UNIVERSIDADE
LUSÓFONA

Data Science na Gestão Eficiente de Sistemas de Abastecimento de Água

Trabalho Final de curso

2ª Entrega Intercalar

João Tomás Prata Rocha, 22303390, LCD

Orientador: Maria Almeida Silva

Co-orientador: Dália Loureiro (Laboratório Nacional de Engenharia Civil)

Departamento de Engenharia Informática e Sistemas de Informação

Universidade Lusófona, Centro Universitário de Lisboa

Direitos de cópia

Resumo

O estudo analisa os fatores que influenciam a água não faturada nos sistemas de abastecimento de água, com o foco nas perdas reais que representam a principal componente deste indicador e um desafio relevante para as entidades gestoras. A metodologia inclui processamento e validação dos dados do Relatório Anual dos Serviços de Água e Resíduos de Portugal, análise exploratória para identificação de padrões e relações, avaliação do desempenho global dos sistemas e teste dos métodos propostos para verificar robustez e capacidade preditiva. Os resultados visam aprofundar a compreensão dos determinantes da água não faturada e apoiar a estimativa dos indicadores em entidades com informação incompleta, contribuindo para a identificação de ineficiências estruturais e para a otimização da gestão dos sistemas de abastecimento.

Palavras-chave: Água não faturada; Machine learning; Modelos preditivos; Perdas reais

Abstract

The study examines the factors influencing non-revenue water in drinking water supply systems, with a focus on real losses, which constitute the main component of this indicator and pose a significant challenge for water utilities. The methodology includes data processing and validation of the Annual Report on Water and Waste Services in Portugal, exploratory analysis to identify patterns and relationships, assessment of the overall performance of the systems, and testing of the proposed methods to verify robustness and predictive capability. The results aim to deepen the understanding of the determinants of non-revenue water and support the estimation of indicators in water utilities with incomplete information, thereby contributing to the identification of structural inefficiencies and the optimization of water supply system management.

Keywords: Non-revenue water; Machine learning; Predictive models; Real losses

Índice

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Siglas

ALC	Active Leakage Control
ANN	Artificial Neural Networks
CRLI	Combined Real Loss Index
CARL	Current Annual Real Losses
DMA	District Metered Area
ERSAR	Entidade Reguladora dos Serviços de Águas e Resíduos
ILI	Infrastructure Leakage Index
IWA	International Water Association
IQR	Interquartile Range
LNEC	Laboratório Nacional de Engenharia Civil
NA	Não Aplicável
NR	Não Reportado
ODS	Objetivos de Desenvolvimento Sustentável
RASARP	Relatório Anual dos Serviços de Águas e Resíduos em Portugal
SVM	Support Vector Machines
TFC	Trabalho Final de Curso
UARL	Unavoidable Annual Real Losses

1 Introdução

As perdas reais constituem um dos principais desafios dos sistemas de abastecimento de água em Portugal, representando a maior fatia da água não faturada e um impacto significativo ao nível operacional, financeiro e ambiental. A importância deste tema justifica a necessidade de aprofundar a sua análise, dada a relevância que assume para a sustentabilidade dos serviços de água.

1.1 Enquadramento

O desperdício de água nos sistemas de distribuição constitui um problema de elevada importância, tanto em Portugal como à escala global. No presente Trabalho Final de Curso, será abordada esta temática, procurando destacar a sua relevância para a gestão eficiente dos recursos hídricos. Para este estudo, serão utilizados dados e indicadores recolhidos, calculados e publicados anualmente pela Entidade Reguladora dos Serviços de Água e Resíduos (ERSAR), entidade responsável pela regulação e supervisão dos serviços de abastecimento público de água, saneamento de águas residuais urbanas e gestão de resíduos urbanos em Portugal [7].

Neste contexto, destaca-se o conceito da água não faturada, que corresponde ao volume de água captada, tratada, transportada, armazenada e distribuída, que não chega a ser faturada aos utilizadores [7]. A água não faturada inclui as perdas reais, como fugas e/ou roturas na rede, as perdas aparentes, resultantes de erros de medição e consumos não autorizados, e o consumo autorizado não faturado, como água utilizada em combate a incêndios e limpeza de condutas. Assim, o indicador da água não faturada constitui um parâmetro essencial para avaliar a eficiência dos sistemas de abastecimento.

Deste modo, pretende-se compreender os fatores que influenciam a água não faturada, com especial enfoque nas perdas reais, que tendem a representar a maior fatia da água não faturada e são, por isso, um dos principais desafios do setor em Portugal.

1.2 Motivação e Identificação do Problema

As perdas de água, compostas pelas perdas aparentes e perdas reais, representam uma parcela considerável da água não faturada. Este fenómeno traduz-se não só em custos económicos elevados para as entidades gestoras, mas também em impactos ambientais e sociais relevantes.

O concelho de Lagoa constitui um exemplo concreto desta realidade. Nos últimos anos, têm sido implementadas várias medidas de reabilitação das redes de abastecimento e de monitorização das perdas, que resultaram numa redução de cerca de 18 % nas perdas reais. De acordo com dados divulgados pela autarquia, as perdas diminuíram de aproximadamente 50 % em 2013 para 32,5 % em 2024 [20]. Apesar desta evolução positiva, o valor continua acima da média nacional de 26,9 % registada nos serviços de distribuição de água [19], o que evidencia que o desperdício de água permanece um dos principais desafios do setor.

Assim, compreender os fatores que contribuem para estas perdas é essencial para melhorar a eficiência dos sistemas de abastecimento e garantir uma gestão sustentável deste recurso fundamental, cuja importância é cada vez mais reconhecida tanto a nível local como global.

1.3 Objetivos

Com o intuito de colmatar a lacuna de conhecimento identificada, este estudo pretende aprofundar a compreensão dos fatores que influenciam a água não faturada, com especial enfoque na componente das perdas reais, que constitui o principal desafio técnico-operacional das entidades gestoras. Procura-se identificar padrões, relações e determinantes que condicionam os indicadores de desempenho Água não faturada e Perdas reais. Esta análise deverá ter em consideração os diferentes sistemas de abastecimento de água existentes: sistemas "em alta", correspondentes às infraestruturas a montante da distribuição, desde a captação até à entrada no sistema "em baixa", e sistemas "em baixa", responsáveis pela distribuição final de água e pela prestação do serviço aos utilizadores. Adicionalmente, o estudo visa desenvolver capacidade preditiva, permitindo estimar estes indicadores (Água não faturada e Perdas reais) em entidades gestoras com informação incompleta.

1.4 Estrutura do Documento

O presente relatório está organizado em seis capítulos, complementados por bibliografia e anexos. A sua estrutura é a seguinte:

- Capítulo 1 – Introdução: Apresenta o enquadramento geral do tema, a motivação e identificação do problema em estudo, os objetivos propostos e uma descrição sintética da organização do documento;
- Capítulo 2 – Pertinência e Viabilidade: Discute a relevância do trabalho e analisa a sua viabilidade técnica, científica e operacional;
- Capítulo 3 – Conceitos Fundamentais: Reúne os conceitos teóricos essenciais, bem como as tecnologias e ferramentas relevantes para o desenvolvimento do trabalho;
- Capítulo 4 – Estado da Arte: Analisa trabalhos e metodologias existentes na literatura associadas ao trabalho em estudo;
- Capítulo 5 – Solução Proposta: Descreve a metodologia proposta e os procedimentos de recolha e pré-processamento dos dados. Inclui ainda a definição da abrangência da solução;
- Capítulo 6 – Método e Planeamento: Apresenta o planeamento inicial do projeto.

2 Pertinência e Viabilidade

2.1 Pertinência

A gestão eficiente dos recursos hídricos é atualmente um dos principais desafios a nível global. A escassez de água, agravada pelas alterações climáticas, pelo aumento da procura e pela degradação das infraestruturas, reforça a necessidade de adotar medidas que assegurem a sustentabilidade dos sistemas de abastecimento. Neste contexto, o estudo da água não faturada assume particular relevância, uma vez que representa a diferença entre a água produzida e a que é efetivamente faturada aos consumidores, refletindo as ineficiências existentes na rede de distribuição.

Em Portugal, a Entidade Reguladora dos Serviços de Águas e Resíduos (ERSAR) tem identificado a água não faturada como um dos indicadores mais críticos do desempenho das entidades gestoras. Segundo o Relatório Anual dos Serviços de Águas e Resíduos em Portugal (RASARP), a média nacional de água não faturada nos serviços em baixa é de cerca de 26,9 %, valor que demonstra a dimensão do problema e a necessidade urgente de medidas de mitigação [19].

Assim, o presente trabalho revela-se pertinente por contribuir para a compreensão dos fatores que influenciam a água não faturada, com especial destaque para a componente das perdas reais, que correspondem às fugas e roturas em qualquer ponto da rede. Através da análise e sistematização da informação disponível, pretende-se fornecer uma visão clara sobre o impacto económico, ambiental e social deste fenómeno, promovendo uma maior consciência sobre a importância da eficiência hídrica e incentivando práticas de gestão mais sustentáveis.

2.2 Viabilidade

A viabilidade deste estudo é garantida pela credibilidade institucional e pela qualidade dos dados utilizados. O trabalho foi desenvolvido no âmbito de um pedido do Laboratório Nacional de Engenharia Civil (LNEC), o que reforça a credibilidade técnica e científica do projeto.

Do ponto de vista técnico, a análise baseia-se em dados concretos e atualizados disponibilizados por entidades competentes (ERSAR) e publicados anualmente, permitindo uma abordagem detalhada e fundamentada à problemática da água não faturada e das perdas reais, algo que nem sempre é possível noutros estudos devido à falta de informação fiável.

No plano económico, a viabilidade é sustentada pelo fato de o estudo recorrer à informação já existente, sem necessidade de investimento adicional em infraestruturas ou equipamentos. Além disso, a aplicação dos resultados poderá contribuir para reduções de custos operacionais nas entidades gestoras, através da identificação de medidas mais eficazes de controlo e deteção de perdas.

Em termos sociais e ambientais, o projeto revela-se igualmente viável e necessário, uma vez que promove a sustentabilidade dos recursos hídricos e contribui para uma maior consciencialização pública sobre o desperdício de água. As conclusões esperadas alinham-se diretamente com os Objetivos de Desenvolvimento Sustentável (ODS) definidos pelas Nações Unidas, nomeadamente o ODS 6 – Água potável e saneamento, que visa garantir a disponibilidade e a gestão sustentável da água para todos.

Por fim, a viabilidade global do estudo é reforçada pela possibilidade de continuação e aplicação prática dos resultados pelas entidades gestoras e em futuras investigações, assegurando que o

trabalho desenvolvido não se esgota no âmbito académico, mas que constitui um contributo real e sustentável para o setor da gestão da água em Portugal.

3 Conceitos Fundamentais

3.1 Conceitos Teóricos

3.1.1 Água Não Faturada

A água não faturada corresponde à diferença entre a água entrada no sistema e a água faturada. Inclui:

- Perdas aparentes;
- Perdas reais;
- Consumo autorizado não faturado [7].

Perdas de Água

As perdas de água correspondem à diferença entre a água entrada no sistema e o consumo autorizado [7]. Podem ser avaliadas ao nível de todo o sistema de abastecimento ou em subsistemas específicos, como a rede de água não tratada, o sistema de adução ou o sistema de distribuição [7]. As perdas de água subdividem-se em perdas reais e perdas aparentes [7].

Perdas Reais

As perdas reais correspondem às perdas físicas de água do sistema pressurizado, desde os órgãos de produção até ao contador do cliente. Incluem fugas em condutas, ramais, acessórios e extravasamentos de reservatórios [7]. O volume perdido depende da frequência, do caudal e da duração da fuga [7].

3.1.2 Indicadores das perdas reais

Current Annual Real Losses (CARL)

As perdas reais anuais atuais (CARL) representam a melhor estimativa disponível do volume médio anual de perdas reais, avaliado de acordo com o **Balanço Hídrico Padrão da IWA** [8].

Este indicador pode ser expresso em volume por ano ou volume por dia [8].
No caso de sistemas com abastecimento intermitente, o CARL deve ser calculado apenas durante os períodos em que o sistema se encontra pressurizado [8].

Unavoidable Annual Real Losses (UARL)

O volume de perdas reais anuais inevitáveis (UARL) representa o menor volume anual de perdas reais que é tecnicamente alcançável num sistema bem gerido e bem mantido [8].

Infrastructure Leakage Index (ILI)

O Índice de Perdas da Infraestrutura (ILI) é uma medida do grau de eficácia com que uma rede de distribuição é gerida, mantida, reparada e reabilitada no que respeita ao controlo das perdas reais, considerando a pressão média operacional atual [8]. O ILI corresponde ao rácio entre o volume anual atual de perdas reais (CARL) e o volume anual inevitável de perdas reais (UARL):

$$ILI = CARL / UARL$$

Perdas reais por ramal e por dia (L/ramal/dia)

As perdas reais por ramal e por dia representam um indicador técnico amplamente utilizado para expressar o volume médio diário de perdas reais por ligação [7]. Este indicador é particularmente

adequado para sistemas com **densidade de ramais igual ou superior a 20 por km de rede**, sendo recomendado pela literatura técnica de referência como medida comparativa de perdas reais em contextos urbanos ou com elevada densidade de ligações [7].

Perdas reais por quilómetro de rede e por dia (m^3/(km·dia))

Este indicador expressa o volume médio diário de perdas reais por quilómetro de rede. A sua utilização é particularmente adequada em sistemas com **densidade de ramais inferior a 20 por km de rede**, nos quais a expressão das perdas por ramal e por dia se torna menos comparável [7]. Nestes casos, a normalização pela extensão da rede permite uma avaliação mais ajustada das perdas reais e uma comparação mais consistente entre entidades com redes pouco densas [7].

Current Real Losses Indicator (CRLI)

O **CRLI** (*Current Real Losses Indicator*) é um indicador composto proposto para combinar, num único valor, duas medidas operacionais das perdas reais: a perda por metro de conduta e a perda por ligação/ramal por dia [21]. A formulação apresentada em materiais técnicos recentes exprime-o como a média geométrica entre **L/m/dia** e **L/prop/dia**, com o objetivo de facilitar a comparação entre sistemas com diferentes densidades de ligação [21]. No presente estudo, este indicador foi operacionalizado através da expressão **CRLI = (L/m/day * L/prop/day)1/**2, de forma consistente com as variáveis disponíveis na base de dados [21].

3.2 Tecnologias e Ferramentas Utilizadas

A realização deste trabalho exigiu a utilização de ferramentas digitais que possibilitaram a organização, análise e interpretação dos dados recolhidos. A escolha das ferramentas teve em consideração a necessidade de garantir precisão, rapidez no processamento de informação e reprodutibilidade dos resultados.

Microsoft Excel

O Microsoft Excel foi utilizado para a organização inicial dos dados e para a realização de tratamentos básicos, incluindo limpeza, filtragem e estruturação das tabelas. Esta ferramenta permitiu uma análise preliminar dos valores e a identificação de padrões gerais nos dados.

Google Colab

Para a análise mais avançada, recorreu-se ao Google Colab, uma plataforma que permite a execução de código Python em ambiente cloud. A sua utilização facilitou o processamento eficiente dos dados, bem como o desenvolvimento de scripts reutilizáveis para automatizar tarefas de análise.

4 Estado da Arte

4.1 Estado da Arte

Introdução

A água é um recurso natural essencial à vida e ao desenvolvimento socioeconómico, mas a sua disponibilidade encontra-se sob crescente pressão devido às alterações climáticas, ao aumento demográfico e à intensificação das atividades humanas [5]. Neste contexto, a sustentabilidade hídrica tornou-se uma prioridade global, exigindo que os sistemas de abastecimento de água minimizem desperdícios e maximizem a eficiência operacional.

A **Água Não Faturada**, definida como a diferença entre a água entrada no sistema e a água faturada, inclui não só as perdas reais e aparentes, mas também o consumo autorizado não faturado [7]. A Figura 1 apresenta a estrutura das componentes do balanço hídrico, detalhando a decomposição da água autorizada, das perdas aparentes, das perdas reais e da água não faturada. Tanto as perdas reais como as aparentes constituem um desafio técnico, económico, ambiental e social de elevada magnitude para as entidades gestoras de sistemas de abastecimento [4, 6, 7].

<table>
<tr><th>A</th><th>B</th><th>C</th><th>D</th><th>E</th></tr>
<tr><td rowspan="10">Água entrada no sistema [m³/ano]</td><td rowspan="4">Consumo autorizado [m³/ano]</td><td rowspan="2">Consumo autorizado faturado [m³/ano]</td><td>Consumo faturado medido (incluindo água exportada) [m³/ano]</td><td rowspan="2">Água faturada [m³/ano]</td></tr>
<tr><td>Consumo faturado não medido [m³/ano]</td></tr>
<tr><td rowspan="2">Consumo autorizado não faturado [m³/ano]</td><td>Consumo não faturado medido [m³/ano]</td><td rowspan="7">Água não faturada (perdas comerciais) [m³/ano]</td></tr>
<tr><td>Consumo não faturado não medido [m³/ano]</td></tr>
<tr><td rowspan="6">Perdas de água [m³/ano]</td><td rowspan="2">Perdas aparentes [m³/ano]</td><td>Uso não autorizado [m³/ano]</td></tr>
<tr><td>Perdas de água por erros de medição [m³/ano]</td></tr>
<tr><td rowspan="4">Perdas reais [m³/ano]</td><td>Perdas reais nas condutas de água bruta e no tratamento (quando aplicável) [m³/ano]</td></tr>
<tr><td>Fugas nas condutas de adução e/ou distribuição [m³/ano]</td></tr>
<tr><td>Fugas e extravasamentos nos reservatórios de adução e/ou distribuição [m³/ano]</td></tr>
<tr><td>Fugas nos ramais de ligação (a montante do ponto de medição) [m³/ano]</td></tr>
</table>

Figura 1– Componentes do balanço hídrico

Os fatores que explicam as perdas são múltiplos e interligados. Entre os determinantes técnicos destacam-se o estado físico da infraestrutura, o nível e variabilidade da pressão, as características dos materiais das condutas, a idade dos sistemas e a densidade de ramais [1, 3].

O primeiro passo para o seu controlo consiste na elaboração rigorosa do balanço hídrico, seguindo metodologias padronizadas como as propostas pela International Water Association (IWA), que permitem quantificar as perdas reais de forma comparável entre sistemas [1, 2, 3]. Neste enquadramento, os indicadores Current Annual Real Losses (CARL), Unavoidable Annual Real Losses

(UARL) e Infrastructure Leakage Index (ILI) assumem particular relevância, constituindo métricas amplamente adotadas para avaliar o desempenho das redes no que respeita às perdas reais [2, 6].

A crescente necessidade de uma gestão hídrica mais eficiente tem impulsionado o uso de abordagens baseadas em dados para a identificação rápida e precisa de fugas. Métodos de **machine learning**, como SVM, redes neuronais e modelos ensemble, têm demonstrado elevada capacidade para analisar dados de sensores em tempo real e identificar padrões não lineares associados a fugas [15, 16]. Técnicas não supervisionadas, como clustering validado por índices de qualidade, reforçam a deteção precoce de anomalias em séries de pressão e caudal [17], enquanto tecnologias avançadas de monitorização, incluindo fibra ótica e redes de sensores sem fios, aumentam a precisão e cobertura da deteção [18]. Em conjunto, estes avanços permitem compreender melhor os fatores que influenciam as perdas reais, incluindo condições operacionais e características das infraestruturas, conforme evidenciado em estudos recentes, inclusive no contexto português [5].

Perdas Reais

As perdas reais correspondem ao volume de água que se perde por fissuras, ruturas ou extravasamentos em sistemas pressurizados, desde a produção até ao contador do cliente [7]. O seu valor depende essencialmente da frequência das fugas, do caudal associado e da sua duração média [7].

Embora não integrem a contabilização oficial do balanço hídrico, as perdas físicas a jusante do contador podem, em alguns casos, assumir relevância e devem ser consideradas na gestão dos consumos e na eficiência global do sistema [7].

As perdas reais correspondem ao volume de água perdido na rede e na infraestrutura sob responsabilidade da entidade gestora. No âmbito do balanço hídrico, estas perdas subdividem-se em diferentes categorias, dependendo do local onde ocorre a fuga, nomeadamente:

- fugas em condutas adutoras e de distribuição;
- fugas nas estruturas dos reservatórios e perdas por extravasamento;
- fugas nas ligações de serviço até ao contador do cliente, geralmente em traçado subterrâneo [13].

Estudos internacionais demonstram que a **maior parte do volume de perdas reais ocorre nas ligações de serviço**, e não nas condutas principais, sobretudo em sistemas com densidade elevada de ligações por quilómetro de conduta [10]. Embora as condutas principais apresentem ruturas mais visíveis e com maiores caudais instantâneos, a duração média das fugas nas ligações é significativamente maior, resultando num impacto anual superior no CARL [10].

A gestão das perdas reais pode ser resumida no modelo dos 4 Componentes, onde o retângulo representa o volume de CARL. Com o envelhecimento da infraestrutura, as perdas tendem a aumentar, contudo, esse crescimento pode ser controlado através dos quatro pilares:

- Gestão de pressões;
- Gestão dos ativos da rede;
- Rapidez e qualidade das reparações;
- Controlo ativo de fugas [13].

A Figura 2 apresenta graficamente estas vertentes principais de atuação, evidenciando a distinção entre perdas inevitáveis, perdas recuperáveis e o nível económico das perdas.

Quando estas componentes são eficazes, o "retângulo" reduz, quando falham, expande-se. Este conceito é conhecido como **"Squeezing the Box"** [11].

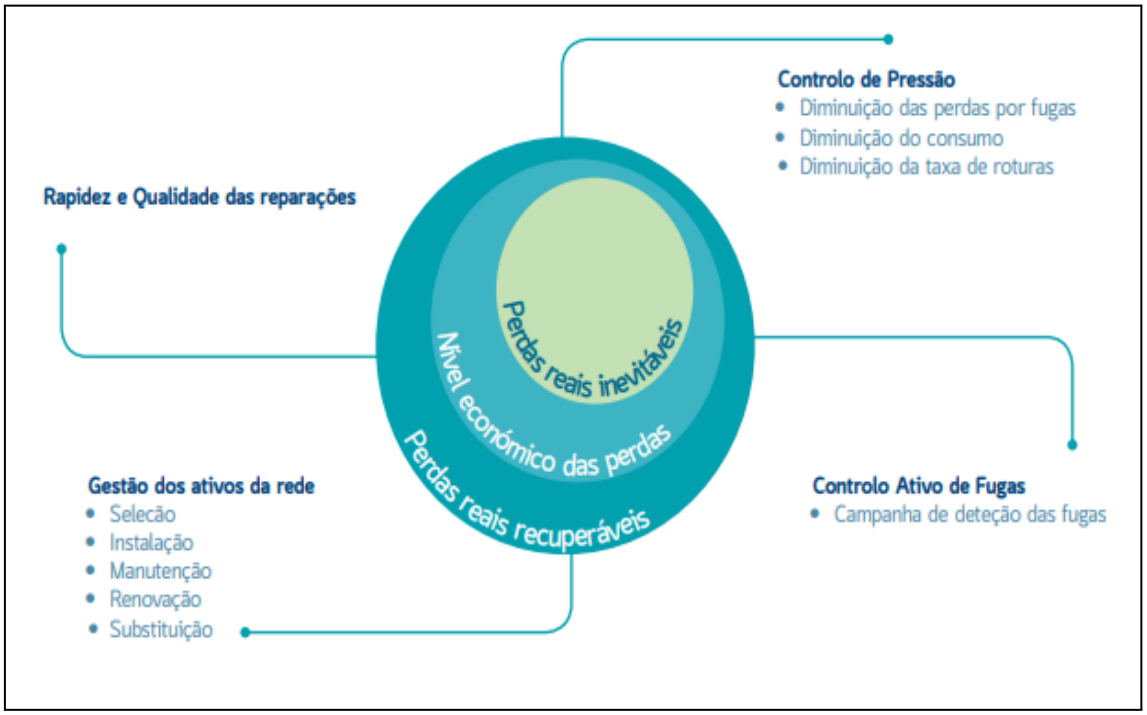

Figura 2 – Vertentes principais da redução das perdas reais (IWA Water Loss Task Force)

Estratégias de redução de perdas reais

Controlo de Pressão

A gestão de pressão é uma das estratégias mais eficazes para reduzir perdas reais, dado que a pressão média de operação influencia diretamente o caudal das fugas e a frequência de roturas [6]. A relação entre pressão e caudal de fuga é amplamente documentada, sendo comum observar-se que pequenas reduções de pressão podem gerar diminuições substanciais no volume perdido [10]. Integrada no modelo da IWA como um dos quatro pilares de controlo de perdas reais, a gestão de pressão contribui também para prolongar a vida útil da infraestrutura e reduzir custos operacionais [9]. Na prática, recorre-se a técnicas como setorização em DMAs e utilização de válvulas ou controladores de pressão, que permitem ajustar a pressão de forma eficiente e minimizar fugas [12].

Gestão dos ativos da rede

A gestão de tubagens e ativos é essencial para assegurar a integridade da infraestrutura, mas a renovação de condutas, quando aplicada isoladamente, apresenta baixa eficiência em termos de custo por volume recuperado [9, 14]. A literatura mostra que apenas uma pequena fração das perdas reais resulta de roturas visíveis em condutas, sendo a maioria associada a fugas nas ligações de serviço [10]. Além disso, a gestão de pressão tende a reduzir de forma mais eficaz a frequência de roturas do que programas extensivos de substituição de tubagens [6].

Rapidez e Qualidade das Reparações

A rapidez e a qualidade das reparações constituem um pilar fundamental na gestão das perdas reais, uma vez que o volume perdido depende sobretudo do tempo em que a fuga permanece ativa. A literatura mostra que as roturas reportadas representam menos de 15% das perdas anuais,

sendo a maior parte do volume associado a fugas não reportadas ou não reparadas, que podem permanecer ativas durante longos períodos [10, 14].

Controlo Ativo de Fugas

O Controlo Ativo de Fugas (ALC) é uma estratégia central na redução das perdas reais, pois permite identificar fugas não reportadas que podem permanecer longos períodos sem deteção. O seu objetivo é reduzir o tempo médio entre o início da fuga e a sua reparação, fator determinante no volume anual perdido [6, 14]. A literatura destaca que a frequência económica das inspeções depende do valor da água perdida, do custo das intervenções e da taxa de crescimento anual das fugas não reportadas. Na prática, o ALC recorre a técnicas como sectorização em DMAs, step testing, sensores acústicos e monitorização contínua de caudal e pressão [6]. Estudos operacionais demonstram que programas estruturados de ALC, combinando inspeção sistemática e tecnologia de deteção, conduzem a reduções significativas e sustentadas das perdas reais [13].

Abordagem de Machine Learning

A utilização de métodos de machine learning supervisionado tornou-se particularmente relevante, destacando-se algoritmos como Support Vector Machines (SVM), Logistic Regression, Artificial Neural Networks (ANN) e ensemble models. Estes métodos demonstraram capacidade para extrair, por exemplo, características não lineares de dados de pressão, permitindo distinguir com maior exatidão situações normais de sinais típicos de fuga [15]. Para além disso, a literatura reporta que estes modelos conseguem ultrapassar problemas frequentes nos sistemas de deteção de fugas, tais como datasets desbalanceados ou incertezas associadas aos requisitos dos utilizadores, que historicamente dificultaram a fiabilidade dos modelos de deteção [15]. No mesmo sentido, investigações anteriores já tinham avaliado a eficácia de SVM, k-Nearest Neighbors, classificadores Bayesian e sistemas neuro-fuzzy para localizar e estimar a dimensão de fugas, reforçando a maturidade destas abordagens no domínio da hidráulica urbana [16].

Em complemento às metodologias supervisionadas, os métodos não supervisionados, particularmente o clustering, ganharam importância na deteção preliminar de comportamentos anómalos. O uso de agrupamentos permite identificar outliers em séries de caudal antes da confirmação da ocorrência de uma rotura, atuando como uma primeira camada de filtragem de eventos [18]. A qualidade destes agrupamentos é assegurada por técnicas de validação como os índices Davies–Bouldin, Dunn, Calinski–Harabasz e Silhouette, que avaliam a coesão interna e a separação entre clusters, garantindo que os padrões detetados representam fenómenos hidráulicos reais e não ruído estatístico [17]. Estes métodos assumem particular relevância em redes complexas, onde a variabilidade operacional pode mascarar sinais de fuga, tornando crucial validar rigorosamente a estrutura dos dados.

Também se identificam metodologias híbridas que combinam diferentes tipos de dados e algoritmos. Exemplos incluem métodos em duas etapas que detetam outliers através de clustering e posteriormente verificam a ocorrência de roturas [18], algoritmos baseados em **partição de grafos**, que determinam zonas críticas e identificam locais adicionais de monitorização necessários para reduzir a área de incerteza durante a procura de fugas [18] e ainda técnicas de processamento de sinal, como **predição linear aplicada a sinais acústicos**, que permitem extrair assinaturas específicas de fugas em contextos ruidosos [18]. Por fim, redes neuronais têm sido aplicadas para

estimar simultaneamente o **tamanho** e a **localização** da fuga, evidenciando a flexibilidade destes modelos em cenários hidráulicos complexos [18].

Conclusão

A redução das perdas reais continua a ser uma prioridade estratégica para assegurar a sustentabilidade dos sistemas de abastecimento de água, exigindo a integração de práticas operacionais consolidadas e de tecnologias emergentes. Os quatro pilares clássicos de controlo são: gestão da pressão, reabilitação de infraestruturas, rapidez das reparações e controlo ativo de fugas. Estas mantêm-se altamente eficazes na redução do CARL e na melhoria do desempenho global, conforme amplamente demonstrado na literatura [6, 10, 13]. Contudo, a crescente complexidade das redes e a necessidade de intervenções mais rápidas e precisas têm impulsionado a adoção de abordagens baseadas em dados, com especial destaque para métodos de machine learning capazes de identificar padrões não lineares, ultrapassar limitações de conjuntos de dados desbalanceados e reforçar a fiabilidade da deteção de fugas [15, 16]. Técnicas não supervisionadas, como o clustering validado por índices de qualidade, reforçam a deteção precoce de anomalias em séries de pressão e caudal [17], enquanto metodologias híbridas e tecnologias avançadas de monitorização, fibra ótica, redes de sensores sem fios, algoritmos baseados em grafos e análise acústica, aumentam significativamente a granularidade e a precisão da localização de fugas [18]. Assim, a convergência entre métodos tradicionais da engenharia hidráulica e soluções inteligentes baseadas em IA constitui atualmente um vetor essencial para melhorar a eficiência, reduzir desperdícios e reforçar a resiliência dos sistemas de abastecimento de água num contexto de crescente pressão sobre os recursos hídricos.

5 Solução Proposta

5.1 Introdução

Nesta secção apresenta-se uma descrição geral da solução desenvolvida, destacando os seus principais objetivos, funcionalidades e componentes implementados. Pretende-se oferecer uma visão clara sobre o trabalho realizado, evidenciando a forma como a solução responde aos requisitos definidos e o valor que acrescenta no contexto do projeto.

Adicionalmente, são disponibilizados os recursos de suporte à análise e validação do trabalho desenvolvido, nomeadamente o link para o repositório Git, que contém o código-fonte produzido e, sempre que aplicável, os dados utilizados, desde que estes sejam públicos:

https://github.com/DEISI-ULHT-TFC-2025-26/DEISI2163-Data-Science-na-Gest-o-Eficiente-de-Sistemas-de-Abastecimento-de-gua.git

5.2 Metodologia

A metodologia adotada neste estudo organiza-se em cinco etapas principais, alinhadas com os requisitos definidos para a análise das perdas reais e da água não faturada:

1. Preparação e pré-processamento dos dados;
2. Análise exploratória dos dados;
3. Proposta de um método para estimativa das perdas reais;
4. Avaliação do desempenho global do sistema no que diz respeito às perdas reais;
5. Teste dos métodos propostos.

A primeira etapa corresponde à **preparação e ao pré-processamento dos dados**, incluindo a recolha, validação, seleção e estruturação da informação e dos indicadores provenientes do RASARP, com especial enfoque nas variáveis associadas à **água não faturada** e às **perdas reais**. Nesta fase, irá proceder-se também ao tratamento de incoerências, normalização das variáveis e verificação da consistência dos registos.

Seguidamente realizar-se-á a **análise exploratória dos dados**, examinando padrões, correlações e distribuição das variáveis selecionadas. Esta etapa permitirá identificar comportamentos anómalos, outliers e potenciais relações relevantes para a identificação de fatores influenciadores da água não faturada e das perdas reais.

A terceira etapa consistirá na **proposta de um método para estimativa das perdas reais**, um dos indicadores de desempenho centrais do estudo, a par da água não faturada, integrando fatores operacionais como pressões médias, frequência de roturas, idade e características da rede, com o objetivo de quantificar de forma mais rigorosa o contributo destes elementos para o aumento ou diminuição das perdas.

A quarta etapa envolve a **avaliação do desempenho global do sistema no que respeita às perdas reais**, agregando os resultados obtidos de modo a identificar ineficiências estruturais e estimar o impacto destas perdas ao nível das entidades gestoras.

Por fim, realizar-se-á o **teste dos métodos propostos**, aplicando-os ao conjunto de dados disponíveis e avaliando a sua robustez, capacidade explicativa e potencial de utilização pelas entidades gestoras.

5.3 Recolha de Dados

A recolha de dados baseou-se exclusivamente na informação pública disponibilizada pela Entidade Reguladora dos Serviços de Águas e Resíduos (ERSAR), através do Relatório Anual dos Serviços de Águas e Resíduos em Portugal (RASARP). Este relatório reúne dados reportados anualmente por mais de duas centenas de entidades gestoras dos serviços de abastecimento de água, tanto em alta como em baixa, constituindo a base estatística mais completa e sistematizada existente em Portugal para a avaliação do desempenho do setor.

No âmbito do presente estudo, os dados foram extraídos de um ficheiro em formato Excel, organizado em diferentes folhas com informação relativa a indicadores e variáveis de caracterização dos sistemas. A seleção incidiu exclusivamente sobre o ramo de atividade do abastecimento de água, identificado pelo código **AA**, tendo sido considerados os registos referentes aos tipos de sistema **Alta** e **Baixa**.

A utilização desta fonte permitiu assegurar a rastreabilidade, consistência institucional e relevância setorial da informação analisada, constituindo uma base adequada para a construção da base de dados de trabalho e para o desenvolvimento das etapas subsequentes de descrição, pré-processamento e análise.

5.4 Descrição dos Dados

Os dados utilizados no presente estudo dizem respeito ao setor do abastecimento público de água em Portugal, com base na informação reportada à ERSAR, no âmbito do RASARP. Este universo inclui entidades gestoras que operam em diferentes níveis do sistema de abastecimento, distinguindo-se entre serviços **em alta**, associados à captação, tratamento e adução de água, e serviços **em baixa**, responsáveis pela distribuição ao utilizador final. No contexto deste trabalho, a análise incide sobretudo sobre os sistemas **em baixa**, dado ser neste segmento que a água não faturada e, em particular, as perdas reais, assumem maior relevância operacional.

A base de dados de trabalho foi construída a partir de duas folhas principais de um ficheiro Excel, ambas referentes ao ramo de atividade do abastecimento de água (**AA**) e, nesta fase da análise, centradas no sistema **em baixa**. A folha **Indicadores_2023** contém **83** indicadores de desempenho reportados pelas entidades gestoras, incluindo métricas associadas ao desempenho dos sistemas de abastecimento, com particular destaque para a água não faturada e para as perdas reais.

Por sua vez, a folha **Dados_2023** integra **108** variáveis de caracterização dos sistemas, abrangendo dimensões estruturais, operacionais e funcionais. Estas variáveis foram utilizadas como suporte à análise exploratória e à identificação de fatores potencialmente explicativos dos indicadores em estudo.

Na sua forma original, os dados encontravam-se organizados em **formato longitudinal**, isto é, cada linha correspondia ao registo de um indicador ou variável associado a uma determinada entidade

gestora. Esta estrutura permitia representar de forma detalhada a informação reportada, embora exigisse posterior reorganização para efeitos de análise estatística.

A distinção entre indicadores e variáveis revelou-se metodologicamente importante, por permitir separar duas dimensões analíticas distintas: por um lado, **os indicadores, que correspondem a métricas de desempenho já calculadas**, e, por outro, **as variáveis, que traduzem características estruturais, operacionais e funcionais dos sistemas suscetíveis de influenciar esse desempenho**. Deste modo, a base de dados construída constitui um suporte adequado para a análise das relações entre as características dos sistemas de abastecimento e os níveis de eficiência observados.

No que respeita às especificidades dos indicadores selecionados, importa salientar algumas particularidades relevantes para a interpretação dos dados. Um conjunto de indicadores, designadamente **água não faturada** (AA08ab), **reabilitação de condutas** (AA09ab), **ocorrência de avarias em condutas** (AA10ab) e **eficiência energética de instalações elevatórias** (AA16ab), partilha uma característica comum, a **classificação da qualidade do serviço** (boa, mediana, insatisfatória). Esta classificação é feita através de valores de referência que permitem atribuir um julgamento ao desempenho de cada indicador [7].

O indicador **acessibilidade física do serviço** (AA01b), apresenta, adicionalmente, uma restrição de domínio, encontrando-se limitado ao intervalo entre 0 e 100%, com um efeito de teto expectável para valores próximos de 100%. Acresce que os limiares de classificação deste indicador variam em função da tipologia da área de intervenção, distinguindo-se entre áreas urbanas, medianamente urbanas e rurais.

Por sua vez, o indicador **adequação dos recursos humanos** (AA14b) distingue-se por apresentar uma classificação não monótona nas áreas urbanas, situação em que tanto valores excessivamente baixos como excessivamente elevados podem ser considerados insatisfatórios.

O indicador de **perdas reais** (**dAA62ab**) merece particular atenção, uma vez que a métrica utilizada na sua avaliação varia em função da densidade de ramais do sistema. Nos sistemas com **densidade igual ou superior a 20 ramais por quilómetro**, as perdas são expressas em **litros por ramal por dia** (*l/ramal·dia*), já nos sistemas com **densidade inferior a esse limiar**, a métrica adotada corresponde ao **volume de perdas por quilómetro de rede por dia** (*m^3/km·dia*). Por essa razão, ao longo da análise, sempre que se recorre ao indicador de perdas reais, a amostra é restringida às entidades com densidade de ramais igual ou superior a 20, de modo a assegurar a comparabilidade entre observações.

5.5 Pré-processamento dos Dados

Numa primeira fase do pré-processamento, procedeu-se à seleção dos atributos a incluir na base de trabalho, com base na sua relevância para os objetivos analíticos do estudo. Esta seleção foi realizada com o **contributo de profissionais qualificados e com conhecimento especializado na área**, assegurando a pertinência das variáveis consideradas. Como resultado desta seleção, foi possível restringir o conjunto de dados a **26 indicadores** e **82 variáveis**, reduzindo a complexidade da base e excluindo elementos sem contributo expectável para a análise da água não faturada e das perdas reais.

Seguidamente, foram aplicados critérios de filtragem relativos ao ramo de atividade e ao tipo de sistema, assegurando a retenção exclusiva dos registos pertencentes ao ramo **AA** e aos sistemas classificados como **Alta** e **Baixa**. Em paralelo, foi necessário garantir robustez na identificação de

determinados campos, atendendo à existência de pequenas variações na designação de algumas colunas no ficheiro original.

Após esta etapa, os dados foram reorganizados para uma estrutura matricial adequada ao tratamento estatístico, passando cada registo a representar uma combinação única entre entidade gestora e tipo de sistema. Posteriormente, a informação proveniente dos dois conjuntos de atributos foi integrada numa única base consolidada, de modo a reunir, num mesmo suporte analítico, variáveis de caracterização e métricas de desempenho.

Por fim, a base consolidada foi segmentada em dois subconjuntos autónomos, correspondentes aos sistemas **em baixa** e **em alta**, permitindo preparar estruturas de dados diferenciadas para as etapas seguintes da análise.

1. **Verificação da consistência e integridade dos dados**

Na primeira etapa de pré-processamento, procedeu-se à **verificação da consistência dos dados**, com especial incidência nas colunas quantitativas, de modo a **identificar entradas incompatíveis com a sua natureza numérica**. Para esse efeito, foram excluídas da análise as colunas categóricas e, nas restantes, os conteúdos foram convertidos para formato *string* e sujeitos à remoção de espaços em branco.

Posteriormente, tentou converter-se o **conteúdo de cada variável para formato numérico**. Sempre que essa conversão não era possível, e desde que o valor não correspondesse a um *NaN* já reconhecido, **o respetivo conteúdo era sinalizado como valor inesperado**. Este procedimento permitiu identificar, de forma sistemática, entradas não numéricas presentes em variáveis que deveriam conter apenas valores quantitativos.

Em complemento, procedeu-se também à **verificação da existência de valores negativos nas variáveis quantitativas**, com o objetivo de assinalar situações potencialmente inconsistentes face à natureza dos indicadores e variáveis em análise. Paralelamente, foi ainda analisada a **existência de duplicados na base de dado**s, quer ao nível de linhas integralmente repetidas, quer ao nível da repetição de entidades gestoras identificadas pela variável **Empresa**.

2. **Tratamento diferenciado dos valores NA e NR**

Após a verificação da consistência dos dados, analisaram-se os códigos especiais NA e NR presentes nas variáveis quantitativas, dado que impediam a sua utilização direta em análises estatísticas.

Os valores NA foram interpretados como situações de não aplicabilidade da variável para um conjunto relevante de entidades. Embora estas variáveis não apresentassem exclusivamente valores NA, a sua presença em elevada proporção indicava que a informação disponível era estruturalmente limitada e pouco comparável entre observações. Acresce que, nas colunas afetadas, o número de registos com valores efetivamente válidos e distintos de NA era reduzido, o que restringia de forma significativa o seu potencial analítico. Por esse motivo, optou-se pela sua **exclusão da base de trabalho**.

Por sua vez, os valores NR foram interpretados como situações de não reporte de informação. Nestes casos, a decisão de exclusão incidiu sobre as variáveis cuja proporção de NR era superior a **25%**, por se considerar que esse nível de ausência comprometia a sua utilização nas etapas subsequentes da análise.

Após esta filtragem, os NR remanescentes foram convertidos em **missing values** nas colunas quantitativas, que posteriormente foram convertidas para formato numérico. Este procedimento permitiu obter uma base de dados mais consistente e adequada às etapas seguintes de análise.

3. **Tratamento dos missing values por variável e por entidade**

Após a **exclusão** das variáveis com valores **NA** e o **tratamento** dos valores **NR**, analisou-se a distribuição destes na base de dados para avaliar a **completude da informação** e apoiar decisões adicionais de filtragem. Numa primeira etapa, a análise incidiu sobre as **variáveis quantitativas**, quantificando-se, para cada atributo, o número e a percentagem de missing values, bem como o número de observações válidas. Seguidamente, a análise foi realizada ao nível das **entidades gestoras**, calculando-se, para cada uma, o número de variáveis quantitativas em falta e a respetiva percentagem face ao total considerado.

Adicionalmente, verificou-se, para cada entidade gestora, a disponibilidade de informação nas **variáveis-alvo centrais do estudo**, água não faturada e perdas reais de água. Com base nestas avaliações, definiram-se como **critérios de exclusão a ausência de valor para água não faturada** e **uma percentagem global de missing values superior a 20%**.

4. **Variáveis com valores em 0.**

Foi analisada a presença de valores iguais a zero nas variáveis quantitativas, com o objetivo de distinguir situações estruturalmente plausíveis de casos pouco informativos para a análise subsequente. Neste âmbito, as variáveis "**Água bruta exportada**" e "**Água tratada exportada**" foram **excluídas da base de trabalho**, por apresentarem uma **elevada proporção de zeros** associada à inexistência de atividade de exportação em muitas entidades gestoras.

Foi igualmente avaliada a variável "**Uso não autorizado**", atendendo à elevada concentração de valores zero, devida à **reduzida capacidade das entidades gestoras em estimar essa variável**. Para esse efeito, analisou-se a sua relação com as **variáveis-alvo**, bem como a sua distribuição segundo o **modelo de gestão** e a **tipologia da área de intervenção**. Com base nesta avaliação, a **variável foi excluída da base de trabalho** por se considerar **pouco informativ**a para os objetivos do estudo.

5. **Análise de outliers nas variáveis quantitativas**

Foi realizada uma análise de outliers nas **variáveis quantitativas**, com o objetivo de identificar observações extremas suscetíveis de influenciar a análise estatística subsequente. Para esse efeito, recorreu-se ao **método de Tukey**, baseado no **intervalo interquartil** (IQR), procedimento habitualmente associado à interpretação de boxplots, quantificando-se o número e a proporção de outliers por variável.

Contudo, apesar da sua identificação, estes valores **não foram removidos da base de dados**. No contexto em estudo, observações extremas podem refletir **características reais das entidades gestoras** e **não necessariamente erros** ou **anomalias de registo**. Assim, considerou-se metodologicamente mais adequado **preservá-los**, garantindo a manutenção da variabilidade inerente aos dados e evitando a eliminação de informação potencialmente relevante.

5.6 Análise Exploratória dos Dados

1. **Análise da variabilidade e da incidência de outliers nas variáveis quantitativas**

Com o objetivo de compreender melhor o comportamento das variáveis quantitativas e identificar possíveis limitações na sua utilização analítica, foi realizada uma análise da sua **variabilidade** e da **incidência de outliers**. Esta avaliação revelou-se importante para distinguir variáveis com elevada

dispersão global daquelas em que a variabilidade resulta sobretudo da presença de valores extremos, permitindo uma leitura mais rigorosa da estrutura dos dados.

Foi realizada uma análise da variabilidade das variáveis quantitativas, recorrendo ao **desvio padrão** como medida de dispersão global. Em paralelo, procedeu-se à identificação de **outliers** por variável, utilizando o critério baseado no **IQR**. A percentagem de observações extremas foi calculada para cada atributo, permitindo comparar a incidência de outliers entre variáveis com diferentes escalas e amplitudes.

Com base nestas duas medidas, procurou-se avaliar a relação entre a **dispersão global** das variáveis e a **frequência relativa de outliers**, de forma a perceber se atributos com maior variabilidade tendiam também a apresentar maior concentração de valores extremos. Para esse efeito, foi construída uma tabela-resumo com o desvio padrão, a percentagem de outliers e a percentagem de *missing values* por variável, complementada por uma análise gráfica da associação entre estas medidas.

2. **Análise das correlações entre variáveis quantitativas**

Numa etapa subsequente da análise exploratória, procedeu-se à **avaliação das relações entre as variáveis quantitativas**, com o objetivo de identificar padrões de associação entre atributos e compreender melhor a estrutura interna dos dados. Atendendo à presença de assimetrias, diferenças de escala e valores extremos, optou-se pela utilização do **coeficiente de correlação de Spearman**, por se tratar de uma medida não paramétrica mais adequada às características da base em análise.

A análise incidiu, numa primeira fase, sobre a **matriz global de correlação** entre as variáveis quantitativas, permitindo observar a intensidade e o sentido das associações entre os diferentes atributos. Numa segunda fase, a atenção foi dirigida especificamente às **variáveis-alvo** do estudo, procurando identificar quais os atributos com maior proximidade estatística a estes indicadores.

Esta abordagem permitiu apoiar a **identificação de variáveis potencialmente mais relevantes** para a interpretação do fenómeno em estudo, bem como sinalizar **redundâncias**, **associações fortes entre grupos de atributos** e possíveis **relações com interesse analítico** para etapas posteriores.

3. **Análise dos padrões de missing values nas variáveis quantitativas**

Complementarmente, foi analisada a distribuição dos ***missing values*** nas variáveis quantitativas, com o objetivo de verificar se a ausência de informação **ocorria de forma independente** ou se **concentrava em determinados atributos**. Para esse efeito, foi construída uma **matriz binária de *missing values***, permitindo identificar as variáveis com maior ausência de informação e avaliar a existência de padrões associados.

A partir desta matriz, procedeu-se à quantificação do número de ***missing values* por coluna** e à **análise da correlação entre padrões de ausência de informação**. Esta etapa permitiu avaliar se determinadas variáveis tendiam a apresentar **falta de dados em simultâneo**, o que poderia indiciar dependências estruturais no processo de reporte ou agrupamentos de variáveis afetadas pelas mesmas limitações de cobertura.

Esta análise revelou-se particularmente útil para compreender se os problemas de completude estavam **dispersos de forma aleatória** ou se, pelo contrário, se **concentravam em subconjuntos específicos de variáveis**, com potencial impacto na interpretação estatística e na seleção de atributos para fases posteriores.

4. Análise exploratória da água não faturada e das perdas reais por modelo de gestão e tipologia

No decurso da análise exploratória dos dados, foi igualmente examinada a relação entre as **variáveis qualitativas** e as **variáveis-alvo** do estudo, com particular incidência sobre o **modelo de gestão** e a **tipologia da área de intervenção**. Esta etapa teve como objetivo perceber de que forma diferentes **perfis de gestão** e **territoriais** poderiam estar associados aos níveis de água não faturada e de perdas reais de água.

A análise foi estruturada em torno do indicador de **água não faturada**, sendo posteriormente replicada para as **perdas reais**, com as devidas adaptações. Numa primeira fase, analisou-se a distribuição da água não faturada em função do modelo de gestão e da tipologia da área de intervenção, recorrendo a representações gráficas do tipo boxplot, complementadas pela ordenação dos grupos com base na mediana. Esta abordagem permitiu comparar a dispersão, a tendência central e a presença de valores extremos entre os diferentes grupos considerados. Seguidamente, identificaram-se os outliers em cada grupo, com base no IQR, e verificou-se se essas observações correspondiam também a outliers nas perdas reais, considerando apenas as entidades com densidade de ramais igual ou superior a 20, conforme definido na secção 5.3.

Ainda no âmbito da água não faturada, construiu-se uma **medida normalizada em m³/km/ano** [22], obtida a partir do volume de água não faturada e do comprimento total da rede, com o objetivo de tornar mais comparáveis entidades com diferentes escalas de infraestrutura. A distribuição desta medida foi analisada segundo o modelo de gestão e a tipologia da área de intervenção, através de boxplots ordenados pela mediana, tendo sido igualmente identificados os respetivos outliers com base no IQR e verificada a sua coincidência com casos extremos nas perdas reais.

Numa fase posterior, procurou-se avaliar se existia alguma associação entre a tipologia da área de intervenção e o modelo de gestão adotado, analisando a distribuição dos modelos de gestão dentro de cada tipologia através das respetivas frequências relativas e medidas descritivas, com o objetivo de perceber se a composição dos grupos ajudava a interpretar as diferenças observadas no indicador.

A mesma estrutura analítica foi aplicada às perdas reais de água, com duas diferenças relevantes. Em primeiro lugar, a análise foi restringida às entidades com **densidade de ramais igual ou superior a 20**, de modo a assegurar a comparabilidade entre observações. Em segundo lugar, foi construída uma medida derivada adicional, designada **CRLI**, calculada segundo a expressão **CRLI = √(L/m/dia × L/ramal/dia)** [21], que combina simultaneamente a extensão da rede e a intensidade das perdas por ramal, permitindo exprimir as perdas numa lógica mais ajustada às características físicas do sistema. A distribuição do CRLI foi analisada segundo o modelo de gestão e a tipologia da área de intervenção, através de boxplots ordenados pela mediana, tendo sido igualmente identificados os respetivos outliers e verificada a sua coincidência com casos extremos na água não faturada.

À semelhança do que foi feito para a água não faturada, analisou-se também, neste contexto, a distribuição dos modelos de gestão dentro de cada tipologia, procurando avaliar se a composição dos grupos poderia ajudar a interpretar as diferenças observadas nas perdas reais entre distintos contextos territoriais.

Em síntese, procurou-se avaliar se as **variáveis qualitativas** associadas à **organização da gestão** e ao **contexto territorial dos sistemas** estavam relacionadas com diferenças na **distribuição das variáveis-alvo**, permitindo aprofundar a compreensão do fenómeno em estudo antes da fase de interpretação dos resultados.

5.7 Abrangência

A solução desenvolvida resulta da integração de conhecimentos adquiridos em várias unidades curriculares, abrangendo áreas essenciais como Programação, Matemática, Estatística, Ciência de Dados, Aprendizagem Automática, Engenharia de Dados, Segurança e Metodologias de Investigação. A Programação sustenta a implementação dos algoritmos e a construção de pipelines eficientes; a Matemática fornece as bases formais para a modelação e otimização; e a Estatística apoia a análise, deteção de outliers e validação dos resultados. A Ciência de Dados estrutura todo o processo de tratamento e exploração da informação, enquanto a Aprendizagem Automática permite treinar e validar modelos para identificar comportamentos anómalos. A Engenharia de Dados assegura a ingestão e gestão de grandes volumes de dados provenientes de sensores, garantindo escalabilidade e integridade. Os princípios de Segurança protegem a informação e reforçam a robustez do sistema. Por fim, as Metodologias de Investigação orientam o enquadramento científico, a abordagem sistemática e a validação do trabalho

6 Resultados e Discussão

6.1 Resultados do pré-processamento dos dados

1. Resultados da verificação de consistência dos dados

Os resultados obtidos mostram que os únicos valores textuais identificados nas colunas quantitativas foram **NA** e **NR**, não tendo sido detetadas outras entradas não numéricas. Em termos de frequência, registaram-se **508 ocorrências de NA** e **1137 ocorrências de NR**, evidenciando uma maior incidência de valores NR na base em análise.

Relativamente à verificação de **valores negativos**, não foram identificados quaisquer casos nas variáveis quantitativas analisadas e quanto à eventual **duplicação de registos**, não se observaram linhas totalmente duplicadas, nem entidades gestoras repetidas.

Estes resultados indicam que, nesta fase, a principal irregularidade identificada na base de dados se relacionava com a presença de **NA** e **NR**, não se tendo verificado problemas associados a valores negativos ou a duplicação de observações.aa

2. Distribuição dos códigos NA e NR nas variáveis quantitativas

A análise inicial à base de dados revelou a presença de valores textuais em colunas de natureza quantitativa, nomeadamente os códigos **NA** e **NR**. No total, foram identificadas **508 ocorrências de NA** e **1137 ocorrências de NR**, num conjunto de 67 variáveis e 218 empresas. A distinção semântica entre ambos os códigos é relevante, enquanto o **NA** traduz situações de não aplicabilidade, isto é, a variável não tem significado para aquela entidade, o **NR** indica informação não reportada, ou seja, a entidade poderia ter fornecido o valor mas não o fez. Esta distinção tem implicações distintas para o tratamento subsequente dos dados.

Como se observa na Tabela 1, os valores **NA** concentravam-se num conjunto restrito de colunas, o que é coerente com a sua natureza de **não aplicabilidade,** nestes casos, os NA não correspondiam a **falhas ocasionais de reporte**, mas antes a situações em que a própria variável não era aplicável, traduzindo-se numa ausência de informação sistemática nas respetivas colunas. As três colunas mais afetadas foram **"Volume de água medido nos 3 meses com menores volumes"**, **"Volume de água medido nos 3 meses com maiores volumes"** e **"Resposta a reclamações, sugestões e pedidos de informação escritos"**, sendo as duas primeiras responsáveis pela grande maioria das ocorrências de NA. Em resultado desta análise, foram removidas **8 colunas** que apresentavam **pelo menos uma ocorrência de NA**, eliminando assim **508** registos potencialmente inaplicáveis e garantindo que apenas variáveis com cobertura universal à amostra foram retidas.

COLUNA	NA	NR	NA+NR	% NA+NR
Volume de água medido nos 3 meses com menores consumos	175	10	185	94.6%
Volume de água medido nos 3 meses com maiores consumos	170	15	185	91.9%
Resposta a reclamações, sugestões e pedidos de informação (AAA0266b)	104	39	143	65.6%
Eficiência energética de instalações elevatórias (AAA072)	27	50	77	44.9%
Fator de uniformização (AAA073b)	0	100	100	44.9%
Valor atual da rede (AAA039b)	0	70	70	31.8%
Custo de substituição (AAA049b)	0	70	70	31.8%
Índice de vida da infraestrutura (AAA04b)	0	68	68	31.1%
Consumo de energia para bombeamento (AAA072b)	0	59	59	30.7%
Ramal afetados por falhas no abastecimento (PAA...)	0	59	59	30.7%
Resposta a reclamações, sugestões e pedidos de informação (AAA096b)	27	19	46	21.1%
Encargo anual com tarifa social (AAA105b)	39	43	43	19.5%
Água captada (AAA0699)	0	41	41	18.1%
Consumo de energia (AAA075b)	0	40	40	17.8%
Ocorrência de falhas no abastecimento (AAA038)	0	26	26	11.8%
Perda média de água (AAA1051)	0	17	17	7.8%
Subsidios aos investimentos (AAA101b)	0	17	17	7.8%
Outros rendimentos (AAA101b)	0	17	17	7.8%
Água não faturada (AAA068b)	0	17	17	7.8%

Tabela 1 – Distribuição de NA e NR por variável quantitativa

Ao contrário do NA, os valores **NR** apresentavam uma **distribuição mais dispersa** pelas variáveis, traduzindo situações de não reporte. Após a remoção das **8 variáveis com NA**, o total de ocorrências de **NR reduziu-se de 1137 para 782**, representando uma diminuição de **31,2%**.

Numa segunda fase, foram excluídas **5 colunas** cuja proporção de NR excedia **25% do total de observações**. Este limiar foi estabelecido com base no critério de que uma taxa de omissão acima deste valor compromete a representatividade estatística da variável e penaliza qualquer método de imputação. Esta eliminação adicional **reduziu o número de ocorrências de NR** para **459**, uma redução de **41,3% face aos 782 registos remanescentes após a primeira fase**. Globalmente, o processo de limpeza reduziu as ocorrências de **NR em 59,6% face ao total inicial de 1137**.

Os restantes **459 valores NR**, distribuídos por variáveis com taxa de omissão inferior a 25%, foram **convertidos em missing values** (NaN) para efeitos de análise e imputação. Esta conversão permite que os algoritmos estatísticos subsequentes reconheçam explicitamente a ausência de informação, em vez de tratar os valores textuais como categorias ou erros de tipo.

3. Distribuição dos *missing values* por variável e por entidade

A análise dos **missing values por variável**, realizada após o tratamento dos códigos NA e NR, revelou que **nenhuma das variáveis quantitativas** ultrapassava o limiar de **20% de missing values**, o que atesta a eficácia do processo de limpeza anterior.

Ao nível das **entidades gestoras**, a distribuição dos missing values revelou-se substancialmente mais desigual do que ao nível das variáveis, como mostra a Tabela 2. Quatro entidades destacaram-se pelo elevado volume de informação em falta, **CM de Estremoz** (71,43%), **CM de Caminha** (65,31%), **CM de Idanha-a-Nova** (46,94%) e **CM de Penedono** (42,86%). Para estas entidades, mais de metade das variáveis, ou quase metade, não dispõe de informação reportada, o que compromete severamente a sua utilidade analítica num modelo multivariado, independentemente da estratégia de imputação adotada.

	EMPRESA	COLUNAS EM FALTA	% MISSING	TOTAL VARS
0	CM de Estremoz	35	71.43%	49
1	CM de Caminha	32	65.31%	49
2	CM de Idanha-a-Nova	23	46.94%	49
3	CM de Penedono	21	42.86%	49
4	CM de Tabuaço	14	28.57%	49
5	CM de Vila Nova de Paiva	14	28.57%	49
6	CM de Avis	13	24.49%	49
7	CM de Cabeceiras de Basto	12	24.49%	49
8	CM de Mourão	12	22.45%	49
9	CM de Vinhais	12	22.45%	49
10	CM de Castro Daire	10	20.41%	49
11	CM de Castro Daire	10	18.37%	49
10	CM de Belmonte	9	18.37%	49
12	CM de Belmonte	9	18.37%	49

Tabela 2 – Entidades com maior percentagem de missing values

A verificação específica das **variáveis-alvo**, Água não faturada (AA08b) e Perdas reais de água (AA15b), identificou **22 entidades** com ausência de informação **em pelo menos uma delas**. Tal como evidenciado na Tabela 3, a análise do padrão de omissão revelou uma assimetria relevante, **17 entidades** não apresentavam valor em **nenhuma das duas variáveis-alvo**, enquanto **5 entidades** reportavam valor para **Água não faturada** mas não para **Perdas reais de água**. Não se observaram casos no sentido inverso, o que é conceptualmente coerente, uma vez que as **perdas reais constituem uma componente da água não faturada**. Assim, a ausência de informação sobre as **perdas reais** não implica necessariamente a **ausência de água não faturada**, embora possa limitar a sua decomposição analítica.

	EMPRESA	ÁGUA NÃO FATURADA	PERDAS REAIS DE ÁGUA	% MISSING EMPRESA
0	CM de Aguiar da Beira	59.3	NaN	12.96%
1	CM de Arganil	NaN	NaN	25.93%
2	CM de Avis	61.5	NaN	22.22%
3	CM de Belmonte	59.7	NaN	16.67%
4	CM de Cabeceiras de Basto	NaN	NaN	22.22%
5	CM de Caminha	NaN	NaN	59.26%
6	CM de Castro Daire	NaN	NaN	16.67%
7	CM de Cuba	62.7	NaN	12.96%
8	CM de Estremoz	NaN	NaN	64.81%
9	CM de Idanha-a-Nova	NaN	NaN	42.59%
10	CM de Monchique	NaN	NaN	12.96%
12	CM de Monforte	NaN	NaN	11.11%
13	CM de Mourão	73.6	NaN	20.37%
14	CM de Penedono	NaN	NaN	36.89%
15	CM de Ribeira de Pena	NaN	NaN	11.11%
16	CM de Terras de Bouro	NaN	NaN	11.11%
17	CM de Vila Nova de Paiva	NaN	NaN	11.11%
19	CM de Vila Viçosa	NaN	NaN	24.07%
21	CM de Vouzela	NaN	NaN	11.11%

Tabela 3 – Missing values nas variáveis-alvo por entidade

Com base nas análises efetuadas, foram aplicados **dois critérios de exclusão sequenciais**. Em primeiro lugar, **foram removidas todas as entidades sem valor válido em pelo menos uma das variáveis-alvo**, reduzindo a base de **218 para 196 observações**. Em segundo lugar, foram **excluídas as entidades com uma percentagem global de missing values superior a 20%**, isto é, com mais de 10 variáveis em falta num total de 49 variáveis disponíveis, o que reduziu a amostra final para **195 entidades e 54 variáveis**. Esta última redução eliminou apenas uma entidade adicional, o que indica que a **maioria das entidades com elevada taxa de omissão já havia sido removida pelo critério das variáveis-alvo**.

4. Análise de valores zero em variáveis quantitativas

A análise da distribuição de valores nulos evidenciou que algumas variáveis apresentavam uma concentração muito elevada de zeros, tal como ilustrado na Tabela 4. Os casos mais extremos foram **"Água bruta exportada"**, com **195 zeros em 195 observações (100%)**, **"Água tratada exportada"**, com **145 zeros em 195 observações (74,4%)**, e **"Uso não autorizado"**, com **140 ocorrências em 195 observações (71,8%)**. Estes resultados indicam que, em alguns casos, os valores nulos refletem a inexistência efetiva de determinada operação, enquanto nos outros sugerem uma reduzida capacidade discriminatória da variável no contexto da análise.

COLUNA	N_TOTAL	N_ZEROS	% ZEROS
Água bruta exportada dAAA65b	195	195	100.0%
Água tratada exportada dAAA67b	195	145	74.4%
Uso não autorizado dAAA65b	195	140	71.8%
Consumo faturado não medido dAAA85b	195	110	56.4%
Índice de gestão patrimonial de infraestruturas dAAA101b	195	105	53.8%
Consumo não faturado não medido dAAA85b	195	73	37.4%
Água captada dAAA65b	195	63	32.3%
Consumo não faturado medido dAAA85b	195	52	26.7%
Subsídios ao investimento dAAA102b	195	45	23.1%
Ocorrência de falhas no abastecimento dAAA103b	195	43	22.1%
Ocorrência de falhas no abastecimento dAAA103b	195	43	22.1%
Falhas no abastecimento dAAA61b	195	43	22.1%

Tabela 4 – Variáveis com maior percentagem de zeros

Dado o interesse conceptual da variável "**Uso não autorizado"** como potencial preditor, o furto de água é teoricamente uma componente da água não faturada, a sua utilidade analítica foi avaliada através da correlação de Spearman com as **variáveis alvo** e da análise visual dos **diagramas de dispersão**. Conforme apresentado na Figura 3, os resultados foram claros, a correlação com **Água não faturada** (AA08b) foi de **−0,377** e com **Perdas reais de água** (AA15b) de **−0,213**, ambas negativas e de magnitude **fraca a moderada**. O sinal negativo é, por si só, contra intuitivo, **seria esperado que maior uso não autorizado se associasse a maior água não faturada**, e não ao contrário. Este resultado pode ser consequência direta da elevada concentração de zeros, que **enviesam o coeficiente de correlação** ao criar um grupo artificialmente homogéneo de observações sem uso não autorizado reportado, sendo que, em muitos casos, esses zeros poderão **não refletir a ausência efetiva do fenómeno**, mas antes a **reduzida capacidade das entidades gestoras para o estimar**, o que os torna potencialmente enganadores e pouco fiáveis.

Uso não autorizado (dAA056b) × Perdas reais de água (AA15b)
Correlação de Spearman

NOME_COLUNA	USO NÃO AUTORIZADO (dAA056b)	PERDAS REAIS DE ÁGUA (AA15b)
Uso não autorizado (dAA056b)	1.000	−0.213
Perdas reais de água (AA15b)	−0.213	1.000

Uso não autorizado (dAA056b) × Água não faturada (AA08b)
Correlação de Spearman

NOME_COLUNA	USO NÃO AUTORIZADO (dAA056b)	ÁGUA NÃO FATURADA (AA08b)
Uso não autorizado (dAA056b)	1.000	−0.377
Água não faturada (AA08b)	−0.377	1.000

Figura 3 — Correlação de Spearman entre a variável "Uso não autorizado" e as variáveis-alvo

A análise complementar por **modelo de gestão** e **tipologia da área de intervenção** teve como objetivo verificar se a presença de valores nulos na variável **uso não autorizado** se comportava de forma distinta entre entidades classificadas como **outliers** e **não outliers**, quer para a **água não faturada**, quer para as **perdas reais**. Os resultados mostraram que a ocorrência de zeros nesta variável era muito frequente em diferentes perfis de entidades, com particular incidência em entidades de **gestão direta** e em áreas **predominantemente rurais**. Esta distribuição não revelou um padrão suficientemente consistente que permitisse considerar o **uso não autorizado** como um fator diferenciador robusto dos casos extremos observados nas variáveis-alvo. Assim, a elevada concentração de zeros sugere uma **capacidade explicativa reduzida**, podendo refletir não apenas **ausência efetiva do fenómeno**, mas também **limitações de deteção**, **quantificação** ou **reporte**.

Com base nestes resultados, optou-se por excluir da base de trabalho as variáveis **"Água bruta exportada"**, **"Água tratada exportada"** e **"Uso não autorizado"**, por apresentarem uma **incidência excessiva de zeros** e **reduzida capacidade informativa** para os objetivos do estudo.

6.2 Resultados da análise exploratória dos dados

1. **Análise da variabilidade e da incidência de outliers nas variáveis quantitativas**

A comparação entre o **desvio padrão** e a **percentagem de outliers** mostrou que existe uma associação positiva moderada entre ambas as medidas, com uma **correlação de Spearman de 0,590**. Este resultado indica que, em termos gerais, variáveis com maior dispersão tendem também a apresentar uma proporção mais elevada de observações extremas, embora essa relação não seja uniforme em todos os casos. O diagrama de dispersão apresentado na Figura 4 , embora confirme a **tendência positiva** sugerida pela **correlação de 0,590**, apresenta **limitações de legibilidade** que condicionam a sua interpretação, a **maioria das variáveis concentra-se junto ao zero no eixo do desvio padrão**, comprimindo os pontos numa zona estreita do gráfico e impossibilitando a identificação individual de cada variável.

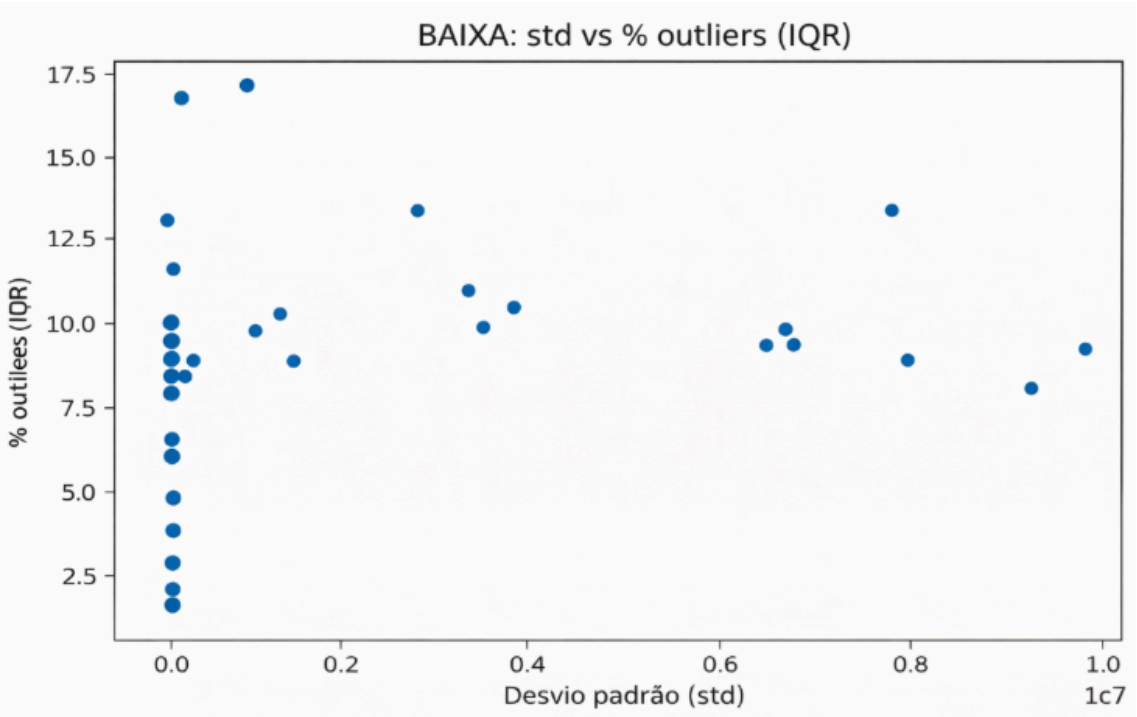

Figura 4 — Relação entre o desvio-padrão e a percentagem de outliers nas variáveis quantitativas

A análise conjunta destas duas ordenações mostra uma **sobreposição parcial** entre variáveis com elevada dispersão e variáveis com maior incidência de outliers. Algumas, como **"Água tratada importada"**, **"Água captada"** e **"Consumo faturado não medido"**, surgem em posições de destaque em ambas as análises, sugerindo que combinam forte variabilidade com presença relevante de valores extremos. No entanto, **outras variáveis com desvio padrão elevado não apresentam necessariamente percentagens de outliers igualmente elevadas**, o que indica que a dispersão global e a ocorrência de valores extremos, embora relacionadas, **não traduzem exatamente o mesmo fenómeno**.

2. Análise das correlações entre variáveis quantitativas

A análise da matriz de correlação de Spearman evidenciou a existência de um número expressivo de **associações positivas elevadas** entre variáveis quantitativas. Tal como se apresenta na Figura 5, a distribuição dos **919 pares positivos** mostra uma concentração nos **intervalos mais baixos**, **169 pare**s entre **10–20%** e **134 pares** entre **0–10%**, mas subsiste um volume relevante de associações fortes, **64 pares** no intervalo **90–100%**, **42 pares** entre **80–90%** e **48 pares** entre **70–80%**, totalizando **154 pares** com correlação superior a **70%**. Este resultado confirma a presença de variáveis com comportamento estatístico muito semelhante e indica **redundância informacional** em múltiplos grupos de atributos.

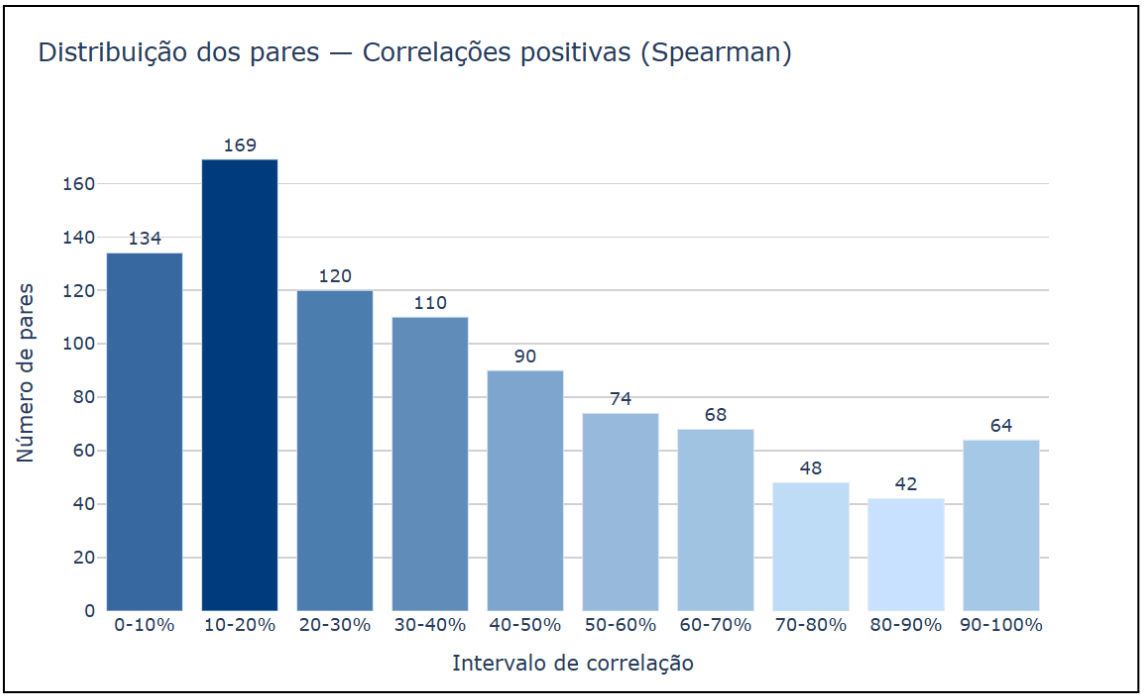

Figura 5 — Distribuição das correlações positivas de Spearman

De um modo geral, as correlações positivas mais elevadas concentram-se em **variáveis relacionadas com consumos**, **volumes de água**, **faturação**, **infraestrutura** e **dimensão operacional**, refletindo o facto de as entidades gestoras de maior dimensão tenderem a apresentar valores mais elevados em todas estas métricas em simultâneo, o que não traduz necessariamente uma relação causal, mas sim uma covariação estrutural associada à escala do sistema.

As correlações negativas são em número claramente inferior, **162 pares** no total, e com menor intensidade global. Como mostra a Figura 6, nenhum par atinge o intervalo **70–100%** em valor absoluto, e a **maioria concentra-se abaixo** dos **30%**. A distribuição é marcadamente assimétrica, com **75 pares** no intervalo **0–10%** e decréscimo progressivo até ao máximo observado de **60–70%**.

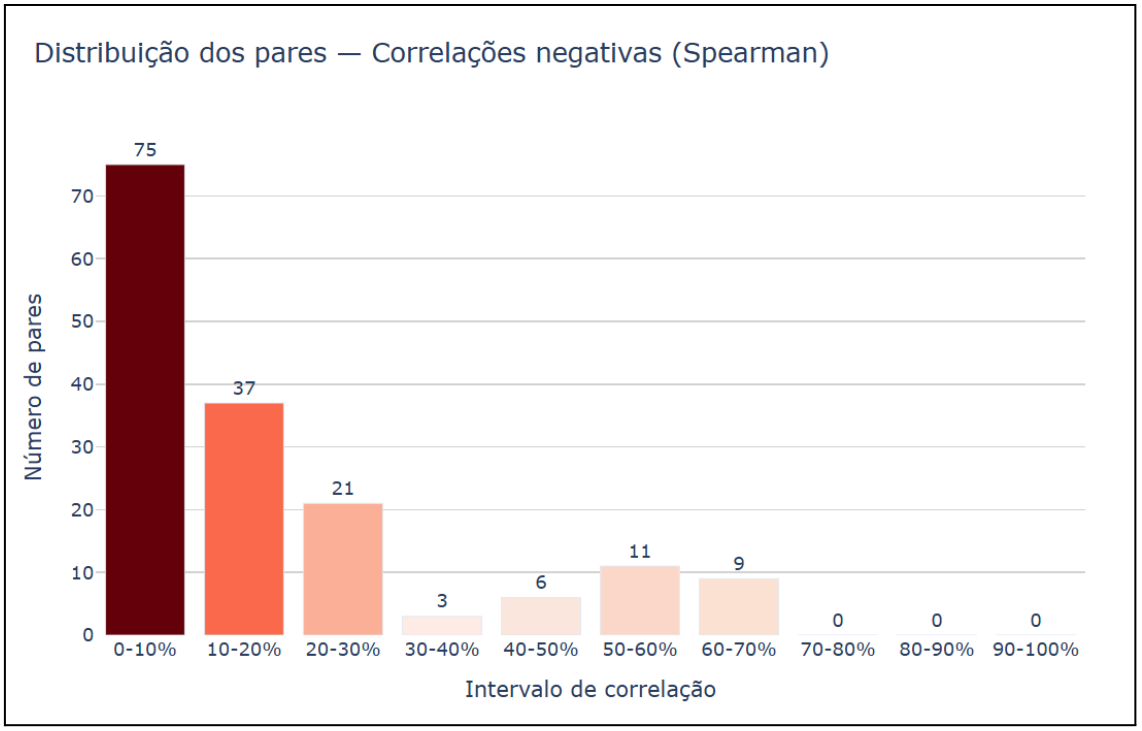

Figura 6 – Distribuição das correlações negativas de Spearman

Em termos gerais, as correlações negativas mais relevantes concentram-se em torno da **Água não faturada**, sugerindo que entidades com **maior volume faturado**, **maior cobertura financeira** e **maior dimensão de consumo autorizado** tendem a registar níveis mais **baixos** deste indicador, o que é conceptualmente coerente com a natureza da água não faturada como reflexo da ineficiência do sistema.

Em síntese, a análise mostra que a estrutura da base é marcada por **muitas correlações positivas fortes** e por um número mais reduzido de **correlações negativas moderadas**, concentradas em torno de variáveis centrais do estudo. Este resultado é relevante para as etapas seguintes, uma vez que confirma a **existência de grupos de variáveis fortemente relacionadas entre si** e reforça a necessidade de atender à **redundância entre atributos** na interpretação dos resultados e em eventuais procedimentos de seleção de variáveis.

3. Análise das correlações entre as variáveis-alvo e as variáveis quantitativas

A análise das correlações com as **variáveis-alvo** revela assimetrias importantes entre os dois indicadores. Como mostra a Figura 7, as **Perdas reais de água** (AA15b) apresentam correlações positivas modestas com o restante conjunto de variáveis: os preditores mais associados são **Perdas reais** (dAA062b) com **ρ = 0,496**, uma variável que mede o mesmo conceito numa **métrica diferente**, o indicador **Água não faturada** (AA08b) com **ρ = 0,478** e o volume de **Água não faturada** (dAA060b) com **ρ = 0,428**.

Este padrão sugere que níveis mais elevados de perdas reais tendem a estar associados a contextos de menor eficiência económica, menor conhecimento infraestrutural e menor capacidade de gestão dos ativos.

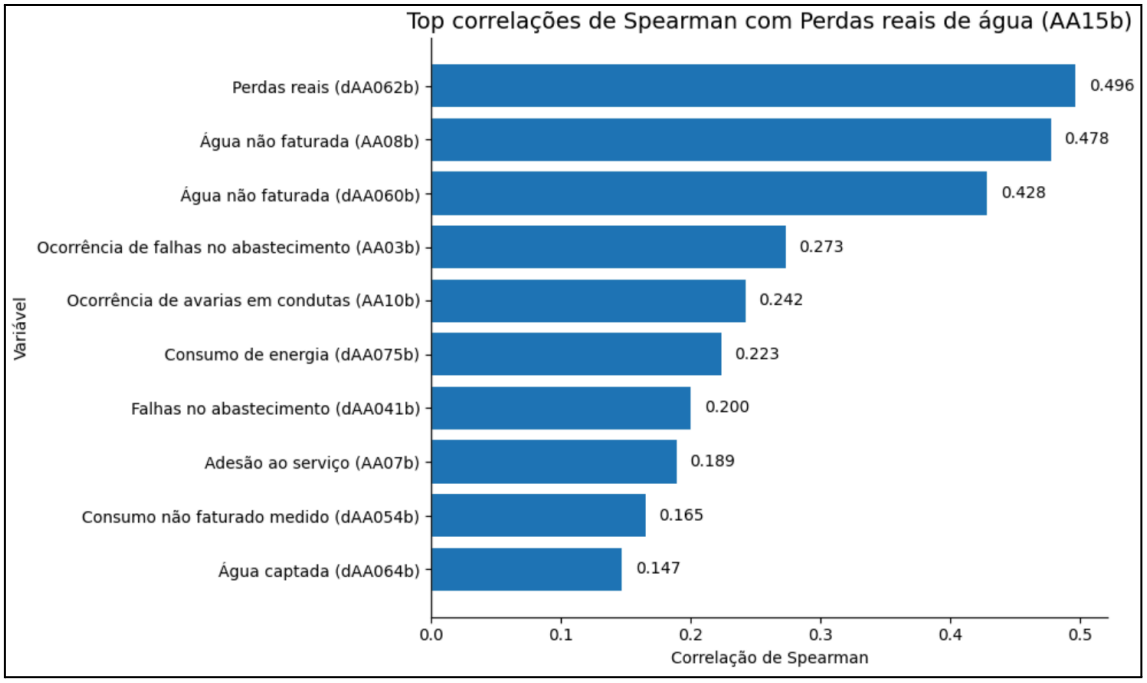

Figura 7 — Principais correlações positivas de Spearman com a variável Perdas reais de água (AA15b)

As **correlações negativas** mais expressivas com as **Perdas reais de água** (AA15b) são, ainda assim, de magnitude moderada/baixa, o que indica a inexistência de relações inversas particularmente intensas, mas sugere associações consistentes com dimensões de desempenho económico e de gestão infraestrutural. A Figura 8 destaca, entre estas associações, o **Encargo anual com tarifário geral** ($\rho = -0{,}292$), a **Acessibilidade económica do serviço** ($\rho = -0{,}278$), o **Índice de conhecimento infraestrutural** ($\rho = -0{,}248$), o **Índice de segurança e resiliência** ($\rho = -0{,}242$) e o **Índice de gestão patrimonial de infraestruturas** ($\rho = -0{,}218$). Em conjunto, estes resultados apontam para uma tendência segundo a qual **entidades com melhor capacidade de gestão dos ativos**, **maior conhecimento sobre o estado das infraestruturas** e **melhores condições de sustentabilidade económica** tendem a apresentar **níveis mais baixos de perdas reais de água**.

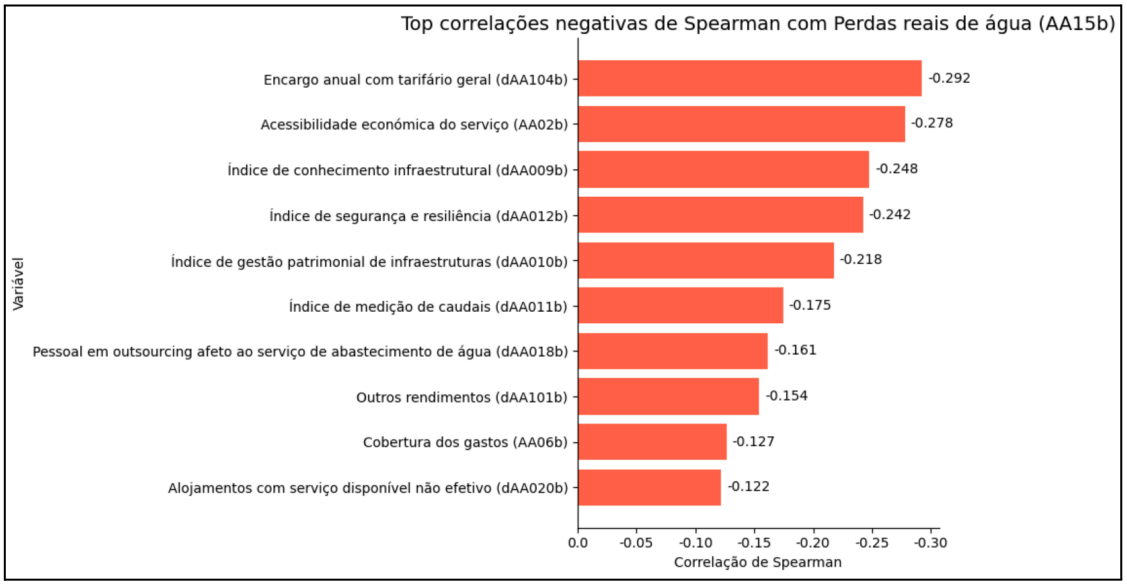

Figura 8 — Principais correlações negativas de Spearman com a variável Perdas reais de água (AA15b)

A **Água não faturada (AA08b)** apresenta **correlações positivas muito fracas** com a generalidade das variáveis, os preditores mais associados, após as próprias **Perdas reais** (**ρ = 0,478**), são o **Consumo não faturado medido** (**ρ = 0,160**) e a **Água captada** (**ρ = 0,144**), tal como mostra a Figura 9.

A variável **Água não faturada** (AA08b) evidenciou um número muito reduzido de associações positivas com as restantes variáveis quantitativas, tendo sido identificadas apenas **7 correlações positivas no total,** como se observa na Figura 9. Este resultado sugere que a água não faturada apresenta uma relação estatística limitada com a maioria dos preditores considerados, o que poderá refletir a natureza mais agregada e multifatorial deste indicador. Entre as associações positivas observadas, destaca-se, em primeiro lugar, a correlação com **Perdas reais de água** (AA15b), com ρ = 0,478, seguindo-se o **Consumo não faturado medido**, com ρ = 0,160, e a **Água captada**, com ρ = 0,144. Ainda que estas duas últimas associações sejam positivas, a sua **magnitude é fraca**.

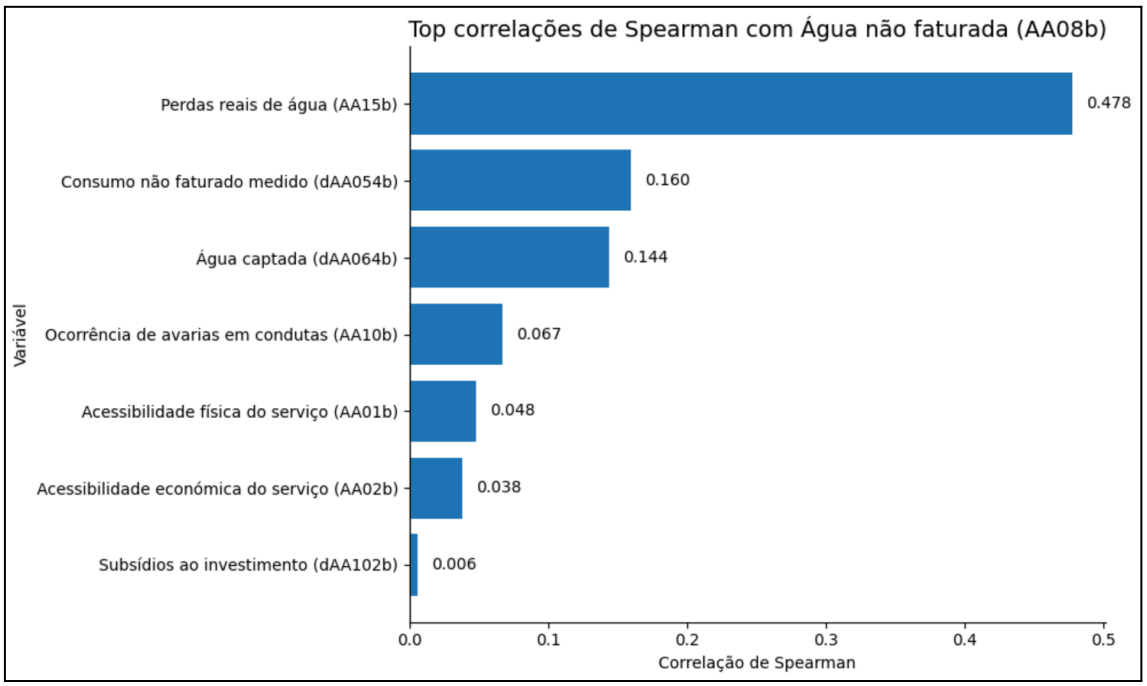

Figura 9 — Principais correlações positivas de Spearman com a variável Água não faturada (AA08b)

Em sentido inverso, a **Água não faturada** (AA08b) evidencia um conjunto mais amplo e mais intenso de **correlações negativas**, revelando um padrão estatístico bastante mais marcado do que o observado nas associações positivas. Tal como evidenciado na Figura 10, as correlações mais elevadas concentram-se sobretudo em variáveis ligadas à **faturação**, à **medição do consumo** e ao **desempenho económico das entidades**, como a **Água faturada não doméstica** (ρ = −0,681), os **Rendimentos tarifários** (ρ = −0,671), o **Consumo faturado medido** (ρ = −0,659) e a **Água faturada** (ρ = −0,658). Este padrão sugere que níveis mais elevados de água não faturada tendem a coexistir com menores volumes faturados, menor capacidade de geração de receita e menor expressão do consumo efetivamente medido.

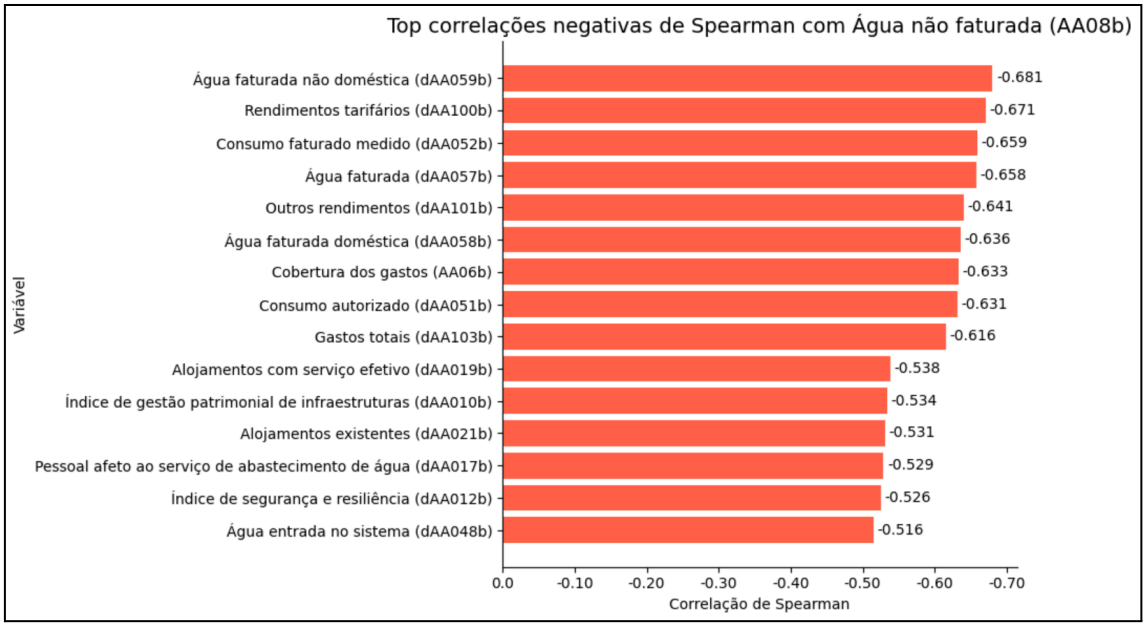

Figura 10 — Principais correlações negativas de Spearman com a variável Água não faturada (AA08b)

Em termos comparativos, os resultados indicam que as **perdas reais de água** apresentam associações mais consistentes com variáveis de natureza **operacional e infraestrutural**, enquanto a **água não faturada** se relaciona mais fortemente, em sentido negativo, com variáveis ligadas à **faturação**, à **medição** e ao **desempenho económico** do sistema. Esta diferença é coerente com a natureza dos dois indicadores, uma vez que as perdas reais representam uma componente física mais específica, enquanto a água não faturada constitui uma medida mais agregada e influenciada por múltiplas dimensões do serviço.

4. Análise dos padrões de missing values nas variáveis quantitativas

A análise dos padrões de *missing values* nas variáveis quantitativas centrou-se na identificação de regularidades na ausência de informação e na avaliação da eventual existência de relações entre variáveis. No conjunto final de variáveis consideradas na análise, apenas 17 apresentavam *missing values*, refletindo o efeito cumulativo das etapas anteriores de seleção, exclusão e tratamento da informação em falta. A Figura 11 apresenta a distribuição do número de *missing values* por variável. Entre elas, a variável com maior número de valores em falta foi o **Consumo de energia (dAA075b)**, com 32, seguida da **Água captada (dAA064b)**, com 22. Destaca-se ainda um segundo grupo constituído por **Cobertura dos gastos**, **Outros rendimentos**, **Rendimentos tarifários**, **Gastos totais** e **Subsídios ao investimento**, cada uma com 13 *missing values*. Este resultado é coerente com o facto de estas variáveis partilharem o mesmo padrão de omissão, tal como já havia sido identificado na análise dos *missing values* por entidade.

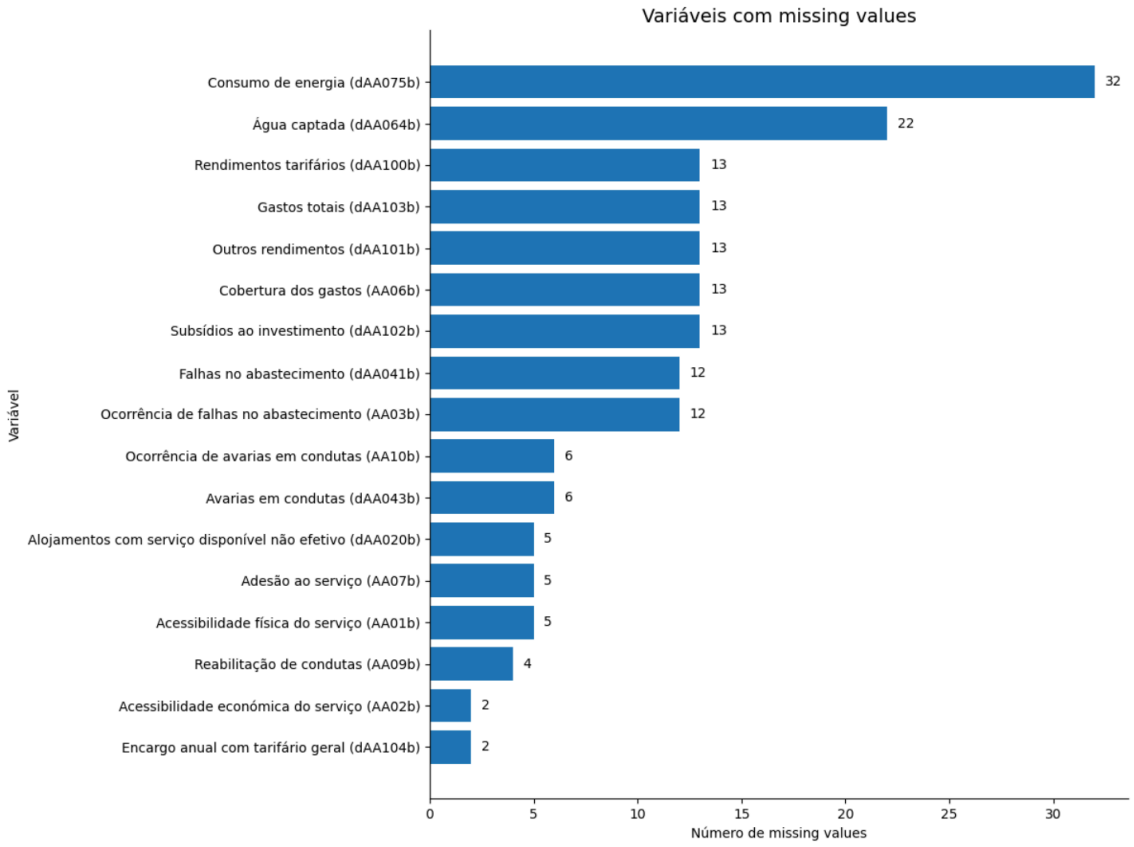

Figura 11 — Número de missing values por variável quantitativa

A análise da correlação entre os *missing values* nas variáveis quantitativas mostrou que a ausência de informação **não ocorre de forma aleatória na base de dados**. Pelo contrário, observam-se grupos de variáveis cujos valores em falta tendem a ocorrer em simultâneo, sugerindo padrões de omissão associados à natureza dos atributos ou ao modo como a informação é reportada pelas entidades. Como se pode observar na Figura 12, a matriz de correlação evidencia **vários pares de variáveis com correlação perfeita** (1,000). Este comportamento é particularmente expressivo em três grupos. No **grupo financeiro**, as variáveis Gastos totais, Outros rendimentos, Rendimentos tarifários, Subsídios ao investimento e Cobertura dos gastos formam um bloco em que a omissão ocorre de forma simultânea. No grupo relativo a **falhas e avarias**, as variáveis Falhas no abastecimento e Ocorrência de falhas no abastecimento, bem como Avarias em condutas e Ocorrência de avarias em condutas, apresentam correlação perfeita entre si. Por fim, no **grupo da acessibilidade**, Acessibilidade física do serviço, Adesão ao serviço e Alojamentos com serviço disponível não efetivo formam igualmente um bloco de omissão conjunta.

Variável 1	Variável 2	Correlação
Avarias em condutas (dAA043b)	Ocorrência de avarias em condutas (AA10b)	1.000
Gastos totais (dAA103b)	Rendimentos tarifários (dAA100b)	1.000
Gastos totais (dAA103b)	Subsídios ao investimento (dAA102b)	1.000
Outros rendimentos (dAA101b)	Rendimentos tarifários (dAA100b)	1.000
Outros rendimentos (dAA101b)	Subsídios ao investimento (dAA102b)	1.000
Rendimentos tarifários (dAA100b)	Subsídios ao investimento (dAA102b)	1.000
Acessibilidade económica do serviço (AA02b)	Encargo anual com tarifário geral (dAA104b)	1.000
Falhas no abastecimento (dAA041b)	Ocorrência de falhas no abastecimento (AA03b)	1.000
Acessibilidade física do serviço (AA01b)	Alojamentos com serviço disponível não efetivo (dAA020b)	1.000
Cobertura dos gastos (AA06b)	Gastos totais (dAA103b)	1.000
Gastos totais (dAA103b)	Outros rendimentos (dAA101b)	1.000
Cobertura dos gastos (AA06b)	Rendimentos tarifários (dAA100b)	1.000
Cobertura dos gastos (AA06b)	Outros rendimentos (dAA101b)	1.000
Adesão ao serviço (AA07b)	Alojamentos com serviço disponível não efetivo (dAA020b)	1.000
Cobertura dos gastos (AA06b)	Subsídios ao investimento (dAA102b)	1.000
Acessibilidade física do serviço (AA01b)	Adesão ao serviço (AA07b)	1.000
Falhas no abastecimento (dAA041b)	Ocorrência de avarias em condutas (AA10b)	0.696
Avarias em condutas (dAA043b)	Ocorrência de falhas no abastecimento (AA03b)	0.696
Ocorrência de avarias em condutas (AA10b)	Ocorrência de falhas no abastecimento (AA03b)	0.696
Avarias em condutas (dAA043b)	Falhas no abastecimento (dAA041b)	0.696
Adesão ao serviço (AA07b)	Subsídios ao investimento (dAA102b)	0.217

Figura 12 — Principais correlações positivas entre os padrões de missing values nas variáveis quantitativas

No que respeita às **correlações negativas** entre os padrões de *missing values*, **não se observaram relações com magnitude relevante**. Todas as correlações negativas apresentaram valores muito reduzidos, **situando-se abaixo de 0,08** em valor absoluto. Este resultado indica a inexistência de padrões inversos consistentes entre variáveis no que diz respeito à ocorrência de informação em falta. Assim, ao contrário do que se verificou nas correlações positivas, **não se identificaram grupos de variáveis com comportamentos de omissão negativamente associados com significado analítico**.

5. Análise exploratória da água não faturada e das perdas reais por modelo de gestão e tipologia

Água não faturada -> Modelo de gestão

A análise da **água não faturada** por **modelo de gestão** evidencia diferenças claras entre os grupos considerados. Como mostra a Figura 13, os sistemas em **concessão** tendem a apresentar os valores mais baixos de água não faturada, seguindo-se os sistemas em **delegação**, enquanto a **gestão direta** surge associada aos valores mais elevados. Este padrão é visível tanto na posição das medianas como na distribuição global dos valores observados.

Em termos de dispersão, a **gestão direta** apresenta também maior variabilidade, refletindo uma maior heterogeneidade entre as entidades deste grupo. Por sua vez, os sistemas em **concessão** revelam uma distribuição mais concentrada em níveis inferiores de água não faturada, enquanto que os sistemas em **delegação** ocupam uma posição intermédia. Quanto aos outliers, identificam-se **dois casos extremos no grupo da concessão** e **um na gestão direta**, não se observando valores atípicos no grupo da **delegação**.

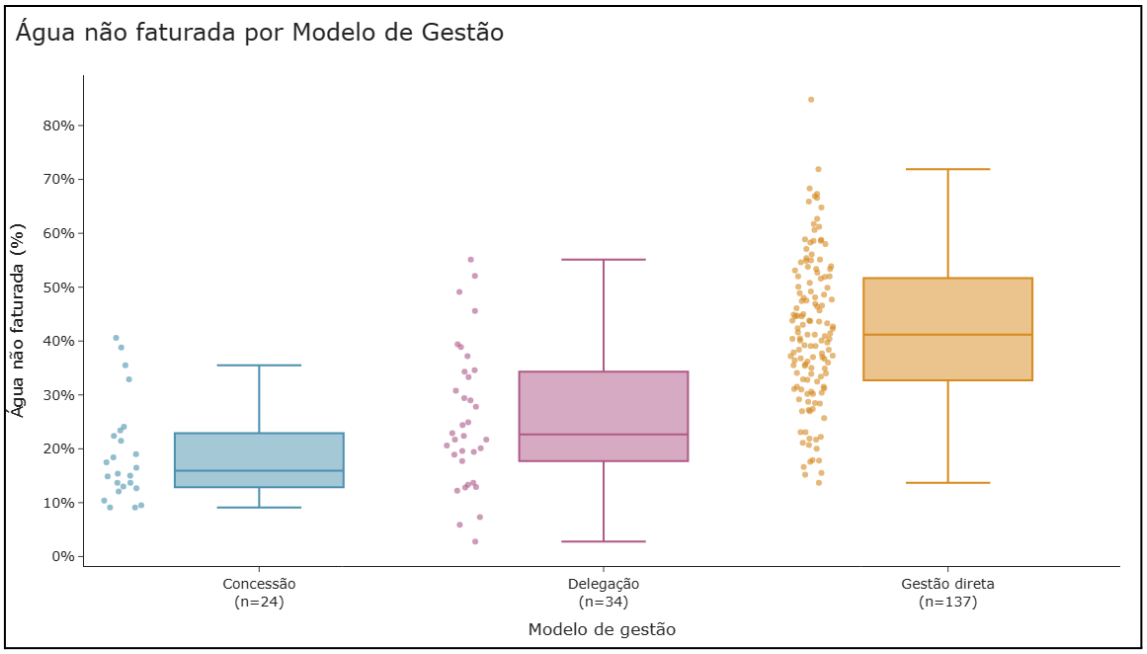

Figura 13 — Distribuição da água não faturada por modelo de gestão

Em termos absolutos, expressos em **m³/km/ano**, este indicador mede o volume de água não faturada por quilómetro de rede e por ano, permitindo comparar entidades de diferentes dimensões com base numa medida normalizada pela extensão da infraestrutura. Como mostra a Figura 14, mantém-se a ordenação observada na análise percentual: os sistemas em **concessão** apresentam, em geral, os valores mais baixos, seguidos dos sistemas em **delegação**, enquanto a **gestão direta** continua a evidenciar os valores mais elevados.

O perfil dos *outliers*, contudo, altera-se de forma relevante. Os dois valores atípicos identificados na **concessão** na análise percentual deixam de se destacar nesta representação, o que sugere que, embora essas entidades apresentassem percentagens relativamente elevadas de água não faturada, operam em sistemas de menor escala e com redes mais curtas. Assim, quando o indicador é expresso em função do comprimento da rede, esses casos perdem expressão relativa. Em sentido inverso, observa-se um outlier particularmente acentuado no grupo da **delegação**, com um valor próximo de 16 mil m³/km/ano, que foi **omitido da Figura 14 para não distorcer a leitura visual da restante distribuição**. Identificam-se ainda vários casos extremos na **gestão direta**, indicando que algumas entidades destes grupos combinam níveis elevados de água não faturada com redes mais extensas, o que se traduz em volumes absolutos por quilómetro substancialmente superiores à mediana do respetivo grupo.

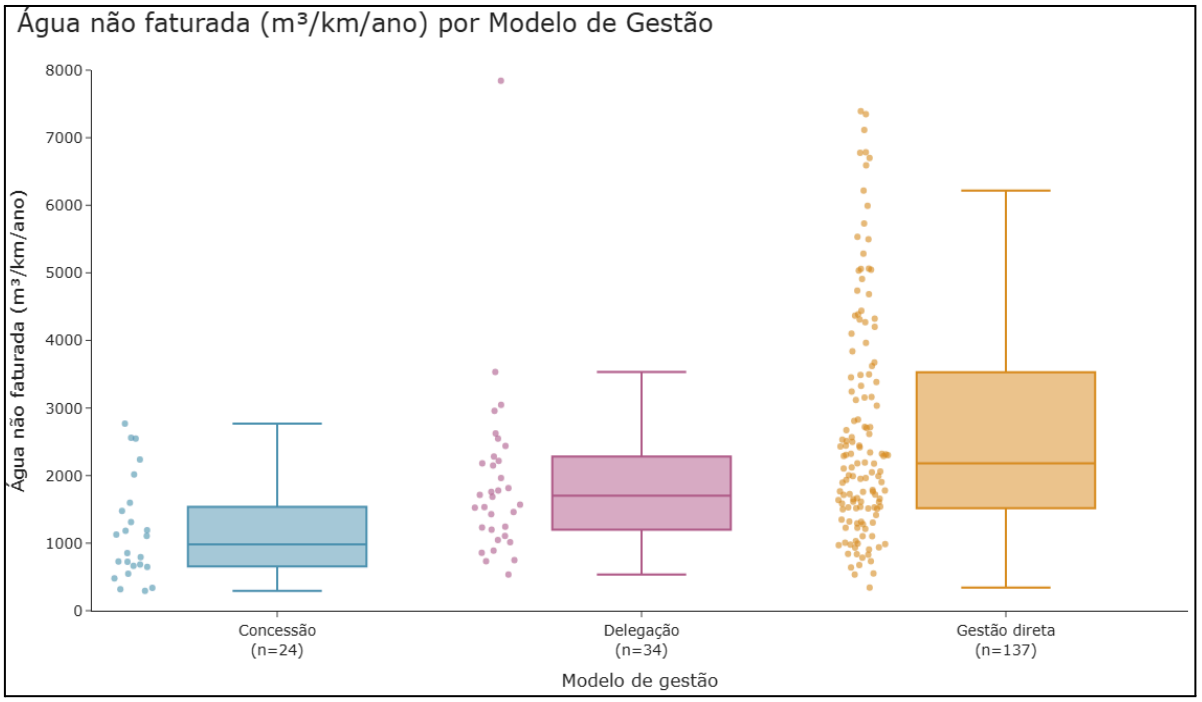

Figura 14 — Distribuição da água não faturada (m³/km/ano) por modelo de gestão

Água não faturada -> Tipologia da área de intervenção

A análise por tipologia da área de intervenção evidencia um padrão claro e consistente em termos percentuais. Como mostra a Figura 15, as áreas **predominantemente urbanas** apresentam os valores mais baixos de água não faturada, seguidas das áreas **mediamente urbanas**, enquanto as áreas **predominantemente rurais** registam os valores mais elevados. Este gradiente é coerente com a expectativa de que contextos rurais, caracterizados por redes mais extensas, menor densidade de ramais e, frequentemente, maiores dificuldades de monitorização e controlo, tendem a apresentar níveis mais elevados de água não faturada.

Em termos de dispersão, as áreas **predominantemente rurais** revelam também maior variabilidade, refletindo uma maior heterogeneidade entre entidades deste grupo. Pelo contrário, as áreas **predominantemente urbanas** apresentam uma distribuição mais concentrada em valores inferiores. Quanto aos outliers, identificam-se **dois casos extremos nas áreas mediamente urbanas** e **um nas áreas predominantemente rurais**.

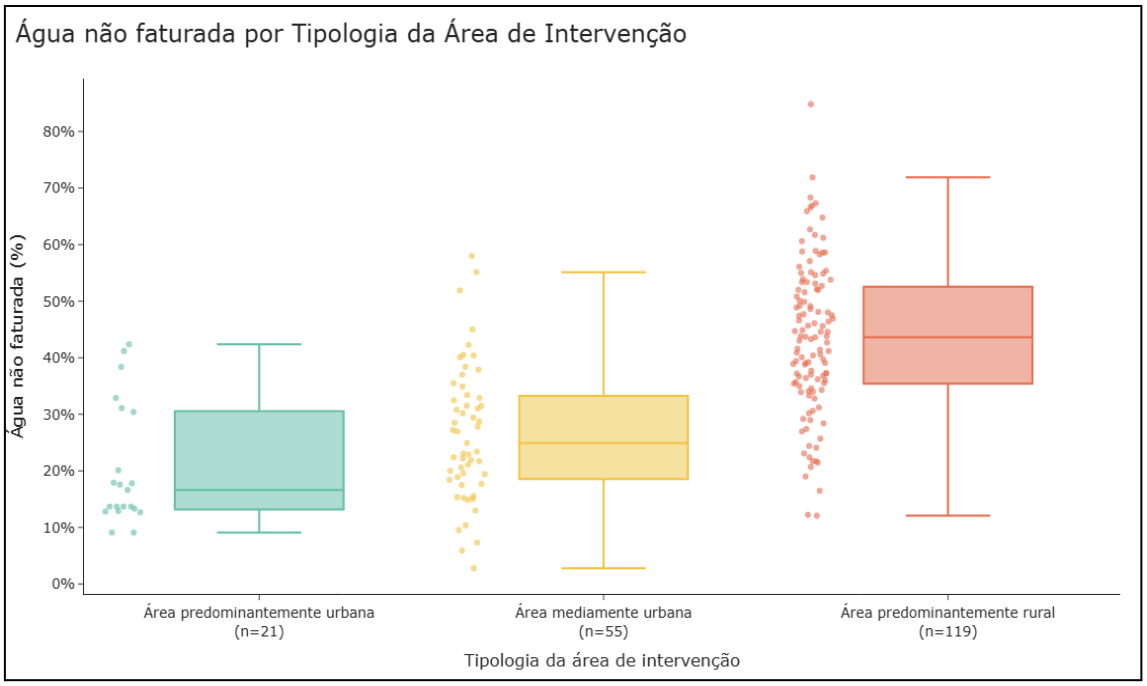

Figura 15 — Distribuição da água não faturada por tipologia da área de intervenção

Em termos absolutos, expressos em **m³/km/ano**, a ordenação dos grupos altera-se de forma relevante. Como mostra a Figura 16, as áreas **predominantemente rurais** passam a apresentar, em geral, os valores mais baixos, seguidas das áreas **mediamente urbanas**, enquanto as áreas **predominantemente urbanas** registam os valores mais elevados. Esta inversão é conceptualmente compreensível, uma vez que os contextos urbanos, embora possam apresentar menores percentagens de água não faturada, concentram volumes de água distribuída substancialmente superiores e redes com maior densidade de ramais, o que se traduz em volumes absolutos mais elevados por quilómetro de rede. Em contrapartida, as áreas rurais, apesar de tenderem a apresentar percentagens mais elevadas, distribuem volumes totais menores em redes mais extensas, originando valores absolutos inferiores neste indicador.

No que respeita aos *outliers*, observam-se vários casos extremos na **área predominantemente rural** e alguns na **área mediamente urbana**, não se identificando valores atípicos na **área predominantemente urbana**. Tal como na análise por modelo de gestão, foi omitido da representação gráfica o mesmo *outlier* de valor muito elevado, que neste caso pertence ao grupo da **área predominantemente rural**, de modo a não comprometer a leitura visual da distribuição dos restantes casos. Esta alteração de padrão é particularmente relevante do ponto de vista analítico, pois mostra que a leitura percentual e a leitura em volume absoluto captam dimensões distintas do problema. Assim, embora as zonas rurais surjam como mais críticas em termos percentuais, são as zonas urbanas que, quando se considera o volume de água não faturada por quilómetro de rede, evidenciam maior impacto operacional e financeiro.

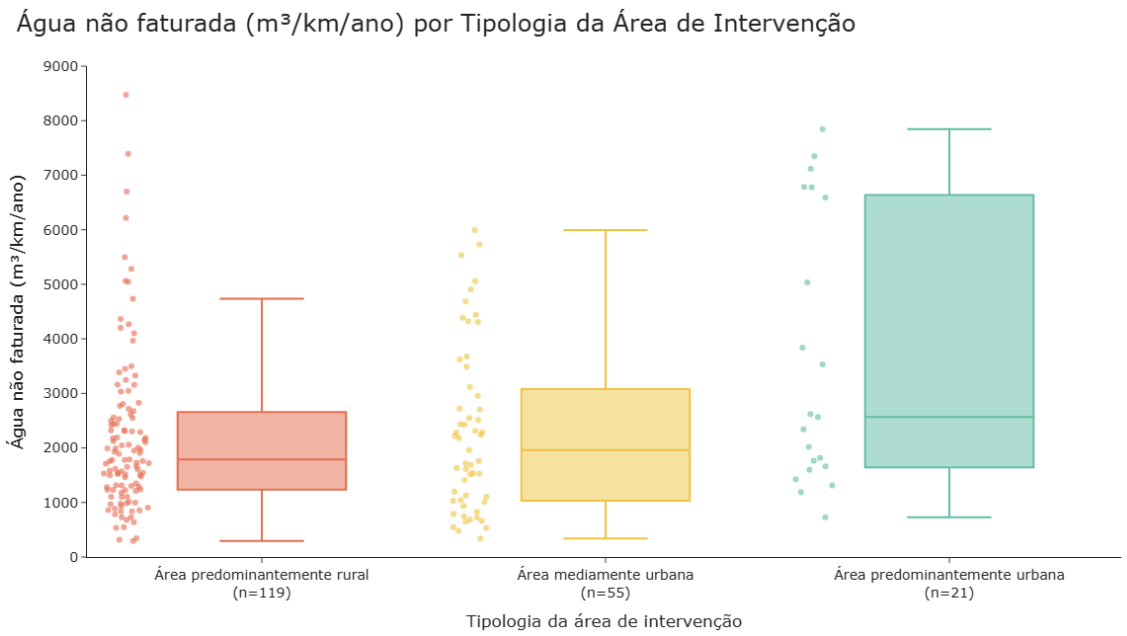

Figura 16 — Distribuição da água não faturada (m³/km/ano) por tipologia da área de intervenção

Água não faturada -> Modelo de gestão e Tipologia de área de intervenção

A análise cruzada da **água não faturada**, em termos percentuais, por **modelo de gestão e tipologia da área de intervenção** permite avaliar de que forma estes dois fatores se combinam na diferenciação do desempenho das entidades. Em vez de considerar separadamente o efeito do **modelo de gestão** ou do **contexto territorial**, esta leitura conjunta procura perceber se determinados perfis institucionais e geográficos estão associados a níveis sistematicamente mais elevados ou mais baixos de água não faturada.

Os resultados sugerem que o desempenho no controlo da **água não faturada** depende da **combinação entre estas duas dimensões**, e não de cada uma de forma isolada. Como mostra a Tabela 5, a **gestão direta em área predominantemente rural** destaca-se como o **perfil de maior risco**, concentrando os valores mais elevados de água não faturada. Este resultado é coerente com a ideia de que a **ausência de pressão contratual**, aliada às dificuldades operacionais associadas a **redes mais extensas**, **menos densas** e **frequentemente mais envelhecidas**, cria condições menos favoráveis ao controlo das perdas.

Em sentido oposto, a **concessão em área predominantemente urbana** emerge como o **perfil de melhor desempenho**, apresentando os valores mais baixos e uma distribuição mais concentrada. Este padrão sugere que a combinação entre um **contexto urbano mais favorável** e um **modelo de gestão sujeito a maior exigência contratual** tende a traduzir-se em melhores resultados. Por sua vez, o facto de a **concessão em área predominantemente rural** apresentar valores substancialmente superiores aos observados nas áreas urbanas indica que **o modelo de gestão, por si só, não é suficiente** para compensar os constrangimentos estruturais associados ao contexto geográfico.

Tipologia	Modelo de gestão	n	percentagem (%)	Média	Mediana	Mínimo	Máximo
Área predominantemente urbana	Concessão	5	23.8	11.7	12.7	9.1	13.7
	Delegação	5	23.8	14.6	13.3	12.8	20.1
	Gestão direta	11	52.4	27.3	30.4	13.7	42.4
Área mediamente urbana	Concessão	11	20.0	17.5	15.4	9.5	32.9
	Delegação	14	25.5	19.3	20.1	2.8	30.8
	Gestão direta	30	54.5	33.2	32.0	15.2	58.0
Área predominantemente rural	Concessão	8	6.7	26.0	22.8	12.1	40.6
	Delegação	15	12.6	35.3	34.6	12.2	55.1
	Gestão direta	96	80.7	46.1	45.3	20.7	84.8

Tabela 5 — Estatísticas descritivas da água não faturada (%) por modelo de gestão e tipologia da área de intervenção

Perdas reais de água -> Modelo de gestão

A análise das perdas reais por modelo de gestão, considerando **apenas as entidades com densidade de ramais igual ou superior a 20**, revela, tal como evidenciado na Figura 17, o mesmo gradiente observado na água não faturada: a concessão apresenta os valores mais baixos, seguida da delegação e com a gestão direta a registar os valores mais elevados e com a maior dispersão.

Em termos de *outliers*, a **gestão direta** e a **delegação** concentram o maior número de casos extremos, com três valores atípicos identificados. Um destes, pertencente à delegação, apresentava um valor de **1209**, substancialmente superior aos restantes e com forte impacto na escala do gráfico, tendo sido excluído da representação para permitir uma leitura mais clara da distribuição dos grupos. A Figura 17 apresentada corresponde, assim, à versão sem esse valor extremo. A concessão apresenta apenas um outlier.

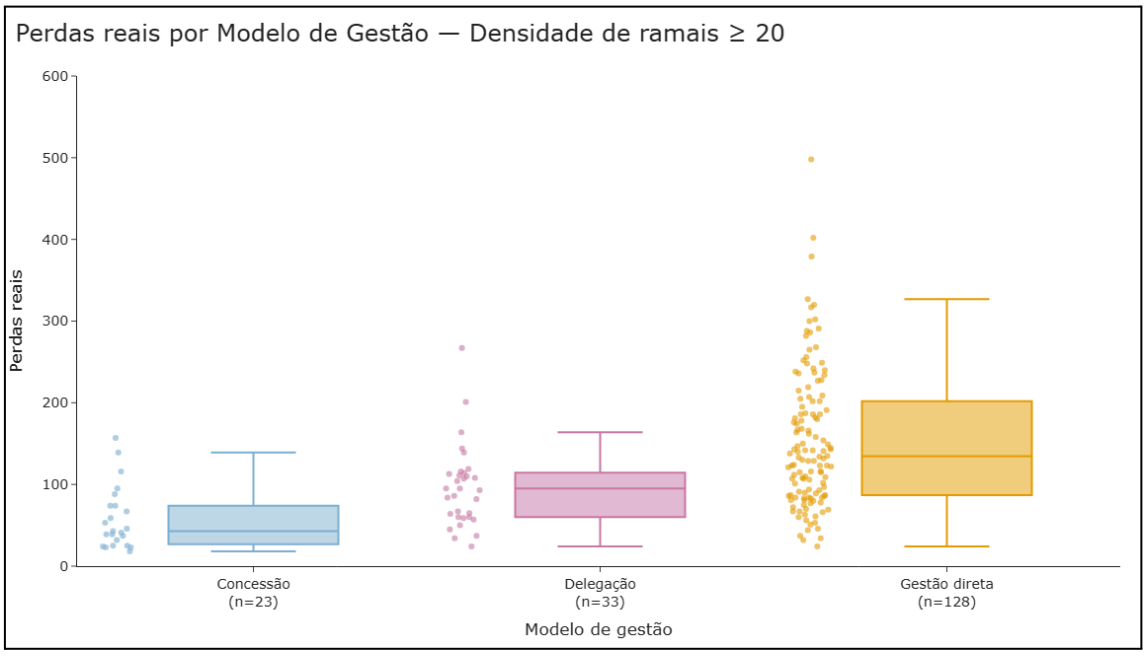

Figura 17 — Distribuição das perdas reais de água por modelo de gestão

O **CRLI** (*Current Real Losses Infrastructure Index*) é um índice que procura avaliar as **perdas reais** a partir de uma medida composta, expressa em (L/m/dia · L/ramal/dia)^1/2. Ao contrário do

indicador expresso apenas em **l/ramal/dia**, que considera unicamente o volume perdido por ramal e por dia, o **CRLI** combina duas dimensões da perda real: a **perda por metro de conduta** e a **perda por ramal e por dia**. Deste modo, permite uma apreciação mais equilibrada do desempenho das entidades, ao incorporar simultaneamente a **extensão da rede** e a **intensidade das perdas ao nível dos ramais**.

Como mostra a Figura 18, a análise por **modelo de gestão** mantém o gradiente já observado, a **concessão** apresenta os valores mais baixos e a distribuição mais concentrada, a **delegação** ocupa uma posição intermédia e a **gestão direta** regista os valores mais elevados e a maior dispersão. A delegação apresenta **três outliers**, sendo que um deles foi removido do segundo gráfico por distorcer significativamente a escala e, por esse motivo, não se encontra visível. A **gestão direta** é o grupo com maior número de casos extremos, refletindo a **heterogeneidade** já observada anteriormente. A **concessão** apresenta apenas um *outlier*, surgindo como o grupo mais homogéneo e mais consistente no controlo das perdas reais. O facto de esta ordenação se manter também neste índice reforça a ideia de que as diferenças observadas entre modelos de gestão não decorrem apenas da **dimensão** ou da **configuração física dos sistemas**.

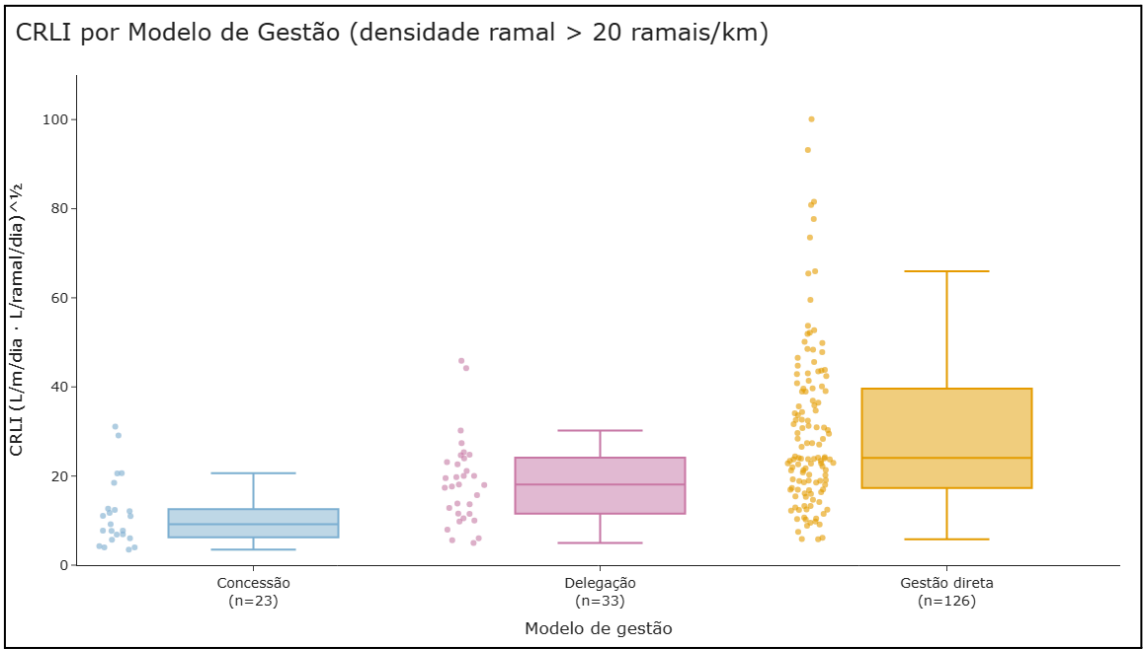

Figura 18 — Distribuição do CRLI por modelo de gestão

Perdas Reais de água-> Tipologia de área de intervenção

Ao contrário do que se observou na água não faturada, a análise das perdas reais por tipologia, considerando apenas as entidades com densidade de ramais igual ou superior a 20, não revela um gradiente claro entre grupos. As medianas das três tipologias são muito próximas entre si, com a área predominantemente urbana a apresentar ligeiramente mais dispersão, e as áreas mediamente urbana e predominantemente rural a comportarem-se de forma muito semelhante. Esta ausência de padrão ordenado sugere que, no caso das perdas reais, a tipologia geográfica não é por si só um fator diferenciador tão relevante como o modelo de gestão.

Ao contrário do que se observou na **água não faturada**, a análise das **perdas reais de água** por tipologia da área de intervenção, considerando apenas as entidades com **densidade de ramais igual ou superior a 20**, não revela um gradiente claro entre grupos. Como mostra a Figura 19, as **medianas das três tipologias são muito próximas entre si**, com a **área predominantemente urbana**

a apresentar ligeiramente mais dispersão, enquanto as áreas **mediamente urbana** e **predominantemente rural** evidenciam comportamentos bastante semelhantes. Este resultado sugere que, no caso das **perdas reais**, a **tipologia geográfica**, por si só, **não constitui um fator diferenciador tão relevante como o modelo de gestão**.

Em termos de *outliers*, os casos extremos identificados concentram-se na **área predominantemente rural**. Tal como já havia sido observado anteriormente, o **outlier mais extremo**, correspondente ao **mesmo caso já assinalado na análise por modelo de gestão**, foi **novamente removido da representação gráfica**, pelas mesmas razões, por **distorcer significativamente a escala** e comprometer a leitura visual dos restantes valores. A Figura 19 corresponde, assim, à versão sem esse valor extremo. As áreas **predominantemente urbana** e **mediamente urbana não apresentam qualquer outlier**.

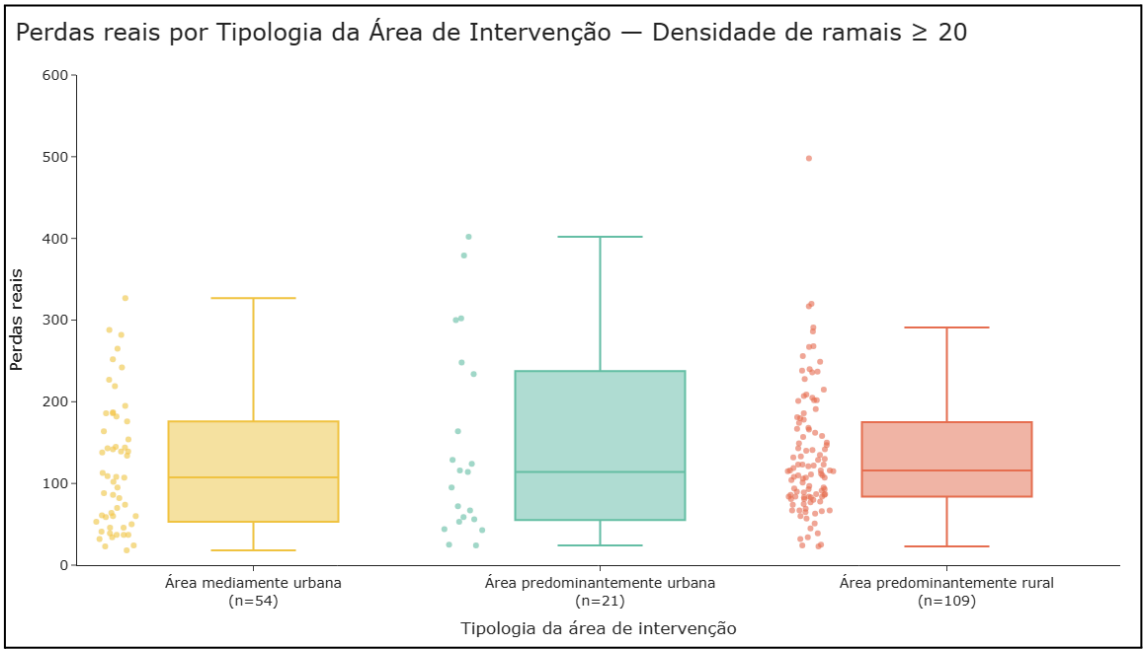

Figura 19 — Distribuição das perdas reais de água por tipologia da área de intervenção

A análise do indicador **CRLI** por tipologia mostrou um padrão ainda mais equilibrado entre grupos. As medianas apresentam-se relativamente próximas, embora as **áreas predominantemente urbanas** revelem maior dispersão e presença de valores mais elevados. As **áreas medianamente urbanas** e as **áreas predominantemente rurais** mantêm distribuições centrais semelhantes, sugerindo que a normalização das perdas em função da rede e dos ramais reduz parte das diferenças observadas na variável original.

A análise do indicador CRLI por tipologia da área de intervenção evidencia um padrão globalmente mais equilibrado entre grupos. Como mostra a Figura 20, as **medianas das três tipologias são relativamente próximas**, não se observando um gradiente claro como o identificado noutros indicadores. Ainda assim, a **área predominantemente urbana** apresenta **maior dispersão** e maior amplitude de valores, enquanto as áreas **mediamente urbana** e **predominantemente rural** mantêm distribuições centrais bastante semelhantes.

Este resultado sugere que a utilização do **CRLI** atenua parte das diferenças observadas nas perdas reais em termos absolutos. Assim, embora as tipologias troquem de posição relativa face à análise anterior, essa inversão perde relevância analítica, uma vez que os valores centrais permanecem

próximos. Em consequência, **não se confirma uma variação sistemática das perdas reais em função do contexto geográfico** quando estas são avaliadas através deste índice.

No que respeita aos **outliers**, foi novamente identificado o caso extremo já assinalado anteriormente na **área predominantemente rural**, com um valor de **CRLI = 207,9**, o qual foi **removido da representação gráfica** para não comprometer a leitura visual da distribuição dos restantes casos. A Figura 20 corresponde, assim, à versão sem esse valor extremo.

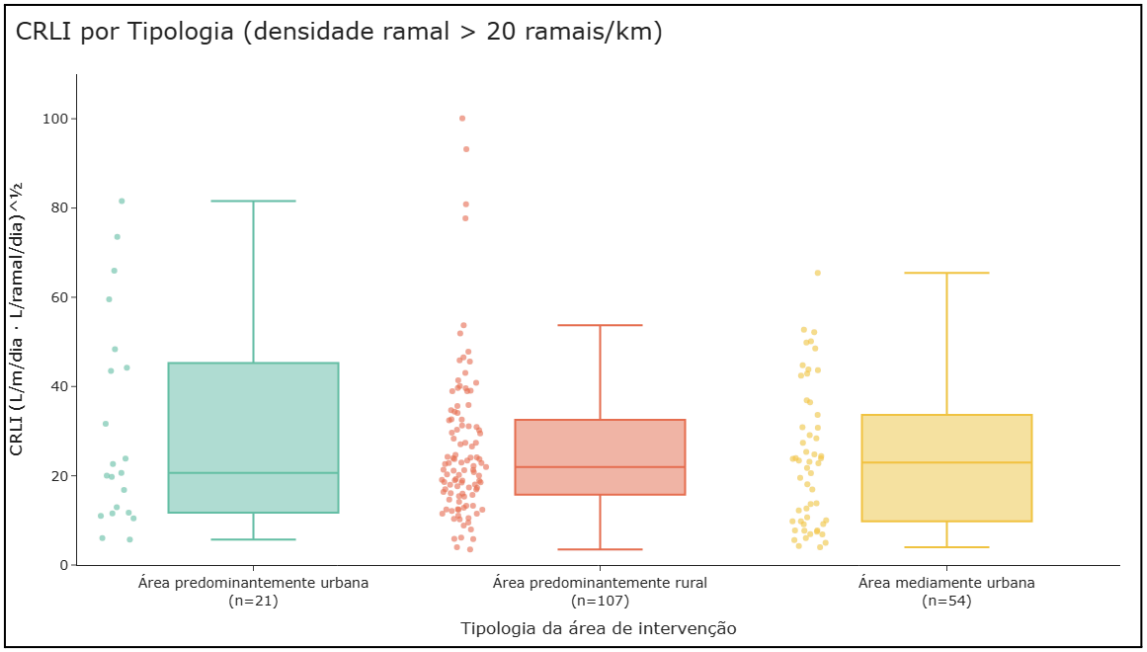

Figura 20 — Distribuição do CRLI por tipologia da área de intervenção

Perdas Reais de água -> Modelo de gestão e Tipologia de área de intervenção

A análise cruzada das **perdas reais de água**, por **modelo de gestão** e **tipologia da área de intervenção**, considerando apenas as entidades com **densidade de ramais igual ou superior a 20**, permite avaliar de que forma estas duas dimensões se combinam na diferenciação do desempenho das entidades. Em vez de considerar separadamente o efeito do **modelo de gestão** ou do **contexto territorial**, esta leitura conjunta procura perceber se determinados perfis institucionais e geográficos estão associados a níveis sistematicamente mais elevados ou mais baixos de perdas reais.

Os resultados sugerem que, também neste caso, o comportamento das **perdas reais** depende da combinação entre estas duas dimensões, embora o **modelo de gestão** surja como o fator mais diferenciador. Como mostra a Tabela 6, em todas as tipologias consideradas, a **concessão** apresenta os valores medianos mais baixos, a **delegação** ocupa uma posição intermédia e a **gestão direta** regista, de modo geral, os valores mais elevados. Este padrão indica que as diferenças entre modelos de gestão se mantêm relativamente estáveis, independentemente do contexto territorial, reforçando a ideia de que o desempenho no controlo das perdas reais não decorre apenas das características físicas ou geográficas do sistema.

Em termos mais específicos, a **gestão direta em área predominantemente urbana** destaca-se como o perfil com pior desempenho, apresentando os valores mais elevados entre os grupos analisados. Em sentido oposto, a **concessão em área mediamente urbana** surge como o perfil com melhor

desempenho, com os valores mais baixos e uma distribuição mais concentrada. Por sua vez, o facto de a **concessão** e a **delegação** apresentarem valores superiores nas **áreas predominantemente rurais** face aos contextos urbanos sugere que os constrangimentos estruturais associados a redes mais extensas e menos densas continuam a penalizar o desempenho, mesmo quando o modelo de gestão é tendencialmente mais favorável. Já no caso da **gestão direta**, a ausência de um agravamento claro nas áreas rurais indica que os problemas de desempenho poderão estar menos associados ao contexto territorial e mais ligados a fragilidades internas de natureza organizacional ou operacional.

Tipologia	Modelo de gestão	n	percentagem (%)	Média	Mediana	Mínimo	Máximo
Área predominantemente urbana	Concessão	5	23.8	52.2	43.0	24.0	116.0
	Delegação	5	23.8	99.8	95.0	59.0	164.0
	Gestão direta	11	52.4	208.2	234.0	44.0	402.0
Área mediamente urbana	Concessão	11	20.4	54.2	41.0	18.0	139.0
	Delegação	14	25.9	81.6	84.0	24.0	144.0
	Gestão direta	29	53.7	164.6	164.0	37.0	327.0
Área predominantemente rural	Concessão	7	6.4	68.6	67.0	23.0	157.0
	Delegação	14	12.8	188.6	107.0	45.0	1209.0
	Gestão direta	88	80.7	142.9	123.0	24.0	498.0

Tabela 6 — Estatísticas descritivas das perdas reais de água por modelo de gestão e tipologia da área de intervenção

7 Método e Planeamento

7.1 Planeamento inicial

Para assegurar o correto desenvolvimento deste Trabalho Final de Curso (TFC), foi necessário planear, de forma estruturada, todas as tarefas a executar ao longo do projeto. Num plano mais geral, incluem-se a revisão bibliográfica necessária para consolidar o enquadramento teórico, a redação do relatório final e as diferentes fases de entrega do TFC. Num plano mais específico, foram consideradas as tarefas técnicas apresentadas no capítulo anterior, essenciais para a implementação, análise e validação da metodologia estudada.

De forma sintetizada e já alinhada com a estrutura apresentada no diagrama de Gantt (Figura 21), foram definidas as seguintes tarefas:

- Revisão Bibliográfica
- Preparação e pré-processamento dos Dados
- Análise Exploratória dos Dados
- Proposta de um Método para a Estimativa dos Parâmetros
- Avaliação do Desempenho Global do Sistema
- Teste dos Métodos Propostos
- Escrita do Relatório
- Entregas do Relatório:
  – 1.ª Entrega do TFC
  – 2.ª Entrega do TFC
  – Entrega Final do TFC

Após a definição destas tarefas, apresenta-se de seguida o respetivo calendário de execução em formato de diagrama de Gantt, permitindo visualizar a distribuição temporal e a interdependência das atividades.

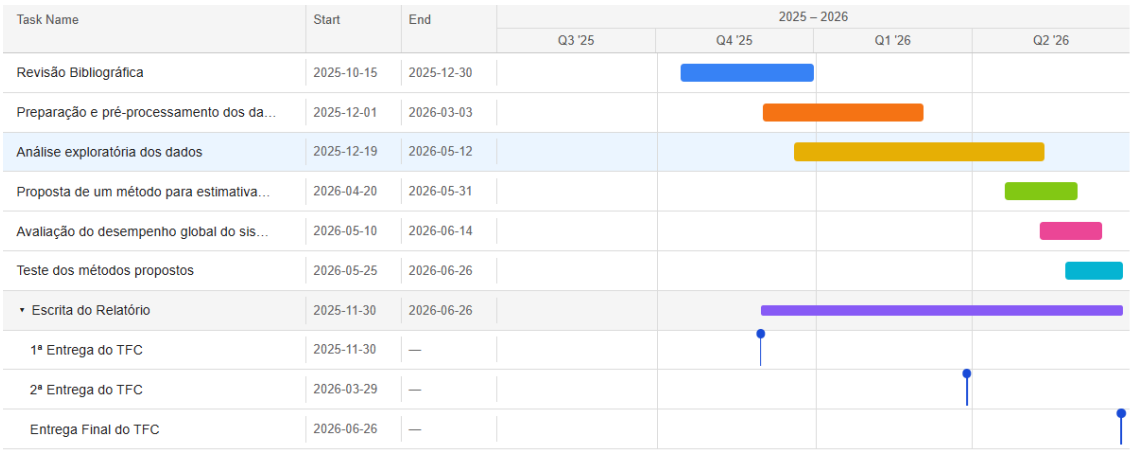

Figura 21 — Diagrama do calendário em formato Gantt 2

7.2 Análise Crítica ao Planeamento

A elaboração do diagrama de Gantt foi realizada numa fase inicial do projeto, antes de existir um contacto direto com os dados e com as ferramentas de trabalho. Esta abordagem implicou que o planeamento fosse construído sem um conhecimento aprofundado da complexidade real de cada etapa, resultando numa calendarização que, embora funcional como referência orientadora, não refletiu com precisão o esforço efetivo exigido por cada tarefa. Ainda assim, a existência de um calendário inicial revelou-se útil na medida em que permitiu estabelecer prazos intermédios que serviram de guias de progressão ao longo do trabalho.

As principais discrepâncias face ao plano original verificaram-se nas duas primeiras etapas metodológicas. A fase de preparação e pré-processamento dos dados revelou-se consideravelmente mais extensa e complexa do que o inicialmente previsto, exigindo um tratamento mais cuidado dos dados e a resolução de problemas que não haviam sido antecipados. De forma semelhante, a análise exploratória dos dados, que se ainda encontra-se em curso à data desta entrega, demonstrou igualmente uma dimensão superior à estimada, justificando o alargamento do seu prazo até meados de maio de 2026.

Em consequência destes desvios, procedeu-se à revisão do calendário das etapas restantes, redistribuindo-as de forma a respeitar a data de entrega final do trabalho, que se mantém inalterada a 26 de junho de 2026.

Bibliografia

[1] Ociepa, E. (2021). Analysis and assessment of water losses reduction effectiveness using examples of selected water distribution systems. Desalination and Water Treatment, 211, 196–209.

[2] Al-Washali, T., Sharma, S., & Kennedy, M. (2016). Methods of assessment of water losses in water supply systems: A review. Water Resources Management, 30, 4985–5001.

[3] Ociepa, E. (2019). Analysis and evaluation of water losses in the collective water supply system. Rocznik Ochrona Środowiska, 21, 1021–1035.

[4] Aclara Technologies. (2023). *Elmhurst Case Study: Water loss reduction*.

[5] Silva, M., et al. (2023). *Determinants of water loss in Portuguese utilities*. Utilities Policy.

[6] Farouk, S., et al. (2021). *Non-revenue water reduction strategies: A systematic review*. Smart and Sustainable Built Environment.

[7] ERSAR – Entidade Reguladora dos Serviços de Águas e Resíduos. (2025). *Guia Técnico 27:* Avaliação *da Qualidade dos Serviços de Águas e Resíduos – 4.ª Geração do Sistema de Avaliação*.

[8] IWA – International Water Association, Water Loss Specialist Group. (2016). Standard Definitions for Water Losses.

[9] Lambert, A. (2002). Assessing non-revenue water and its components: a practical approach. IWA Water Loss Task Force.

[10] Lambert, A. O. (2008). Infrastructure Leakage Index (ILI) as Water Losses Indicator.

[11] Vermersch, M., Carteado, F., Rizzo, A., Johnson, E., Arregui, F., & Lambert, A. (2016). Guidance *Notes on Apparent Losses and Water Loss Reduction Planning*.

[12] Adedeji, K. B., Hamam, Y., Abe, B. T., & Abu-Mahfouz, A. M. (2017). Pressure Management Strategies for Water Loss Reduction in Large-Scale Water Piping Networks: A Review.

[13] EPAL – Empresa Portuguesa das Águas Livres (2015). Active Water Loss Control. 2.ª Edição. Lisboa.

[14] Pacific Water (2015). Four methods of real loss management.

[15] Gaurav & Rathi (2025). Machine learning-based leakage identification in water distribution system. Water and Environment Journal.

[16] Javadiha, M.; Blesa, J.; Soldevila, A.; Puig, V. (2019). Leak Localization in Water Distribution Networks using Deep Learning.

[17] Sousa, D. P.; Du, R.; Silva Jr., J. M. B.; Cavalcante, C. C.; Fischione, C. (2023). *Leakage detection in water distribution networks using machine-learning strategies*. Water Supply, 23(3), p. 1115-1126.

[18] Lučin, I.; Lučin, B.; Čarija, Z.; Sikirica, A. (2021). Data-Driven Leak Localization in Urban Water Distribution Networks Using Big Data for Random Forest Classifier. Mathematics, 9(6), 672.

[19] ERSAR – Entidade Reguladora dos Serviços de Águas e Resíduos (2023). Relatório Anual dos Serviços de Águas e Resíduos em Portugal (RASARP) – Edição 2024.

[20] Algarve Primeiro (2024). Lagoa inaugurou Centro de Combate às Perdas de Água.

[21] Pearson, D., Wyatt, A., Koelbl, J., & Trow, S. (2025). *IWA Global Water Loss KPI Initiative: Meeting with ERSAR and LNEC*. IWA Water Loss Specialist Group & IWA Benchmarking and Performance Assessment Specialist Group.

[22] Leemans, B. (2025). *Workshop on the evaluation of water leakage levels: European perspective and its evolution*. European Commission, DG ENV C2 – Marine Environment & European Commission Clean Water Services.

Anexos

Anexo I:

Construção do primeiro gantt-chart

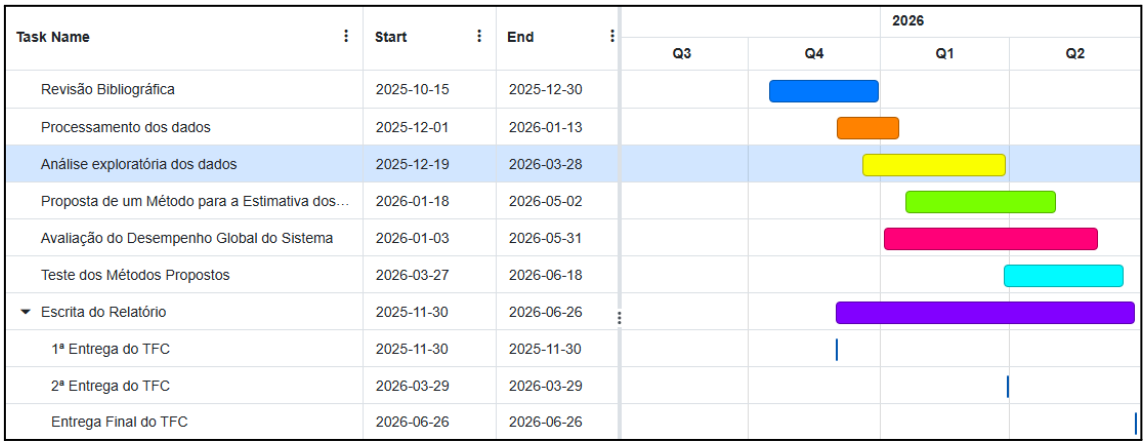

Figura 22 — Diagrama do calendário em formato Gantt 1

Anexo II:

Formulário de declaração de uso de ferramentas de Inteligência Artificial a anexar a relatório

Todos os relatórios deverão incluir anexo com cópia, devidamente preenchida, do formulário abaixo.

Assinalar as opções aplicáveis e completar os campos solicitados.

1. Utilização de IA

[ ] Não foram utilizadas ferramentas de IA na realização deste trabalho.
[x] Foram utilizadas ferramentas de IA na realização deste trabalho.

---

2. Ferramentas utilizadas

Assinalar todas as que se aplicam.

Assistência geral à escrita, análise ou ideação

[x] ChatGPT
[ ] Microsoft Copilot
[ ] Gemini
[x] Claude
[ ] Perplexity
[ ] Outras. Quais? _____

Assistência à programação / desenvolvimento

[ ] GitHub Copilot
[x] Claude
[x] OpenAI Codex
[ ] Cursor
[ ] Tabnine
[ ] Amazon CodeWhisperer / Amazon Q
[ ] Outras. Quais? _____

Geração de imagem / design / multimédia

[ ] DALL·E
[ ] Midjourney
[ ] Stable Diffusion
[ ] Canva AI / Magic Design
[ ] Outras. Quais? _____

Outros usos

[ ] Contexto: Ferramentas? _____

Figura 23 — 1ªpag. do formulário de declaração de uso de ferramentas de Inteligência Artificial

3. Fases do trabalho em que foi utilizada IA

[ ] Planeamento do trabalho
[ ] Pesquisa exploratória / levantamento inicial de informação
[ ] Documentação técnica
[ ] Redação do relatório
[ ] Desenho / modelação / arquitetura
[ ] Design / prototipagem / interface
[x] Geração de código
[x] Revisão / refatoração / debugging de código
[ ] Criação de testes / casos de teste
[ ] Análise de resultados
[ ] Preparação de apresentação ou materiais auxiliares
[ ] Outros. Quais? _____

4. Tipo de utilização

Descrever sucintamente como a IA foi utilizada.

Exemplos: brainstorming, estruturação de secções, revisão linguística, sugestão de arquitetura, geração de exemplos, explicação de conceitos, geração parcial de código, correção de erros, criação de casos de teste, apoio ao design.

A inteligência artificial foi usada como recurso complementar ao longo da realização deste trabalho, sobretudo no apoio à organização de ideias, clarificação de conteúdos teóricos, suporte na construção e ajuste de código, e auxílio na resolução de dificuldades técnicas pontuais. Também serviu de apoio na leitura e compreensão dos resultados obtidos, bem como na melhoria da redação e da coerência do relatório. Em todos os casos, coube-me analisar, validar e adaptar à informação gerada, bem como decidir sobre a sua integração no trabalho.

5. Partes do trabalho afetadas

Indicar as secções, componentes, módulos, ficheiros, entregáveis ou atividades que foram influenciados pelo uso de IA.

As partes do trabalho mais influenciadas pelo uso de IA foram os notebooks em Python, nomeadamente no apoio à estruturação, revisão e correção de código, bem como o relatório escrito, sobretudo ao nível da organização do texto, clarificação de ideias e melhoria da redação.

Figura 24 — 2ºpag. do formulário de declaração de uso de ferramentas de Inteligência Artificial

6. Exemplos de *prompt*

Inserir exemplos de *prompt*, diferenciando por âmbito (enquadrado na questão 2) e fase (enquadrado na questão 4)

Exemplos de prompts utilizados incluem perguntas como:

Podes ajudar-me a interpretar estes resultados estatísticos?

Consegues rever este código em Python e identificar possíveis erros?

Podes reformular este parágrafo em português académico, tornando-o mais claro, coeso e adequado ao contexto do relatório, sem alterar o conteúdo técnico?

7. Validação, revisão e intervenção dos autores

Descrever que verificação, revisão, correção, adaptação ou reescrita foi realizada pelos autores.

Nota: se a IA tiver sido usada em código, testes, scripts, modelos, consultas, configurações ou outros artefactos técnicos, deve ser indicado de que forma os autores validaram o funcionamento e confirmaram a sua compreensão.

Toda a informação gerada com apoio de IA foi sujeita a verificação, revisão e adaptação da minha parte antes de ser integrada no trabalho. No caso do código em Python, procedi à leitura, teste e validação do seu funcionamento, confirmando que compreendia a lógica implementada e ajustando o que foi necessário ao contexto da análise. Relativamente à interpretação de resultados e ao conteúdo escrito do relatório, revisei criticamente as sugestões obtidas, corrigi formulações inadequadas e reescrevi textos sempre que necessário, de forma a garantir rigor, coerência e adequação académica.

8. Grau de utilização

[ ] Residual
[x] Moderado
[ ] Extensivo

[ ] Utilização homogénea
[ ] Grau de uso diferenciado por fase ou componente de trabalho

Descrever sucintamente os diferentes usos.

A utilização de IA foi moderada e limitada a tarefas de apoio, como a clarificação de conceitos, a revisão e melhoria de código em Python, o apoio à interpretação de resultados e a reformulação de excertos do relatório. A IA não foi utilizada para definir a orientação do trabalho, determinar os passos seguintes ou tomar decisões metodológicas, uma vez que essas decisões foram feitas por mim em articulação com as minhas professoras. O seu uso ocorreu apenas em momentos pontuais de desenvolvimento, análise e redação, sem caráter determinante na condução global do trabalho.

Figura 25 — 3ªpag. do formulário de declaração de uso de ferramentas de Inteligência Artificial

9. Trabalhos em parceria

Protecção de dados confidenciais e recursos proprietários de parceiros

[x] O trabalho foi realizado em parceria com entidade externa ao DEISI

No caso da resposta anterior ser verdadeira, responder às seguintes questões:

[ ] O parceiro tem regras para restringir submissão de dados
[ ] As submissões validam aplicação de regras de tratamento de dados
[ ] Foram implementados mecanismos para restringir a partilha de recursos proprietários

10. Declaração de responsabilidade

Ao assinarem a presente declaração, os autores declaram que:

- a informação acima é verdadeira e reflete o uso efetivo de ferramentas de IA na realização do trabalho;
- compreendem que a IA não substitui autoria nem responsabilidade académica;
- verificaram a validaram e veracidade das referências bibliográficas incluídas no relatório
- assumem integralmente a responsabilidade técnica, científica, ética e académica por todo o conteúdo submetido, incluindo texto, código, modelos, testes, imagens, diagramas e restantes artefactos entregues.

11. Identificação dos autores

Nome(s): João Tomás Brata Rodr

Número(s): 22303390

Data: 12 / 04 / 2026

Assinatura(s): João Brata

Figura 26 — 4ªpag. do formulário de declaração de uso de ferramentas de Inteligência Artificial