



UNIVERSIDADE
LUSÓFONA

Aplicação de técnicas de análise de dados
para deteção de fugas e roturas em sistemas
de distribuição de água

Trabalho Final de curso

Relatório Intercalar 1º Semestre

Nome do Aluno: Wilson Silva

Nome do Orientador: Maria Almeida Silva

Nome do Orientador: Dália Loureiro

Trabalho Final de Curso | LEI | 01/12/2024

www.ulusofona.pt

Direitos de cópia

(Aplicação de técnicas de análise de dados para deteção de fugas e roturas em sistemas de distribuição de água), Copyright de (*Wilson Silva*), Universidade Lusófona.

A Escola de Comunicação, Arquitetura, Artes e Tecnologias da Informação (ECATI) e a Universidade Lusófona (UL) têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

Resumo

A água é um elemento fundamental à vida e, com as alterações climáticas existe cada vez mais escassez deste elemento, sendo necessário controlar o uso dela. Nos sistemas de distribuição de água, a redução de perdas de água é essencial para garantir a disponibilidade e gestão sustentável da água. A criação de zonas de medição e controlo (ZMC) é uma das medidas mais eficazes para controlo de perdas de reais e aparentes. Para um controlo adequado das perdas de água em sistemas de distribuição de água potável, estes devem encontrar-se divididos em ZMC, sendo que, nestas zonas, todos os pontos de entrada e saída de água devem ser medidos.

Propõe-se neste trabalho uma metodologia para processamento dos dados de caudal medidos em ZMC, que permita detetar fugas e roturas em sistemas de distribuição de água. Existem diversos artigos e teses sobre perdas e roturas de água e também algumas metodologias para análise de dados de caudal. Contudo, em geral, estas são focadas apenas num problema das séries de caudal, não os analisando de uma forma mais global, tal como se pretende neste estudo. É um trabalho pertinente tendo em conta os dias de hoje, onde as alterações climáticas são cada vez mais visíveis e quando em Portugal se vê cada vez mais escassez de água. É então essencial dispor de dados fiáveis para identificar rapidamente fugas e roturas, solucionando estes problemas, evitando grandes perdas de água e custos adicionais associados.

Abstract

Water is a fundamental element for life, and with climate changes, there is increasingly more scarcity of this resource. It is necessary to control its use. In water distribution systems, reducing water losses is essential to ensure the availability and sustainable management of water. The creation of measurement and control zones (ZMC) is one of the most effective measures for controlling real and apparent losses. For proper control of water losses in potable water distribution systems, they should be divided into ZMCs, where all water inlet and outlet points must be measured.

A methodology for processing flow data measured in ZMC is proposed in this study, which allows for the detection of leaks and breaks in water distribution systems. There are various articles and theses on water losses and breaks, as well as some methodologies for flow data analysis. However, in general, these focus only on a specific flow series problem, rather than analyzing them in a more comprehensive manner as intended in this study. This work is relevant considering today's climate change challenges, where water scarcity is increasingly visible in Portugal. It is essential to have reliable data to quickly identify leaks and breaks, thereby solving these issues and avoiding significant water losses and associated costs.

Conteúdo

Resumo.....	iv
Abstract	v
Conteúdo.....	vi
Lista de Figuras	vii
Lista de Tabelas	viii
1 Introdução	1
1.1 Enquadramento	1
1.1.1 Motivação e Identificação do Problema.....	1
1.1.2 Objetivos.....	2
1.1.3 Estrutura do Documento	2
2 Viabilidade e Pertinência.....	3
2.1 Pertinência	3
2.2 Viabilidade	3
3 Conceitos Fundamentais	4
3.1 Conceitos Teóricos.....	4
3.1.1 Séries Temporais	4
3.1.2 Dados Anómalos (Outliers).....	4
3.1.3 Eventos Anómalos	4
3.2 Modelos e Algoritmos Relevantes	4
3.2.1 Algoritmos de fácil implementação.....	4
3.2.2 Algoritmos de implementação mais avançada	5
3.3 Tecnologias e Ferramentas Utilizadas	5
4 Estado da Arte	7
5 Solução a Desenvolver	8
5.1 Introdução	8
5.2 Metodologia.....	8
5.3 Abrangência	8
Bibliografia	10
Glossário.....	11

Lista de Figuras

Figura 1 - Água Potável e Saneamento [16]	3
Figura 2 - Gráfico de Gantt [17].....	9

Lista de Tabelas

1 Introdução

1.1 Enquadramento

Este projeto foi apresentado pelo Laboratório Nacional de Engenharia Civil (LNEC), um instituto público de ciência e tecnologia (C&T), com o estatuto de Laboratório do Estado, que desenvolve investigação em todos os domínios da Engenharia Civil, o que lhe confere uma perspetiva única e multidisciplinar.

Os sistemas de distribuição de água são meios de transporte de água desde a origem, as estações de tratamento de água (ETA), até ao consumidor final, antes do contador dos clientes, e estes devem assegurar uma qualidade de serviço adequada em termos de eficiência e eficácia.

A adequada gestão das perdas de água constitui um dos principais desafios das entidades gestoras de serviços de abastecimento de água. Estas podem ser classificadas em:

- perdas reais (inclui todas as perdas físicas de água na rede de abastecimento);
- perdas aparentes (inclui os usos não autorizados e erros de medição dos contadores).

Por sua vez, os sistemas de distribuição de água podem estar divididos em zonas de medição e controlo (ZMC). A criação de ZMC é uma das medidas mais eficazes para controlo de perdas de reais e aparentes. Estas zonas têm todos os pontos de entrada e saída com medidores de caudal, de modo que se sabe a quantidade de água que passa por cada ponto, sabendo assim o caudal. Estas zonas integram a tecnologia de telemedição que permite recolher dados sobre o caudal da água em curtos espaços de tempo, por exemplo, 5, 10, 15 minutos entre recolhas.

1.1.1 Motivação e Identificação do Problema

A motivação principal do trabalho é melhorar a operação e gestão dos sistemas de distribuição de água através da deteção de anomalias em medições de caudal. Isso é fundamental para identificar ruturas de condutas, ramais e acessórios ou consumos atípicos, garantindo dados históricos fiáveis para o planeamento e conceção desses sistemas. Controlar as perdas de água é crucial para a sustentabilidade económica e ambiental, reduzindo a necessidade de novas fontes de água e reduzindo o consumo de energia. Além disso, a metodologia proposta visa ser flexível e aplicável a diferentes tipos de eventos e séries temporais de fluxo, contribuindo para a eficiência e eficácia na gestão dos recursos hídricos.

O problema da escassez de água, agravado pelas alterações climáticas, torna ainda mais urgente a adoção de soluções tecnológicas que otimizem o uso desse recurso vital. Estes eventos anómalos representam desafios significativos à gestão técnica das entidades gestoras de sistemas de abastecimento de água. Estes eventos podem ter um impacto significativo no desperdício de água tratada, aumento de custos operacionais e na qualidade do serviço de abastecimento de água. Tecnologias emergentes, como a telemedição das redes e novos equipamentos para medição de caudal, técnicas estatísticas e de *machine-learning*, oferecem soluções inovadoras para identificar de forma mais célere estes eventos anómalos e otimizar a gestão técnica destes sistemas.

1.1.2 Objetivos

O principal objetivo deste projeto será detetar eventos anómalos em sistemas de distribuição de água, através de dados de caudal referentes a zonas de medição e controlo.

1.1.3 Estrutura do Documento

O presente relatório está organizado da seguinte forma:

Na Secção 1 é indicado o enquadramento deste trabalho.

Na Secção 2 é discutida a pertinência e viabilidade do trabalho proposto.

Na Secção 3 são explicados alguns conceitos fundamentais para a interpretação do trabalho.

Na Secção 4 descreve estudos e trabalhos feitos na área abrangida por este TFC.

Na Secção 5 é apresentada a solução a desenvolver.

Na Secção 6 é exposto o planeamento inicial.

2 Viabilidade e Pertinência

2.1 Pertinência

Sendo a água potável um bem essencial em escassez, deve ser dado foco à redução do seu desperdício atualmente. Neste projeto, os dados fornecidos pelo Laboratório Nacional de Engenharia Civil irão ser analisados para verificar a água que entra na rede de distribuição e, assim, utilizando técnicas de análise de dados, poder-se verificar se existe alguma rotura na tubagem da rede ou se existem consumos atípicos (e.g., furtos de água significativos em que água é ilegalmente utilizada pelas pessoas e não é faturada).

A vantagem de poder ser feita esta análise é que podemos reduzir a quantidade de recursos essenciais em escassez que são desperdiçados. Com esta análise, caso seja necessária uma reparação, poderá ser mais fácil determinar a sua localização e enviar uma equipa para o local correto, reduzindo os custos de manutenção. Além disso, se houver um furto poderá ser mais fácil determinar a sua localização e enviar uma equipa para fiscalizar o lugar.

2.2 Viabilidade



A viabilidade deste projeto é garantida por ter sido proposto pelo Laboratório Nacional de Engenharia Civil e também porque alinha com uma das metas identificadas nos Objetivos de Desenvolvimento Sustentável, Água Potável e Saneamento. Sendo bem-sucedido trará poupanças às entidades gestoras, porque conseguem localizar problemas nos sistemas de distribuição de água, e trará lucro, pois permite faturar água que anteriormente não era faturada.

Figura 1 - Água Potável e Saneamento [16]

3 Conceitos Fundamentais

3.1 Conceitos Teóricos

3.1.1 Séries Temporais

As séries temporais são um conjunto de observações feitas ao longo do tempo. Neste tipo de análise, a variável fundamental é o tempo, o que permite verificar se as observações vizinhas são dependentes ou não.

As séries temporais podem ser usadas para prever o futuro com base nos dados recolhidos anteriormente.

3.1.2 Dados Anómalos (*Outliers*)

Dados anómalos são valores num conjunto de dados que são muito diferentes dos restantes valores observados. Estes podem ser maiores ou menores do que a maioria dos dados observados.

Um problema na medição e comunicação de dados através de sensores ou uma inserção de dados incorretos pode dar origem a dados anómalos.

3.1.3 Eventos Anómalos

Eventos anómalos são um conjunto de dados que indicam um comportamento irregular ao longo do período. Por exemplo, uma rotura nos sistemas de distribuição de água ou eventos de grande superfície como espetáculos, concertos e festivais pode originar um evento anómalo numa série de caudal.

3.2 Modelos e Algoritmos Relevantes

3.2.1 Algoritmos de fácil implementação

Lista-se a seguir os métodos mais comuns e de fácil implementação para deteção de dados anómalos.

- **Método de Deteção de Anomalias de Grubbs:**
É um método estatístico utilizado para identificar dados anómalos num conjunto de dados que segue uma distribuição normal.
Este teste é útil para detetar um único dado anómalo em dados amostrais e pode ser aplicado repetidamente para detetar múltiplos dados anómalos. [1]
- **Método de Deteção de Anomalias de Tukey:**
O método de deteção de anomalias de Tukey, também conhecido como Boxplot de Tukey ou Teste do Bigode, é uma técnica estatística simples e eficaz para identificar *outliers* em conjuntos de dados. Este método utiliza a distribuição dos dados para determinar limites além dos quais os dados são considerados anómalos. [2]
- **Método Z-Score:**

É uma técnica estatística utilizada para medir o quão longe um valor está da média em relação ao desvio padrão.

Calcula-se para cada valor do caudal. Valores com Z-score absolutos maiores que 3 são considerados dados anómalos. [3]

3.2.2 Algoritmos de implementação mais avançada

Lista-se a seguir os métodos que recorrem a técnicas de *machine-learning* para detecção de dados anómalos. Incluem-se também métodos clássicos baseados na análise estatística de dados (e.g. Modelos ARIMA).

- **Modelo ARIMA (AutoRegressive Integrated Moving Average):**
É o modelo mais usado para prever dados de séries temporais. Ele é especialmente útil para dados que mostram tendências e sazonalidades.
Trata-se de um modelo de séries temporais que pode ser utilizado para detetar dados anómalos através da análise de resíduos. [4]
- **Modelo One-Class SVM (Support Vector Machine):**
O One-Class SVM (Support Vector Machine) é uma variante do SVM que é usada principalmente para detecção de dados anómalos.
É utilizado para identificar dados anómalos em dados complexos, treinando um modelo de SVM para reconhecer a "classe" normal dos dados, ou seja, conjunto de dados em que se conheça que não existam dados anómalos. Este modelo não é aplicável com séries temporais, mas pode ser adaptado os dados em pequenos segmentos. [5]
- **Modelo Complex Event Processing (CEP):**
É uma técnica de análise de dados em tempo real que permite identificar padrões, tendências e relações dentro de fluxos de eventos, esta pode ser usada com séries temporais. [6]
- **Modelo Isolation Forest:**
É um modelo usado principalmente para a detecção de anomalias. Este modelo é eficaz na identificação de dados anómalos em conjuntos de dados grandes e complexos, mas para séries temporais é necessário dividir os dados em segmentos para utilizar este modelo na forma mais eficiente. [7]
- **Modelo Stream Mining:**
Técnicas para análise de dados em fluxo contínuo, são úteis em aplicações onde o fluxo de entrada de dados é contínuo e devem ser processados rapidamente. Utilizado para detecção de anomalias em tempo real. Ideal para séries temporais. [8]

3.3 Tecnologias e Ferramentas Utilizadas

Para desenvolver este projeto vai ser utilizada linguagem de programação Python que é uma linguagem muito usada para análise de dados. A mesma contém diversas bibliotecas que iremos

usar para manipulação e visualização de dados como por exemplo: Matplot, seaborn, Matplot, pandas e numpy.

Eventualmente iremos utilizar também a linguagem de Programação R que é uma linguagem usada para análise estatística, análise de gráficos e também análise de dados.

4 Estado da Arte

O problema da escassez de água, agravado pelas alterações climáticas, torna ainda mais urgente a adoção de soluções tecnológicas que otimizem o uso desse recurso vital. Estes eventos anómalos representam desafios significativos à gestão técnica das entidades gestoras de sistemas de abastecimento de água. Por isso em 2010 começou-se a abordar a questão de como gerir as fugas e como as classificar, como detetá-las, como localizá-las e como controlá-las, porque as perdas de água no sistema de distribuição, sem serem controladas, podem resultar numa grande perda económica. Para melhorar foi proposta a criação de modelos em tempo real que integram estes métodos de classificação de fugas. [9]

Por volta de 2015 e 2021, foi feita uma melhoria de gestão e operação dos sistemas de distribuição de água através da deteção de anómalos em medição de fluxo desses sistemas, o que pode indicar uma rutura de tubagem no sistema ou consumos incomuns como festivais, eventos, entre outros. [10][11]

Mais tarde, em 2017, foi proposto o melhoramento na classificação de fugas e foi feita uma revisão dos métodos a utilizar para tal classificação. Os métodos estatísticos têm por base só estatística para determinar se existe alguma fuga. Os métodos de previsão-classificação usam dados obtidos anteriormente para verificar se existe alguma fuga e os métodos de classificação utilizam redes neuronais para identificar fugas. [12]

Por volta de 2019, foi feita uma comparação nas técnicas/métodos utilizados para a deteção de fugas nos sistemas de distribuição de água onde foi concluído que análise probabilística bayesiana é melhor quando poucos dados de medição de caudal estão disponíveis e que SVM e redes neurais artificiais (ANN) têm dificuldades em detetar fugas só quando existe dados de medição de caudal. [13]

Em meados de 2022, foi publicado um estudo onde era possível detetar fugas em tempo real usando apenas emissões acústicas da tubagem dos sistemas de distribuição de água. Essas emissões eram capturadas por dispositivos colados ao longo do sistema e a seguir os sinais capturados eram processados por algoritmos de *machine-learning*. O estudo foi muito promissor e estável conseguindo detetar falhas nos sistemas de distribuição de água. [14]

No início de 2023, foi publicado um estudo que pretende identificar alterações no caudal/pressão na rede para identificar fugas usando *machine-learning*. [15]

5 Solução a Desenvolver

5.1 Introdução

Nesta secção será apresentada uma descrição de uma metodologia ser desenvolvida com os vários passos propostos. Também são identificadas as unidades curriculares importantes para o desenvolvimento do trabalho.

5.2 Metodologia

A metodologia a desenvolver tem os seguintes passos:

- Processamento de dados
 - Recolha de dados
 - Análise exploratória dos dados
 - Normalização
 - Preenchimento de Falhas
 - Identificação de eventos anómalos
- Análise Exploratória Avançada dos Dados
 - Identificação de padrões de consumo, identificação de sazonalidades
- Aplicação de Métodos
 - Método Tukey (Boxplot),
 - Possibilidades de utilização de outros dados para além do caudal a entrar no sistema (como temperatura)

5.3 Abrangência

Data Science – Os conhecimentos usados desta unidade curricular serão a utilização da linguagem Python, como também as bibliotecas Matplotlib para mostrar gráficos, seaborn que é uma biblioteca que corre por cima da biblioteca Matplotlib para fácil criação de gráfico, pandas para analisar os dados, numpy que é uma biblioteca matemática.

6 Método e Planeamento

6.1 Planeamento inicial

Para realizar este relatório temos de definir datas para cada passo da metodologia.

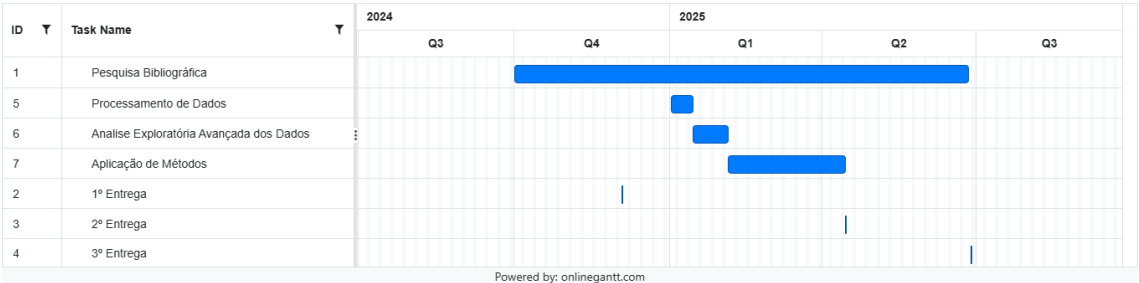


Figura 2 - Gráfico de Gantt [17]

Bibliografia

- [1] Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, 11(1), 1-21.
- [2] Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.
- [3] Iglewicz, B., & Hoaglin, D. C. (1993). *How to Detect and Handle Outliers*. ASQC Quality Press.
- [4] Box, G. E. P., & Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.
- [5] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7), 1443-1471.
- [6] Luckham, D. C. (2002). *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*. Addison-Wesley Longman Publishing Co., Inc.
- [7] Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In 2008 Eighth IEEE International Conference on Data Mining (pp. 413-422). IEEE.
- [8] Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A Survey on Concept Drift Adaptation. *ACM Computing Surveys*, 46(4), 1-37.
- [9] R. Puust, Z. Kapelan, D. A. Savic & T. Koppel. (2010). A review of methods for leakage management in pipe networks.
- [10] Dália Loureiro, Conceição Amado, André Martins, Diogo Vitorino, Aisha Mamade & Sérgio Teixeira Coelho. (2015). Water distribution systems flow monitoring and anomalous event detection: A practical approach.
- [11] José Carlos Carrasco-Jiménez, Filippo Baldaro e Fernando Cucchiatti. (2021). Detection of Anomalous Patterns in Water Consumption: An Overview of Approaches.
- [12] Yipeng Wu & Shuming Liu. (2017). A review of data-driven approaches for burst detection in water distribution systems.
- [13] J. C. van der Walt, P. S. Heyns & D. N. Wilke. (2019). Pipe network leak detection: comparison between statistical and machine learning techniques.
- [14] Ali Fares, I. A. Tijani, Zhang Rui & Tarek Zayed. (2022). Leak detection in real water distribution networks based on acoustic emission and machine learning.
- [15] Nishant Sourabh, P.V. Timbadiya & P. L. Patel. (2023). Leak detection in water distribution network using machine learning techniques.
- [16] <https://museudoamanha.org.br/sites/default/files/6%20-%20%C3%81gua%20Pot%C3%A1vel%20e%20Saneamento.%20Vetorizada.%20PNG.%20Com%20t%C3%ADtulo.%20Colorida.png>
- [17] <https://www.onlinegantt.com/#/gantt>

Glossário

LNEC	Laboratório de Engenharia Civil
TFC	Trabalho Final de Curso
ZMC	Zonas de Medição e Controlo
ODS	Objetivos de Desenvolvimento Sustentável
Boxplot	Método de Deteção de Anomalias de Tukey
ARIMA	<i>AutoRegressive Integrated Moving Average</i>
SVM	<i>Support Vector Machine</i>
CEP	Complex Event Processing