISE EXPLORATÓRIA DA SÉRIE TEMPORAL DO CASO DE ESTUDO 2

Similarmente ao CE1, são apresentadas na Tabela 6 algumas medidas descritivas, representativas da série temporal de caudal do CE2, referente ao ano de 2018, validada e normalizada temporalmente para um período de 10 minutos. Das 52.560 medições contabilizadas, cerca de 28,5% são medições em falta, ou seja, existem 14.967 intervalos de 10 minutos sem medição, o que perfaz aproximadamente, 105 dias com dados em falta no ano de 2018. A média de caudal no ano de 2018 foi de 23,6 m^3/h, sendo que o desvio padrão tomou o valor de 9,7 m^3/h. O valor máximo de caudal registado foi de 82,0 m^3/h enquanto o mínimo foi de 1,7 m^3/h.

Como as séries temporais dos sistemas de abastecimento de água evidenciam sazonalidades diárias e semanais, apresenta-se na Figura 18 o valor médio de caudal horário para os diferentes dias da semana (i.e., dias úteis, sábados e domingos) da série temporal do CE2.

Tabela 6 - Medidas descritivas da série temporal do caso de estudo do CE2.

Medida descritiva	**CE2 (2018)**												
	Jan.	**Fev.**	**Mar.**	**Abr.**	**Mai.**	**Jun.**	**Jul.**	**Ago.**	**Set.**	**Out.**	**Nov.**	**Dez.**	**Total**
Medições	4.464	4.032	4.464	4.320	4.464	4.320	4.464	4,464	4.320	4.464	4.320	4.464	52.560
Medições em falta	969	1.155	760	857	1.545	1.778	1.676	1.694	1.664	1.173	822	874	14.967
Média (m^3/h)	21,2	20,5	19,7	20,2	22,3	25,0	27,8	29,8	29,6	24,4	21,3	21,5	23,6
Desvio Padrão (m^3/h)	10,0	8,8	8,8	9,2	9,3	9,8	9,7	7,9	7,7	8,9	9,1	9,1	9,7
Mínimo (m^3/h)	5,2	1,7	4,9	4,9	5,4	5,4	5,5	5,4	10,7	5,7	5,8	5,6	1,7
P25 (m^3/h)	11,3	11,7	10,5	11,0	14,4	17,9	21,6	24,3	24,0	17,3	12,3	12,5	15,6
P50 (m^3/h)	22,3	21,5	21,3	21,7	23,3	25,5	27,3	28,8	28,9	25,2	22,4	22,4	24,3
P75 (m^3/h)	28,2	27,4	26,6	27,3	29,2	31,4	34,0	35,1	34,6	30,7	28,5	28,8	29,9
Máximo (m^3/h)	82,0	43,8	42,0	45,9	48,1	58,0	59,4	58,0	54,6	55,6	56,1	45,1	82,0

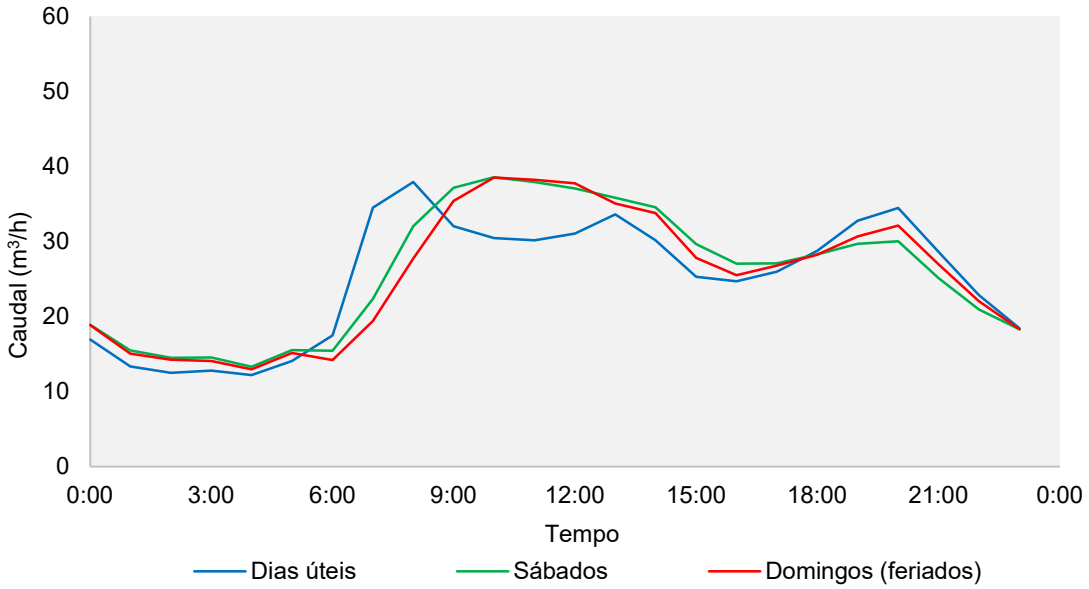

Figura 18 - Caudal horário médio para os diferentes dias da semana da série temporal do CE2

Como esperado, através da Figura 18 podemos concluir que a série temporal do CE2 tem uma alta sazonalidade diária e semanal. Existe um padrão de distribuição específico para os dias úteis e um padrão para os dias de fim de semana e feriados. Similarmente ao CE1, observa-se um baixo consumo no período noturno e um alto consumo no período da manhã e da tarde. No entanto, no CE1 existe um aumento no valor de caudal médio horário nos fins de semanas e feriados, enquanto no CE2 esse aumento não se verifica. A justificação prende-se com o facto de, no CE1, a rede de abastecimento de água estar inserida no meio urbano, em que a população tem por hábito sair de manhã dos seus lares para trabalhar e regressar apenas ao final do dia. Em oposição, nos fins de semana a população passa mais tempo nos seus lares. No CE2, a rede de abastecimento de água está localizada no interior de Portugal, num meio rural, onde existe uma alta taxa de população envelhecida e os hábitos da população diferem do meio urbano, não existindo tanta flutuação da população, dai não se verificar o aumento do valor de caudal horário médio nos fins de semana e feriados.

Tendo como objetivo analisar a sazonalidade da série temporal do CE2, apresenta-se agora na Figura 19 o valor médio de caudal diário para os dias úteis, sábados e domingos em diferentes meses do ano de 2018. A representação gráfica da totalidade dos meses não será exibida devido ao elevado número de medições que dificulta a leitura da informação. Sendo assim, recorreu-se aos meses de abril, agosto e outubro para exemplificar as sazonalidades da série temporal do CE2, com a segunda feira como dia inicial e o domingo como o dia final.

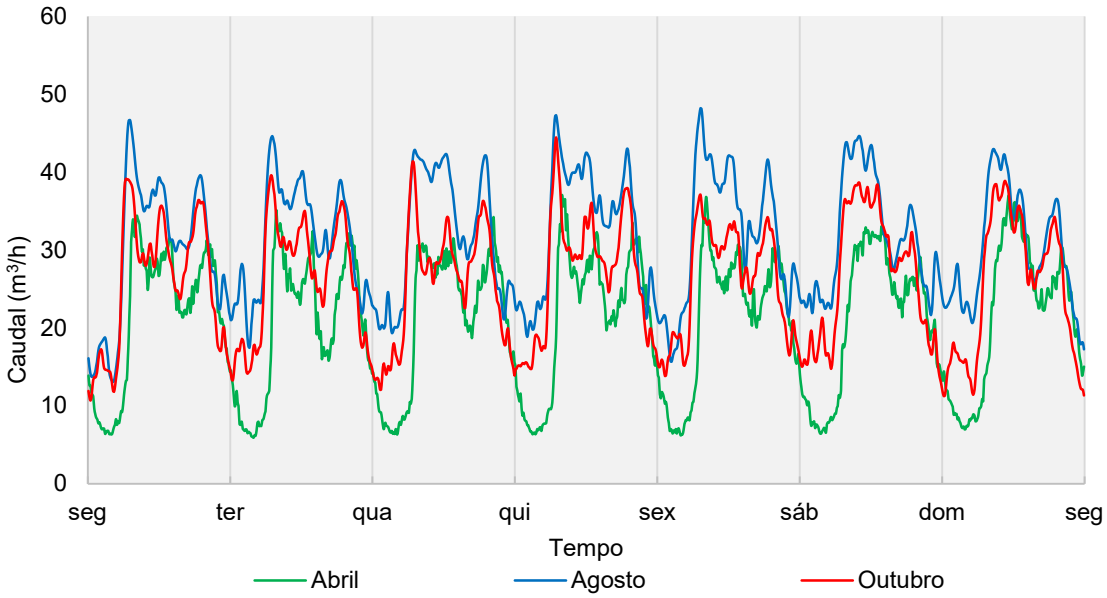

Figura 19 - Caudal diário médio da série temporal do CE2 em m^3/h

O valor médio de caudal diário, representado na Figura 19 demonstra que existe um padrão de consumo muito idêntico para os dias úteis. Este padrão altera-se nos fins-de-semanas, onde também existe um padrão de consumo especifico para estes dias. Observe-se que o padrão de consumo para diferentes dias da semana não sofre variações significativas ao longo do ano.

Em oposição à série temporal do CE1, a série temporal do CE2 não apresenta uma grande discrepância entre os valores médios de caudal diário ao longo do ano. No entanto, e em concordância com a análise feita anteriormente no CE1 foi considerado que os meses do semestre de inverno iam de novembro até abril, e os meses do semestre verão iam de maio a outubro, com o objetivo de representar o valor médio caudal mensal do ano de 2018 da série temporal do CE2. Assim sendo, apresenta-se nas Figura 20 e Figura 21 o valor médio de caudal mensal para os meses de inverno e para os meses de verão, respetivamente. Em ambas as figuras, também é representado a média diária anual do caudal

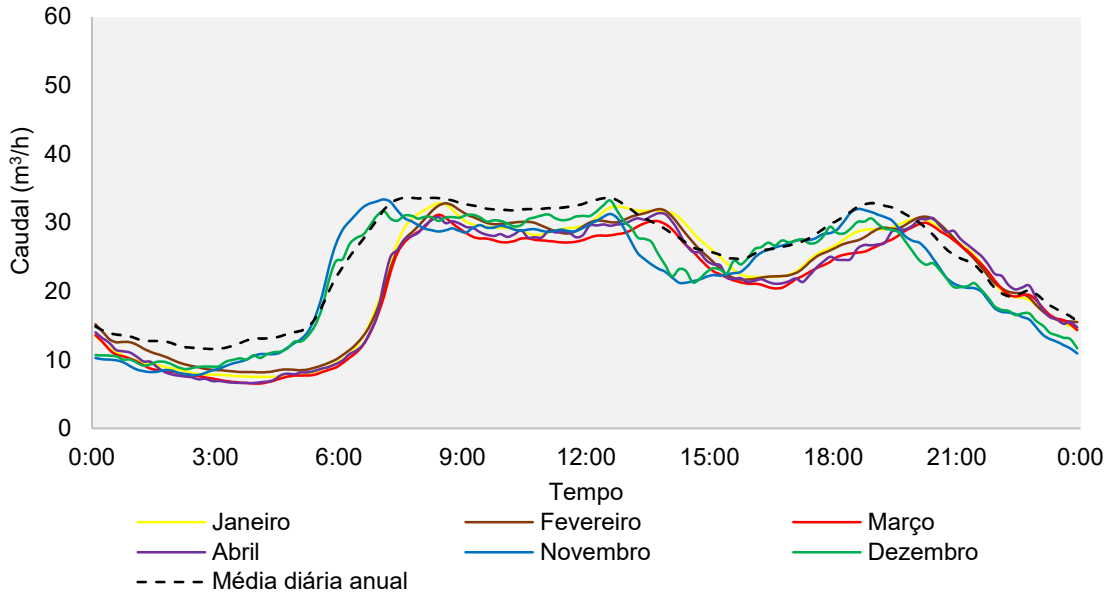

Figura 20 - Caudal mensal médio da série temporal do CE2 em m³/h para os meses de inverno

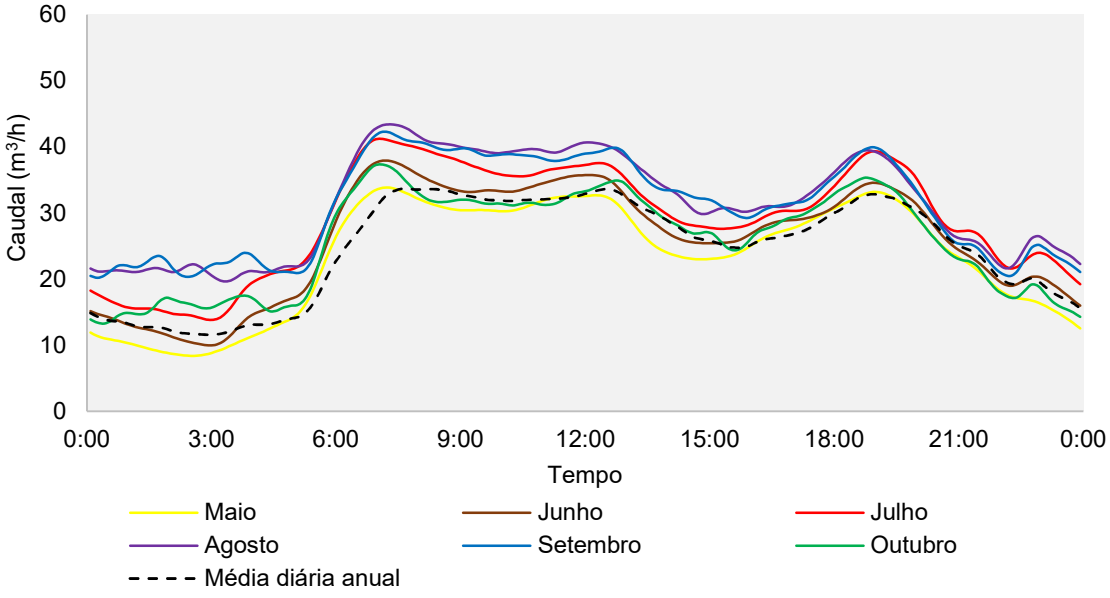

Figura 21 - Caudal mensal médio da série temporal do CE2 em m³/h para os meses de verão

Como esperado, os valores médios de caudal mensal apresentados nas Figura 20 e Figura 21, demonstram que os padrões de distribuição sofrem alterações pouco significativas entre os diferentes meses. Nos meses do semestre de inverno, observa-se um pico de consumo matinal entre as oito e nove da manhã, enquanto que nos meses do semestre de verão o mesmo pico dá-se por volta das seis e sete da manhã. Da mesma forma, observa-se que o pico de consumo noturno se dá mais cedo no verão, do que no inverno. Consegue-se justificar estas alterações nos horários da população, pelo aumento do calor nos meses considerados

de verão e também, por o CE2 referir-se a uma rede do interior de Portugal, em que o sustento da maioria da população provém da agricultura. Tradicionalmente, nos meses de maior calor os trabalhadores agrícolas iniciam os trabalhos mais cedo para evitar as horas em que o calor é mais intenso. Para além da alteração dos horários da população, a partir das Figura 20 e Figura 21 observa-se um aumento do valor médio de caudal mensal nos meses de maior calor, como expectável.

De maneira a compreender se existe variabilidade entre as medições ao longo dos meses do ano de 2018, apresenta-se na Figura 22 os diagramas de extremos e quartis para cada mês da série temporal do CE2.

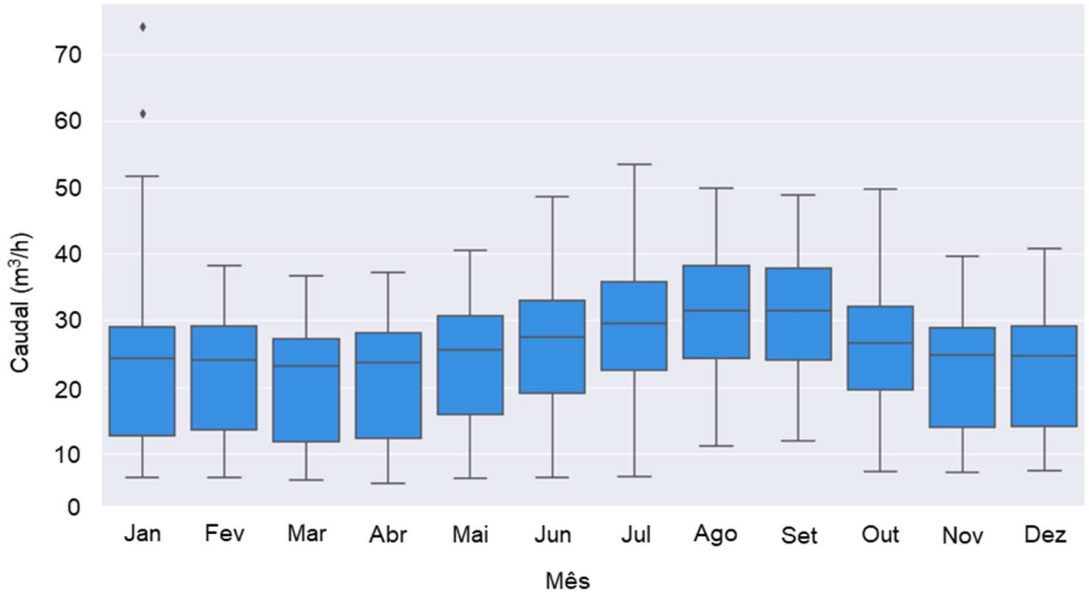

Figura 22 - Diagramas de extremos e quartis para cada mês da série temporal do CE2

Nos diagramas de extremos e quartis para cada mês do CE2 apresentados na Figura 22, demonstram uma alta dispersão dos dados (i.e., variabilidade entre medições) e uma distribuição assimétrica dos dados nos meses de inverno, enquanto para os meses de verão os diagramas apresentam uma baixa amplitude com uma distribuição simétrica dos dados.

4.2.3. RESULTADOS E DISCUSSÃO

Na presente subsecção apresentam-se os resultados de três testes comparativos entre os modelos de reconstrução anteriormente apresentados. No primeiro teste considerou-se a série temporal do CE2 com espaçamentos entre medições de 1 hora, tendo-se como objetivo avaliar o desempenho dos modelos e o seu tempo de computação na previsão de um dia da semana completo. Similarmente, no segundo teste o objetivo também passou por avaliar o desempenho dos modelos e o seu tempo de computação na previsão de um dia da semana completo. No entanto, considerou-se a série temporal do CE2 com espaçamentos de 10 minutos. No terceiro teste, também se considerou a série temporal com intervalos de 10

minutos para prever um feriado e avaliar o seu desempenho, bem como o seu tempo de computação.

No Teste 1 (avaliação do desempenho e tempo de computação dos modelos de reconstrução na previsão de um dia da semana), foi considerado como histórico o mês de setembro completo, mais os três primeiros dias do mês de outubro, da série temporal do CE2, com intervalos de 1 hora. Assim sendo, os modelos preveem 24 períodos de 1 hora, completando um dia. Na Figura 23, apresentam-se os resultados obtidos para cada modelo bem como as medições reais de caudal.

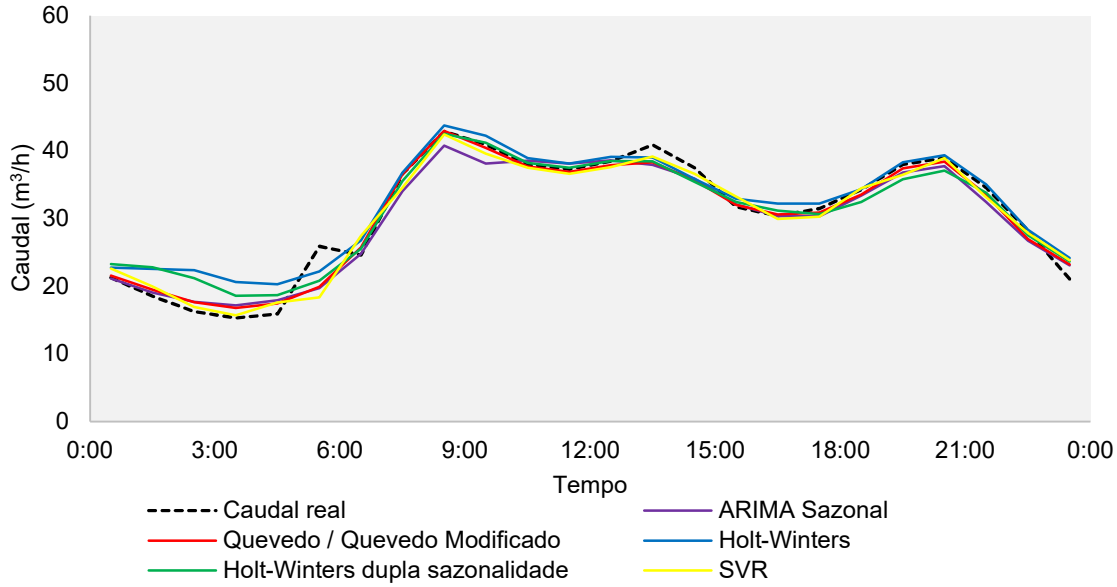

Figura 23 - Comparação dos cinco modelos de reconstrução considerando um dia da semana e intervalos de 1 hora da série temporal do CE2

A abordagem desenvolvida por Quevedo e Quevedo modificado apresentam o melhor resultado (RMSE=1,64), seguida do modelo clássico ARIMA sazonal (RMSE=1,99), do modelo SVR (RMSE=2,04), e por último dos modelos de suavização exponencial Holt-Winters de dupla sazonalidade e Holt-Winters (RMSE=2,28 e RMSE=2,47), respetivamente.

Praticamente todos os modelos são relativamente rápidos na previsão de um dia completo para a série temporal do caso de estudo com intervalos de 1 hora, demorando menos de 40 segundos até obter uma previsão, à exceção do modelo ARIMA sazonal (t=163s). Na Tabela 7, apresenta-se um resumo do tempo de computação (em segundos) e o valor de RMSE para todos os modelos.

Tabela 7 - Comparação do desempenho dos modelos e o seu tempo de computação considerando um dia da semana e intervalos de 1 hora da série temporal do CE2

Modelos	Dia da semana (1h)	
	RMSE (m^3/h)	Tempo de computação (s)
ARIMA sazonal	1,99	163,0
Quevedo	1,64	0,5
Holt-Winters	2,47	32,3
Holt-Winters dupla sazonalidade	2,28	12,0
Quevedo modificado	1,64	0,5
SVR	2,04	1,5

No Teste 2 (avaliação do desempenho e tempo de computação dos modelos de reconstrução na previsão de um dia da semana), foi considerado como histórico o mês de setembro completo, mais os três primeiros dias do mês de outubro, da série temporal do CE2, com intervalos de 10 minutos. Assim sendo, os modelos preveem 144 períodos de 10 minutos, completando 1 dia. Na Figura 24, apresenta-se os resultados obtidos para cada modelo bem como as medições reais de caudal.

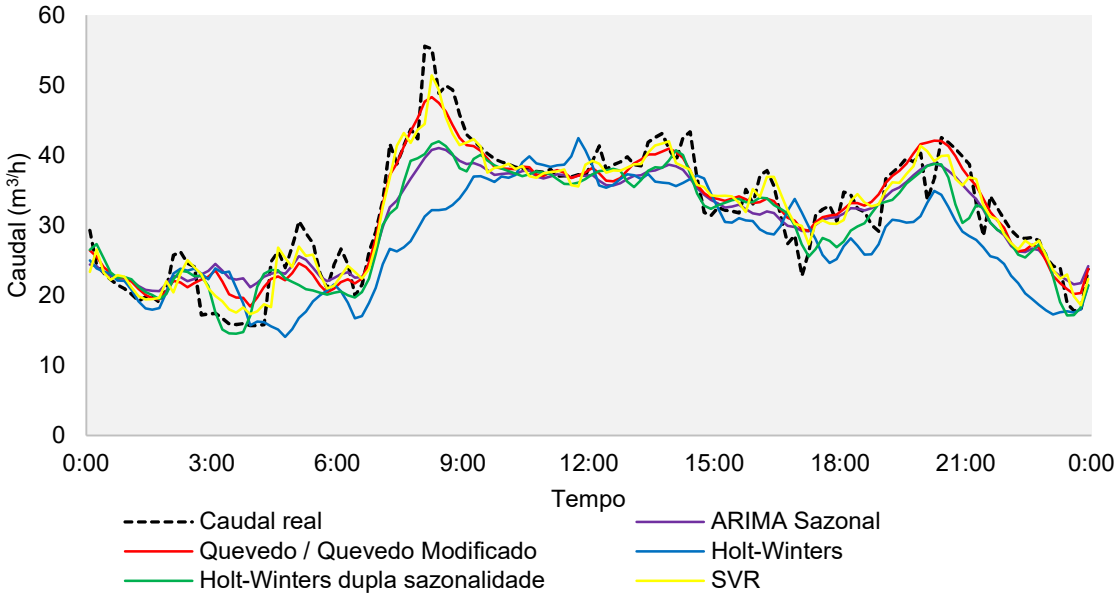

Figura 24 - Comparação dos cinco modelos de reconstrução considerando um dia da semana e intervalos de 10 minutos da série temporal do CE2

Na previsão de um dia completo com intervalos de 10 minutos, o modelo SVR apresenta o melhor desempenho (RMSE=2,70), seguido da abordagem preconizada por Quevedo (RMSE=2,79). Os restantes modelos falham a previsão no decorrer do período da manhã, sendo que o modelo ARIMA sazonal apresenta o terceiro melhor resultado (RMSE=3,98),

seguido do Holt-Winters de dupla sazonalidade (RMSE=4,04) e por último do Holt-Winters (RMSE=7,14).

Contudo a diferença no tempo de computação entre os modelos é bastante significativa, sendo a abordagem preconizada por Quevedo a que apresenta resultados mais rápidos (t=0,7s), seguido do modelo SVR (t=3,9s), dos modelos de suavização exponencial, Holt-Winters e Holt-Winters de dupla sazonalidade (t=77,8s e t=281s), respetivamente. Em último, o modelo ARIMA sazonal (t=803s). Na Tabela 8 apresenta-se um resumo do tempo de computação (em segundos) e o valor de RMSE para todos os modelos.

Tabela 8 - Comparação do desempenho dos modelos e o seu tempo de computação considerando um dia da semana e intervalos de 10 minutos da série temporal do CE2

Modelos	Dia da semana (10 minutos)	
	RMSE (m^3/h)	Tempo de computação (s)
ARIMA sazonal	3,98	803,0
Quevedo	2,79	0,7
Holt-Winters	7,14	77,8
Holt-Winters dupla sazonalidade	4,04	281,0
Quevedo Modificado	2,79	0,7
SVR	2,70	3,9

No Teste 3 (avaliação do desempenho e tempo de computação dos modelos de reconstrução na previsão de um feriado), foi considerado como histórico o mês de setembro completo, mais os quatro primeiros dias do mês de outubro, da série temporal do CE2, com intervalos de 10 minutos. Assim sendo, os modelos preveem 144 períodos de 10 minutos, completando 1 dia. A diferença dos testes anteriores prende-se com o facto de o último dia da série temporal do CE2 ser um feriado (i.e., 5 de outubro). Na Figura 25, apresentam-se os resultados obtidos para cada modelo bem como as medições reais de caudal.

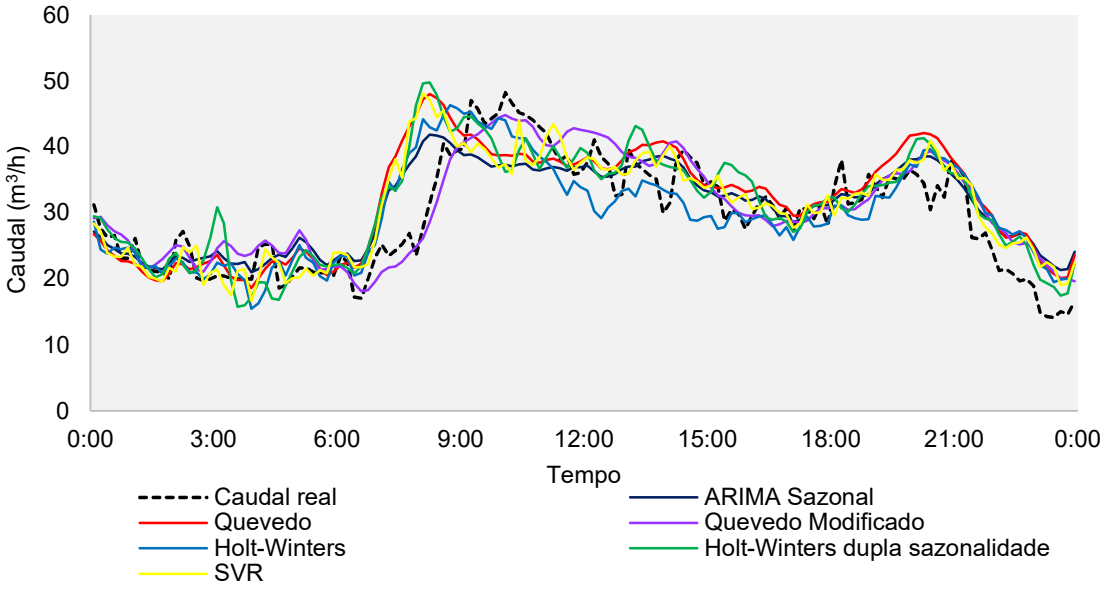

Figura 25 - Comparação dos cinco modelos de reconstrução considerando um feriado e intervalos de 10 minutos da série temporal do CE2

Na previsão de um feriado que se apresenta na Figura 25 é possível concluir que apenas a abordagem Quevedo Modificado tem um desempenho razoável (RMSE=3,55) e os restantes modelos falharam na previsão do feriado. Os modelos autorregressivos (i.e., ARIMA sazonal) e os modelos de suavização exponencial (i.e., Holt-Winters e Holt-Winters de dupla sazonalidade) realizam previsões baseados nos períodos sazonais e não permitem "sair" fora destes períodos. Similarmente, a abordagem original de Quevedo realiza as previsões do caudal diário agregado utilizando um modelo autorregressivo e o modelo SVR, efetua as suas previsões com base nas regressões com as observações anteriores. Na Tabela 9, apresenta-se um resumo do tempo de computação (em segundos) e o valor de RMSE para todos os modelos.

Tabela 9 - Comparação do desempenho dos modelos e o seu tempo de computação considerando um feriado e intervalos de 10 minutos da série temporal do CE2

Modelos	**Feriado (10 minutos)**	
	RMSE (m^3/h)	Tempo de computação (s)
ARIMA sazonal	4,75	682,0
Quevedo	5,66	0,7
Holt-Winters	4,97	77,5
Holt-Winters dupla sazonalidade	5,48	203,0
Quevedo modificado	3,55	0,6
SVR	5,18	4,1

4.3. CASO DE ESTUDO 3

4.3.1. DESCRIÇÃO DAS SÉRIES TEMPORAIS DO CASO DE ESTUDO 3

O caso de estudo 3 (CE3) é uma rede de distribuição de água localizada no sul de Portugal, na região do Algarve, na freguesia de Almancil, no município de Loulé. A ZMC, tem como objetivo a medição e controlo do fornecimento de água a um empreendimento turístico de luxo, com alta sazonalidade. Além de uma população flutuante de cerca de 3.000 habitantes no inverno e de 14.000 no verão, a rede de distribuição de água também abastece alguns campos de golfe. Esta EG é considerada uma das concessionárias de água mais experiente digitalmente no país, uma vez que possui medição remota em tempo real para todos os seus consumidores. A EG usa no seu processo de aquisição de dados um medidor de caudal de impulso, na entrada da rede de distribuição de água, que faz o registo das medições a cada 2 m^3 de água, tornando o intervalo entre medições não igualmente espaçado.

Tendo em conta que os modelos de reconstrução requerem dados validados, tratados e normalizados temporalmente, os dados disponibilizados pela EG são referentes ao ano de 2017 e foram previamente validados e normalizados temporalmente em intervalos de 10 minutos e 1 hora. Para tal, recorreu-se à ferramenta computacional apresentada anteriormente desenvolvida por Ferreira *et al.* (2022), para a validação de séries temporais de caudal.

Nos dados disponibilizados pela EG observou-se: valores anormalmente baixos, períodos sem medição e patamares estáticos, que na totalidade representou em média 6,7% de valores anómalos identificados. Deste processo de validação resulta a série temporal com um espaçamento normalizado e com falhas de longa duração.

4.3.2. ANÁLISE EXPLORATÓRIA DA SÉRIE TEMPORAL DO CASO DE ESTUDO 3

Da mesma maneira que os casos de estudo anteriores, apresenta-se na Tabela 10 algumas medidas descritivas. Estas medidas são representativas da série temporal de caudal do CE3 referente ao ano de 2017, validada e normalizada temporalmente para um período de 10 minutos. Das 52.560 medições contabilizadas, cerca de 5,22% são medições em falta, ou seja, existem 2.746 intervalos de 10 minutos sem medição, o que perfaz aproximadamente, 20 dias com dados em falta no ano de 2017.

A média de caudal no ano de 2017 foi de 58,5 m^3/h, sendo que o desvio padrão foi de 43,4 m^3/h. O valor alto do desvio padrão pode ser explicado pela diferença de consumos ao longo do ano, caracterizado por um baixo consumo nos primeiro e quarto trimestres, e um consumo elevado no terceiro trimestre. O valor máximo de caudal registado foi de 273,3 m^3/h enquanto o mínimo foi de 2,5 m^3/h.

Tabela 10 - Medidas descritivas da série temporal do CE3

Medida descritiva	CE3 (2017)												
	Jan.	**Fev.**	**Mar.**	**Abr.**	**Mai.**	**Jun.**	**Jul.**	**Ago.**	**Set.**	**Out.**	**Nov.**	**Dez.**	**Total**
Medições	4.464	4.032	4.464	4.320	4.464	4.320	4.464	4.464	4.320	4.464	4.320	4.464	52.560
Medições em falta	303	369	178	146	518	176	162	168	227	155	167	177	2.746
Média (m^3/h)	25,2	17,2	31,9	55,5	63,0	88,7	97,2	93,1	87,0	65,4	45,7	31,8	58,5
Desvio Padrão (m^3/h)	14,6	11,5	18,1	27,5	35,0	43,0	47,6	44,9	47,7	38,9	26,9	18,0	43,4
Mínimo (m^3/h)	4,1	2,6	4,0	9,0	2,5	11,5	20,1	20,2	12,2	10,7	8,4	6,7	2,5
P25 (m^3/h)	15,2	9,1	19,0	34,3	34,0	55,5	60,3	56,5	49,2	34,2	23,9	17,1	25,0
P50 (m^3/h)	21,5	14,4	27,4	50,8	56,4	82,3	87,2	84,0	75,0	58,0	38,3	26,9	47,6
P75 (m^3/h)	30,9	20,7	40,6	73,9	88,0	113,7	125,8	123,7	116,1	86,6	61,3	43,3	81,1
Máximo (m^3/h)	94,2	101,3	117,0	167,5	188,7	248,4	273,3	234,5	253,3	220,3	170,8	108,7	273,3

Similar aos casos de estudo anteriormente apresentados e com objetivo de explorar as sazonalidades diárias e semanais, apresenta-se na Figura 26 o valor médio de caudal horário para os diferentes dias da semana (i.e., dias úteis, sábados e domingos) da série temporal do CE3.

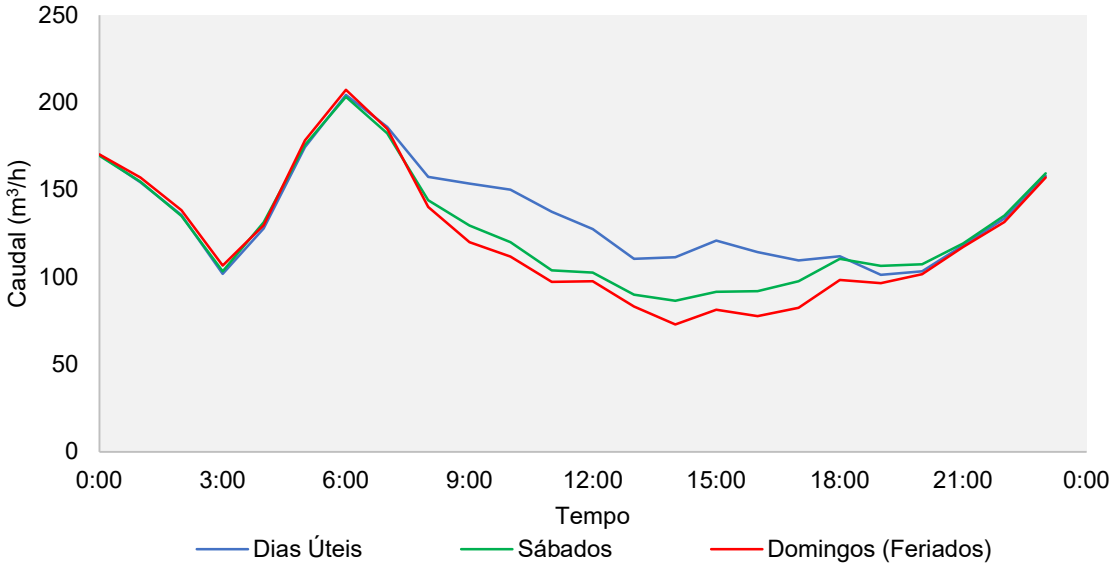

Figura 26 - Caudal horário médio para os diferentes dias da semana da série temporal do CE3

A série temporal do CE3 não difere dos restantes casos de estudo. Na Figura 26 observa-se uma alta sazonalidade diária e semanal. Existe um padrão de distribuição específico para os dias úteis e um padrão para os dias de fim de semana e feriados. No entanto, o CE3 apresenta diferenças significativas em comparação aos casos de estudos anteriormente apresentados, no que diz respeito aos picos de consumo. Na Figura 26, observa-se que o pico de maior consumo se dá no período noturno, enquanto o pico de menor consumo se dá no período da tarde, situação completamente diferente do CE1 e CE2. As diferenças apresentadas prendem-se com o facto de a rede de abastecimento do CE3 estar inserida num empreendimento com campos de golfe e as regas dos mesmos serem realizadas no período da noite.

Da mesma forma que se analisou a sazonalidade das séries temporais do CE1 e CE2, apresenta-se agora na Figura 27 o valor médio de caudal diário para os dias úteis, sábados e domingos em diferentes meses do ano de 2017. A representação gráfica da totalidade dos meses não será exibida devido ao elevado número de medições que dificulta a leitura da informação. Sendo assim, recorreu-se aos meses de abril, agosto e outubro para exemplificar as sazonalidades da série temporal do CE3, com a segunda-feira como dia inicial e o domingo como o dia final.

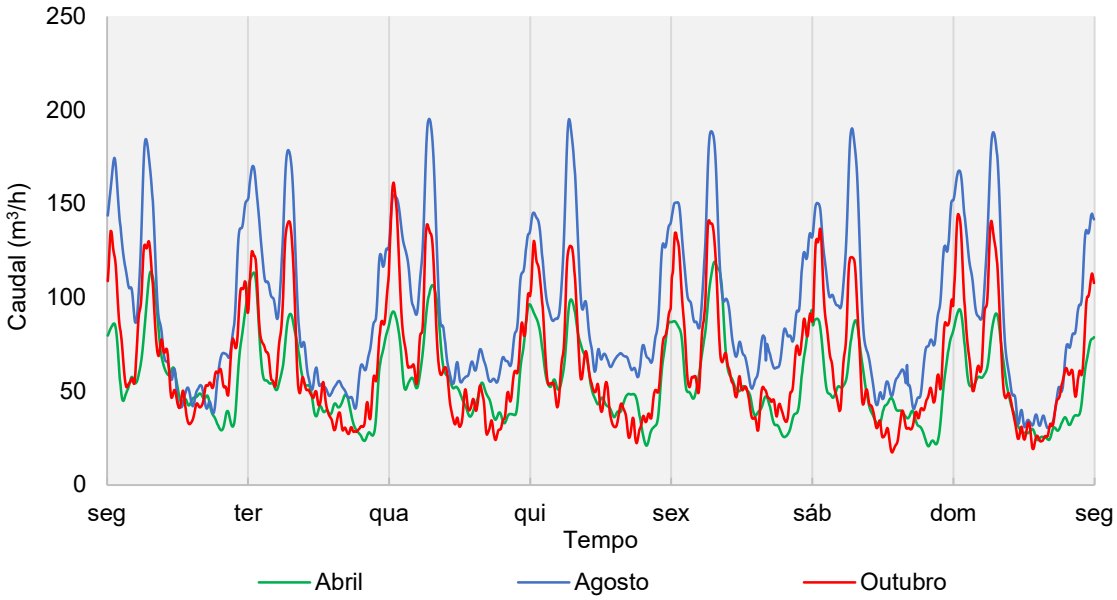

Figura 27 - Caudal diário médio da série temporal CE3 em m^3/h

Na Figura 27, representa-se o valor médio de caudal diário e observa-se que existe um padrão de consumo muito idêntico para os diferentes dias da semana (i.e., dias úteis, fim de semana e feriados). Similarmente ao CE1 e em oposição ao CE2, a série temporal do CE3 apresenta grande discrepância entre os valores médios de caudal ao longo do ano. Nos casos de estudo anteriormente apresentados pode-se observar um aumento do consumo nos meses de maior calor, o que também se verifica no CE3, comparando a linha do mês de abril e do mês de agosto.

Em concordância com as análises feitas anteriormente, será considerado que os meses do semestre de inverno vão de novembro a abril, e os meses do semestre de verão vão de maio a outubro, com o objetivo de representar o valor médio caudal mensal do ano de 2017 da série temporal do CE3. Assim sendo, apresentam-se nas Figura 28 e Figura 29 o valor médio de caudal mensal para os meses de inverno e para os meses de verão, respetivamente. Em ambas as figuras, também é representado a média diária anual do caudal.

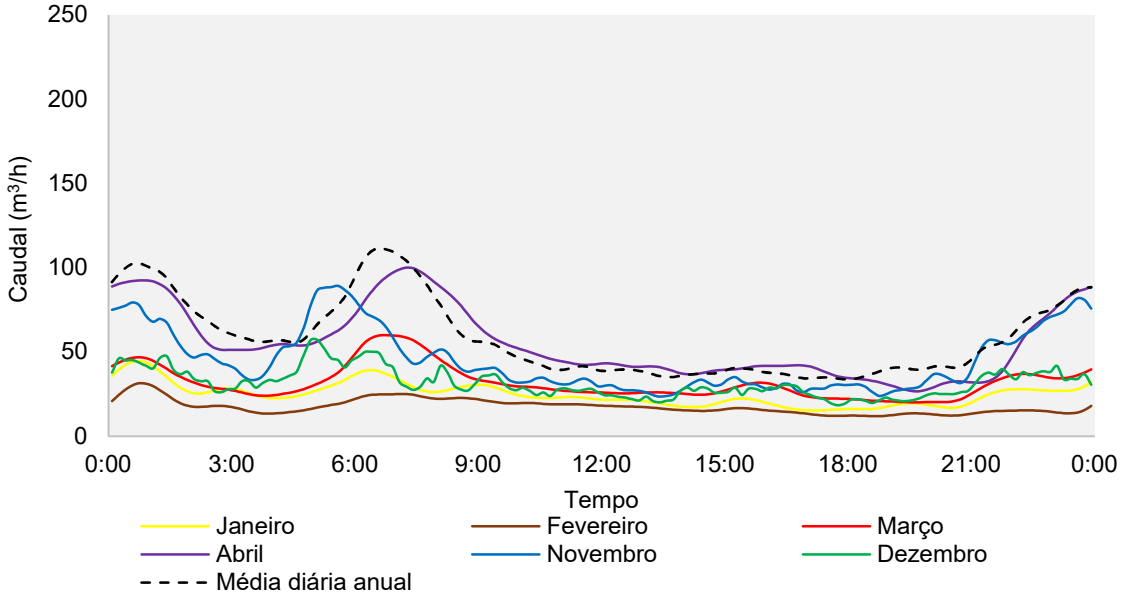

Figura 28 - Caudal mensal médio da série temporal do CE3 em m³/h para os meses de inverno

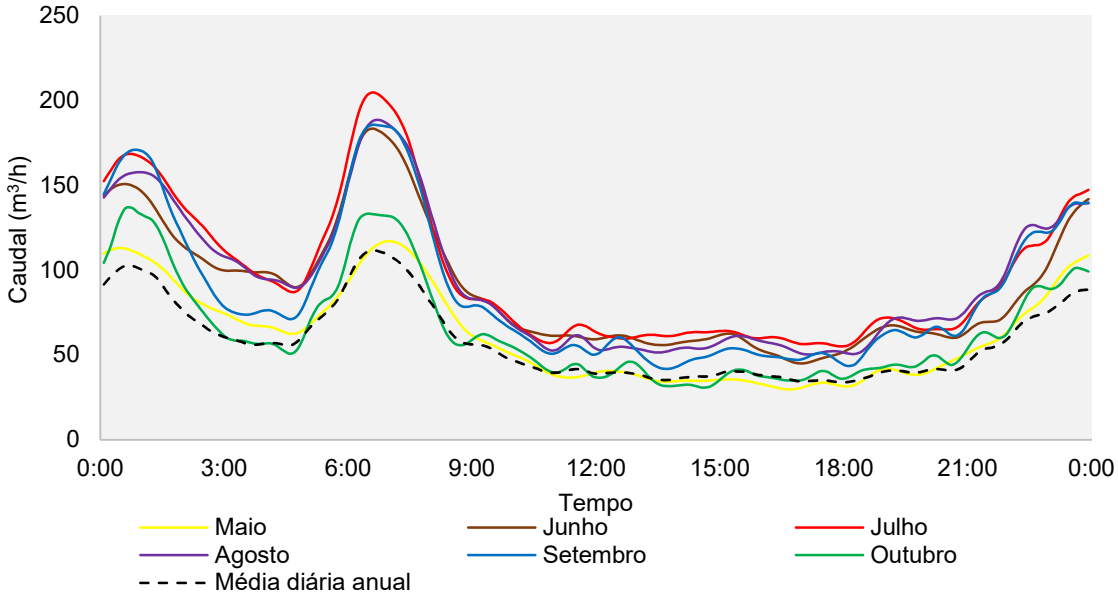

Figura 29 - Caudal mensal médio da série temporal do CE3 em m³/h para os meses de verão

Como esperado, os valores médios de caudal mensal, apresentados nas Figura 28 e Figura 29, demonstram que os padrões de distribuição sofrem alterações pouco significativas entre os diferentes meses. Observe-se que nos meses de inverno o pico de consumo matinal nem sempre se dá à mesma hora, em oposição aos meses de verão que o pico acontece sempre no mesmo horário. Como esperado, nos meses de verão observa-se um aumento do valor médio de caudal mensal.

Por forma a compreender a existência de variabilidade entre medições ao longo dos meses do ano de 2018, apresenta-se na Figura 30 os diagramas de extremos e quartis para cada mês da série temporal do CE3.

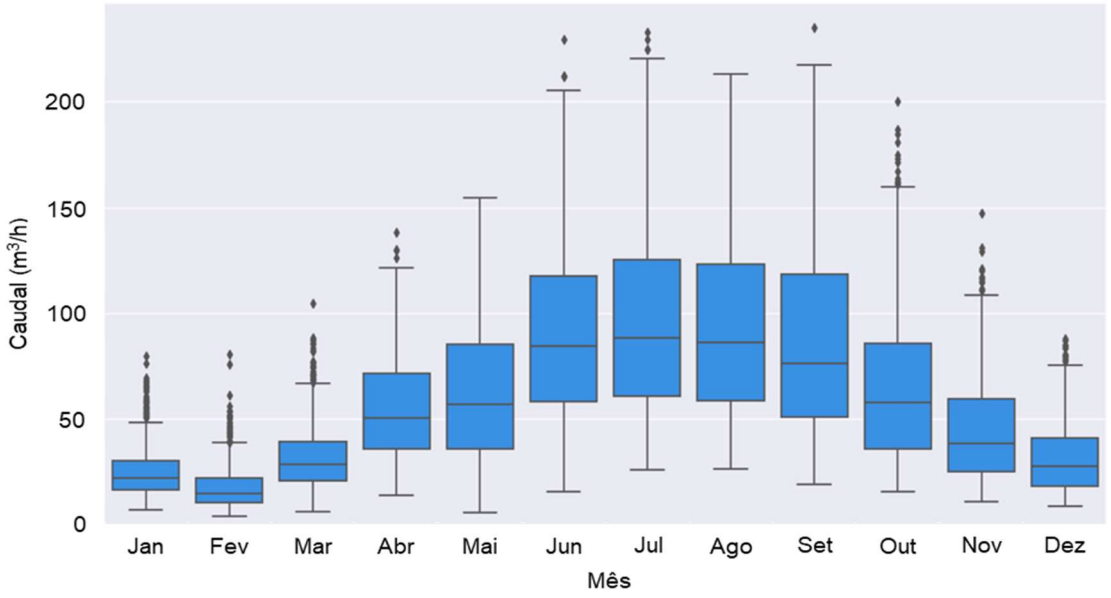

Figura 30 - Diagramas de extremos e quartis para cada mês da série temporal do CE3

Os diagramas de extremos e quartis para cada mês do CE3 apresentados na Figura 30, evidenciam uma alta dispersão dos dados (i.e., alta variabilidade entre medições), uma distribuição simétrica nos meses de maior calor e alguns dados discrepantes (i.e., *outliers*). Nos meses de inverno, os diagramas demonstram uma amplitude muito baixa dos dados e a presença de muitos dados discrepantes.

4.3.3. RESULTADOS E DISCUSSÃO

Na presente subsecção apresentam-se os resultados de três testes comparativos entre os modelos de reconstrução anteriormente apresentados. No primeiro teste considerou-se a série temporal do caso de estudo com espaçamentos entre medições de 1 hora, tendo-se como objetivo avaliar o desempenho dos modelos e o seu tempo de computação na previsão de um dia da semana completo. Similarmente, no segundo teste o objetivo também passou por avaliar o desempenho dos modelos e o seu tempo de computação na previsão de um dia da semana completo. No entanto, considerou-se a série temporal do caso de estudo com espaçamentos de 10 minutos. No terceiro teste, considerou-se a série temporal com intervalos de 10 minutos para prever um feriado e avaliar o seu desempenho, bem como o seu tempo de computação.

No Teste 1 (avaliação do desempenho e tempo de computação dos modelos de reconstrução na previsão de um dia da semana), foi considerado como histórico o mês de setembro completo, mais os três primeiros dias do mês de outubro, da série temporal do CE3, com

intervalos de 1 hora. Assim sendo, os modelos preveem 24 períodos de 1 hora, completando 1 dia. Na Figura 31, apresentam-se os resultados obtidos para cada modelo bem como as medições de caudal.

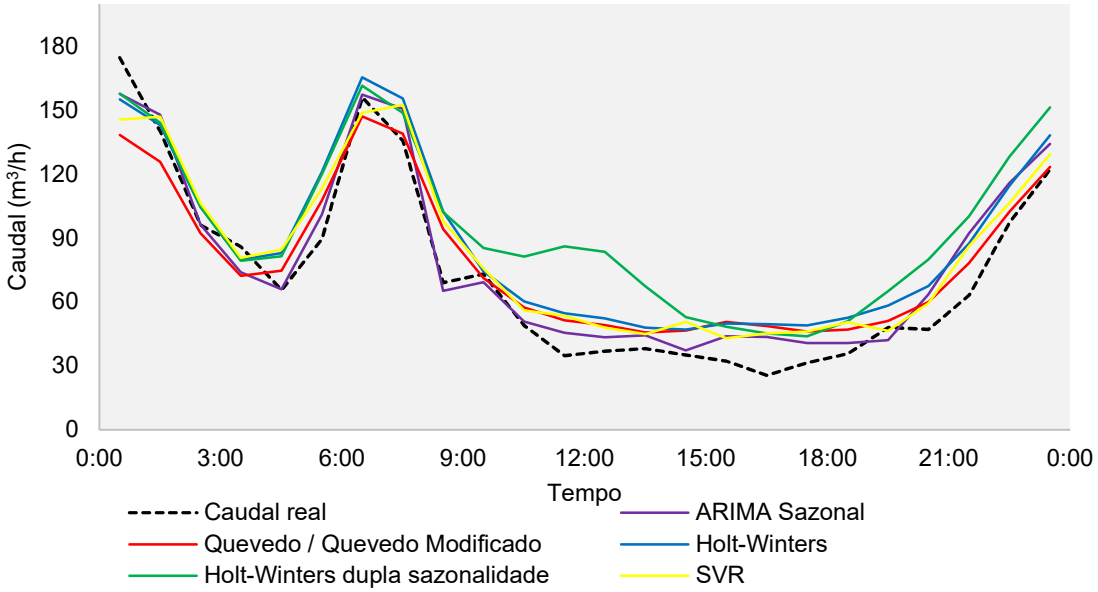

Figura 31 - Comparação dos cinco modelos de reconstrução considerando um dia da semana e intervalos de 1 hora da série temporal do CE3

Na previsão de um dia completo da série temporal do CE3 com intervalos de 1 hora, os modelos apresentados não conseguem acertar na totalidade da previsão, sendo que no período da manhã e da tarde todos os modelos falham. O modelo que consegue ter um melhor desempenho é o ARIMA sazonal (RMSE=11,70), seguido do modelo SVR (RMSE=15,38), da abordagem desenvolvida por Quevedo (RMSE=14,69) e por último dos modelos de suavização exponencial Holt-Winters e Holt-Winters de dupla sazonalidade (RMSE=17,67 e RMSE=25,50), respetivamente.

Todos os modelos são relativamente rápidos na previsão de um dia completo para a série temporal do caso de estudo com intervalos de 1 hora, demorando menos de 40 segundos até obter uma previsão. Na Tabela 11 apresenta-se um resumo do tempo de computação (em segundos) e o valor de RMSE para todos os modelos.

Tabela 11 - Comparação do desempenho dos modelos e o seu tempo de computação considerando um dia da semana e intervalos de 1 hora da série temporal do CE3

Modelos	Dia da semana (1h)	
	RMSE (m^3/h)	Tempo de computação (s)
ARIMA sazonal	11,70	38,0
Quevedo	14,69	0,4
Holt-Winters	17,67	26,5
Holt-Winters dupla sazonalidade	25,50	12,0
Quevedo Modificado	14,69	0,4
SVR	15,38	1,0

No Teste 2 (avaliação do desempenho e tempo de computação dos modelos de reconstrução na previsão de um dia da semana), foi considerado como histórico o mês de setembro completo, mais os três primeiros dias do mês de outubro, da série temporal do CE3, com intervalos de 10 minutos. Assim sendo, os modelos preveem 144 períodos de 10 minutos, completando 1 dia. Na Figura 32, apresentam-se os resultados obtidos para cada modelo bem como as medições de caudal.

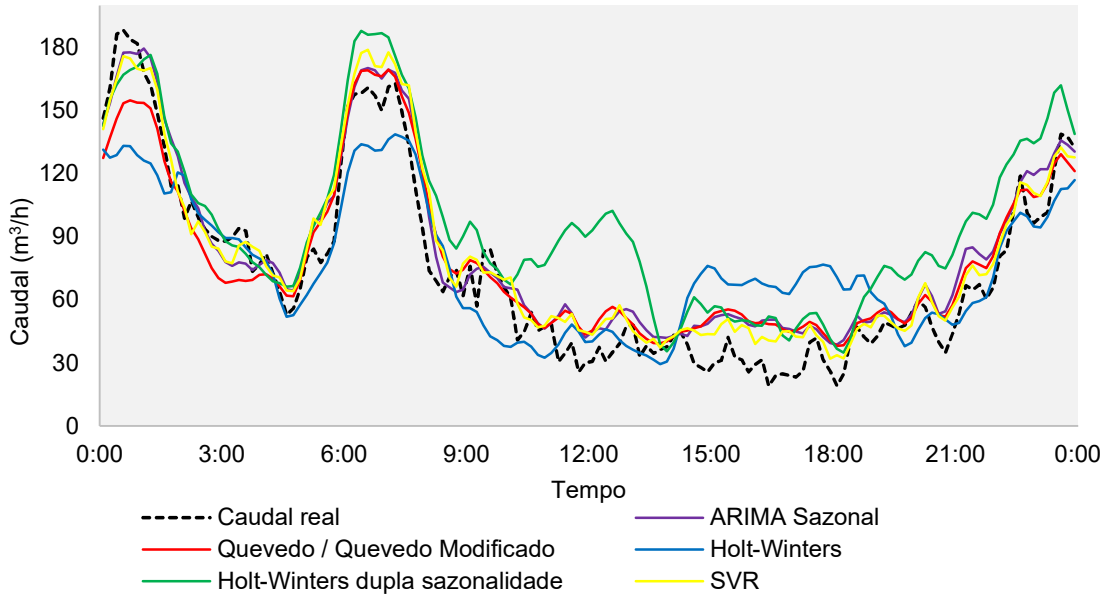

Figura 32 - Comparação dos cinco modelos de reconstrução considerando um dia da semana e intervalos de 10 minutos da série temporal do CE3

Dos modelos apresentados, nenhum consegue acertar na totalidade do feriado. O SVR aproxima-se mais do valor real de caudal na previsão de um dia completo da série temporal do CE3 com intervalos de 10 minutos (RMSE=12,62), seguido do modelo clássico ARIMA sazonal (RMSE=13,92), da abordagem desenvolvida por Quevedo e Quevedo modificado

(ambos com RMSE=14,86) e por último dos modelos de suavização exponencial, Holt-Winters de dupla sazonalidade e Holt-Winters (RMSE=23,42 e RMSE=23,95), respetivamente.

Contudo a diferença no tempo de computação entre os modelos é bastante significantiva, sendo a abordagem preconizada por Quevedo a que apresenta resultados mais rápidos (t=0,7s), seguida do modelo SVR (t=1s), dos modelos de suavização exponencial, Holt-Winters e Holt-Winters de dupla sazonalidade (t=73,2s e t=281s), respetivamente. Em último, o modelo ARIMA sazonal (t=706s). Na Tabela 12, apresenta-se um resumo do tempo de computação (em segundos) e o valor de RMSE para todos os modelos.

Tabela 12 - Comparação do desempenho dos modelos e o seu tempo de computação considerando um dia da semana e intervalos de 10 minutos da série temporal do CE3

Modelos	Feriado (10 minutos)	
	RMSE (m^3/h)	Tempo de computação (s)
ARIMA sazonal	13,92	706,0
Quevedo	14,86	0,7
Holt-Winters	23,95	73,2
Holt-Winters dupla sazonalidade	23,42	281,0
Quevedo Modificado	14,86	0,7
SVR	12,62	1,0

Normalmente, os problemas surgem na previsão de feriados quando estes ocorrem durante os dias úteis, uma vez que o padrão de distribuição de água de um feriado está geralmente relacionado com o padrão de distribuição dos domingos (em oposição aos padrões de distribuição dos dias úteis).

No Teste 3 (avaliação do desempenho e tempo de computação dos modelos de reconstrução na previsão de um feriado), foi considerado como histórico o mês de setembro completo, mais os quatro primeiros dias do mês de outubro, da série temporal do CE3, com intervalos de 10 minutos. Assim sendo, os modelos preveem 144 períodos de 10 minutos, completando 1 dia. A diferença para os testes anteriores prende-se com o facto de o último dia da série temporal do CE3 ser um feriado (i.e., 5 de outubro). Na Figura 33 apresentam-se os resultados obtidos para cada modelo bem como as medições reais de caudal.

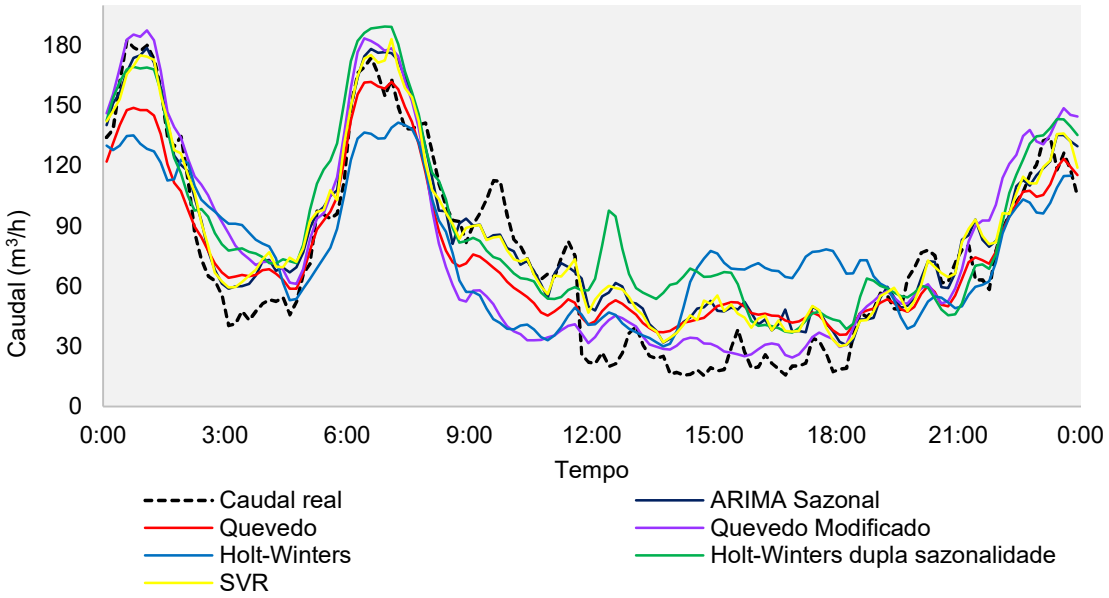

Figura 33 - Comparação dos cinco modelos de reconstrução considerando um feriado e intervalos de 10 minutos da série temporal do CE3

Na previsão de um feriado que se apresenta na Figura 33 é possível concluir que todos os modelos falharam na previsão do feriado. Os modelos autorregressivos (i.e., ARIMA sazonal) e os modelos de suavização exponencial (i.e., Holt-Winters e Holt-Winters de dupla sazonalidade) realizam previsões baseados nos períodos sazonais e não permitem "sair" fora destes períodos. Similarmente, a abordagem original de Quevedo realiza as previsões do caudal diário agregado utilizando um modelo autorregressivo e o modelo SVR, efetua as previsões tendo como base, as regressões das observações anteriores. Na Tabela 13 apresenta-se um resumo do tempo de computação (em segundos) e o valor de RMSE para todos os modelos.

Tabela 13 - Comparação do desempenho dos modelos e o seu tempo de computação considerando um feriado e intervalos de 10 minutos da série temporal do CE3

Modelos	Feriado (10 minutos)	
	RMSE (m^3/h)	Tempo de computação (s)
ARIMA sazonal	16,46	787,0
Quevedo	18,92	0,7
Holt-Winters	32,18	83,9
Holt-Winters dupla sazonalidade	25,42	272,0
Quevedo modificado	18,92	0,7
SVR	16,57	1,0

5. SÍNTESE E CONCLUSÕES

A presente dissertação teve como objetivo o estudo de modelos de reconstrução de séries temporais de sistemas de abastecimento de água, aplicado a três casos de estudo reais, que se consideram representativos da maioria das entidades gestoras portuguesas. No Capítulo 1, apresenta-se um enquadramento geral do tema, os objetivos e a estrutura desta dissertação.

O Capítulo 2, apresenta a síntese de conhecimentos, nomeadamente, nos seguintes temas: 1) Séries temporais dos sistemas de abastecimento de água; 2) Técnicas de validação de séries temporais de caudal; 3) Técnicas de reconstrução de séries temporais de caudal. Começa-se por apresentar numa primeira fase, um enquadramento geral das séries temporais em sistemas de abastecimento de água (SAA), abordando o processo de aquisição de dados por parte das entidades gestoras, as principais anomalias encontradas nos dados adquiridos e alguns exemplos gráficos de consumo de caudal, com o objetivo de demonstrar a grande evidência de ciclos diários e semanais das séries temporais dos SAA, que deve ser tida em conta na seleção dos modelos de reconstrução. Devido à existência de dados anómalos nas séries temporais de SAA foram abordadas algumas técnicas de validação de dados de séries temporais, destacando-se a metodologia desenvolvida por Ferreira *et al.* (2022). Esta metodologia foi aplicada no desenvolvimento de uma ferramenta computacional que tem como finalidade a validação de séries temporais de caudal dos SAA. Esta ferramenta foi desenvolvida para o sistema operativo Windows e permite o processamento de séries temporais de caudal (carregadas através de ficheiros no formato CSV ou TXT). A ferramenta computacional, a documentação de apoio e o código base estão disponíveis num repositório GitHub (https://github.com/Ferreira-B/Flowrate-time-series-processing) para livre utilização por entidade gestoras e pela comunidade científica. Por último, foram descritos modelos de previsão de séries temporais que permitem a reconstrução das séries. Dessa forma, apresentaram-se algumas aplicações de modelos de previsão em SAA, tais como modelos autorregressivos, modelos de suavização exponencial, modelos de aprendizagem automática e modelos híbridos.

No Capítulo 3 apresentam-se as técnicas implementadas na reconstrução das séries temporais de caudal de SAA, com o principal objetivo de realizar a comparação recorrendo a métodos que permitam avaliar o seu desempenho. Apresenta-se a formulação de cinco métodos que permitem a reconstrução das séries temporais de caudal dos SAA e uma melhoria de um desses métodos. Na primeira secção é apresentado o modelo autorregressivo (i.e., ARIMA sazonal), na segunda a técnica de reconstrução desenvolvida por Quevedo *et al.* (2010) e a melhoria à sua abordagem, na terceira apresentam-se os métodos de suavização exponencial (i.e., Holt-Winters simples e Holt-Winters de dupla sazonalidade). Na quarta secção é apresentado um modelo de aprendizagem automática (i.e., SVR). Por último, apresenta-se a métrica adotada na avaliação do desempenho das diferentes técnicas de reconstrução.

No Capítulo 4, apresenta-se a implementação dos modelos de reconstrução de séries temporais de caudal. Descrevem-se os diferentes casos de estudo, realiza-se uma análise exploratória da série temporal para cada caso de estudo e por último, apresentam-se os resultados e discute-se os testes comparativos dos modelos de reconstrução de séries temporais. No primeiro caso de estudo (CE1), o SAA está localizado na área metropolitana Lisboa e abastece uma população com cerca de 3.300 habitantes. O segundo caso de estudo (CE2) tem a sua rede localizada no interior de Portugal na região sul, abastece uma população de 110 habitantes e um aeroporto, atualmente desativado. No terceiro caso de estudo (CE3), o SAA é localizado na região sul de Portugal, numa zona afetada pela sazonalidade turística, no inverno abastece cerca de 3.000 habitantes e no verão chega a abastecer 14.000 habitantes. A análise exploratória das séries temporais de caudal, para cada caso de estudo, realizou-se tendo como base um ano de dados históricos. Com esta análise, pretendeu-se mostrar as diferentes características de cada série temporal (i.e., sazonalidades).

Os resultados e discussões apresentadas no Capítulo 4 são relativos a três testes comparativos entre os modelos de reconstrução apresentados. No primeiro teste (i.e., Teste 1) considerou-se as séries temporais dos casos de estudo com espaçamentos entre medições de 1 hora, tendo-se como objetivo avaliar o desempenho dos modelos e o seu tempo de computação na previsão de um dia útil completo. Similarmente, no segundo teste (i.e., Teste 2) o objetivo também passou por avaliar o desempenho dos modelos e o seu tempo de computação na previsão de um dia útil completo. No entanto, considerou-se as séries temporais dos casos de estudo com espaçamentos de 10 minutos. No terceiro teste (i.e., Teste 3), considerou-se as séries temporais com intervalos de 10 minutos para prever um feriado e avaliar o seu desempenho, bem como o seu tempo de computação. A comparação entre os modelos, foi realizada tendo como base que os modelos de reconstrução recorrem a dados históricos para realizar previsões de curta duração, sendo que, para os presentes casos de estudo, foi considerado, no máximo, um mês e quatro dias de registos históricos com intervalos de 1 hora e 10 minutos, para a previsão do dia da semana e do feriado, respetivamente.

No Teste 1, os modelos apresentados conseguiram prever o dia da semana da série temporal do CE2 com intervalos de 1 hora, no entanto, a previsão do dia da semana com intervalos de 1 hora da série temporal do CE1, apresentou diferenças significativas no seu desempenho e na série temporal do CE3 os modelos apresentados não conseguem acertar na previsão. Todos os modelos são relativamente rápidos na previsão de um dia da semana completo das séries temporais dos casos de estudo, com intervalos de 1 hora, demorando menos de 40 segundos até obter uma previsão, à exceção do modelo ARIMA sazonal. No entanto, as séries temporais com intervalos de 1 hora têm um uso muito limitado na operação em tempo real dos SAA. Por esse motivo, no Teste 2 o mesmo dia da semana foi previsto com intervalos de 10 minutos.

Na previsão de um dia da semana, das séries temporais dos casos de estudo em intervalos de 10 minutos, podemos concluir que o modelo SVR destacou-se dos restantes, apresentando um desempenho superior nas previsões de todas as séries temporais dos casos de estudo e obteve resultados em menos de 4 segundos. Contudo, os problemas surgem na previsão de feriados quando estes ocorrem durante os dias de semana, uma vez que o padrão de distribuição de água de um feriado está geralmente relacionado com o padrão de distribuição dos domingos (em oposição aos padrões de distribuição dos dias úteis).

No Teste 3, os modelos não conseguiram prever o feriado das séries temporais dos casos de estudo, com intervalos de 10 minutos. Os modelos autorregressivos (i.e., ARIMA sazonal) e os modelos de suavização exponencial (i.e., Holt-Winters e Holt-Winters de dupla sazonalidade) realizam previsões baseados nos períodos sazonais e não permitem "sair" fora destes períodos. Similarmente, a abordagem original de Quevedo realiza as previsões do caudal diário agregado utilizando um modelo autorregressivo e o modelo SVR efetua as previsões tendo como base um conjunto de dados composto pelas observações anteriores ao período do dia em causa para vários dias, do mesmo tipo, em histórico. A abordagem de Quevedo modificada, no CE1 consegue prever o pico de consumo matinal associado aos dias de feriados, no entanto, falha a previsão no período da tarde. No CE2, consegue acertar na previsão do feriado da série temporal e no CE 3 não acerta a totalidade da previsão do feriado.

As técnicas de reconstrução de séries temporais de caudal têm de permitir a previsão de todos os cenários possíveis. No estudo comparativo entre os diferentes modelos de reconstrução de séries temporais de caudal, realizado na presente dissertação, pretendeu-se demonstrar a fragilidade dos modelos ao prever um dia de feriado quando este ocorre durante um dia de semana. No entanto, no presente estudo não foi avaliado qual o impacto na previsão de um dia de feriado quando este ocorre em diferentes dias de semana, mas sim o impacto na previsão quando o dia de feriado ocorre num dia de semana aleatório.

Em suma, podemos concluir que o modelo autorregressivo, os modelos de suavização exponencial, a abordagem original de Quevedo e o modelo de aprendizagem automática não são adequados para a aplicação em técnicas de reconstrução de séries temporais de caudal, pois falham na previsão de um dia de feriado quando este ocorre durante um dia útil. Abordagem de Quevedo modificada, proposta na presente dissertação, apresentou os resultados mais próximos dos valores reais de caudal na previsão de um feriado, das séries temporais dos casos de estudo, com intervalos de 10 minutos. No entanto, abordagem Quevedo modificada apresenta fragilidades, devido a certas falhas apresentadas nas previsões de um feriado. Dessa forma, recomenda-se um estudo comparativo dos modelos na previsão do caudal diário agregado de um feriado.

Como trabalhos futuros recomenda-se a inclusão da previsão de caudal dos sistemas de abastecimento de água, quando existe ausência de dados históricos de confiança, por exemplo, devido a mudanças discrepantes nos padrões de consumo, motivadas pelas restrições impostas durante o combate à pandemia COVID-19.

No decurso do presente trabalho de mestrado, foi elaborado um artigo em conferência internacional, nomeadamente:

1. Ascensão, C., Ferreira, B., Barreira, R., & Carriço, N. (2021). Comparison of reconstruction methods for water supply systems flow rate time series. In Proceedings of the 1st International Conference on Water Energy Food and Sustainability (ICoWEFS 2021) (pp. 851–858). Springer International Publishing.

No Anexo I apresenta-se cópia do artigo.

.

REFERÊNCIAS BIBLIOGRÁFICAS

Akaike, H. (1974) A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, 19 (6), pp. 716–723. DOI:10.1109/TAC.1974.1100705.

Alvisi, S., Franchini, M. and Marinelli, A. (2007) A short-term, pattern-based model for water-demand forecasting, *Journal of Hydroinformatics*, 9 (1), pp. 39–50. DOI:10.2166/hydro.2006.016.

Antunes, A., Andrade-Campos, A., Sardinha-Lourenço, A. and Oliveira, M. S. (2018) Short-term water demand forecasting using machine learning techniques, *Journal of Hydroinformatics*, 20 (6), pp. 1343–1366. DOI:10.2166/hydro.2018.163.

Bakker, M., Vreeburg, J. H. G., van Schagen, K. M. and Rietveld, L. C. (2013) A fully adaptive forecasting model for short-term drinking water demand, *Environmental Modelling and Software*, 48, pp. 141–151. DOI:10.1016/j.envsoft.2013.06.012.

Barrela, R., Amado, C., Loureiro, D. and Mamade, A. (2017) Data reconstruction of flow time series in water distribution systems- a new method that accommodates multiple seasonality, *Journal of Hydroinformatics*, 19 (2), pp. 238–250. DOI:10.2166/hydro.2016.192.

Box, G. E. P. and Jenkins, G. M. (1976) *Time Series Analysis: Forecasting and Control (Revised Edition)*, Enders, R. (ed.) . Revised. Oakland, California: Holden-Day.

Box, G., Jenjins, G., Reinsel, G. and Ljung, G. (2018) *Time Series Analysis: Forecasting and Control*, *Wiley*. Vol. 1.

Boyle, T., Giurco, D., Mukheibir, P., Liu, A., Moy, C., White, S. and Stewart, R. (2013) Intelligent metering for urban water: A review, *Water (Switzerland)*, 5 (3), pp. 1052–1081. DOI:10.3390/w5031052.

Brentan, B. M., Luvizotto Jr., E., Herrera, M., Izquierdo, J. and Pérez-García, R. (2017) Hybrid regression model for near real-time urban water demand forecasting, *Journal of Computational and Applied Mathematics*, 309, pp. 532–541. DOI:10.1016/j.cam.2016.02.009.

Caiado, J. (2010) Performance of Combined Double Seasonal Univariate Time Series Models for Forecasting Water Demand, *Journal of Hydrologic Engineering*, 15 (3), pp. 215–222. DOI:10.1061/(asce)he.1943-5584.0000182.

Cugueró-Escofet, M. À., García, D., Quevedo, J., Puig, V., Espin, S. and Roquet, J. (2016) A methodology and a software tool for sensor data validation/reconstruction: Application to the Catalonia regional water network, *Control Engineering Practice*, 49, pp. 159–172. DOI:10.1016/j.conengprac.2015.11.005.

de Marinis, G., Gargano, R. and Tricarico, C. (2008) Water Demand Models for a Small Number of Users, in: *Water Distribution Systems Analysis Symposium 2006*. Reston, VA: American Society of Civil Engineers, pp. 1–14.

Donkor, E. A., Mazzuchi, T. A., Soyer, R. and Alan Roberson, J. (2014) Urban Water Demand Forecasting: Review of Methods and Models, *Journal of Water Resources Planning and Management*, 140 (2), pp. 146–159. DOI:10.1061/(asce)wr.1943-5452.0000314.

Dudek, G. (2013) Forecasting Time Series with Multiple Seasonal Cycles, pp. 52–63.

Ferreira, B., Carriço, N., Barreira, R., Dias, T. and Covas, D. (2022) Flowrate Time Series Processing in Engineering Tools for Water Distribution Networks, *Water Resources Research*, 58 (6), pp. 1–20. DOI:10.1029/2022WR032393.

Firat, M., Turan, M. E. and Yurdusev, M. A. (2010) Comparative analysis of neural network techniques for predicting water consumption time series, *Journal of Hydrology*, 384 (1–2), pp. 46–51. DOI:10.1016/j.jhydrol.2010.01.005.

Gagliardi, F., Alvisi, S., Kapelan, Z. and Franchini, M. (2017) A probabilistic short-termwater demand forecasting model based on the Markov chain, *Water (Switzerland)*, 9 (7), pp. 7–14. DOI:10.3390/w9070507.

Galvas, G. (2016) Time series forecasting used for real-time anomaly detection on websites author: Georgios Galvas, (October).

Ghalehkhondabi, I., Ardjmand, E., Young, W. A. and Weckman, G. R. (2017) Water demand forecasting: review of soft computing methods, *Environmental Monitoring and Assessment*, 189 (7), pp. 313. DOI:10.1007/s10661-017-6030-3.

Groppo, G., Costa, M. A. and Libânio, M. (2019) Predicting water demand: A review of the methods employed and future possibilities, *Water Science and Technology: Water Supply*, 19 (8), pp. 2179–2198. DOI:10.2166/ws.2019.122.

Han, J., Micheline, K. and Jian, P. (2012) *Data Mining: Concepts, Models, Methods, and Algorithms*, *IIE Transactions*. Vol. 36. DOI:10.1080/07408170490426107.

Hassan, S. N., Ahmad, M. H., Suhartono and Mohamed, N. (2012) A comparison of the forecast performance of double seasonal ARIMA and double seasonal ARFIMA models of electricity load demand, *Applied Mathematical Sciences*, 6 (133–136), pp. 6705–6712.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning_ Data Mining, Inference, and Prediction, Second Edition*. 2nd ed. 20. Stanford, California: Springer International Publishing.

Henriques, J., Palma, J. and Ribeiro, Á. (2006) Medição de caudal em sistemas de abastecimento de água e de saneamento de águas residuais urbanas, *Série Guias Técnicos Nº 9*.

Herrera, M., Torgo, L., Izquierdo, J. and Pérez-García, R. (2010) Predictive models for forecasting hourly urban water demand, *Journal of Hydrology*, 387 (1–2), pp. 141–150. DOI:10.1016/j.jhydrol.2010.04.005.

Hyndman, R. J. and Athanasopoulos, G. (2012) *Forecasting : Principles and Practice*.

Hyndman, R. J., Koehler, A. B., Ord, J. K. and Snyder, R. D. (2008) *Springer Series in Statistics Forecasting with Exponential Smoothing*.

Jowitt, P. W. and Xu, C. (1992) Demand forecasting for water distribution systems, *Civil Engineering Systems*, 9 (2), pp. 105–121. DOI:10.1080/02630259208970643.

Kirstein, J. K., Høgh, K., Rygaard, M. and Borup, M. (2019) A semi-automated approach to validation and error diagnostics of water network data, *Urban Water Journal*, 16 (1), pp. 1–10. DOI:10.1080/1573062X.2019.1611884.

Li, W. and Huicheng, Z. (2010) Urban water demand forecasting based on HP filter and fuzzy neural network, *Journal of Hydroinformatics*, 12 (2), pp. 172–184. DOI:10.2166/hydro.2009.082.

Loureiro, D., Amado, C., Martins, A., Vitorino, D., Mamade, A. and Coelho, S. T. (2016) Water distribution systems flow monitoring and anomalous event detection: A practical approach, *Urban Water Journal*, 13 (3), pp. 242–252. DOI:10.1080/1573062X.2014.988733.

Mamade, A. (2013) *Profiling consumption patterns using extensive measurements*, *Tecnico de Lisboa*.

Mohamed, N., Ahmad, M. H. and Ismail, Z. (2010) Double Seasonal ARIMA Model for Forecasting Load Demand, *Matematika*, 26 (2), pp. 217–231.

Mounce, S. R., Boxall, J. B. and Machell, J. (2010) Development and Verification of an Online Artificial Intelligence System for Detection of Bursts and Other Abnormal Flows, *Journal of Water Resources Planning and Management*, 136 (3), pp. 309–318. DOI:10.1061/(asce)wr.1943-5452.0000030.

Mounce, S. R., Mounce, R. B. and Boxall, J. B. (2011) Novelty detection for time series data analysis in water distribution systems using support vector machines, *Journal of Hydroinformatics*, 13 (4), pp. 672–686. DOI:10.2166/hydro.2010.144.

Puig, V., Ocampo-Martínez, C., Pérez, R., Cembrano, G., Quevedo, J. and Escobet, T. (2017) *Real-time Monitoring and Operational Control of Drinking-Water Systems*, Puig, V., Ocampo-Martínez, C., Pérez, R., Cembrano, G., Quevedo, J., and Escobet, T. (eds.) . Cham: Springer International Publishing. DOI:10.1007/978-3-319-50751-4.

Quevedo, J., Puig, V., Cembrano, G., Blanch, J., Aguilar, J., Saporta, D., Benito, G., Hedo, M. and Molina, A. (2010a) Validation and reconstruction of flow meter data in the Barcelona water distribution network, *Control Engineering Practice*, 18 (6), pp. 640–651. DOI:10.1016/j.conengprac.2010.03.003.

Quevedo, J., Puig, V., Cembrano, G., Blanch, J., Aguilar, J., Saporta, D., Benito, G., Hedo, M. and Molina, A. (2010b) Validation and reconstruction of flow meter data in the Barcelona water distribution network, *Control Engineering Practice*, 18 (6), pp. 640–651. DOI:10.1016/j.conengprac.2010.03.003.

Romano, M. and Kapelan, Z. (2014) Adaptive water demand forecasting for near real-time

management of smart water distribution systems, *Environmental Modelling and Software*, 60, pp. 265–276. DOI:10.1016/j.envsoft.2014.06.016.

Schwarz, G. (1978) Estimating the Dimension of a Model, *The Annals of Statistics*, 6 (2), pp. 461–464. DOI:10.1214/aos/1176344136.

Smola, A. J. and Scholkopf, B. (2004) A tutorial on support vector regression, pp. 199–222. DOI:10.1023/B:STCO.0000035301.49549.88.

Spyros, M., Steven C, W. and Rob J, H. (1997) *Forecasting: Methods and Applications*, *Forecasting methods and applications*. 3rd ed. Wiley.

Taylor, J. W. (2003) Short-term electricity demand forecasting using double seasonal exponential smoothing, *Journal of the Operational Research Society*, 54 (8), pp. 799–805. DOI:10.1057/palgrave.jors.2601589.

Taylor, J. W., de Menezes, L. M. and McSharry, P. E. (2006) A comparison of univariate methods for forecasting electricity demand up to a day ahead, *International Journal of Forecasting*, 22 (1), pp. 1–16. DOI:10.1016/j.ijforecast.2005.06.006.

Val, H. (2013) *Evolução do Sistema de Telegestão da Empresa Evolução do Sistema de Telegestão da Empresa Portuguesa das Águas Livres , S . A .* Instituto Superior de Engenharia de Lisboa.

Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory*. 1st ed. Springer International Publishing.

Xu, W., Zhou, X., Xin, K., Boxall, J., Yan, H. and Tao, T. (2020) Disturbance Extraction for Burst Detection in Water Distribution Networks Using Pressure Measurements, *Water Resources Research*, 56 (5), pp. 1–17. DOI:10.1029/2019WR025526.

Zhou, S. L., McMahon, T. A., Walton, A. and Lewis, J. (2002) Prolegomenon on theory and applications of tables of marks, *Match*, 259 (46), pp. 7–23.

ANEXO I

Comparison of reconstruction methods for water supply systems flow rate time series

Carlos Ascensão [1[0000-0002-3227-890X]], Bruno Ferreira [1[0000-0002-2863-7949]], Raquel Barreira [1[0000-0002-8326-1593]], Nelson Carriço [1[0000-0002-2474-7665]]

[1] INCITE, Barreiro School of Technology, Polytechnic Institute of Setúbal, Setúbal, Portugal
carlosfpascensao@gmail.com

Abstract. The purpose of this paper is to compare the performance of five univariate models for the reconstruction of flow rate time series. Errors in the measurements may occur due to problems in the sensor or in the communication system with data logger, thus generating missing data in the flow rate time series. The presence of missing values in flow rate data restricts its use in network operation processes. The performance of seasonal ARIMA, Standard and double seasonality Holt-Winters, and original and improved Quevedo approach are assessed. The analysis is made considering a real Portuguese case study and 1-month of flow rate data at 1-hour and 10-minute period. The holidays compared to the weekdays show great differences in consumption patterns. For this reason, the effect of forecasting holidays is assessed. Obtained results evidence that the improved Quevedo model can cope with different time step intervals and type of day being forecasted, with a reduced computation time.

Keywords: Flow rate, forecasting, reconstruction methods, time series, water supply systems.

1 Introduction

Flow rate monitoring is an increasingly recurring practice in water utilities, due to the larger accessibility and availability of telemetry equipment and remote management systems. The stored time series data can be used for many tasks in operating and monitoring systems (e.g., demand forecasting, burst detection). The measured data are acquired by sensors and stored in the data logger, which communicates remotely to the management system [1]. Errors in the measurements can happen and may be caused by problems in the sensor or in the communication system with the data logger due to power failures, storage limitations or working outside the operational range generate the missing data in time series [2].

The treatment of flow rate time series is a challenging task for water utilities. The validation processes are based on simple heuristics. Usually non-validated data is replaced using reconstruction methods by predicting the measures with multivariate or univariate statistical models for flow rate time series [3]. More advanced techniques

such as machine learning may be applied to forecast water demand in water supply systems which consider air temperature, precipitation, and flow rate. However, multivariate models that require many variables represent a greater challenge in their application and data acquisition, thus making the operationalization of the models a hard task. Also, the application of advanced techniques, in terms of system monitoring, requires real-time operation and multivariate models do not provide good results when applied in real-time [1]. For these reasons water utilities search for simple forecasting models with low difficulty and complexity in required data, model application and operationalization.

Flow rate time series may show great evidence of daily and weekly cycles that must be considered by forecasting models [4]. Literature shows that autoregressive models and exponential smoothing models with components and parameters that express seasonality are able to obtain reasonable results [1, 9]. Caiado [6] assessed the performance of three different univariate models for water demand forecasting, namely, Holt-Winters, ARIMA and GARCH model and results suggest that all the univariate time series models can be quite useful for short-term forecasting. Quevedo *et al*. [2] developed a short-term forecasting methodology with the purpose of reconstructing missing data in water supply systems. This methodology considers an aggregate daily flow model based on ARIMA models and a 10-minute model based on distributing the daily flow using a 10-minute demand pattern. Cugueró-Escofet *et al*. [7] applied a methodology for reconstruction of missing flow rate data using a double seasonal Holt-Winters.

This paper compares the performance of five forecasting models, namely, a seasonal ARIMA, a seasonal and double seasonal Holt-Winters, a method proposed in Quevedo *et al.* [2] and our improvement of Quevedo approach. The performance assessment is carried out for a complete forecasted day using the root-mean-squared-error (RMSE) and was applied to a real Portuguese case study.

2 Reconstruction methods

2.1 Seasonal ARIMA

The ARIMA models are derived from the family of Auto Regressive Moving Average (ARMA) models. Their difference is the integration component that allows differentiating the series to be possible to apply to non-stationary time series. In order to forecast, ARIMA models use a polynomial of the previous values together with the previous prediction errors. Seasonal ARIMA models considers an additional polynomial for the seasonal component [1].

The function of the ARIMA models can be represented by the degrees of the model (p,d,q), where *p* represents the number of autoregressive terms, *d* represents the number of differentiations and *q* the number of lagged forecast errors in the prediction equation. The polynomial function dedicated to the seasonal component works only with a periodicity. Similar to the polynomial of the regular component it can also be represented by (P,D,Q)s where P, D and Q represent the degrees of the model and *s* represents the number of seasonal periods.

The degrees of the seasonal ARIMA model (p,d,q)(P,D,Q)s can be selected based on the Bayesian information criterion and assessing the fitted values. For the current study, and since 1-hour and 10-minute time intervals are considered, we used the seasonal ARIMA parameters $(2,0,0)(2,0,0)_{24}$ and $(1,0,0)(1,0,1)_{144}$, respectively.

2.2 Quevedo approach

The present approach shows the implementation and improvements to the model for the reconstruction of time series data from water supply systems presented in Quevedo *et al.* [8].

The procedure for reconstructing missing data consists of two modules. The first module gives the prediction of aggregated daily flow based on the seasonal ARIMA models, denominated as the aggregated daily flow model. This model requires a historic record of daily aggregate flow to be able to predict the daily volume taking advantage of the main components of the ARIMA model. For the selection of the polynomial degrees (p,d,q)(P,D,Q)s it was based on the Bayesian information criterion [9] evaluating the set of models generated by $0\leq p\leq 3$, $0\leq P\leq 1$, $0\leq q\leq 3$ e $0\leq Q\leq 1$.

The second module determines a set of flow distribution patterns at 10-minute intervals, consisting of 144 average flow rate values for each pattern. Distribution patterns consider the variation in measurements between weekdays and weekends. For this reason, patterns must be determined for the days of the week (Monday to Friday), and of the weekend (Saturdays and Sundays). Holidays should also be considered for the impact they have on the analysis. Quevedo *et al.* [8] determined that consumption habits for holidays are the same as on Sundays. However, by considering ARIMA as the aggregate daily flow forecast model it is not possible to take into account the effect of a holiday during a weekday.

Improvements to the Quevedo approach are proposed to estimate the daily aggregate flow of a holiday. When initializing the model, it is verified that the date to be forecasted is within the subset with holiday date. If the date does not coincide, the model runs according to the initial approach and estimates the aggregate daily flow with the seasonal ARIMA model. If the date coincides, the aggregate daily flow estimation starts with a simple exponential smoothing model, for which the input is a subset with the aggregate daily flow for Sundays. Estimation for the aggregate daily flow of a holiday is carried as a new Sunday.

To reconstruct flow rate time series, Quevedo *et al.* [8] distributes the aggregate daily flow estimate by the distribution pattern of the day to be reconstructed.

2.3 Exponential smoothing

Holt Winters method is an exponential smoothing method considering trend and seasonal components [10]. In the urban water sector, exponential smoothing methods are well known and have been used in automatic forecasting models [1]. The main characteristic is its simplicity, considering it can be optimized only with the least squares.

Holt-Winters method is based on level, trend and seasonality [11] and models can be divided into two versions based on seasonality patterns, namely, additive or

multiplicative seasonality. Depending on the type of seasonal pattern presented in the data, one of the reference versions can be chosen [12]. In additive seasonality, the difference in seasonal fluctuation between successive is constant, while in multiplicative seasonality the variation is a percentage [12]. In this article, only the Holt Winters models with multiplicative seasonality are considered.

Standard Holt-Winters

Initial values of the components (i.e., level, trend and a seasonal index) are required to start the Standard Holt-Winters multiplicative seasonality model. According to [10] initial level is obtained by averaging the observations from the first seasonal period. The estimated initial trend uses a moving average of the first seasonal period and the seasonal indices are estimated using the average of first seasonality period.

The model components are based on three smoothing parameters: α, β and δ. These parameters represent the level parameter, trend parameter, the seasonal parameter for the seasonal cycle (daily in our case), respectively. These parameters can be estimated by minimizing the RMSE and are usually restricted to lie between 0 and 1.

Double seasonal Holt-Winters

The double seasonal Holt-Winters accommodates two seasonal periods, in this case daily and weekly. To consider the effect of weekly seasonality, the double seasonal Holt-Winters model requires one more seasonality component than the standard model.

The double model requires initial values for initializing – trend, level, daily seasonality, and weekly seasonality index. Taylor [13] formulation was used to represent the initial values, using a s1-period cycle for the daily seasonality and s2-period cycle for weekly seasonality. According to Taylor [13] the initial trend, was chosen as the average of (1) 1/s2 of the difference between the mean of the first s2 and second s2 observations and (2) the average of the first differences for the first s2 observations. The initial level was chosen as the mean of the first two s2 observations minus s2 and half times the initial trend. The initial values for the daily seasonal index are defined by the average of the ratios of actual observation to s1-point centered moving average, taken from the corresponding half-hour period in each of the first 7 days of the time series. The initial values for the weekly seasonal index were set as the average of the ratios of actual observation to s2-point centered moving average, taken from the corresponding half-hour period on the same day of the week in each of the first 2 weeks of the demand series, divided by the initial value of the smoothed within-day seasonal index.

The model components are based on four smoothing parameters α, β, Υ and δ. The first three parameters are similar to the ones previously presented for standard Holt Winters. In addition, a seasonal parameter for bigger seasonal cycle (weekly in our case) is considered. Similarly, these parameters can be estimated by minimizing the RMSE and are usually restricted to lie between 0 and 1.

2.4 Performance assessments

A complete day of missing data (i.e., long gap duration) was assumed to assess the performance of the presented models. For this reason, models will be trained with historical data with a duration no smaller than a month. The last day will be forecasted using the trained models to assess its performance.

The different forecasting models' parameters were calibrated by minimizing the RMSE of the fitted model. The performance of the models' predictions was assessed using the RMSE between real and predicted measurement.

3 Applications and discussion

In this section, an analysis to the performance of the five flow rate time series reconstruction models is carried.

The flow rate time series of the case study is collected using an impulse flow rate meter installed at the inlet of the water distribution network of a residential area with up to 3,000 inhabitants and was provided by a water utility located in Lisbon metropolitan area. This case study considers a 1-month of historical flow rate data recorded in intervals of 1-hour and 10-minute. As such, the model with 1-hour intervals will predict 24-steps and the model with 10-minutes intervals the model will predict 144-steps.

Initially, forecasts were carried considering a weekday and 1-hour intervals. Figure 1 presents the obtained results for each model as well as the real flow rate data.

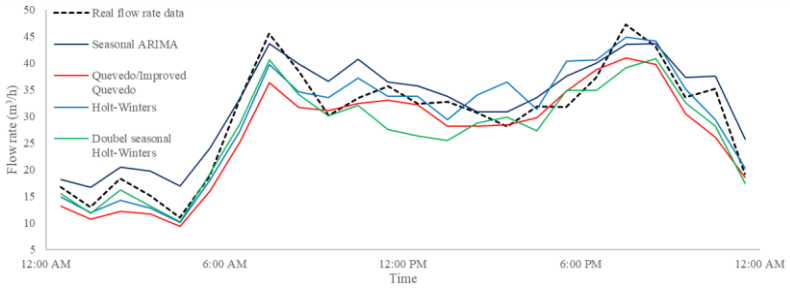

Fig. 1. Comparison for five forecasting techniques considering weekday with hourly intervals.

All models performed reasonably well predicting a day in the series at 1-hour intervals for all models. The seasonal ARIMA model presented the best values (RMSE=3,78), followed by the Holt-Winters (RMSE=3,89), double seasonal Holt-Winters (RMSE=4,09), and lastly the Quevedo (RMSE=4.41). The computation time of each model was assessed, and it was concluded that all models were relatively fast (i.e., less than 40 seconds). Table 1 shows a summary of computational time (in seconds) and the RMSE obtained for all models.

Time series with 1-hour intervals have a very limited use in real-time water supply systems operation. Application of advanced techniques, such as machine learning requires time series with shorter time intervals to operate water supply system in real-

time. As such, the same day of the week is predicted with the series of 10-minute intervals. The obtained results for each model are presented in Figure 2 and Table 1.

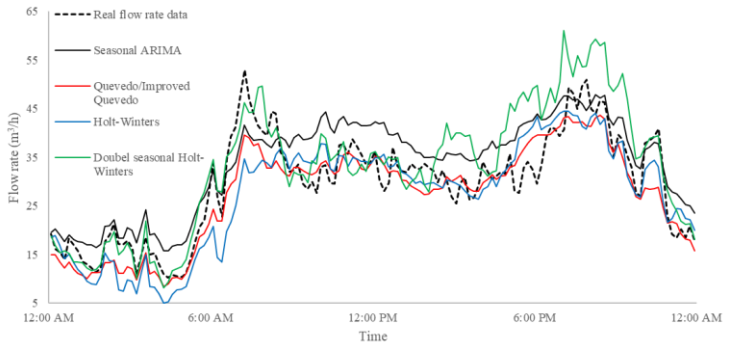

Fig. 2. Comparison for five forecasting techniques considering weekday (10-minute intervals).

All models predicted reasonably well, with the Quevedo presenting the best overall results (RMSE=4,69), followed by Holt-Winters (RMSE=5,97), seasonal ARIMA (RMSE=6,10) and lastly double seasonal Holt-Winters (RMSE=6.76). Nonetheless, the difference in computation time amongst methods is quite significant, with the Quevedo (t=1s), Holt-Winters (t=108s), Double seasonal Holt-Winters (t=286s) and lastly seasonal ARIMA (t=591s).

Problems may arise when forecasting holidays during weekdays since the expected behavior of the water demand of a holiday is usually related to the behavior of the Sundays (in opposition to the behavior of weekday) [14]. As such, a holiday during a weekday is predicted with the series of 10-minute intervals. In addition to the four methods already compared, a fifth is considered with the Quevedo improvement. The obtained results for each model are presented in Figure 3 and Table 1.

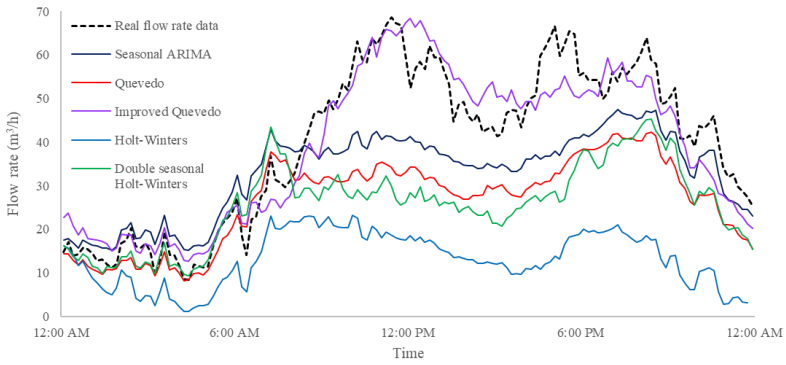

Fig. 3. Comparison for five forecasting techniques considering holiday with 10-minute intervals.

Based on Figure 3 and Table 1 it is possible to conclude that, overall, the improved Quevedo performed reasonably well (RMSE=5,84). The remaining models failed to capture the holiday variations. The autoregressive and exponential smoothing models make predictions based on seasonality periods and cannot "look" outside the seasonality periods. Similarly, the original Quevedo approach forecast the aggregate daily flow using an autoregressive model (i.e., seasonal ARIMA).

Table 1. Comparison of model's performance and computational time

Models	Weekday (1-hour)		Weekday (10-minute)		Holiday (10-minute)	
	RMSE	Computation time (s)	RMSE	Computation time (s)	RMSE	Computation time (s)
Seasonal ARIMA	3.78	12	6.10	591	12.50	624
Quevedo	4.41	1	4.69	1	16.91	1
Holt-Winters	3.89	38	5.97	108	30.53	101
Double seasonal Holt-Winters	4.09	12	6.76	286	19.21	284
Improved Quevedo	4.41	1	4.69	1	5.84	1

Conclusions

Missing data from flow rate time series resulting from a validation process must be reconstructed to apply advanced techniques that requires validated data. Usually, the reconstruction of the flow rate time series is performed by forecasting models. This paper presents a comparison of five methods to the reconstruction of flow rate time series, namely, the seasonal ARIMA, the standard and double seasonal Holt-Winters, the original and improved Quevedo approach. The comparison is based on 1-month of historical flow rate time series at 1-hour and 10-minute intervals of a real Portuguese case study. A complete day was forecasted and analyzed for weekdays and holidays. In weekdays, forecasts with 1-hour intervals obtained a reasonable RMSE result for all methods. Similarly, reasonable RMSE results were obtained for all models considering weekdays with the 10-minute intervals. Nonetheless, great difference in computation time were observed amongst methods. On the other hand, and when forecasting a holiday, only the improved Quevedo approach produced reliable results.

Future research may include the forecast of flow rate data when in absence of reliable historical data, for instance, due to changes in patterns motived by recent lockdowns.

Acknowledgement

The authors want to acknowledge Fundação para a Ciência e a Tecnologia, (grant number DSAIPA/DS/0089/2018) through the Data Science and Artificial Intelligence in Public Administration Programme for supporting WISDom project. The authors also acknowledge the participating water utilities for their contribution.

References

1. Puig V., Ocampo-Martínez C., Pérez R., Cembrano G., Quevedo J., and Escobet T.: *Real-time Monitoring and Operational Control of Drinking-Water Systems*. Springer International Publishing, Cham (2017).
2. Quevedo J., Puig V., Cembrano G., Blanch J., Aguilar J., Saporta D., Benito G., Hedo M. and Molina A.: Validation and reconstruction of flow meter data in the Barcelona water distribution network. *Control Eng. Pract. 6(18)*, 640–651 (2010).
3. Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B. & Sliusarieva, A.: Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resour. Res.* 48, 1–14 (2012).
4. De Marinis, G., Gargano, R. & Tricarico, C.: Water demand models for a small number of users. In: *8th Annu. Water Distrib. Syst. Anal. Symp* (2007)
5. Taylor, J. W., de Menezes, L. M. & McSharry, P. E.: A comparison of univariate methods for forecasting electricity demand up to a day ahead. *Int. J. Forecast.* 22, 1–16 (2006).
6. Caiado, J.: Performance of combined double seasonal univariate time series models for forecasting water demand. *J. Hydrol. Eng.* 15, 215–222 (2010).
7. Cugueró-Escofet M., García D., Quevedo J., Puig V., Espin S., and Roquet J.: A methodology and a software tool for sensor data validation/reconstruction: Application to the Catalonia regional water network, *Control Eng. Pract.*, vol. 49, pp. 159–172, (2016)
8. Quevedo J., Puig V., Cembrano G., Blanch J., Aguilar J., Saporta D., Benito G., Hedo M., Molina A.: Validation and reconstruction of flow meter data in the Barcelona water distribution network. *Control Eng. Pract.* 18, 640–651 (2010).
9. Schwarz, G.: Estimating the Dimension of a Model. *Ann. Stat.* 6, 461–464 (1978).
10. Spyros, M., Steven C, W. & Rob J, H.: *Forecasting: Methods and Applications*. Wiley, (1997).
11. Hyndman, R. J., Koehler, A. B., Ord, J. K. & Snyder, R. D. *Springer Series in Statistics Forecasting with Exponential Smoothing*. (2008).
12. Galvas G.,: Time series forecasting used for real-time anomaly detection on websites. Vrije Universiteit, Amsterdam (2016).
13. Taylor, J. W.: Short-term electricity demand forecasting using double seasonal exponential smoothing. *J. Oper. Res. Soc.* 54, 799–805 (2003).