UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Engenharia
Programa de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos

Daniel Bezerra Barros

LEAK DETECTION AND LOCATION BASED ON HYDRAULIC AND QUALITY MONITORED DATA

Belo Horizonte
2024

Daniel Bezerra Barros

LEAK DETECTION AND LOCATION BASED ON HYDRAULIC AND QUALITY MONITORED DATA

Thesis presented to the Postgraduate Program in Sanitation, Environment and Water Resources of Federal University of Minas Gerais as a partial requirement to obtain the title of Master in Sanitation, Environment and Water Resources.

Concentration area: Water resources

Research Line: Physical and mathematical modeling in hydraulics

Supervisor: Prof. Ph.D. Bruno Brentan

Belo Horizonte
2023

Barros, Daniel Bezerra.
B277l Leak detection and location based on hydraulic and quality monitored data / Daniel Bezerra Barros. – 2023.
1 recurso online (207 f.: il., color.) : pdf.

Orientador: Bruno Melo Brentan.

Tese (doutorado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Bibliografia: f. 179-207.
Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia sanitária - Teses. 2. Recursos hídricos - Desenvolvimento - Teses. 3. Agua - Distribuição - Teses. 4. Detectores de vazamento - Teses. 5. Teoria dos grafos - Teses. I. Brentan, Bruno Melo. II. Universidade Federal de Minas Gerais. Escola de Engenharia. III. Título.

CDU: 628(043)

Ficha catalográfica elaborada pela Bibliotecária Roseli Alves de Oliveira CRB/6 2121
Biblioteca Prof. Mário Werneck, Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM SANEAMENTO, MEIO AMBIENTE E RECURSOS HÍDRICOS

FOLHA DE APROVAÇÃO

"LEAK DETECTION AND LOCATION BASED ON HYDRAULIC AND QUALITY MONITORED DATA"

DANIEL BEZERRA BARROS

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Bruno Melo Brentan - Orientador

Profa Maria Mercedes Gamboa Medina

Prof. Cristovão Vicente Scapulatempo Fernandes

Profa Dídia Isabel Cameira Covas

Prof. Iran Lima Neto

Aprovada pelo Colegiado do PG SMARH	Versão Final aprovada por
Profa. Priscilla Macedo Moura	Prof. Bruno Melo Brentan
Coordenadora	Orientador

Belo Horizonte, 26 de janeiro de 2024.

Documento assinado eletronicamente por **Dídia Isabel Cameira Covas**, **Usuária Externa**, em 26/01/2024, às 13:39, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Maria Mercedes Gamboa Medina**, **Usuário Externo**, em 26/01/2024, às 13:39, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Iran Eduardo Lima Neto**, **Usuário Externo**, em 26/01/2024, às 14:25, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Cristovao Vicente Scapulatempo Fernandes**, **Usuário Externo**, em 29/02/2024, às 17:39, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Bruno Melo Brentan**, **Professor do Magistério Superior**, em 11/03/2024, às 10:07, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Priscilla Macedo Moura**, **Coordenador(a) de curso de pós-graduação**, em 11/03/2024, às 13:14, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2947196** e o código CRC **A92197AA**.

Referência: Processo nº 23072.200816/2024-79 SEI nº 2947196

ACKNOWLEDGEMENTS

Agradeço em meu idioma nativo para que minha família e amigos possam ler e entender o mais intrínseco sentimento que as palavras possam ter.

Quando entrei no doutorado, sabia dos desafios a enfrentar: a distância da família, a pesquisa a desbravar, pouco tempo com os amigos e ainda uma pandemia a assombrar. Mesmo assim, tive vitórias e momentos especiais com as pessoas que me acolheram, às quais vou agradecer.

Agradeço ao Bruno Brentan, que se tornou meu amigo sem deixar de ser orientador. Agradeço por estar comigo nas minhas batalhas, trazer realidade aos meus devaneios e tornar realidade meus sonhos. Vou deixar isso bem claro: sou seu primeiro aluno de doutorado, espero ter estado à altura, porque sei que seus sonhos são altos.

Agradeço à Cinthia Santos, uma amiga especial que dividiu conquistas e que me faz sempre sonhar mais longe. Agradeço também a Alan de Gois, Jonas Martins, Jordana Allagio, Luana Siebra e Luiza Virginia por tornarem a vida e a sala de estudo mais acolhedoras.

Agradeço à minha família pelo amor incondicional, por estar comigo sempre, mesmo que eu esteja muito distante. Agradeço pelas orações, pelo cuidado e pelo amparo. Hoje eu posso voar alto e longe, mas a coragem vem porque sei que há um lar que me ama e onde posso pousar.

Agradeço a Deus por sempre zelar por mim.

Agradeço também ao apoio financeiro das Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior e a Fundação de Amparo à Pesquisa do Estado de Minas Gerais ao projeto APQ-04483-18 em que este trabalho foi desenvolvido, financiado no âmbito da Chamada 503/2018, CONFAP Water JPI 2017.

Viver é um desafio
Desafiar é viver
Por isso eu vou vivendo
Sempre buscando aprender
Para não ser devorado
Pela falta do saber

Se posso dou um sorriso
Se não posso, um lamento
Mas não fico esperando
Sonhando sou avarento
E busco sonhar meus sonhos
Até no sopro do vento

Nas gotas fracas da chuva
Que a terra vai borrifando
E faz levantar o cheiro
De chuva que vou cheirando
Eu sonho dias melhores
E levo a vida cantando

Nildo Cordel

RESUMO

As perdas nas redes de distribuição de água (RDA) representam uma porcentagem significativa da água captada e distribuída. Além do valor diretamente relacionado a perda física da água, a identificação, reparo e manutenção geram custos significativos o que reflete diretamente no valor repassado aos consumidores. Por isso, essa tese apresenta métodos para a detecção e localização de vazamentos. Para isso, são primeiramente estudados os comportamentos hidráulicos durante vazamentos e como esses eventos afetam parâmetros hidráulicos como pressão e vazão e parâmetros de qualidade como a idade da água e a concentração de cloro. Os resultados do estudo indicam que os dados de qualidade da água podem ser uma fonte eficaz de informações para futuros sistemas de detecção de vazamentos. Outro aspecto abordado é a estratégia de posicionamento ideal de sensores de pressão nas redes de distribuição de água. A abordagem apresentada utiliza a teoria de processamento de sinal em grafos e algoritmos de agrupamento espectral para identificar os nós mais adequados para a instalação dos sensores. Uma métrica baseada na taxa de cobertura dos sensores é apresentada para avaliar o sistema de monitoramento de pressão e relacionam o número de sensores a taxa de cobertura, assim os operados sabem quais são as vantagens para instalação de mais sensores e quantos sensores são viáveis para um melhor monitoramento. Além disso, duas propostas para detecção de vazamentos são apresentadas, a primeira utiliza como tratamento dos dados monitorados a Análise de Componentes Independentes (IQR) e esses dados tratamos são analisados por dois algoritmos, Intervalo Interquartil e Perfil de Matriz (MP), respectivamente. A metodologia é avaliada utilizando dados de benchmark e demonstra ser eficaz na detecção de vazamentos, identificando alguns casos em poucos minutos após o início do vazamento. O desempenho varia de acordo com as características do vazamento, com o método IQR sendo mais eficiente em vazamentos com início abrupto e o método MP se destacando em vazamentos com aumento gradual no fluxo. A segunda proposta apresentada para detecção de baseia-se no processamento de sinal em grafos, onde as mudanças na pressão e qualidade da água provocadas pelos vazamentos modificam a estrutura de um grafo temporal, e a análise da estrutura do grafo permite a detecção de novos vazamentos. A metodologia é validada com sete vazamentos, apresentando uma precisão de 86% na identificação desses eventos. A pesquisa ainda apresentas as vantagens em utilizar dados monitorados de qualidade da água para a detecção de vazamentos em comparação com a utilização dos dados de pressão. Por fim, uma metodologia para a localização de vazamentos é proposta na qual utiliza um grafo multicamadas na determinação das áreas com provável vazamento. O grafo tem como camadas a representação da topologia da rede e a correlação entre os dados monitorados, sendo a conexão intercamadas dada pela área de cobertura dos sensores. O estudo apresentou uma proposta robusta e promissora para a detecção e localização de vazamentos detectando vazamentos 15 minutos após o início e com menos de 50 metros do real local com vazamento.

Palavras-chave: Redes de distribuição de água; Detecção de vazamentos; Localização de vazamentos; Teoria de grafos.

ABSTRACT

Losses in water distribution networks (WDN) represent a significant percentage of the water catchment and distributed. In addition to the value directly related to the physical loss of water, detection, repair, and maintenance generate significant costs, which directly reflect the value passed on to consumers. Therefore, this thesis presents methods for detecting and localization leaks. To this end, hydraulic behaviours during leaks are first studied and how these events affect hydraulic parameters such as pressure and flow and quality parameters such as water age and chlorine concentration. Study results indicate that water quality data can be an effective source of information for future leak detection systems. Another aspect addressed is the ideal placement strategy for pressure sensors in WDN. The presented approach uses graph signal processing theory and spectral clustering algorithms to identify the most suitable nodes for installing sensors. A metric based on the sensor coverage rate is presented to evaluate the pressure monitoring system and relate the number of sensors to the coverage rate, so operators know what the advantages are for placing more sensors and how many sensors are viable. better monitoring. Furthermore, two proposals for detecting leaks are presented, the first uses Independent Component Analysis to process the monitored data and this data is analysed by two algorithms, Interquartile Range and Matrix Profile. The methodology is evaluated using benchmark data and proves to be effective in detecting leaks, identifying some cases within a few minutes after the start of the leak. Performance varies according to the characteristics of the leak, with the IQR method being more efficient in leaks with an abrupt onset and the MP method excelling in leaks with a gradual increase in flow. The second proposal presented for detection is based on signal processing in graphs, where changes in water pressure and quality caused by leaks modify the structure of a temporal graph, and the analysis of the graph structure allows the detection of new leaks. The methodology is validated with seven leaks, presenting an accuracy of 86% in identifying these events. The research also presents the advantages of using monitored water quality data to detect leaks compared to using pressure data. Finally, a methodology for locating leaks is proposed in which a multilayer graph is used to determine areas with probable leaks. The graph has as layers the representation of the network topology and the correlation between the monitored data, with the interlayer connection being given by the sensors' coverage area. The study presented a robust and promising proposal for detecting and locating leaks, detecting leaks 15 minutes after the beginning and less than 50 meters from the actual leak location.

Keywords: Water distribution networks; Leak detection; Leak localization; Graph theory.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1

Introduction and Contextualization

1.1 Introduction

The continuous population's growth, coupled with challenges posed by climate change and water crises, added complexity to the management and control of Water Distribution Networks (WDNs). In response to the increasing water demands, WDNs are becoming more interconnected and complex, facilitating the occurrence, and spread of faults. Situations that increase the urgency in identifying, repairing, and maintaining these faults, particularly due to interconnections that not only impact water distribution but also affect road and energy infrastructures (GUIDOTTI et al., 2016; XU et al., 2019). A relevant category of faults in this context relates to water losses, which can be classified as physical losses resulting from leaks and bursts, and apparent losses related to measurement errors, fraud, and clandestine connections (STRAMARI et al., 2023).

The sum of physical and apparent losses, referred to as total losses, represents a significant proportion of the total volume of water collected, treated, and distributed in supply systems (XIN et al., 2014). According to Trata Brasil Institute in the National Information System on Sanitation (SNIS) revealed that in 2019, the average water losses in Brazilian distribution networks reached approximately 38% of abstracted and treated water (TRATA, 2020). Globally, Liemberger and Wyatt (2019) estimated annual losses of around 126 billion m³ in water systems worldwide. These estimates underscore the relevance and growing need for the development and implementation of methods that can quickly and efficiently identify and locate these losses.

Methods aimed at mitigating losses, especially in the context of leaks in the water distribution phase, have been the subject of ongoing research. A commonly employed method uses microphones and sensors to detect ground noises and correlate them with potential leaks, classified as the acoustic methods (MARTINI et al., 2016). Acoustic approaches are applied locally especially for localization, requiring on-site inspection by a specialized technical team that, through their expertise, performs leak detection and location (HUNAIDI et al., 2004). Automatic approaches, utilizing acoustic information for detection and localization, rely on correlating noises generated by water flow, with the system comparing the arrival times of these noises at sensors connected to the pipelines (HU et al., 2021). However, the effectiveness of these methods has been impaired by interference from noises emitted from other underground networks, such as sewage, gas, and electricity. Additionally, the need for regular physical inspections adds time and cost to the overall process (SAGNARD et al., 2016).

Another area of research focused on early automated leak detection and location is based on the analysis of hydraulic data from continuous monitoring, such as flow, pressure, and reservoir level data (MOUNCE and MACHELL, 2006; HU et al., 2021). The application of mathematical-statistical approaches to analyse this data has shown notable progress in this context. Effectiveness in detecting and locating leaks through data monitoring depends on optimizing sensor placement. It is essential that these devices are strategically placed to acquire data that contributes significantly to comprehensive monitoring of the entire network (MYSOREWALA, 2019, CARDOSO et al., 2021). Furthermore, the selection of ideal locations for installing sensors at certain points in the network considers the challenge associated with the high costs involved in purchasing and installing these devices (ALMAZYAD et al., 2014). The strategic selection of WDN monitoring locations is a frequent topic of studies, mainly due to the importance of monitored data in the control and operation of the entire network. Many strategies are proposed for this selection, aiming to maximize network coverage (TANYIMBOH and CZAJKOWSKA, 2018; HANH et al., 2019; LI et al., 2019), while others focus on choosing points more sensitive to anomalous events, such as leaks, pump failures, water contamination (SARRATE et al., 2014; CHRISTODOULOU, 2019; SHAHRA and WU, 2020). Additionally, clustering methods are employed to identify areas in the network with similar information, such as demand and coordinates, and, based on this clustering, choose a representative monitoring point (GIUDICIANNI et al., 2020; PENG et al., 2022).

Sensor placement studies highlight the importance of efficient and economically viable monitoring point selection strategies, emphasizing the complexity involved in finding a trade-off between network coverage, sensitivity to leak detection, and costs associated with implementing these strategies (ZHAO et al., 2020; SANTOS-RUIZ et al., 2022). Several methods and approaches have been proposed for the strategic selection of sensor placement to meet specific objectives (CASILLAS et al., 2015; HU et al., 2017). One such approach is the use of Genetic Algorithms (GA), notably developed for solving optimization problems. The GA application stands out for its ability to find optimal solutions in challenging contexts (CARRICK and MACLEOD, 2013).

Researchers have explored the application of GA in different contexts to meet specific objectives (ABOWD and MYNATT, 2000). An example is the methodology presented by Cassilas (2013), which uses GAs to optimize sensor placement, aiming to minimize the number of undetected leaks. For this, the authors used computational hydraulic simulation processes of leaks at each demand node and evaluate how many of these nodes the sensor set detects. The authors simulate leaks of different dimensions and create pressure sensitivity matrices[1] for each case of simulated leak, so GA selects a set of sensors that are more sensitive at the different flow rates tested and a higher number of nodes in each scenario. The author's methodology demonstrates the effectiveness of GAs in solving practical problems related to monitoring infrastructure in water systems.

GAs addresses problems considering a specific objective, but there are approaches that incorporate multiple objectives, and multiobjective algorithms have received considerable attention. The use of the Nondominated Sorting Genetic Algorithm II (NSGA-II) is particularly notable in this context. Illustrating this application, Cardoso et al. (2021) use NSGA-II to place quality sensors considering three objective functions that minimize detection time, maximize detection probability, and minimize the number of sensors. Ferreira et al. (2023) employs NSGA-II for the strategic selection of water quality monitoring points. The authors aim to maximize two objectives simultaneously: monitoring sensitivity to changes in pipe roughness and sensitivity to leak events detection. The authors highlight the applicability of NSGA-II in situations where optimizing multiple objectives is crucial for decision-making.

Although approaches to sensor placement have shown satisfactory results, approaches using GAs and NSGA-II require post-processing for the choice of result sets generated by these methods. In this context, Brentan et al. (2021) proposed a multi-criteria decision-making approach, which involves aggregating result sets and selecting the most efficient monitoring strategy. Additionally, it is essential to note that these methods depend on extensive computational hydraulic simulations for data acquisition and result validation, which entails a time-consuming process and significant computational effort. The subsequent need for post-processing, coupled

[1] The sensitivity matrix represents how variables (such as water flow and pressure) in different parts of the network respond to changing conditions at specific points (DEUERLEIN *et al*., 2017).

with the requirement for calibrated WDS models, has contributed to restricting the application of these methods.

Considering the limitations generated when applying approaches that require post-processing and calibrated networks, methodological proposals from other fields of research have been applied. Giudicianni (2020), for example, presents an approach that dispenses with the use of hydraulic simulations and GAs for placing water quality sensors. Instead, this approach explores the topological connectivity of the WDS, identifying and selecting central points in node clusters. The authors adopt a representation of the network as a graph[2], where vertices correspond to network nodes (demand nodes, reservoirs, and tanks), and edges represent pipes, pumps, and valves. The representation of WDS as graphs has favoured the application of methods related to other research areas, particularly those related to data analysis and object relationships. Giudicianni et al. (2022) represents the WDN as a weighted graph[3] and determines points for sensor placement through graph theory metrics that determine which pipes are most central in the network. The application of this approach can also benefit other research branches, for example, in the stages of leak detection and localization using data monitored by sensors.

The analysis of hydraulic data is increasingly employed in anomaly detection and localization, often associated with physical problems in water distribution networks. Various approaches, such as the use of data prediction methods through machine learning algorithms, have been proposed for leak detection (AYATI et al., 2022; FAN and YU, 2022; FARES et al., 2023). In these methods, algorithms estimate the values of monitored data, and any significant deviation between monitored and estimated values is interpreted as an anomaly. Another common approach involves mathematical and statistical analyses of monitored data (BUCHBERGER and NADIMPALLI, 2004; ROMANO et al., 2011; SOLDEVILA et al., 2022). In these methods, algorithms are employed to estimate average values and deviations in time series, identifying anomalies when the data does not conform to expected patterns.

---

[2] A graph is a mathematical structure that consists of a set of vertices connected by edges that represent the relationships between the vertices.
[3] A weighted graph consists of vertices connected by edges that have an associated numerical value, called a weight. This weighting adds information such as the intensity or cost of relationships between vertices.

Despite the promising application of these approaches, their effectiveness is intrinsically linked to the availability of high-quality historical data in sufficient quantity, including anomalous conditions for training and validating models (HU et al., 2021). Seeking to mitigate data shortcomings, data treatment measures, such as filling missing data (EMMANUEL et al., 2021), data segmentation to facilitate the identification of seasonality (CAI et al., 2020), and noise reduction (YUNJUN et al., 2019), are essential to ensure the robustness of these models. In this context, data preprocessing techniques, such as filtering, normalization, logarithmic transformation, and correlation analysis, have been widely adopted. These techniques aim to prepare the data for the application of mathematical-statistical algorithms in anomaly detection and localization, improving the reliability and effectiveness of the analytical process (ZHANG et al., 2019).

Approaches related to graph theory have stood out as a subject of study in data preprocessing. Research in the field of signal analysis in graphs has shown significant results in detecting anomalies in complex systems, particularly in interconnected data networks. Xie (2021) proposed a methodology to predict failures and improve satellite maintenance efficiency through anomaly detection in temporal graphs. The authors established a graph with spatial and temporal dependence, using wavelet variance to predict the structural values of the graph. The application of signal analysis in graphs in WDNs may be viable through the correlation between monitored data in creating temporal graphs. Barros (2023a) presented a methodology for leak detection, in which temporal pressure monitoring data are used to construct a temporal graph. The analysis of vertex importance in this graph is employed as data preprocessing, considering each monitoring point as a vertex, and using the correlation between data monitored by each sensor as the weight of the edges between vertices. The authors concluded that this approach enhances the efficiency of the leak detection process, highlighting that leaks substantially influence the correlation between monitored data and, consequently, the graph structure.

After the detection of anomalies, the precise localization of anomalous points is essential for repairs and maintenance. Different approaches are proposed to locate anomalies, including recognizing behavioural patterns, data classification, and prediction. Genetic algorithms and machine learning algorithms, such as the Random

Forest algorithm, have been applied to classify contamination data and associate anomalous behaviours with leak localization (Barros, 2022; Grbčić, 2021). Despite the advantages observed with the use of the Random Forest algorithm, as indicated by previous research, this method still has limitations, especially related to data acquisition. For its application, extensive hydraulic simulations and calibrated water distribution networks are required.

In this context, graph theory continues to offer advantages over these methods, especially in network representation, eliminating the need for extensive simulations. Researchers have explored graph theory to locate anomalies in various areas of study. Liang (2021) proposed a methodology to locate temporal anomalous data using multi-temporal graphs, utilizing cross-correlation between time series of satellite, energy, and steam turbine data to create the graph. Herrera (2023) presented a proposal to identify and highlight critical elements in a complex multilayered network, using distinct internet, router, and metro networks as layers of a multilayered graph. The analysis between elements of the layers, mainly in relation to the graph, identifies critical interlayer vertices and edges, prioritizing management measures.

As evidenced, the application and robustness of anomaly detection and localization methods in WDNs can be enhanced by incorporating different signal sources. In addition to conventional hydraulic data, other information sources play a crucial role in the leak detection and localization process. Recent studies propose integrating water quality sensors, along with specific analytical approaches, mainly to identify and locate contaminations.

However, it is noteworthy that water quality data, such as chlorine concentration, turbidity, and conductivity, constitute additional sources of information that may be more sensitive to leaks than traditional pressure and flow data. The study conducted by Barros (2023) exemplifies the impacts of simulated leaks on water age and chlorine concentration. The results indicate that, due to changes in water flows necessary to meet demands arising from leaks, water quality parameters can be affected more significantly than monitored pressure data.

Considering the above, the effectiveness in leak detection and localization demands the implementation of different stages and sources of information. In this context, this thesis addresses proposals related to the strategic placement of monitoring sensors, the impacts of leaks on water quality, specific methods for leak and anomaly detection in data, and, finally, a methodology for the precise localization of leaks. Each of these approaches is detailed in separate chapters of the thesis. Chapter 2 presents leak modelling methods and an analysis of water quality behaviour in simulated leak scenarios. Chapter 3 describes a methodology for leak detection through graph signal processing, using pressure data. Chapter 4 proposes a methodology for the strategic placement of pressure monitoring sensors, using graph signal processing techniques and signal sampling methods. Chapter 5 addresses leak detection, also based on pressure data, applying statistical methods for anomaly identification in the data. Chapter 6 presents an approach for leak localization using multilayered graphs and a process for determining similarity between simulated and monitored data. Finally, Chapter 7 discusses the contributions presented throughout the body of the thesis.

1.2 Objectives

1.2.1 General objective

The objective of this thesis is to develop a methodology for the detection and localization of leaks in water distribution networks through signal processing on graphs and patter recognition. For such development, simulations of leaks in the water distribution network studied are used to obtain data that will be used throughout the process. Therefore, the work is composed of subprocesses that can be characterized by the following specific objectives:

1.2.2 Specific objectives

• Evaluate the equations proposed in the literature for leak modelling, as well as assess the impacts of each leak model on hydraulic and water quality parameters, specifically, free chlorine concentration, and water age.

• Water pressure sensor placement: develop a database containing hydraulic and water quality behaviour through computational simulations and, based on this data, develop a sensitivity analysis, and apply signal sampling techniques to place sensors, aiming to maximize the coverage of the water distribution network.

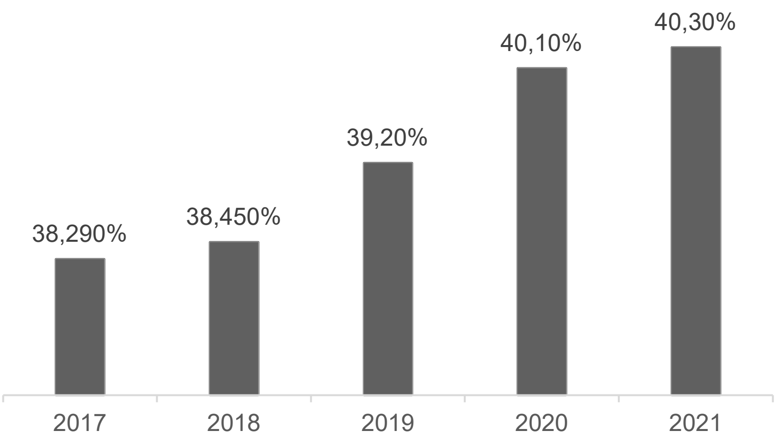
• Detect leaks by identifying anomalies on pressure and quality sensor information using graph theory and statistical approaches.

• Localize leaks using graph theory approach and data similarity algorithm, to associate them with leak locations and pinpoint them with maximum precision.

1.3 Justification

The efficient management of water resources is a crucial demand in response to the growing challenges associated to water scarcity and the need to preserve and maximize the efficiency existing infrastructures (XIANG et al., 2021). Brazilian water networks in which, even with the development of techniques for detection, localization, and repairing leaks, there is still an increasing increase in losses in the distribution stage (Figure 1.1), of the water that was collected and treated (TRATA, 2023).

Figure 1.1 - Losses in water distribution



Precise and effective detection, as well as the localization of leaks in water distribution systems, are fundamental elements to ensure the sustainability of these infrastructures, minimizing water losses and environmental impacts. Conventional approaches, such as the use of acoustic equipment and geophysical methods, often face limitations in terms of precision and efficacy, especially in complex and extensive systems. Therefore, monitoring data is increasingly being explored, employing mathematical-statistical methods to analyse this data and mathematical methods for network representation.

However, the continuous expansion and constant changes in water distribution networks have made the computational representation of these networks more complex. This complexity is also reflected in monitored hydraulic data, which exhibit noise and anomalous behaviours. Seeking to address these challenges, graph theory has been applied at various stages in leak detection and localization processes. Network representation through graph theory is achieved using weighted graphs, incorporating hydraulic and physical information such as flow rate, diameter, and pipe length (GIUDICIANNI et al., 2021), while considering connectivity information and the importance of nodes in the network. In this context, part of this thesis aims to analyse the representation of water distribution networks as graphs, exploring hydraulic and connectivity information as weights of graph edges. Additionally, graph theory is applied as a pre-processing method for monitored data, creating temporal graphs, and analysing the temporal connectivity relationship to detect anomalies.

In addition to the application of this theory, detection methods using data analysis algorithms are employed to identify anomalies in pressure data and associate them with leaks in the network. Monitored hydraulic data, including flow, pressure, and reservoir levels, are widely used in leak detection and localization processes. However, as an additional data source, the inclusion of water quality data may represent an innovative potential and enhance leak detection and localization. Variations in water characteristics can be correlated with the presence of leaks, contributing, in conjunction with hydraulic data, to mathematical-statistical analysis methods. Continuous monitoring of these parameters provides information that enriches the analysis, contributing to the precision in leak detection and localization.

This research, therefore, aims to fill existing gaps in the literature by exploring the synergy between Graph Theory, statistical methods, and water quality data, offering a comprehensive and innovative approach to leak detection and localization. The expectation is that the results obtained will not only improve the operational efficiency of hydraulic systems but also contribute to the conservation of water resources and environmental sustainability. The study aims to provide a significant contribution to advancing knowledge in this area, promoting more effective practices in hydraulic system management.

2

An Investigation on the Effect of Leakages on the Water Quality Parameters in Distribution Networks

This chapter is an adapted version of Barros, D., Almeida, I., Zanfei, A., Meirelles, G., Luvizotto Jr, E., Brentan, B. An Investigation on the Effect of Leakages on the Water Quality Parameters in Distribution Networks. **Water**, v. 15, n. 2, p. 324, 2023.

Abstract

Leakages in distribution networks can reach more than 30% of the water supplied, entailing important risks for the water infrastructure with water contamination issues. Therefore, it is necessary to develop new methods to mitigate the amount of water wastes. This study proposes to seek for new sources of information that can help for a more sustainable water use. Hence, an analysis of the network is presented showing the hydraulic behaviour during leaks occurrence, putting emphasis on how these events affect and modify water quality parameters, like water age and chlorine concentration. The study enhances that water quality data can be an effective source of information in case of leaks, being a possible source of information for future detection systems. In addition, this study proposes to use graph theory on the water network. The results highlight how an analysis of the shortest path between the leak location and the reservoir could provide meaningful information for future detection systems.

2.1 Introduction

Water is a fundamental resource for life, which is why, in 2000, at the United Nations General Assembly, 191 countries agreed on seven main Millennium Development Goals (MDG) (MARINHO et al., 2020). Among them, the fundamental target of MDG-7 was to halve the proportion of people without sustainable access to safe drinking water and basic sanitation by 2015 (HUNTER et al., 2010). According to the website MDG Brazil (2013), this goal was achieved in advance by the country, but the lack of water, sanitation and hygiene is closely linked with other MDGs, such as MDG-4, which aims to reduce child mortality. According to Howard et al. (2003) the lack of water, sanitation and hygiene is the cause of several diseases that are behind the deaths of children around the world. Deaths from diarrheal attributed to insufficient water supply and sanitation is ranked as the 6th disease that most kills children. At this point, Brazil still registers an infant mortality rate (under one year old) of 15.6 and a child mortality rate (under 5 years old) of 19 deaths per thousand births (BRASIL, 2013)

Although in 2000 1.1 billion people in the world did not have access to quality water and its countless benefits, paradoxically, the waste and losses of treated water in Water Distribution Network (WDN) in several countries are a real problem. According to the World Health Organization (WHO) large cities in Africa, Asia, Latin America, the Caribbean and North America have an average loss of water in the distribution system of 35%, with Latin America leading this front with 42% (WHO, 2009). Rebouças (2003) estimates that this loss may be even greater in Brazil, reaching 60% in some cities, against the benchmark values of 5% to 15% in developed countries.

In the report produced by Trata Brasil (BRASIL, 2019) on water losses caused by leaks in pipes, lack of flow meters, measurement errors, clandestine connections, and water theft in Brazil, only in 2022, the losses represented 40.25% of water volume in the distribution. The study is obtained from the database of the National Sanitation Information System (SNIS) and showed that losses are equivalent to 6.5 billion cubic meters. This value represents an average loss in total revenue of 39.02%, about 12 billion reais for service providers (water and sewage). It is worth noting the great social impact of these losses, whose wasted volume could supply the 13.5 million people who currently reside in favelas for approximately 2 years.

It is worth mentioning the measured physical losses, which are losses caused by leaks in the network and branches, structural leaks, overflows, and discharges during the processes of raw water adduction, treatment, storage, treated water adduction and distribution in the network, which directly affect the relation between demand and production of water, which resulted in values close to 3.4 billion cubic meters of water lost.

Physical losses represent an increase in the energy cost of pumping water, overuse of production and distribution systems and higher cost for managing the environmental impact of the activity (BRASIL, 2019). Detecting these leaks is essential to ensure reduction of these impacts. For this, different methods are used, one of these for example is the ground penetration radar. This method consists of analysing cross-sectional soil profiles around the pipes to detect water leaks (GOULET, COUTU e SMITH, 2013).

The ground penetration radar method is named as external detection approach, as well as the acoustic methods that aim to identify leaks by anomalies in sound waves that can travel through pipes and/or surrounding surface when pressurized water leaks through an orifice (ASLAM et al., 2018; KHAN et al., 2021). These techniques can identify even small leaks, but they have several disadvantages: high time consumption, difficult application in large areas, dispersion of the acoustic signal, etc. (KHULIEF et al. 2012; GOULET, COUTU e SMITH, 2013).

The second major category of leak detection uses techniques that are based on the continuously monitoring of internal parameters of the pipe, such as water speed or pressure, using monitoring sensors (ABDULSHAHEED, MUSTAPHA e GHAVAMIAN, 2017; SADEGHIOON et al., 2018). Taking advantage of the large number of measurements that can be provided, the Inverse Transient Analysis (ITA) method uses data to simulated transient events in pipes looking to improve leak detection (CAPPONI et al., 2017; DIAO et al., 2019). A variation of the ITA is the Pressure Flow Bypass Method. A leak is considered identified if there are pressure deviations between the two edges of the pipe. The efficiency of these models, as well as of several other internal detection techniques, depends on the efficient location of the sensors and mainly on the level detail of models to reduce the errors between simulations and the actual measured values (ABDULSHAHEED, MUSTAPHA e GHAVAMIAN, 2017).

Continuous monitoring generates a large amount of data, and this allows the application of techniques related to data mining. These techniques are called Data-driven approaches and look for outliers that deviate from data behaviour patterns and associate them with anomalies. Wu and Liu (2017) classify them into three categories: classification methods, predictive classification methods, and statistical methods.

Classification methods generally use data-driven approaches to learn the normal behaviour of the network, and thus being able to recognize and classify abnormal events like leaks (ZANFEI et al., 2022). Predictive classification methods differ from the classification and rely on the availability of normal hydraulic data (i.e. not affected by anomalies). The normal data are therefore used to create a model for making predictions and are used techniques like Kalman filter (JUNG and LANSEY, 2015) and Support vector machine (XU et al., 2021) to predict the data value and if they have a difference greater than a given threshold an anomaly is detected. Finally, statistical methods directly use the discrepancies caused by leakage in the measured data to detect anomalies. For this, methods such as standard deviations (ROMANO, KAPELAN and SAVIĆ, 2014) and independent component analysis (PEREIRA et al., 2021) have been used in literature.

The monitoring methods presented so far for leak detection purposes only considers hydraulic data, like pressures and flow rates. However, there is also a growing concern about the quality of distributed water, such as water contamination by organic and non-organic pollutants such as arsenic, copper, pesticides and trihalomethanes (HUNTER et al., 2010). Bangladesh had the biggest fight against contaminated water in history. The country was plagued with the contamination of its surface waters by microorganisms, which caused acute gastrointestinal disease, a range of other diseases and even death, affecting mainly children. In 1970, the United Nations Children's Fund (UNICEF) together with the Department of Public Health Engineering, mobilized to build tube wells. In 1983 the first arsenic contamination patients were identified and several studies to identify the magnitude of the situation were carried out in universities and laboratories. It is estimated that about 42 million people have been exposed to concentrations above 10 μ g/L, which is the maximum level recommended by the WHO. Symptoms of arsenic contamination include Skin lesions and different types of cancer (bladder, lung, liver, and kidneys) (SMITH, LINGAS and RAHMAN, 2000).

Although the mass contamination in Bangladesh was not caused by a direct intrusion into a WDN, the risk cannot be underestimated. Fox et al. (2016) demonstrates that an external contamination can enter the network and remain at the point of generating quality drops during short-term transient pressure events. In Fox et al. (2016) and Collins and Boxall (2013) show that the presence of contaminants also below the network should not be neglected, as the influence zone does not depend on its position. The authors emphasize a greater dependence on the distance between the contaminant and the entry orifice, the porous medium around the pipe and the orifice' size.

The contaminants detection in a network is based on the premise that a contaminant injected into a WDN, whether deliberately, accidentally, or naturally, will affect at least one of the monitored parameters (ARAD et al., 2013). In their work, Perelman et al. (2012) demonstrates, using as a parameter the total chlorine, electrical conductivity, pH, temperature, total carbon, and turbidity, that even when a parameter does not detect the presence of contamination, other parameters detected it and therefore the model obtained a satisfactory result. However, to avoid false positives, they assumed that at least two parameters should detect the contamination. Because of this, efficient sensors placement that monitor these parameters can be a key factor for probability and detection time to ensure the fewest number of affected consumers. Therefore, sensor placement is usually addressed by multi-objective formulation (ARAD et al., 2013).

Given the risks of contaminant intrusion and its massive health consequences, quality must be a factor when analysing leaks in WDN. According to Kumar et al. (2010), due to the chlorination carried out during the adduction and throughout the system, it is possible to correlate the loss of water caused by leaks with the loss of chlorine injected into the network. Therefore, obtaining data from monitoring stations can serve as a basis for identifying any anomaly in chlorine levels. Additionally, due to the dynamic hydraulic flow observed in a network, leaks at different points produce different effects on water quality, reinforcing the potential of this parameter in identifying leaks.

Even if much research is proposed at the literature for detecting leaks and other physical anomalies in water systems, most part of them are focused on processing hydraulic data (e.g., flow, pressure, tank level) instead using water quality data (e.g., free chlorine concentration, pH, turbidity). Since the water quality parameters is not often monitored as the hydraulic one and the modelling of mass transportation is much more complex than hydraulic equations, water quality parameter are not explored to their full potential. For better understanding the water quality changes due to the presence of leaks in water distribution systems, this work investigates the effects of leaks on water quality parameters and the possibility of using water quality monitoring data for detecting leaks. Thus, this work simulates computationally leaks and conceptually proves that quality data can be an additional source of information for monitoring not only water quality, but also in cases of leaks. For this, Epanet 2.2 and the python package Water Network Tool for Resilience (WNTR) (KLISE et al., 2018) are used in the simulation processes. Finally, the article also presents a methodology to determine the shortest path travelled by water between the reservoirs and the leak site. This methodology uses concepts linked to graph theory and analyses changes in flows in pipes due to leaks.

2.2 Material and Methods

To carry out the water quality study as a leak indicator, EPANET software and the WNTR package (KLISE et al., 2018) in Python environment are used. EPANET is widely used as a support tool for the analysis of water distribution systems, allowing the execution of steady and extended period simulations of the hydraulic behaviour and water quality of pressurized distribution systems (ROSSMAN, 2000). WNTR, in turn, is based on the EPANET program, but its application programming interface is more flexible, allowing for changes in the structure and operations of the network (KLISE et al., 2018).

Knowing that water quality changes due to water loss in cases of leaks (KUMAR et al., 2010), this research investigates which are the impacts on quality parameters when simulated leaks occur. Thus, to perform the simulations, two orifice equations are used to modelling the leaks. Both equations are used in leak simulations to determine which one best represents a real leak and its uses in the different stages of this research. The simulations are performed varying parameters of the leak, such as its orifice size, the duration of the leak and form of start. These parameters variations allow to have more simulation scenarios and to observe the influence of different leakages on the water quality changes.

2.2.1 Leakage mathematically modelling

The simulations are performed using the standard equation of the EPANET 2.2 Software. For the leak simulation in WDN with EPANET, emitter devices are used. According Rossman (2000) such devices are associated to junctions that model the flow through orifices or nozzles with direct discharge to the atmosphere. The flow through these devices varies depending on the pressure at the junction, according to a flow law of the type:

$$q = C_e . P^y \qquad (2.1)$$

where q is the flow rate, P is the pressure head, C is the discharge coefficient and y is the pressure exponent.

After the standard simulations, one other equation is used to compare their influence on leak detection through quality. The equation is the Standard Orifice Equation (COLLINS and BOXALL, 2013):

$$q = C_e \cdot A \sqrt{2gP} \quad . \qquad (2.2)$$

where A is the orifice area and g is the gravity acceleration. The orifice equation is chosen since the simulation in EPANET mode using the WNTR package does not allow the inclusion of leaks in specific periods of the simulation. Therefore, the orifice equation is used to determine the leak flow at specific times of the simulation, using the hourly pressure and including with additional demand on the network nodes.

2.2.2 Simulation process

Leakage simulations are made proposing three different scenarios where the location and duration of the simulated leaks changes. To obtain a more realistic and to perform a more robust investigation, all leakage scenarios are considered at 4 different nodes during the simulation period. Scenarios A and B presented overlapping leaks with a duration of 7 days, while leaks in scenario C does not overlap and have a duration of 14 days. For all simulations, the same orifice diameters are used: 10 mm, 8 mm, 15 mm and 9 mm respectively, for the four nodes selected in each scenario. These nodes are named Junction 1 to Junction 04 in each scenario and parameters assigned to each node are shown in Table 2.1.

Table 2.1 Leakage scenarios and parameters.

Scenario A				
	Junction 01	Junction 02	Junction 03	Junction 04
Nodes	188	122	50	45
	110	263	241	49
Start leakage (days)	1	2	2	5
Duration leakage (hour)	120	144	72	30
Start form	Three days of raise	Five days of raise	Abrupt	Two days of raise
Scenario B				
	Junction 01	Junction 02	Junction 03	Junction 04
Nodes	255	137	29	136
Start leakage (days)	3	2	1	4
Duration leakage (hour)	24	3	72	No end
Start form	One days of raise	Abrupt	Two days of raise	Seven days of raise
Scenario C				
	Junction 01	Junction 02	Junction 03	Junction 04
	213	45	150	156
Start leakage (days)	2	5	7	12
Duration leakage (hour)	60	21	96	67.2
Start form	Two days of raise	Abrupt	Three days of raise	One days of raise

All parameters in Table 2.1 are chosen randomly and will be used as they appear, only changing the nodes with leaks. Scenarios can also be repeated, but with the change of nodes where leaks occur. These variations in sites and parameters allow a variability to allow certain range of simulated leaks and improve the investigation process.

2.2.3 Nodes sensitivity to leakage

A sensitivity analysis is performed to determine which nodes or regions in the network are most sensitive to leaks. This will allow the determination of possible monitoring points, which will be used to expose the results in this research. For this, initially two simulation processes are performed. The first lasts 24 hours considering the network without leakages and saving hourly pressure and quality data. The second process adds a leak to a specific junction and studies its influence on the other network junctions. This process is repeated until all junctions are simulated with a leak. Each simulation also lasts 24 hours and the pressure and quality data from each junction are used to create the hourly sensitivity matrices S_{ij} determined by Equation:

$$S_{ij} = \frac{D_i - D_i^*}{q_j} \tag{2.3}$$

where D_i is the pressure or quality data of the node i without leaks, D_i^* is the pressure or quality data with leakage and q_j is the leak flow rate in node j. This process results in two square matrices, one for pressure and another for quality. Through the sensitivity matrices are possible to determine the nodes with pressure and quality most influenced by leaks. Thus, the analysis of data from the most sensitive nodes are enable a better understanding of the behaviour of pressure and quality.

After checking the sensitivity of each node, a new simulation process is performed with a different leak, following the parameters depicted in Table 2.1.

In the present paper, the water age is used initially as a quality parameter to be observed. The age of water can be determined by the EPANET software and has supported computational modelling research in water networks (SEYOUM and TANYIMBOH, 2017). The pressure and quality of the most sensitive nodes are monitored to identify the behaviour of these parameters under different leak scenarios. Thus, using mainly the daily average of pressure and quality, it is possible to mathematically identify its alteration.

2.2.4 Graph theory and shortest path

To study in more detail the relationships between leakage events and quality variations it is possible to use the graph theory, which is a mathematical approach which identifies the interactions between objects. A graph is represented by $G = (V, E, W)$ where V portrays the vertices of the graph, and these vertices are the representation of the objects; E are the edges of the graph, representing the connections between vertices; and W are the edge weights, characterizing stronger or weaker connections between vertices. A WDS can be represented by a graph considering the nodes as the vertices and the edges as the pipes. The weights of the edges can be the flow rates, head losses or roughness of the pipes (ZANFEI et al., 2022; BARROS et al., 2023).

In the present work, the graph theory is used to determine the shortest paths between the nodes and reservoirs, considering the maximum flow rates of the pipes as edge weight. Hence, a graph structure is created starting from the network topology, where the graph vertices represent the network nodes, while the graph edges represent the pipes. Furthermore, at each edge it is assigned a weight that is the maximum flow rate of the related pipe. This approach uses the Network Analysis in Python (NetworkX) Python package (HAGBERG, SWART and CHULT, 2008), which allows the creation, manipulation, and study of dynamic structures of complex networks and graphs.

The shortest path determination in a graph starts by evaluating a starting point and a destination point. After this calculation, the distances between the edges to the neighbouring vertices of the starting point are calculated. This distance can be considered using the smallest number of edges to the destination vertex, the edge weights or other factors attached to the edges. The method determines which neighbour vertex has the shortest distance to the starting point, and then performs the same process for the determined neighbour vertex, always directing it to the destination vertex. At the end of the process is obtained the edges and vertices belonging to the shortest path (MAO and ZHANG, 2013).

2.2.5 Evaluation method

To quantify the effect of different leaks on water quality parameters it is proposed to use two evaluation methods. Firstly, the use of maps to associate variation in the analysed parameters with the spatial characteristics of the network. Secondly, to calculate variation in both hydraulic parameters and quality parameters from the scenarios affected by leaks, and the normal behaviour of the network (i.e., without leaks). Particularly, to emphasize the variations in the different parameters analysed, it is proposed to calculate a $Delta$ parameter defined as the percentage difference between the value in the scenario affected by leaks, and the normal condition. This parameter is defined as:

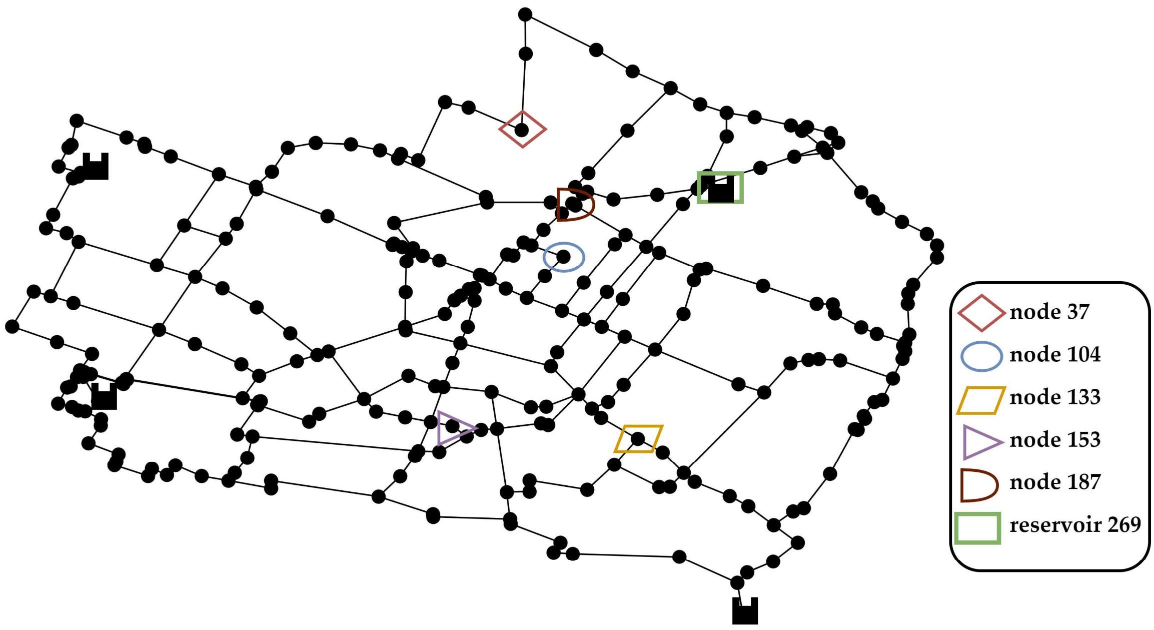
$$Delta = \frac{D-D^*}{D} x100 \quad (2.4)$$

where D is the pressure or quality data without leaks, D^* is the pressure or quality data in the scenario with leakages. Thus, it is possible to highlight and quantify the modification in the various parameters with respect to the presence of a leak or not.

2.3 Case study

The network used for this study Figure 2.1 is presented in the work of Bragalli er al. (2012), based on the city's distribution network in the Emilia-Romagna region of Italy, Modena. It has a total of 272 nodes, of which 268 are junctions and 4 are reservoirs. In addition, it features 317 pipes and has no valves or pumps.

Figure 2.1 - Modena Network.



The Figure 2.1 presents the Modena network and highlights some nodes and a reservoir, which will be explored during this paper. This network has been widely used in a plethora of problems, like sensor placement (MANKAD et al., 2022), leak detection (BARROS et al., 2023), energy efficiency (LENZI et al., 2013).

Table 2.2 reports the maximum flow losses caused by leaks in the different scenarios of the proposed investigation.

Table 2.2 - Leakage nodes and flow rates.

Scenario A				
Nodes	188	122	50	35
Flows (L/s)	2.55	1.45	4.34	1.31
Nodes	110	263	241	49
Flows (L/s)	2,36	1,38	4,47	0,98
Scenario B				
Nodes	255	137	29	136
Flows (L/s)	2.00	1.59	4.78	0,83
Scenario C				
Nodes	213	45	150	156
Flows (L/s)	2,23	1,52	3,39	1,89

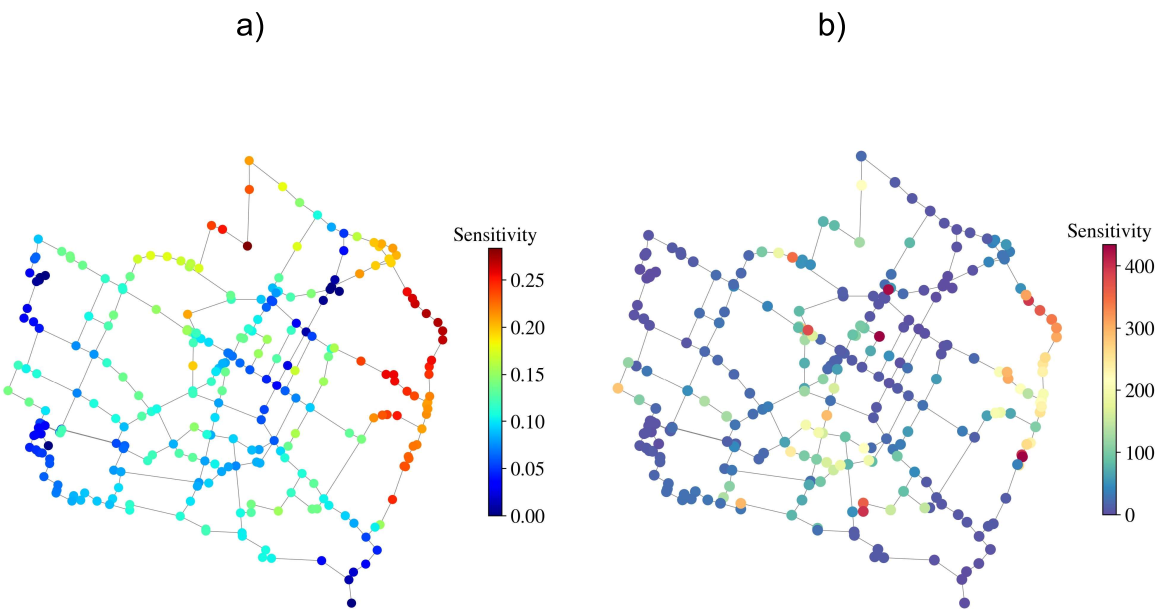
The maximum leakage flows range from 0,89 to 4,78 L/s, which represent small and large leaks according to Quinones-Grueiro et al. (2021), this represents the inclusion of different number of leakages in terms of magnitude. Comparing these values to total consumption of the network (406,94 L/s), this would mean leakage at 0,2% to 1,17% of total consumption.

2.4 Results and discussions

2.4.1 Sensitivity analysis

To evaluate the simulation results graphically, as mentioned above, two average sensitivity matrices are created, one for the water quality and another for the pressure. These matrices are constructed being the rows representing the node with simulated leak and the columns with the sensitivity data of the other network nodes, this is possible through Equation 3. Then, the average value for each column of the matrix is calculated, resulting in an average value for each node for both quality and pressure. These values are shown through the map in Figure 2.2.

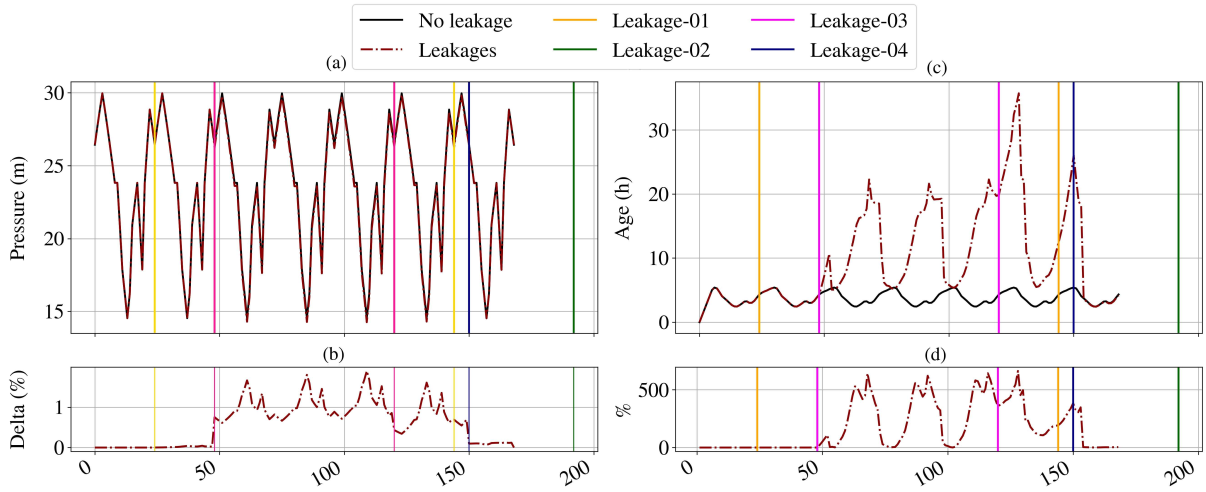
Figure 2.2 - Sensitivity map: (a) Pressure and (b) Quality.



The sensitivity for both parameters (Fig. 2.2) is lower at nodes around the reservoirs. This happens because the reservoirs are at a fixed level (constant hydraulic grade) since they supply water directly to these nodes (low head loss influence). Furthermore, the map allows to highlight some nodes that exhibit particularly different sensitivity value. A notable point is junction 104, which is highlighted in red in Fig. 2.2 This node exhibits high sensitivity for all scenarios, with or without overlapping leaks during the simulation period.

It is proposed to assign junctions 188, 122, 50 and 45 to the parameters of junctions 01, 02, 03 and 04 of scenarios A to make an additional test. Figure 2.3 show the variation of pressure and quality in node 104.

Figure 2.3 - Node 104 - Monitoring data: (a) Quality variation, (b) Quality variation delta, (c) Pressure variation and (d) Pressure variation delta.



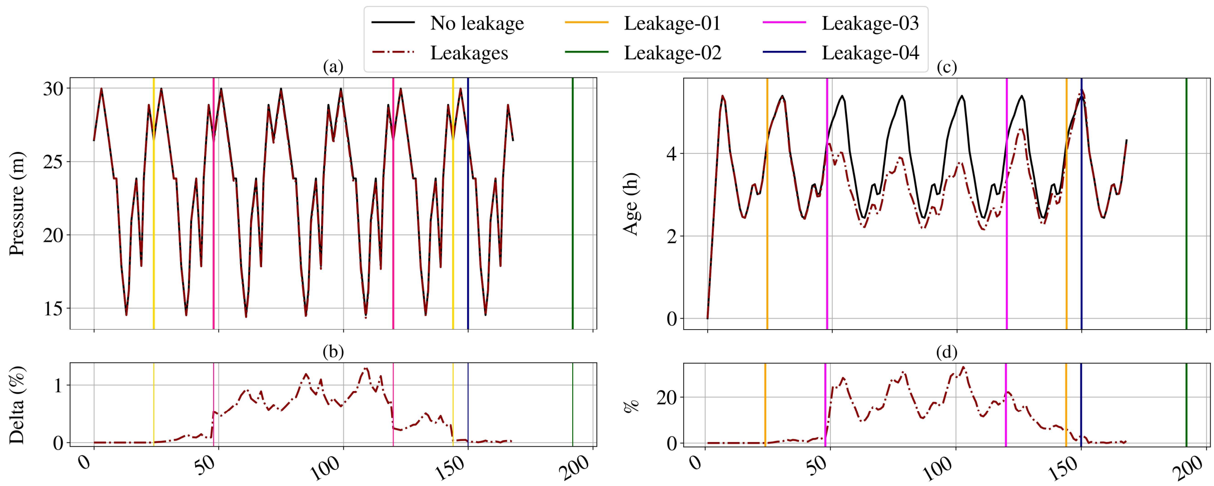
For the simulated leak scenario, the nodes had leakage flows following the Table 2.2. Even with the highest tested leak flows (4,34 L/s) the pressures in the sensor nodes exhibit a low variability. If it is much higher flows than those tested, the entire network behaviour is modified and can be easily identified by pressure and quality data. In contrast, the water age in the sensor node was significantly affected by small or big leakages. It is observed that the Delta parameter for the pressure data (Figure 2.3b) changes slightly more than 1%, while the Delta for the quality data have changes greater than 500% (Figure 2.3d). This shows that although the pressure is effective on the network coverage to detect leaks, water quality data can help to increase this value or reduce the sensors number. The study highlights that the use of the quality sensor can provide important support in leak detection, given the high variability shown during leak events.

2.4.2 Pressure and Quality

After the general analysis of the entire network, shown in Figure 2.2, simulations are carried out considering the leakage in specific nodes following the Table 2.2. To analyse the influence of simulated leaks in the network, some nodes are also chosen to monitor their quality and pressure values. In particular, the nodes 104 and 187, stand out due to their higher quality sensitivity and, therefore, are used as sensor nodes in the next stages of this research.

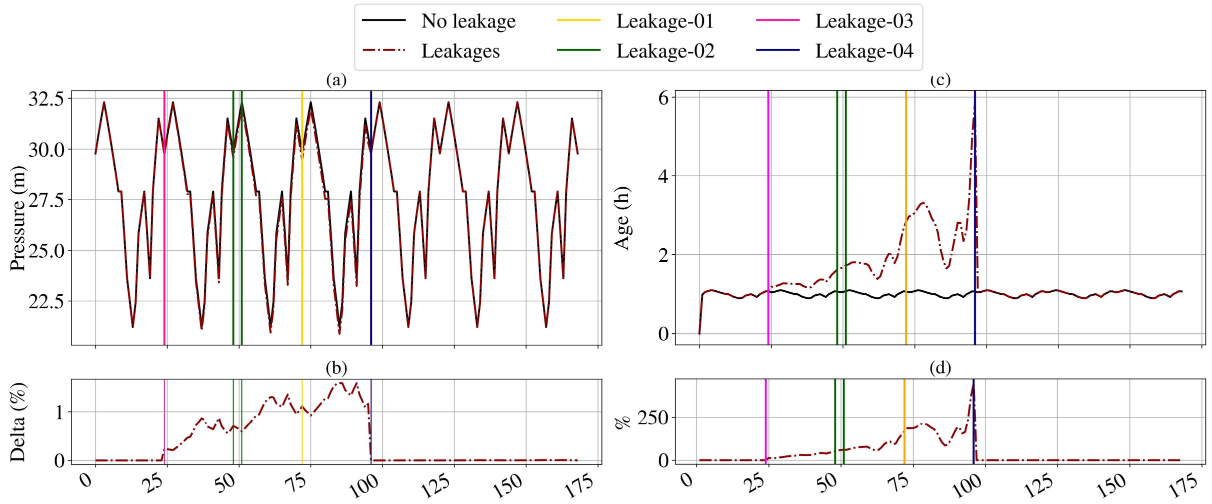
Considering the parameters of scenario, A, and assigning them the properties of "Junction 01", "Junction 02", Junction 03" and "Junction 04", to nodes 110, 263, 241 and 49, respectively, Figure 2.4 highlight the variations of quality and pressure for the monitoring node 104.

Figure 2.4 - Sensitivity of node 104 for leaks in scenario A: (a) Pressure variation, (b) Pressure variation delta, (c) Quality variation and (d) Quality variation delta.



It is worth to note that in Figure 2.4 the pressure behavior also does not exhibit significant variations. On the other hand, quality varies consistently, showing fluctuations of almost 40% its value (Fig. 2.4d), reinforcing its possible role in helping to detect leaks. Other scenarios reinforce the role of quality as a future ally for detecting leaks in networks. As further confirmation, it shown in Figure 2.5 the results related to Scenario B, linking the parameters of junctions 01, 02, 03 and 04 to nodes 255, 137, 29 and 136, respectively. Figure 2.5 shows the variations of the monitoring node 187.

Figure 2.5 - Sensitivity of node 187 for leaks in scenario B: (a) Pressure variation, (b) Pressure variation delta, (c) Quality variation and (d) Quality variation delta.



Finally, Figure 2.6 proposes the results for scenario C that assigned to nodes 213, 45, 150 and 156 the parameters of junctions 01, 02, 03 and 04 presented in Table 1. Node 153 is used as monitoring node:

Figure 2.6 - Sensitivity of node 153 for leaks in scenario C: (a) Pressure variation, (b) Pressure variation delta, (c) Quality variation and (d) Quality variation delta.

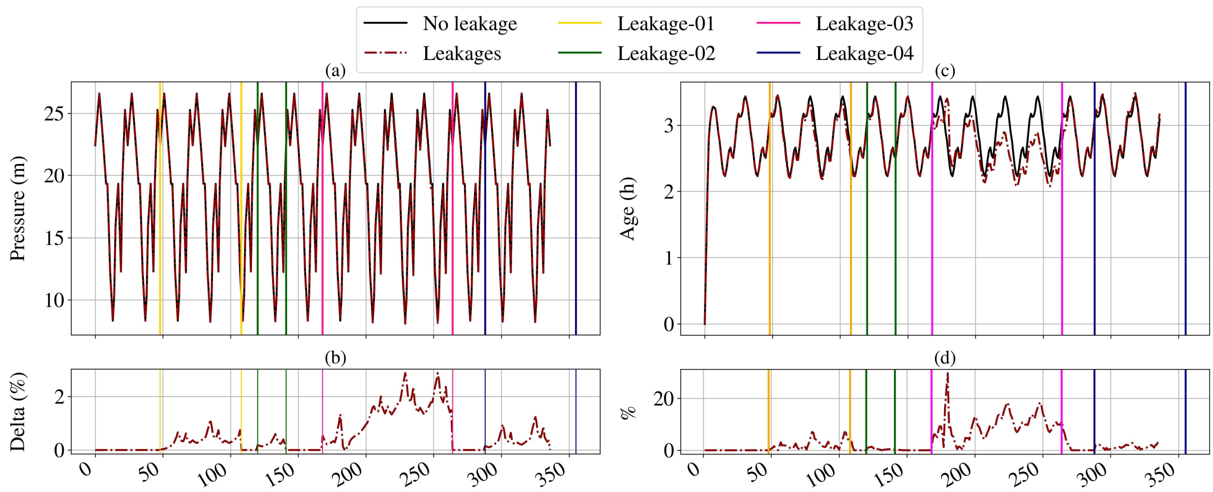


Figure 2.6b highlight that the water age changes due to leaks because, to meet the demand and leakage, the water travels different paths, in some cases increasing the water velocities in the pipes or receiving water from different reservoirs. However, water age is not a real monitoring parameter, so a simulation process observing chlorine is proposed afterward.

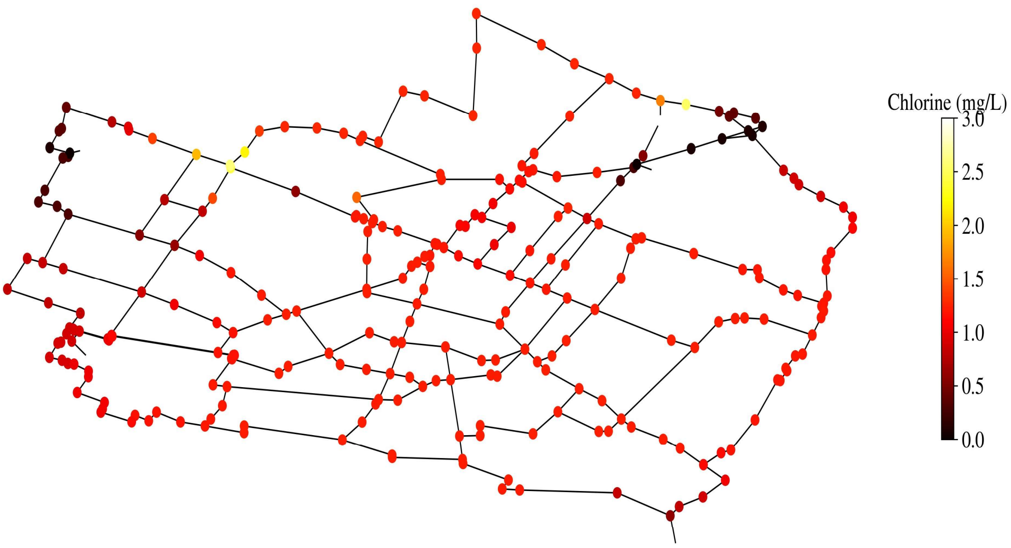
2.4.3 Chlorine simulations

Although the water age is a known parameter due to its sensible variations, it does not have real monitoring way. On the other hand, Chlorine might be well monitored, and it has been used for multiple scopes. For instance, to support the location of contamination points, to determine the water age and to ensure that the potability standards in terms of concentrations are respected. In addition, Gobet et al. (2001) present a wireless monitoring sensor that have an accuracy of 0,02mg/L, that promotes leakage detection with a slight change in the concentration of the monitored parameter (CHENG et al., 2015; CARDOSO et al., 2021).

Hence, to identify the variations of the chlorine concentration during leak events, it is proposed to perform simulations considering the reservoirs as sources of chlorine with a continuous supply of water at a concentration of 3,0 mg/L of chlorine. The simulations had the flow reaction coefficient equal to -2,5, following the default value for chlorine simulations (ROSSMAN, 2000), and the pipe wall reaction coefficient equal to 0,15 (FISHER et al., 2017).

As node 104 proved to be dominant for water age sensitivity in the previous scenarios, it is proposed to analyze this node once more. This node has a chlorine concentration that is very sensitive to leakages at almost all nodes of the network. Figure 2.7 highlight the maximum absolute differences of the chlorine concentration values at each node with the node 104. These maximum differences are evaluated considering the simulations with a leak in each node and for their whole duration. For this, a new process of simulations is performed, in this case a 24-hour simulation was performed individually at each node and the maximum chlorine concentration was observed at node 104.

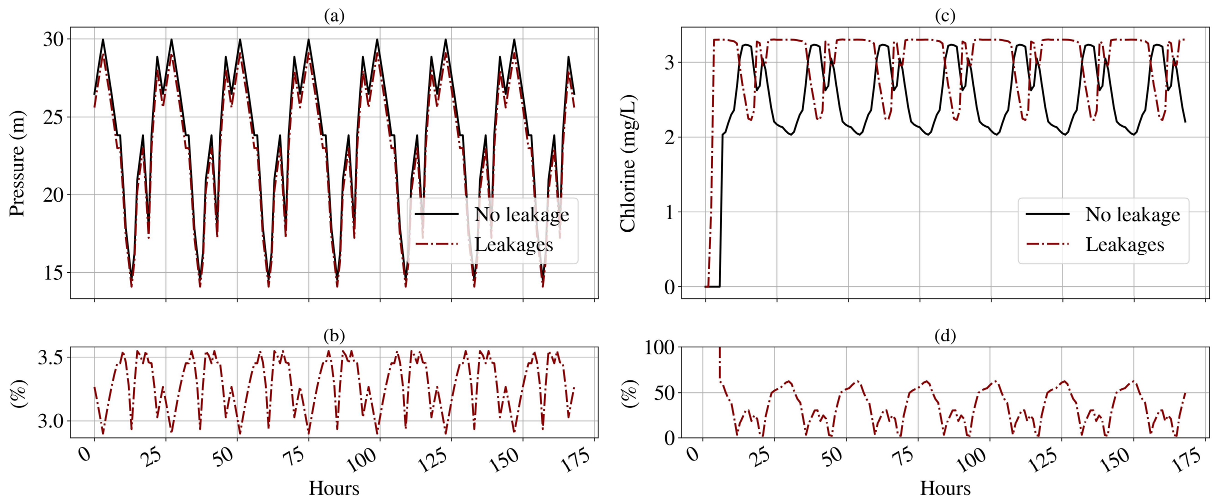
Figure 2.7- Maximum difference in chlorine concentration.



The numerical a chlorine variation shown in Figure 2.7 demonstrates that most of the network nodes leakage scenarios (86.6%) affect approximately 1,0 mg/L in the chlorine concentration at node 104. With 4 scenarios (1,5%) with values close to 2,0 mg/L and 3 nodes (1,12%) with approximate values at 3,0 mg/L. Finally, 10,78% of the nodes affect up to 1mg/L.

Considering, for example, a leak at node 37, the chlorine concentration at node 104 is completely changed (Figure 2.8). The Figure 2.8 shows that the chlorine concentration at node 104 assumes a completely different behavior depending on the presence of a leakage at node 37 or not. Although the maximum value changes slightly, passing from a 3,00 mg/L without leakage to a 3,05 mg/L with the leakage, the signal exhibits consistent variation during time. These variations may provide a significant source of information. For example, it can be observed that the concentration difference is 2.5 times greater than the sensitivity of the monitoring equipment presented by Gobet et al. (2001) (0,02mg/L). This would result into a possible detection by the opportune sensors.

Figure 2.8 - Pressure and chlorine behavior at note 104 with leakage at node 37: (a) Pressure variation, (b) Pressure variation delta, (c) Quality variation and (d) Quality variation delta.



Given that the purpose of this study is to support leak detection tasks, and if a sensor with a sensitivity of 0,02 mg/L could be placed at node 104, the inducted changes in the chlorine concentration due to a leakage could be theoretically detected in 99,5% of network nodes. These monitoring data can be used coupled with data-driven techniques and hydraulic simulations so that, through them, information such as leak location, determination of leak demand and chlorination failure may be rapidly and accurately identified and corrected.

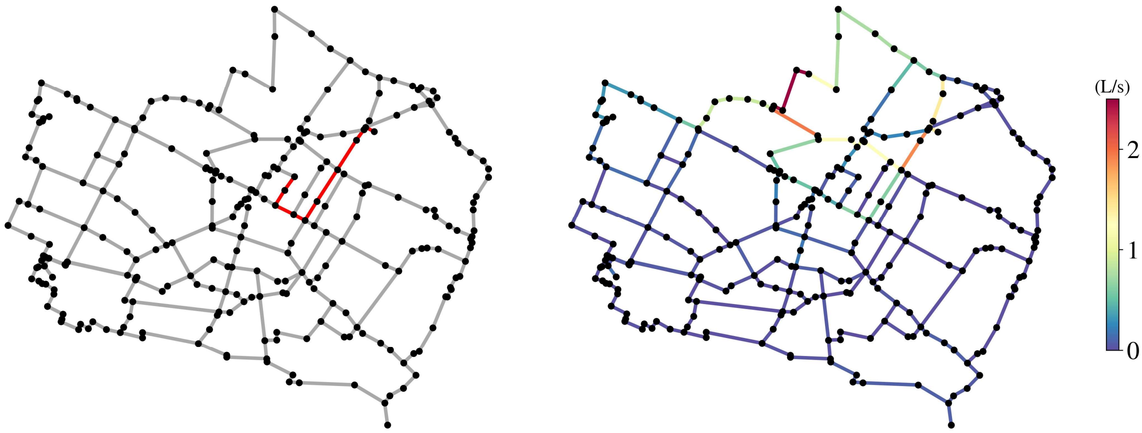
The important variability highlighted in Figure 2.7 and Figure 2.8 was expected also for chlorine, because the mass transfer is altered when changing the pipe flow rates to meet the leak demands. To assess the changes inducted by leak events in the chlorine concentration values during the simulations, it is proposed to use the graph representation of the network. In particular, the following section provide an analysis of shortest path between the reservoir and node sensors.

2.4.4 Shortest path and Flow changes

For this application the Modena network is represented as a graph where the nodes are represented by vertices, the pipes are represented by edges and the maximum flow rates were used as edge weights. With these assumptions, Figure 2.9 show the shortest path results, as well as the maximum flow rates differences for each pipe due to a leakage.

Figure 2.9 - Network behavior flow – (a) Shortest path – reservoir 269 to node 104 and (b) Flow differences – leakage node 37.

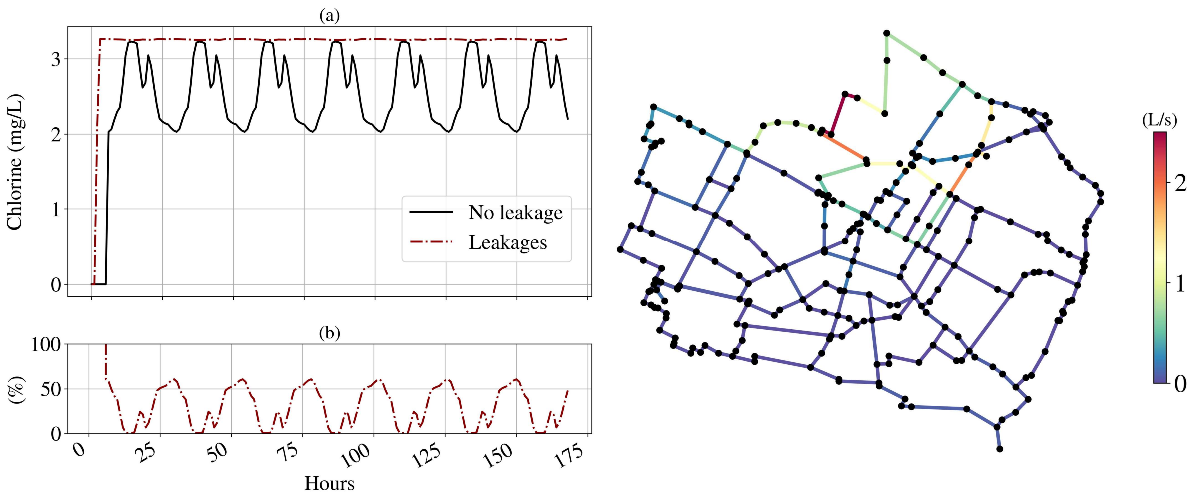
a) b)



The Figure 2.9a show the shortest path found through between reservoir 269 and node 104, with the entire total amount of water needed to supply the node's demand coming from reservoir 269. This total amount is determined through the Trace function of the Epanet software. Figure 2.9b shown the differences, at each pipe, between the maximum flow values without leakages and the maximum flow values with a leak in node 37. It can be noticed that there are changes of up to 2,5 L/s, and that there are changes in the flow rates in the pipes present in the shortest path between the reservoir and node 104. Thus, these changes in the flow rates in the pipes that belong to the shortest path explains the variations highlighted in the quality parameters analyzed in Section 2.4.2. Another leak is afterward simulated individually at node 133, and Figure 2.10a) shows the variations of chlorine concentration in node 104, as well as the map of the variations of the flow rates during this different leak.

Figure 2.10 - Behavior of quality and flow with leakage 133 – (a) S Chlorine – node 104 and (b) Flow differences – leakage node 133.

a) b)



It can be observed in Figure 2.10 that, once again, the behavior of the chloride concentration at node 104 is altered, even with a leak simulation in a more distant node and that receives a smaller contribution of water from reservoir 269. This behavior supports more the use of quality data for leak detection, since it was proven that points in the network can suffer great changes in quality, even with distant leaks, which perhaps wouldn't affect the pressure, for example, of this node.

2.5 Discussions and partial conclusion

The water loss and contamination in the Water Distribution Networks is a global problem that needs attention for the techniques development that help to reduce such damages. Water quality data provides an abundant source of information for the leak location and optimization of the sensor placement in the Water Distribution Network. In addition to the high sensitivity of such a parameter, the equipment for its monitoring is already on the market with a high accuracy, confirming the importance of exploring this aspect.

This research is a proof of concept on water quality monitoring also for leak situations. The results show, through computer simulations and the observation of water age and chlorine concentration, that the monitored parameters of water quality undergo changes greater than pressure (data usually used for detection) and proved to be an additional and effective source of information in case of leaks. The change of water paths in the pipes to meet the leak demand modifies the entire behaviour of the quality parameters change, and in most cases a significant change is observed. This change only does not occur mainly in nodes immediately connected to reservoirs or very close. Even so, it stands out in this proof of concept that water quality data can be used for various purposes, such as for detecting and locating leaks and failures in the operation of pumps and valves. The behaviour of flows was proven by evaluating the shortest path, by representing WDS as a graph. Through a simulation of leaks in all nodes, it was possible to observe that water quality is strongly affected in some nodes. Thus, a small number of sensors placed in the WDN at optimized points, as happened with node 104, could mean a great efficiency in detection leaks and a great saving for service providers.

How to monitor the chlorine concentration is effective for detecting leaks, other factors may be explored, such as chlorination failures, pump, and valve operation. An additional point that can also be explored is the quality sensors placement to detect and locate leaks.

3

Novel pressure sensor placement method for leakage detection in water distribution networks using graph signal processing and sampling theory

This chapter is an adapted version of Barros, D., Giudicianni, C., Creaco., E. Meirelles, G., Brentan, B. Novel pressure sensor placement method for leakage detection in water distribution networks using graph signal processing and sampling theory. **Water Resources Research**, 2023 (Under review).

Abstract

A suitable strategy for the optimal pressure sensor placement in water distribution networks (WDNs) allows water managers and engineers to better monitor and control the system also in terms of leakage detection and reduction. This paper addresses the task by adopting the innovative graph signal processing theory. First, Spectral clustering algorithm is applied to separate nodes with similar sensitivity and coordinates and creating a graph for each cluster. Then, three metrics from signal sampling on graphs (SSG) theory are employed for selecting nodes as potential sensor position. A novel metric, based on coverage rate of placed sensors, is presented for the evaluation of the pressure monitoring system, which also allows the definition of the most appropriate number of sensors to be installed. The proposed approach is applied to two WDNs and compared with other methods known in the scientific literature. Results show the effectiveness of the method in placing pressure sensors ensuring an extensive coverage to detect leakages with a smaller number of sensors.

3.1 Introduction

Due to water crisis, climate changes, population growth and more and more interconnected and complex systems, efficient management of water distribution networks (WDNs) has become a crucial factor in ensuring the provision of clean water to the citizens in ever-larger cities (GIUDICIANNI et al., 2020). Current WDN management strategies seek to monitor and control the main hydraulic variables, such as pressure, flow, reservoir levels as well as the settings of pumps and valves (PÉREZ-PADILLO et al., 2020; BONILLA et al., 2022). Pressure monitoring is important to guarantee the quantity and quality of supplied water, operating the system under economic constraints and detect abnormal functioning condition (e.g., leaks, control device failure or chemical intrusion) (JUN and KWON, 2019; ZHOU et al., 2019; RAMOS et al., 2020). Santos-Ruiz et al. (2022) stated that pressure sensors are more attractive, especially for detecting leaks, because they are cheaper and easier to install and maintain than flow sensors.

Nevertheless, sensors have limitations. Qi et al. (2018) presented metrics to assess the capacity of the sensors and highlights five main problems: nodes whose leaks are undetected by the sensors; undetected demands, derived from the undetected nodes; detection dimension, the number of sensors that detect the anomaly; the leaky region detected by a sensor; and the minimum detectable leak flow. Also in this sense, Huang et al. (2020) presented an assessment of the impacts of leaks on hydraulics and water quality. The authors conclude that the degree of impact of a leak depends not only on the diameter of the pipe, but also on its location, the time it occurs and that the effects of leaks can affect areas far from the actual location of the leak.

For helping to improve the management of water systems, several studies proposed strategies to monitoring the WDNs (GALUPPINI et al., 2020). Genetics algorithms have been applied to select pressure monitoring points addressing different objectives. Casillas et al. (2013) used a genetic algorithm to place sensors in a way that minimizes the number of non-isolated leaks. Soroush and Abedini (2019) used a geostatistical and genetic algorithm tool with the function the minimize the pressure data variance. In this sense, Romero-Ben (2022) presented a methodology for placing sensors using a custom genetic algorithm in which it seeks to minimize a metric based on distance from the pipe to possible points monitored.

Multiple objectives optimization is also applied to determine pressure monitoring points in WDN. Ferreira et al. (2022) presented a methodology for placing pressure sensors using the NSGA-II algorithm and two objective functions: maximizing the sensitivity of pressure monitoring network related to roughness changes in pipes and maximizing the sensitivity of the sensor for pipe burst events. The authors present three optimal solutions taken from the Pareto front that individually maximizes each objective function and a trade-off solution. Hu et al. (2022) also presented a methodology for placing pressure sensors using the NSGA-II algorithm but addresses five objective functions. The authors also present a multi-criteria approach for the selection of the most appropriate sensor sets presented at the Pareto front.

However, when using multi-objective algorithms this post-processing is necessary to select the alternative solution presented that best meets the needs of the operator. In this sense, Brentan et al. (2021) presented a multi-criteria decision-making technique to cluster solutions from Pareto front to simplify the selection of the most efficient monitoring strategy. Approaches for the sensor placement that do not employ multi-objective algorithms are also widely studied. Sarrate et al. (2014) used the k-means algorithm to partition WDNs following the sensitivity to faults of the nodes and choose the node closest to the cluster center as the monitoring point. Peng et al. (2022) applied the Structural Deep Clustering Network algorithm to partition WDNs following topological structure and hydraulic characteristics under multiple operating conditions. For each cluster, the highest sensitive point was selected. The authors concluded that the number of monitored points is more sensitive to the level of pipe rupture than to changes in the number of sensors.

This question highlights another widely studied aspect, the number of sensors to be placed in a network. Indeed, since it is impossible to place sensors at all points in the network, mainly due to budget constraints, the optimal number of sensors to be installed is still a field for improvement. However, due to difficult application and high computational value, approaches that use multi-objective algorithms, greedy algorithms, and exhaustive analysis in conjunction with WDN simulations are falling into disuse (RAEI et al., 2019; QUINTILIANI et al., 2020; KHORSHIDI et al., 2020). In this sense, Perelman et al. (2016) presented an approach using a greedy algorithm and a failure simulation process to define which monitoring points are possible to detect the greatest number of failures. However, due to the large computational effort, the algorithm resulted unsuitable for large WDNs. In this regard, some research have tried to work with different methods that have been successfully applied in other areas (XU et al., 2008). Using graph theory tools, Giudicianni et al. (2020) presented a methodology for water quality sensor placement without carrying out hydraulic simulations, but only exploiting the topological properties of the graph associated to the WDN. After a preliminary clustering of the WDN, sensors were placed in the topologically most central nodes. Giudicianni et al. (2022) used a weighted graph for reducing the computational burden of the optimization phase by defining potential sensor locations on the hydraulic/topological-wise most central pipes.

Graph theory is a mathematics branch that can model and identify complex interactions in data by modeling a system as a graph, in which vertices and edges represent the objects and the interactions between them (GUTIÉRREZ-PÉREZ et al., 2013). Di Nardo et al; (2018) used graph spectral techniques to simplify the management and the monitoring of WDNs. Torres et al. (2017) combined graph theory and statistical inference to characterize hydraulic performance and water quality. The authors used different graph analysis metrics to determine correlations between network elements. Di Nardo et al; (2018b) proposed a novel topological based approach for identifying primary pipes on which focus field investigation and maintenance in case of limited budgets. Wei et al. (2019) used the Graph Fourier Transform operator to sample WDN nodes for the monitoring of water quality to recover the dynamics of the system with a limited number of detection points. However, computational problems can be generated in very large graphs, making it difficult to analyse and process the information. To get around this situation, Signal Sampling on Graphs (SSG) was developed to reduce the size of graphs without having large losses in their information (TANAKA et al., 2020). SSG tools seek to find conditions for the recovery of signals in the graph, through a subset of vertices (TSITSVERO et al., 2016). Wei et al. (2019) used this theory to select sampling points to retrieve quality data from a WDN. The authors concluded that by monitoring about 30 to 40% of the nodes, it is possible to fully recover the network dynamics.

The SSG can be an effective tool for sensor placement in WDN, as the approach indicates the vertices to be monitored so that the graph can be well represented and the data from the other vertices retrieved (TANAKA et al., 2020). This is a process analogous to placing sensors in WDN that consider pressure, flow, or sensitivity data from nodes. The SSG analyses the data of the vertices, and through them seeks to recover or evaluate the situation of the other vertices data. Following this line of sensor placement through the SSG (SAKIYAMA et al., 2016), working on a literature graph, placed sensors on more informative vertices to predict the values of unmonitored vertices. Zhou et al. (2022) proposed a novel graph-based hydraulic grade reconstruction method to estimate unknown pressures of WDNs by using signal processing theory. The method showed to be very efficient in case of limited number of sensors since it does not require precise determination of WDN physical parameters.

Different metrics can be used in the SSG process, looking for vertices or edges that best represent the graph. Seshadhri et al. (2014) used the wedge sampling metric, which performs the grouping of vertices that best represents the graph. This grouping occurs following three concepts: transitivity, local grouping, and degree grouping. Chen et al. (2016) used uniform sampling method, projected sampling, and active sampling to select vertices in known graphs. The sampling process can be applied to select a set of monitoring vertices, to retrieve signals at all unmonitored vertices (TSITSVERO et al., 2016; LORENZO et al., 2018). Pesenson et al. (2008) used shape and nomenclature, in which the authors found conditions for the recovery of signals from the graph through a set of vertices, calling them an exclusivity set. However, previous research already worked with metrics that sought to reduce the complexity in manipulating data in large graphs. Adamic et al. (2001) introduced a signal sampling approach that considers vertices with many connections to other vertices.

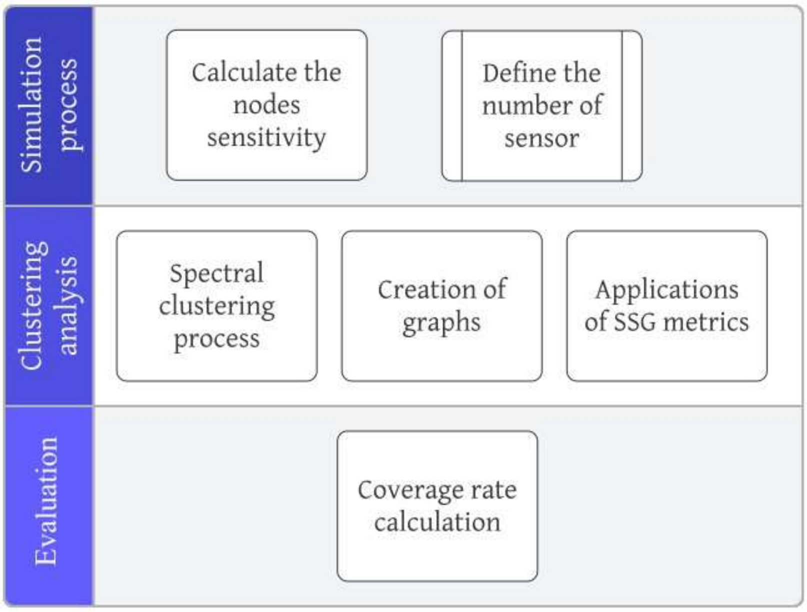
The development of fast and effective techniques for the selection of monitoring points without the need for post-processing is still a crucial task to address (BRENTAN et al., 2021; CARDOSO et al., 2021). Based on the presented text, the current methodology addresses the use of graph theory tools and the SSG technique to optimize the selection of locations for pressure sensors in WDNs. This innovative approach allows for an efficient distribution of sensors based on topological features and centrality metrics in WDNs, simultaneously reducing computational complexity compared to traditional approaches heavily reliant on hydraulic simulations.

In this regard, using the SSG to place monitoring sensors in a WDN can be an efficient strategy. Therefore, the present research uses metrics linked to SSG to select locations for pressure sensor placement. To guarantee an even distribution of them, the spectral clustering process is used to partition the network, and then, separately, the SSG tools are applied on each cluster to select the monitoring nodes. An innovative metric for evaluating the quality of the monitoring layout is also defined, which relies on the simulations of leaks of different intensities and consequent induced pressure variations in sensor locations. Finally, a method based on the leakage coverage rate, for the selection of the optimal number of pressure sensors to be installed, is proposed.

3.2 Materials and methods

This research combines graph theory and signal sampling on graphs for defining an optimal pressure sensor placement in a WDN. First, the areas of the WDN are modelled as an undirected weighted graph. Then, a signal sampling method is applied to determine monitoring points. The description of the methodological steps is presented in Figure 3.1.

Figure 3.1 - General methodology flowchart.



Thereafter, the proposed methodology consists of three main steps:

- Simulation process: Epanet software application in a programming environment to carry out hydraulic simulations and obtain node sensitivity values.
- Clustering analysis: application of Spectral clustering method (VON et al. 2007) to define areas with similar characteristics. Application of metrics of SSG to select potential nodes for pressure sensor placement: Centrality, weights, and shorter paths-based sampler; betweenness centrality-based sampler; and PageRank based sampler.
- Evaluation process: application of a novel leakage simulation-based metric for the assessment of the monitoring layout.

The methodology is implemented by dividing the WDN into clusters, where nodes with similar sensitivity are grouped. The core concept revolves around treating each cluster as a stand-alone graph, enabling the application of SSG metrics to determine optimal sensor positioning. This approach is designed to improve sensor distribution across the network, as each cluster corresponds to a dedicated sensor. The goal is to achieve a more comprehensive coverage of sensors, with each cluster having a sensor, improving overall system efficiency.

3.2.1 Clustering analysis

Cluster analysis is a set of techniques for data mining and pattern recognition which allows understanding the natural structure of a dataset or finding hidden organizations in unlabelled datasets, by revealing intrinsic similarities in the data (HAN et al., 2001). Clustering algorithms divide a dataset into meaningful groups, leading to groups where belonging data are as similar as possible according to a similarity measure or proximity relationships. In this work, the spectral clustering is used to define clusters in WDN to place sensors in each cluster. To apply the method is mainly used data represented in the form of graphs, and the relationships of proximity are represented by the edges of the graph. Graph theory base is a mathematics branch with the ability to identify and model complex interactions (BIGGS et al., 1986).

In this work, the WDN is modelled as an undirected weighted graph, considering its physical topology as $G = (V, E, W)$ (GIUDICIANNI et al., 2018), with V are objects (vertices) corresponding to the network nodes (junction nodes, tanks and reservoirs); E is the interaction between the objects and corresponding to the pipes, valves and pumps in a WDN; and W, in turn, are the weights assigned to the edges, representing stronger connections between certain vertices. W can be associated for instance with pipe flows and head losses or another physical-chemical parameter (KLISE et al., 2020). Due to the difficulty on analysing large graphs, it is important to reduce the amount of data while keeping as much important information as possible Tanala et al. (2020) and Qi et al. (2018). In this way, and seeking to reduce the size of the graphs, the theory of sampling signals on graphs was developed.

To cluster the network based on node similarity, a graph weighted by the maximum flow of pipes is used, and sensitivity to leaks is assigned as information to the graph vertices. This sensitivity value is determined through computer simulations of leaks at each node and the analysis of the impact of the simulated leaks on other nodes in the network, following the approach presented by Pérez et al. (2011). The edge weight information and vertex sensitivity are used as a feature matrix in the application of the Spectral Clustering algorithm (SÁNCHEZ-GARCÍA et al., 2014; VON et al., 2007). The spectral clustering process can be described in 5 steps, the first is the calculation of a Gaussian affinity matrix (Gm) that represents the similarity between vertices information. This matrix is determined by:

$$Gm_{\,ij} = \frac{\exp\,(-\,\|x_i - x_j\|}{2\sigma^2} \tag{3.1}$$

where x_i and x_j are the feature vectors of vertices i and j, respectively. σ is the parameter that controls the width of the Gaussian distribution that influences the affinity of neighbouring vertices (CHANG and YEUNG, 2008). After this first step, the Laplacian matrix (L) is calculated to detain the structure of the connections between the vertices, being determined by:

$$L = D - Gm \tag{3.2}$$

where, D is a diagonal matrix resulted from the sum of the rows of Gm.

The next step is the decomposition of L to calculate the eigenvalues and eigenvectors:

$$L\,x\,\wedge = \wedge\,x\,\vee \tag{3.3}$$

where $\wedge$ is a eigenvalues matrix and $\vee$ is a diagonal matrix containing the eigenvectors. Thus, the first k eigenvectors associated with the smallest eigenvalues higher than zero built are selected to form a new matrix (U). The matrix U is used assigning to a cluster based on its coordinates in the space generated by the selected eigenvectors (CHANG and YEUNG, 2008; BERAHMAND et al., 2022).

3.2.2 Signal sampling process

Following Leskovec and Faloutsos (2006) guidance on vertex sampling, this article uses three vertex sampling metrics to select monitoring points in the subgraphs generated by the clustering analysis. The first metric uses a combination of three measures: centrality, sum of edge weights, and counting how many shortest paths pass through each node. The second metric prioritizes central vertices related to information flow. And finally, the third metric is based on the PageRank vertex ranking metric.

3.2.2.1 Centrality, weights, and shorter paths-based sampler (CWSBS)

The main concept of SSG is to reduce the size of the original graph while preserving important properties and features of it. Therefore, graph information can be a determining factor for the selection of points for sampling. To meet these factors is developed a metric that uses information of vertices centrality, the edge weights, and the shortest paths. Thus, each cluster is considered a subgraph (G_{gp}) being first determined the subgraph centrality (c) (ESTRADA et al., 2005) for each vertex through a series of weighted sums:

$$c(i) = \sum \frac{1}{(1+d_{i,j})} \tag{3.4}$$

where, i is the node for which the centrality is calculated. j indicates a node belonging to the subgraph, excluding the i node itself. $d_{i,j}$ is the shortest path length between i and j .This process identifies and ranks vertices with greater centrality. After this step, the vertices are sorted based on their higher edge connections. The edge weights are represented by matrix W. Thus, the sum of the edge weights that vertex has is considered (HAGBERG et al., 2008; HAGBERG and CONWAY, 2020).

The third step calculates how many shorter paths pass through each vertex. For each vertex v_i in the graph G_{gp}, we determine the shortest path from all other vertices to v_i and increment a counter for each intermediate node encountered along the way. This process can be mathematically represented as follows:

For each shortest path, increment the centrality count for each vertex that lies on the path. This is mathematically expressed as: $C(v_j) = C(v_j) + 1$ for all v_j on the shortest path to v_i.

Finally, we combine the subgraph centrality scores for all vertices in G_{gp}. The vertices that are most important in the context of our analysis are those with the highest centrality scores. The overall subgraph centrality (S_{gc}) score can be computed by adding the centrality scores of the vertices. Additionally, we give more weight to vertices connected to edges with larger weights and shorter paths. The combination of these measures can be expressed as:

$$S_{gc}(v_i) = \sum_j C(v_j) + \alpha \cdot Weight(v_i) + \beta \cdot PathLength(v_i) \tag{3.5}$$

where $S_{gc}(v_i)$ is the subgraph centrality score of vertex v_i. $C(v_j)$ is the centrality count of vertex v_j as computed in the previous steps. α and β are weighting factors for the edge weights and path lengths, respectively. $Weight(v_i)$ is the weight of edges connected to vertex v_i. $PathLength(v_i)$ is the length of the shortest paths to vertex v_i These formulas help us identify and rank the most important vertices within the graph G_{gp} based on their subgraph centrality scores.

3.2.2.2 Betweenness centrality- based sampler (BCBS)

Betweenness centrality vertices sampling approach is applied and uses the measure of current flow between vertices to determine the relative importance of each vertex in a graph (BRANDES, 2001). The equation for calculating the centrality of approximate current flow between vertices is a combination of several measures:

$$Bc(v) = \sum \frac{\mu(i,j\,|v)}{\mu_{i,j}} \tag{3.6}$$

where, v is the vertex for which the betweenness centrality is determined. $\mu_{i,j}$ is the total number of shortest paths between vertex i and vertex j. $\mu(i, j\ |v)$ is the number of shortest paths between i and j that pass-through vertex v.The higher the value of the betweenness centrality for a vertex, the more central and important it is considered in the graph in terms of connections and communication between other vertices (BRANDES, 2001; HAGBERG et al., 2008; FENG and WANG, 2022).

3.2.2.3 PageRank Based Sampler (PrBS)

PageRank metric assigns a score to a vertex based in connectivity, counting the quality and quantity of that vertex in terms of connected edges. This metric is introduced by Larry Page in 1998 (PAGE, 1998) and uses the random path method. This metric was based on users connected to a web browser without an address bar or 'back' option. Thus, to access the web page, the user would have to access the page through an edge on the previous page or through a button that chooses a page that will be opened randomly. The probability of randomly accessing the web page (vertex) is determined by:

$$PR(i) = \frac{(1-\tau)}{N+\tau\sum_{j=1}^{t}\frac{PR(T_j)}{k(i)}} \tag{3.7}$$

where T_j represents all edges from vertex i to j; $PR(i)$ and $PR(T_j)$ are the PageRank values from vertex i to vertex j; $k(i)$ is the vertex; τ is a damping factor usually equal to 0.5 (ZHANG and ZHANG, 2019). Finally, the metric selects the vertices with the highest values of PR as the sampling point. Zheng et al. (2017) applied this concept and an adaptation in a graph with one million vertices to reduce information search time and increase search accuracy. Therefore, the use of this metric seems a promising field to explore for WDN graphs.

3.2.3 Assessment process

The methodology presented in this work is based on leak coverage terms. Notably, leak simulations are performed individually in the WDN nodes and if this leak changes the pressure of at least one monitored node, the leak is considered detected. Leakage simulations rely on the addition of a leakage flow rate, as a percentage of the total network demand value (q), individually on each node by using the emitter equation E_c (KLISE et al., 2020):

$$E_c = \frac{q}{\sqrt{P}} \tag{3.8}$$

where P is the average daily pressure head. In general, the literature considers that an anomaly is detected when changes appear on monitored data, but do not specify what value of change in the data. An example of this situation is presented by Cardoso et al. (2021) that considers the detection of a contamination in network since one of the sensors is affected in any change in the concentration of the monitored behaviour.

The coverage rate is determined for each monitoring system layout as a function of the number of sensors and presented on a Pareto front. A similar methodology is presented by Zhao et al. (2020) to evaluate the performance of sensors. The present research uses the different quantities of sensors in the termination of the coverage rate. Thus, the methodology allows to define the proper number of sensors for each percentage of leak detection.

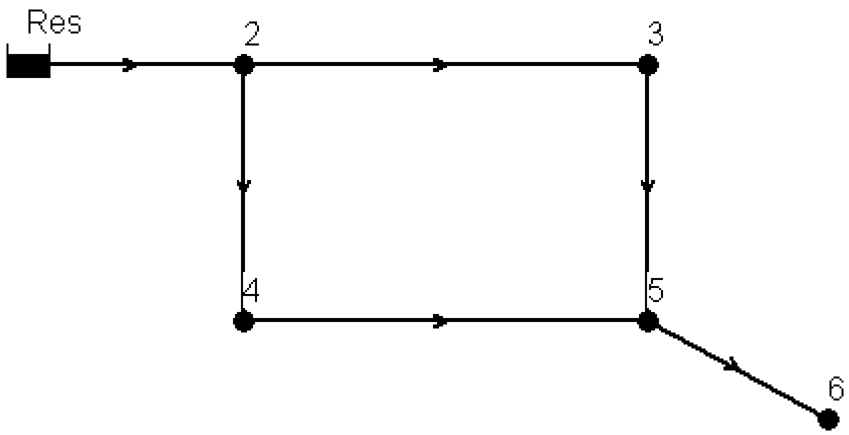
The coverage rate considers the sensitivity of the sensors to pressure variations. For this, two pressure load change limits are considered, respectively 0.5m and 1m, simulating the possibility of choosing sensors with different sensitivity. Then, pressure changes in the sensors are observed, if a leak in a certain node result in a pressure drop greater than the threshold (0.5m or 1m), the leaking node is considered covered. The number of sensors to be placed by the method follow a maximum number about 4% of the total number of WDN nodes, following the indication of Vrachimis et al. (2020).

To evaluate the effectiveness of sensor placement, the present research compares the sensor layouts proposed by other researcher working on the tested case studies. In this regard, an exhaustive random selection process of sensors will be performed, from 1 to the maximum number of sensors placed in this research, and among the selected monitored nodes. For each number of sensors, the process will calculate the corresponding coverage rate. A Monte Carlo analysis is used to estimate the possible results of coverage rate and the choice of nodes.

3.2.4 Explanatory application

For a better understanding of the proposed methodology, a simple case study is shown. (Figure 3.2) with 5 demand nodes (with demand ranging from 1 to 5 L/s), a reservoir and 6 pipes (with length of 230m, diameter of 150mm, and roughness of 90 according to Hazen-Williams).

Figure 3.2 - Topology of the explanatory network used to explain the methodological process.



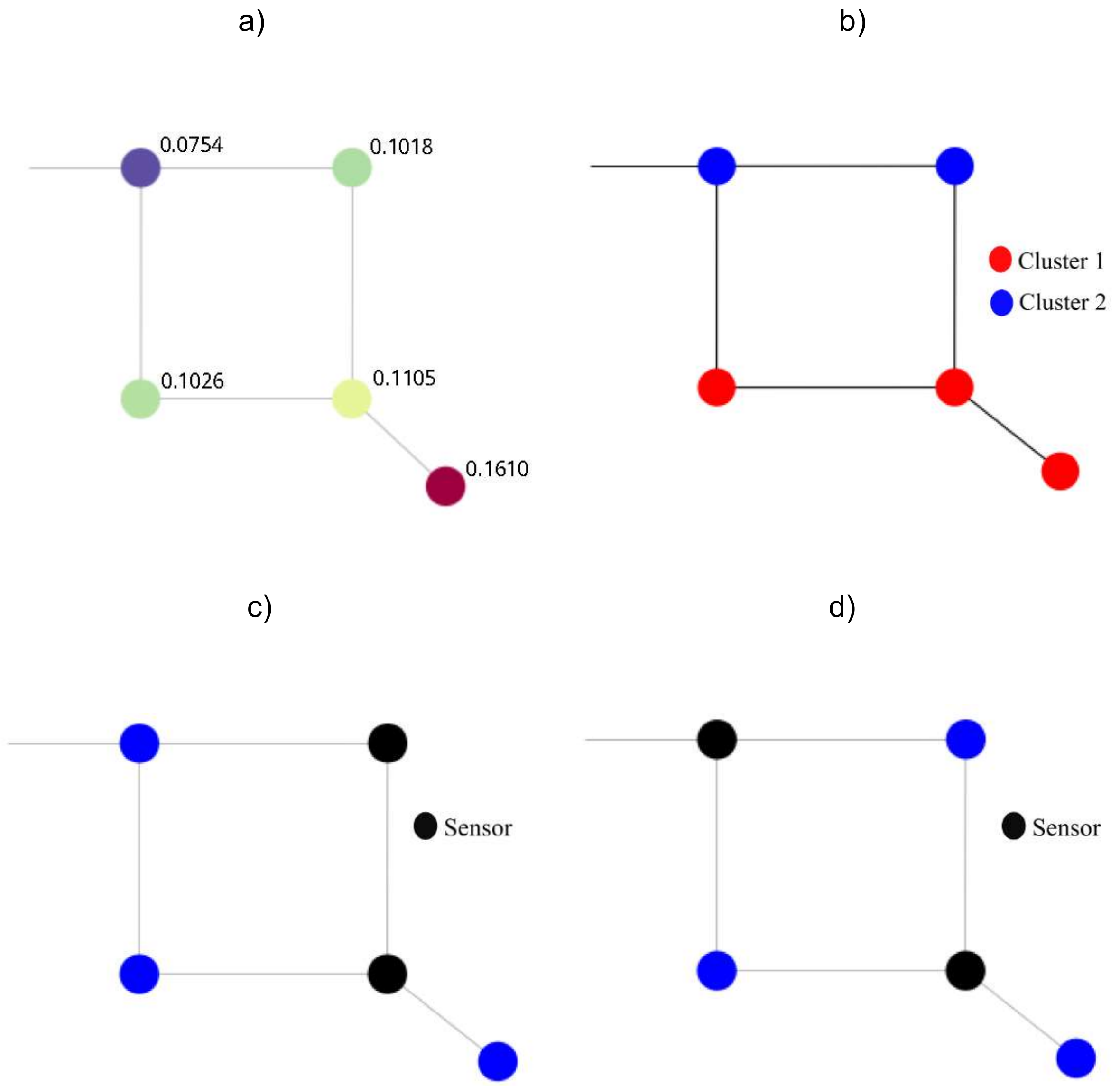
As a first step, a 24-hour simulation is performed, and the pressure data of all nodes are saved. Then, a leak per simulation is added individually at each node, with a leak flow rate of approximately 1.2 L/s and the pressure is also saved for all nodes. Then, sensitivity matrix is calculated, as shown at Table 3.1.

Table 3.1- Sensitivity matrix calculated to explanatory network.

Nodes	2	3	4	5	6
2	0.0754	0.0737	0.0735	0.0730	0.0711
3	0.0754	0.1018	0.0864	0.0918	0.0894
4	0.0754	0.0868	0.1026	0.0943	0.0920
5	0.0754	0.0963	0.0992	0.1105	0.1076
6	0.0754	0.0963	0.0992	0.1105	0.1610

Node 6 has higher sensitivity to leaks, since it is at a dead-end node and far from the reservoir. Thus, the leak at other nodes changes water flows and consequently the pressures. For a better understanding of the sensitivity of the nodes, the maximum value is used in the creation of the clusters. The nodes sensitivity, the resulting clusters and the nodes selected for monitoring are shown in Figure 3.3.

Figure 3.3 - Application and results to explanatory network. a) Sensitivity. b) Clusters. c) Sensors – CWSBS. d) Sensors - BCBS and PrBS



The maximum values shown in Figure 3.3a corroborate the values presented in Table 3.1, where it is possible to observe node 6 with greater sensitivity. It is also possible to observe the low sensitivity of node 2, since it is directly connected to a fixed level reservoir, which makes the pressure less susceptible to changes. The maximum sensitivity values are also used to determine the network clusters and are shown in Figure 3.3b.

The nodes are then clustered with the Spectral clustering process, being nodes 4, 5 and 6 for cluster 1 and the other nodes for cluster 2. The maximum flow of the pipes is used as edge weights, resulting in weighted matrices for each cluster, as shown in Table 3.2.

Table 3.2 - Weight matrix calculated to explanatory network.

Nodes	4	5	6
4	0	0.00259	0
5	0.00259	0	0.00469
6	0	0.00469	0

a) Cluster 1

Nodes	2	3
2	0	0.00359
3	0.00359	0

b) Cluster 2

These matrices are used to create graphs using the NetworkX package (HAGBERG et al., 2008). The Water Network Tool for Resilience (WNTR) package (KLISE et al., 2020) is used for the hydraulic simulation process and to acquire data related to WDN, such as pressures, flows and node coordinates. After creating the graphs, a signal sampling process is carried out on graphs using the Network analysis with Python package (HAGBERG and CONWAY, 2020) for calculating the vertex sampling metrics.

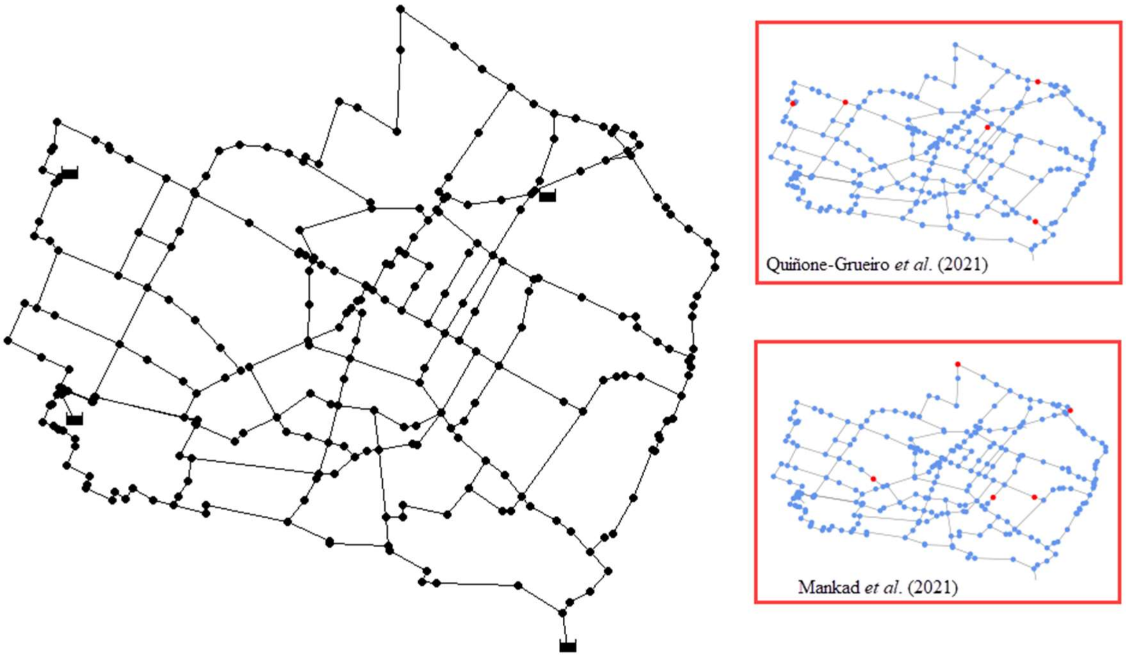
The application of three vertex sampling metrics on the generated clusters sought to find two monitoring points in the explanation network, one for each cluster. Nodes 2 and 5 are selected as monitoring points by the CWSBS metric (Fig. 3.3c). The BCBS and PrBS metrics selected the same nodes, 3 and 5, and are shown in Figure 3.3d.

3.3 Case Studies

The research proposal is applied on two benchmark networks. The first is the Modena network (BRAGALLI et al., 2012). This system model has 268 demand nodes (with an elevation ranging from 32m to 42m above sea level), 317 pipes (with diameter ranging from 150mm to 350mm, and length ranging from 220m to 1000m) and 4 source nodes (with total head of 73.0m, 73.80m, 72.0m, 74.5m). A pattern was used for the hourly demand multiplier to represent the typical daily variation in the users' demand in the system, with multiplier values ranging from 0.75 to 1.25.

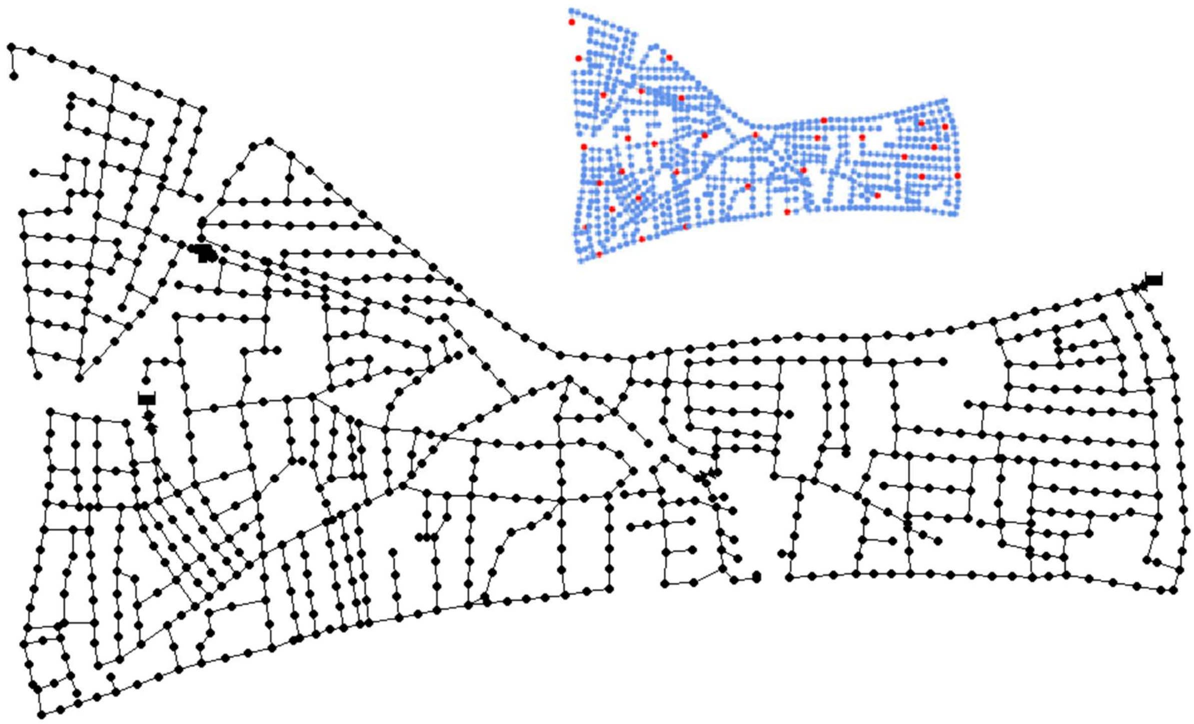
Figure 3.4 shows the topology of the Modena network, as well as the pressure sensors placed by Quinones-Grueiro et al. (2021) (upper-right panel) and Mankad et al. (2021) (lower-right panel). Quinones-Grueiro et al. (2021) performed several leak simulations of different sizes to obtain data to train a neural network. The authors aimed to place pressure sensors to detect and estimate the size of leaks in the network. On the other hand, Mankad et al. (2021) aimed to place sensors to obtain sufficient hydraulic information for monitoring the WDN and estimating pressure values in place. They performed simulations of random leaks by increasing the demand of the chosen nodes by between 0 and 200% of the original value.

Figure 3.4 - Modena network and details showing sensors placed from previous works.



The second case study is L-Town WDN (Figure 3.5) (VRACHIMIS et al., 2022). This WDN was used during the Battle of the Leakage Detection and Isolation Methods (BattLeDIM) and contains 782 demand nodes (with an elevation ranging from 1.5m to 75.0m above sea level), 905 pipes (with a diameter ranging from 75mm to 200mm, and length ranging from 10 to 80m), 2 source nodes (with total head of 100m). A pattern was used for the hourly demand multiplier to represent the typical weekly variation in the users' demand in the system, with multiplier values ranging from 0.2 to 1.4 for residential demand and from 0.5 to 1.2 for commercial demand. Originally this network already has 33 pressure monitoring sensors placed by the organizers of BattLeDIM.

Figure 3.5 - L-town network and detail showing sensors placed on previous work.



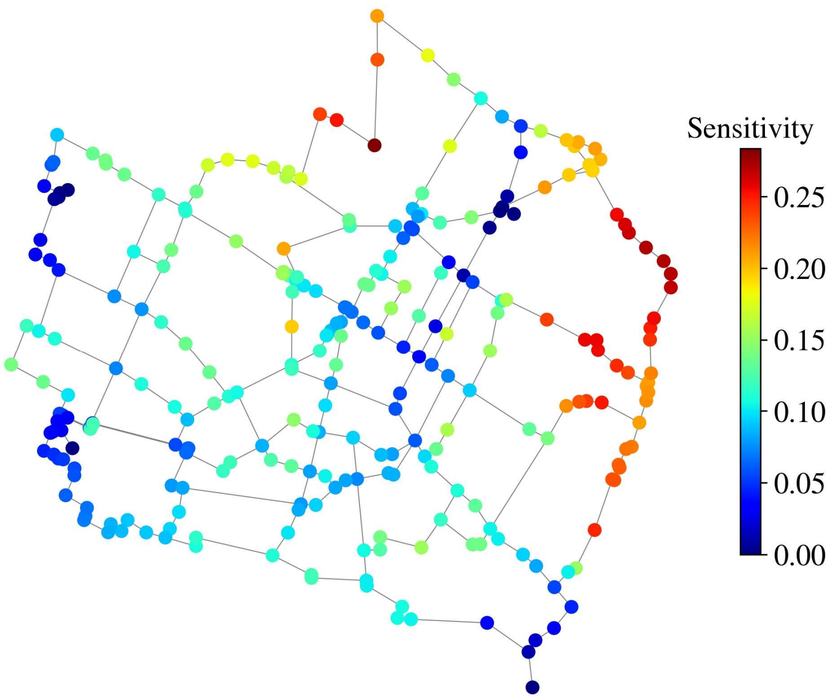
Pressure sensors layouts defined in the above-mentioned works: Quinones-Grueiro et al. (2021) and Mankad et al. (2021) for the Modena WDN and Vrachimis et al. (2022) for the L-Town WDN, have been used for the comparison with those obtained through the proposed methodology.

3.4 Results and discussions

3.4.1 Application to the Modena network

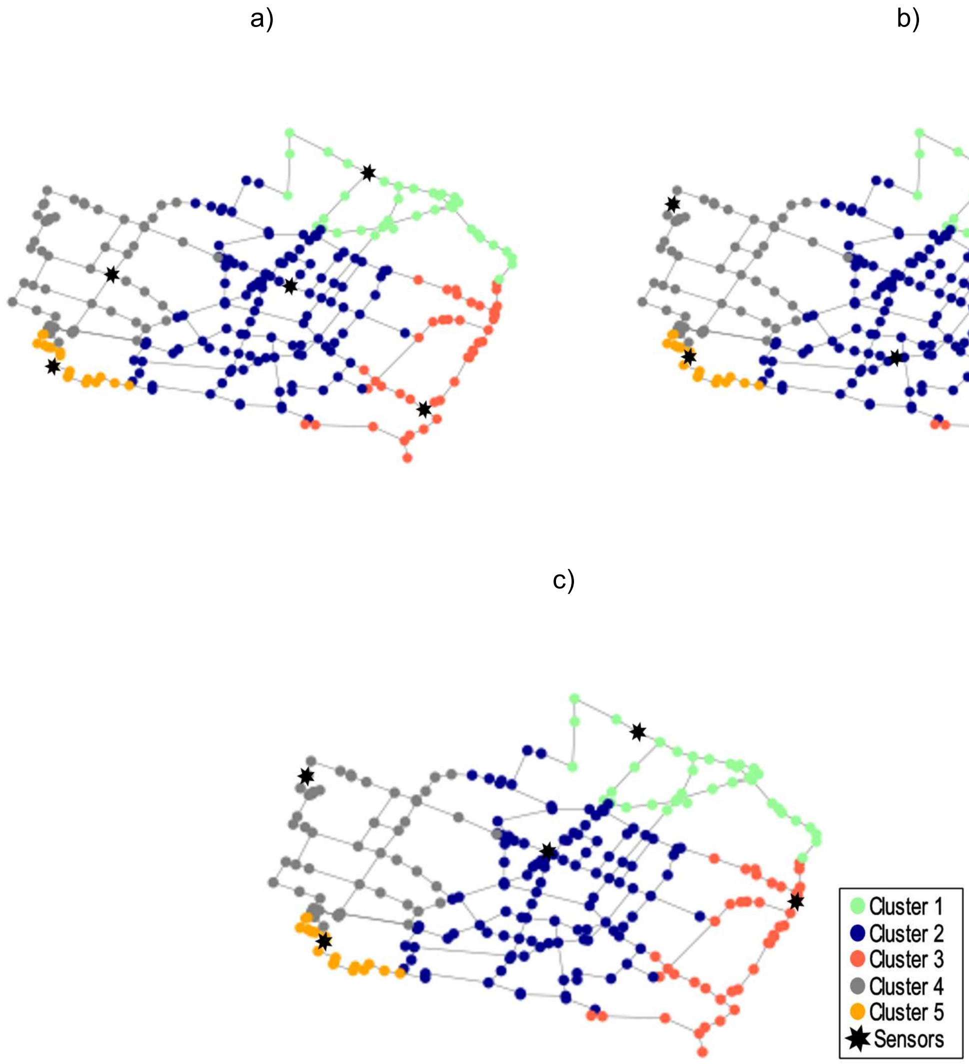
The first step is the application of the Spectral clustering process on the sensitivity matrix calculated at maximum flow time step to partition the network from 2 to 10 clusters. As the nodes sensitivity is determined under normal and anomalous circumstances, the first simulation process occurs without leaks lasting 24 hours. The second process run a leak with a flow rate of 2 L/s individually at each node of the network and check the pressure at the other nodes. After these simulations, the sensitivity value to leaks of all nodes is determined and the maximum value is used in the clustering process. These sensitivity values can be seen in Figure 3.6.

Figure 3.6 - Maximum sensitivity - Modena network.



The maximum quantity of clusters refers to twice the number of sensors placed by other researches (5 sensors). Indeed, the aim is also to understand what the additional advantage would be provided by the addition of a further sensor, and consequently, building up a benefit trend as a function of the number of installed sensors. After clustering the network, the SSG metrics were applied to select the potential location for the sensor in each cluster. Figure 3.7 shows the cluster and the nodes selected for sensor placement.

Figure 3.7- Maximum sensitivity - Modena network. a) CWSBS. b) BCBS. c) PrBS.



It is observed in Figure 3.7 that the proposed methodology guaranteed a greater spreading of sensors over the WDN. These sensors scattering ensures a better probability of event detection (LIU and AUCKENTHALER, 2014). The coverage rate of the sensors is calculated to determine the effectiveness of the sensor placement process.

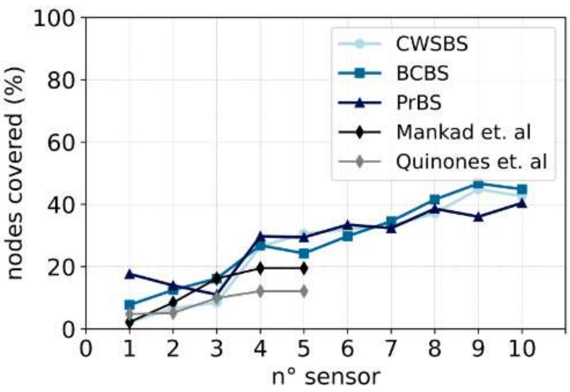
This coverage rate is calculated by simulating leakage in all network nodes individually. The leaking node is considered covered if the leak in it changes the pressure on at least one sensor node by 1m (Fig. 3.8a and Fig. 3.8b) or 0.5m (Fig. 3.8c and Fig. 3.8d). Leakage flows considering 0.5% and 1% of the network demand resulted in leakages with approximately 2.64 L/s and 5 L/s. The simulations with these two flow rates allow identifying the sensor coverage to low/medium and high intensity leaks, i.e., whether the sensor set is best suited for smaller or larger leaks.

The researches Quinones-Grueiro et al. (2021) and Mankad et al. (2021) present a methodology of sensor placement with a fixed quantity of sensors and do not present the monitoring points if a smaller number of sensors is placed. Therefore, and believing that the indicated monitoring points are the best even individually, an exhaustive process of random choice is carried out between the sensors placed by authors. Thus, 100 random selection processes are performed among the sensors, and the maximum coverage rate is shown in Figure 3.8.

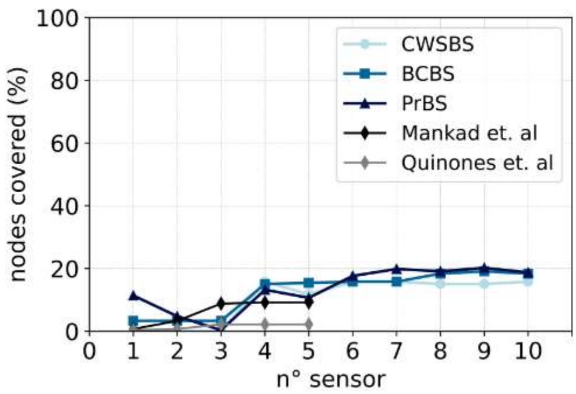
Figure 3.8 - Coverage rate by number of sensors - Modena network. a) Leak demand: 1% (5 L/s). b) Leak demand: 0.5% (2.64 L/s). c) Leak demand: 1% (5 L/s). d) Leak demand: 0.5% (2.64 L/s).

1 meter

a)

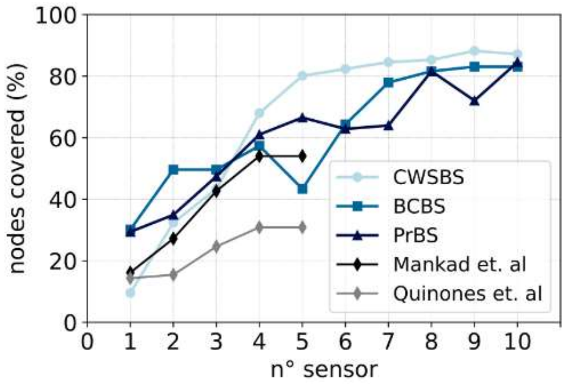


b)

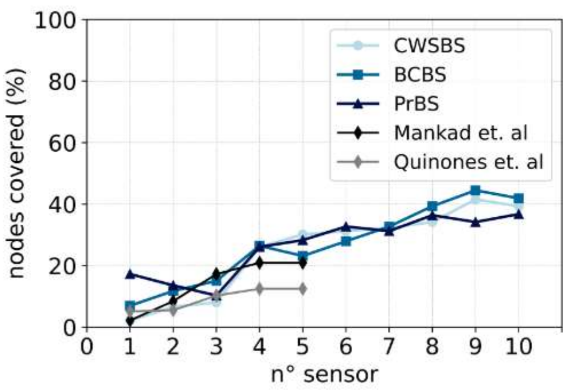


0.5 meter

c)



d)



As expected, these trends show increasing values of the node covered as the number of sensors increases up to 10. However, the trend is not strictly monotone, and sometimes the addition of a sensor can produce a reduction in the coverage. In Figure 3.8 it is possible to see that the PrBS achieves equal coverage value for 4 and 5 sensors, mainly for 1m alteration and with leaks of 5 L/s (Fig. 3.8a). It can also be seen that the 4 sensors placed by BCBS reached a coverage rate (29%) higher than the 5 sensors placed by the same metric (25%). This difference in coverage rate may be related to the locations where the sensors are installed. In fact, if points close to reservoirs or with low sensitivity to leaks are chosen, the coverage rate in these situations will be lower. Even so, this shows that it is possible to have a broad coverage rate with a smaller number of sensors, thus reducing installation costs.

Still on the results shown in Figure 3.8 it is possible to assess the significant improvement if more accurate sensors are placed. This can be better seen in Fig 3.8a and 3.8c, where the best coverage rate is 45% achieved by BCBS with 9 sensors for 1m change and 84% for the same number of sensors, but with a 0.5m change in pressure. For smaller leaks (i.e., 2.64 L/s), there is an increase of about 20% in coverage rate in all metrics if placed 10 sensors that identify a change of 0.5 m compared to sensors that identify a change of 1 m.

3.4.2 Application to the L-town network

The methodology is also tested on the L-town network, which has greater hydraulic complexities, such as residential, industrial, and commercial demands, in addition to having areas isolated by tanks and pressure reducing valves. Detection in these isolated areas can be more difficult, since a leak in these areas may not affect pressure elsewhere in the network, perhaps requiring sensors placed in the area itself to ensure leak detection. Firstly, the Spectral clustering process was applied on the maximum hourly sensitivity (Figure 3.9) The maximum number was set according to the number of sensors placed on the WDN during the BattLeDIM (VRACHIMIS et al., 2022), roughly equal to the 4% of the number of nodes.

Figure 3.9 - Maximum sensitivity - L-town Network.

The sensor placement occurred using the metrics of signal sampling, first considering the entire network as a graph, and placing one sensor and then using each cluster as a graph and placing one sensor per cluster. Figure 3.10 shows the L-town network with 33 clusters and the sensors placed by each metric.

Figure 3.10 - Sensors placed and clusters - L-town network (33 sensors). a) CWSBS. b) BCBS. c) PrBS.

a) b)

c)

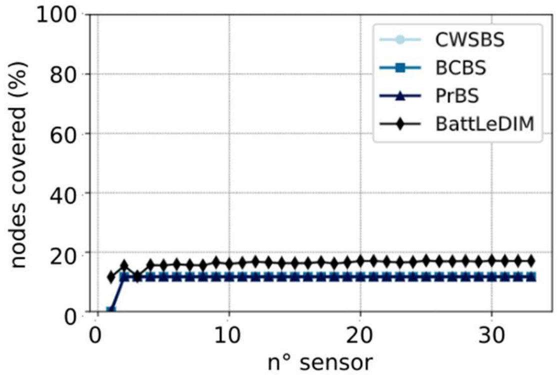
Following the results of placing sensors in the Modena network, the three methods for selecting the monitoring points ensured a greater spread of the sensors (Fig. 3.10). To check the sensor coverage rate, the leakage simulations on the L-TOWN network also used 5 and 10% of the total network demand, which represents leakages with 3.3 L/s. Following the results of placing sensors in the Modena network, the three methods for selecting the monitoring points ensured a greater spread of the sensors (Fig. 3.10). To check the sensor coverage rate, the leakage simulations on the L-TOWN network also used 5 and 10% of the total network demand, which represents leakages with 3.3 L/s and 6.6 L/s respectively. Furthermore, a higher leak flow rate was tested with about 10 L/s (about 15% of demand) due to BattLeDIM having leaks with high flow rates, in some cases with 20 L/s. and 6.6 L/s respectively. Furthermore, a higher leak flow rate was tested with about 10 L/s (about 15% of demand) due to BattLeDIM having leaks with high flow rates, in some cases with 20 L/s.

To compare the number of sensors placed by the sampling metrics, an exhaustive process of random choice is carried out among the sensors placed by BattLeDIM and the maximum coverage values for each number of sensors are shown in Figure 3.11. This process analyzed the combinations between 1 and 33 among the monitoring nodes presented by BattLeDIM and exposes the best maximum coverage rates achieved for each quantity. This will help the process of comparing the sensors placed by the present research and assess whether a coverage rate like that achieved with 33 sensors placed with a smaller number of sensors would be possible. concentration of 20 mg L^{-1}.

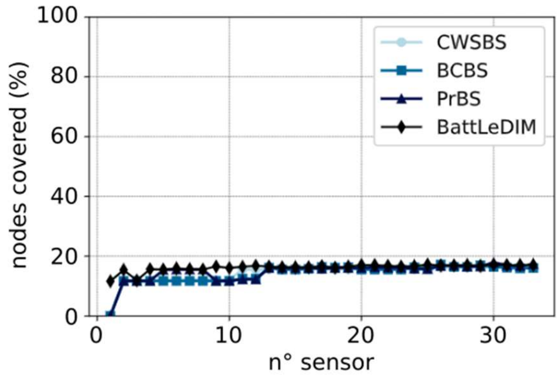
Figure 3.11- Coverage rate by number of sensors – L-Town network. a) Leak demand: 0.5% (3.3 L/s). b) Leak demand: 1.0% (6.6 L/s). c) Leak demand: 1.5% (10L/s). d) Leak demand: 0.5% (3,3 L/s). e) Leak demand: 1.0% (6,6 L/s). f) Leak demand: 1.5% (10 L/s).

1 meter

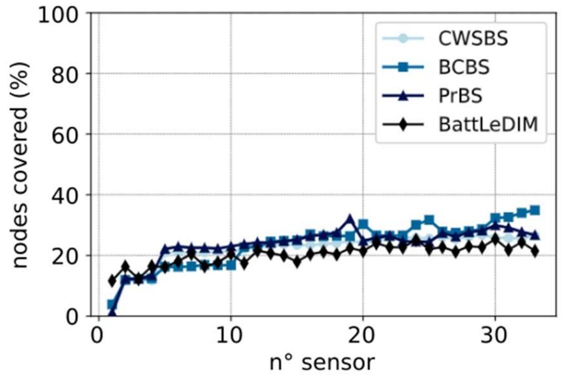
a)



b)

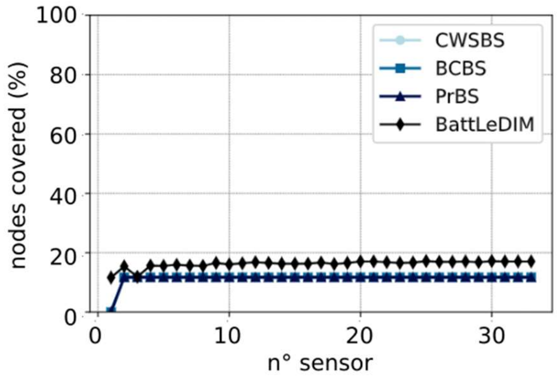


c)

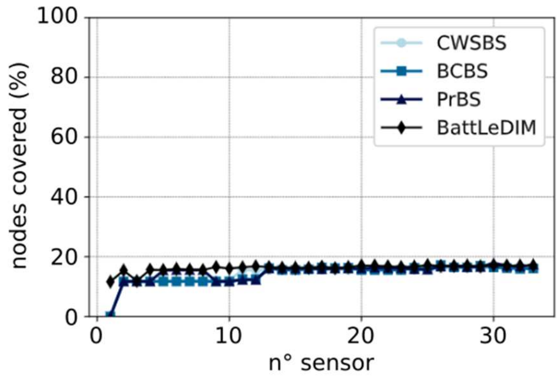


0.5 meter

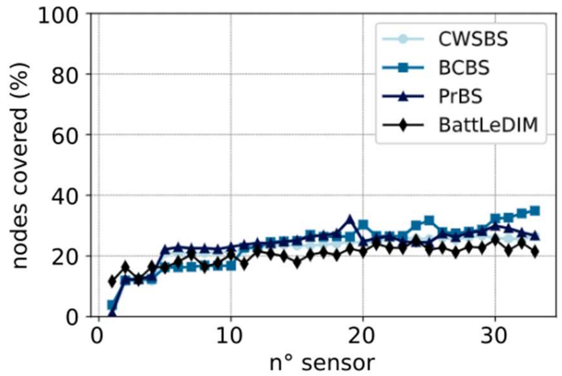
d)



e)



f)



It can be seen in Figure 3.11 that the coverage rate is similar between the metrics and the sensors placed by BattLeDIM for leaks with 5% of the total network demand, both for changes in the sensors pressure of 1m (Fig3.11a). Sensors placed by battle achieve higher coverage rate (17%) for all sensor quantities. The sensors placed by the proposed approaches reach 12.5% with two sensors placed and this value remains for all quantities. These 12.5% are referenced to nodes in area C, whose leaks are always covered by the sensors and the area is filled by a tank, which can generate significant change in pressures in case of leaks.

If the leaks are with 10% of the total demand and observing a variation of up to 1m in the pressure data of the sensors (Fig. 3.11b) the coverage rate achieved by BattLeDIM sensors is almost constant. Meanwhile, the PrBS metric achieves the same results as BattLeDIM for 5 to 8 sensors. When placed 13 sensors or more, all metrics and sensors placed by BattLeDIM achieve the same results, about 17%. However, the superiority of the sensors placed by BattLeDIM is overcome for the larger leak, with 10 L/s. It is observed in Figure 3.11c that from 5 sensors placed the metrics CWSBS and PrBS have higher results than the BCBS and BattLeDIM. However, with 33 sensors placed BCBS achieve the highest coverage rate of 36%, while BCBS and PrBS reaches 23% and BattLeDIM reaches 21%.

Considering the number of sensors placed by BattLeDIM and their coverage rates, with 22 sensors placed there is a coverage rate of 79% (Fig. 3.11f) and with 33 sensors this rate increases only 4%, a small value due to the large addition of sensors. This situation can also be observed for the sensors placed by the sampling metrics (Fig. 3.11f). Notably, for the RNS metric and with 22 sensors a coverage rate of 77% was obtained equal to that for 33 sensors, while for the PrBS metric with 25 and 30 sensors (Fig. 3.11f) a coverage rate of 77% was achieved. This proves that analysing the performance trend of the monitoring system can be a determining factor in terms of economic savings and leakage detection, since it is possible to obtain similar coverage rates with a smaller number of sensors.

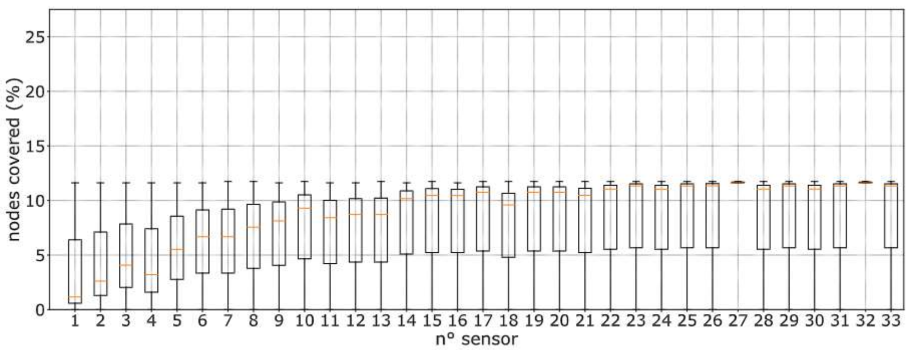
This same behaviour is repeated if more sensitive sensors are used, which detect changes in 0.5 meters or more in pressure. For leaks with 0.5% of demand (Fig. 3.11d). the sensors placed by CWSBS and BCBS reach a maximum coverage rate of 13%

and the sensors placed by PrBS and BattLeDIM reach about 15%. For leaks with 1% of network demand (Fig. 3.11e), the sensors placed by the battle achieve higher coverage rate (57% with 32 sensors). Among the approaches presented, the one with the highest coverage rate is BCBS with 31 sensors and 55% coverage. However, the highest coverage rates for leakage with 10 L/s (Fig. 3.11f) are achieved by the signal sampling metrics, reaching 92% for the CWSBS and BCBS metrics. While metric achieves its highest coverage rate with 30 sensors placed (91%). For the sensors placed by BattLeDIM the maximum coverage value is reached with 33 sensors and about 82%.

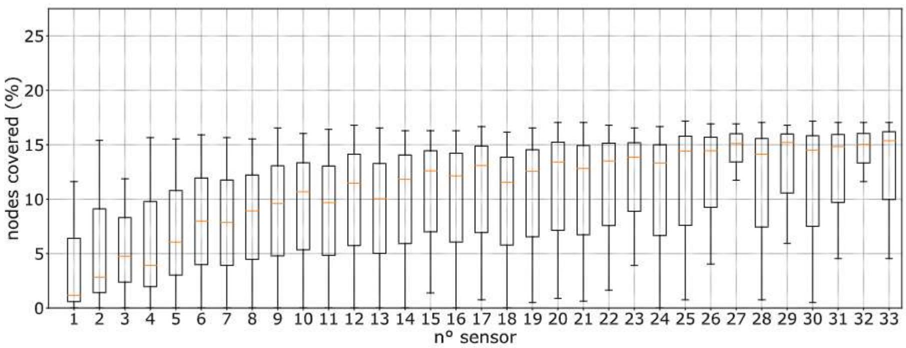
To identify the coverage rate achieved by BattLeDIM sensors in the exhaustive process of choosing the nodes, Figure 3.12 and 3.13 shows this rate by the number of sensors and can help in comparing the results with the used signal sampling metrics.

Figure 3.12- Maximum, mean and minimum coverage rate by 1 meter - Sensors BattLeDIM). a) Leak demand: 0.5% (3.3 L/s). b) Leak demand: 1.0% (6.6 L/s). c) Leak demand: 1.5% (10L/s).

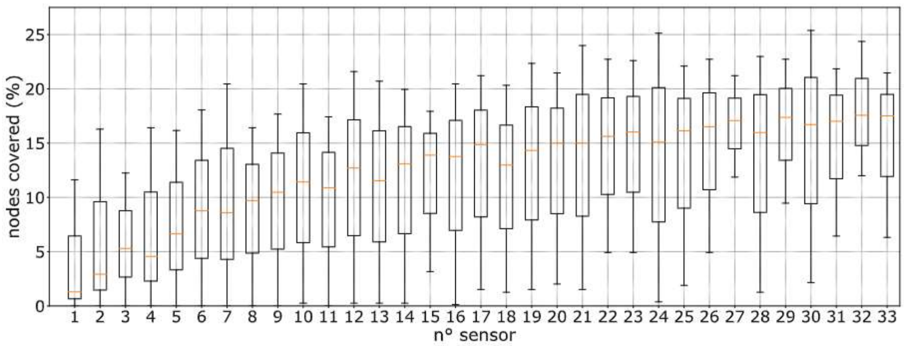
a)



b)



c)

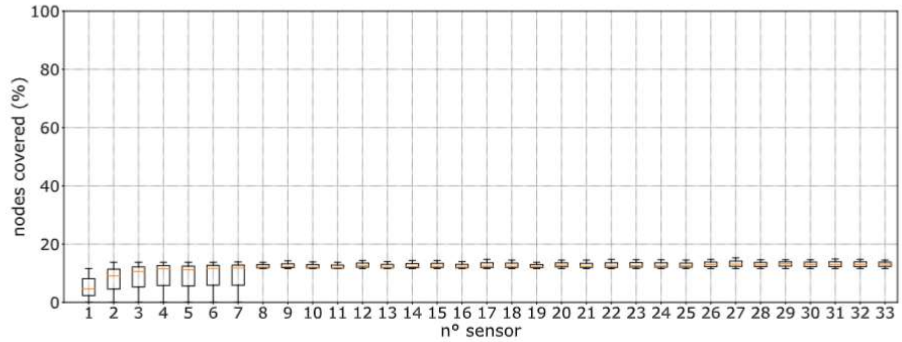


For a better visualization of the results, it is decided to reduce the scale of the y-axis (covered nodes (%)) of Figure 3.12. It is confirmed in Figure 3.12 that even with higher leak flow rates, there is not a very significant impact on the pressure of the sensor nodes. In all cases shown in Figure 3.12a the maximum coverage rate is 12% being the detection of nodes in area C. Leaks in this area are easily identified due to their supply being exclusively by a tank connected to a pump. Thus, when a leak occurs, the level of the tank reduces and affects the pressure on all nodes in the area. For leaks with 10% of demand (Fig. 3.12b) it reaches the maximum detection with 9 sensors (17%) and keeping the value close to this percentage for other quantities of sensors. It can be observed that the best coverage rate when analyzing changes equal to or greater than 1m occurs for leaks of 10 L/s, reaching 25% coverage with 24 and 30 sensors placed. It is observed that in all cases shown in Figure 3.12 the boxes remain with a constant size for the different numbers of sensors. It turns out that the behaviour for all cases is asymmetric, which means that there is a big difference between the minimum and maximum values as the locations of the sensors are changed.

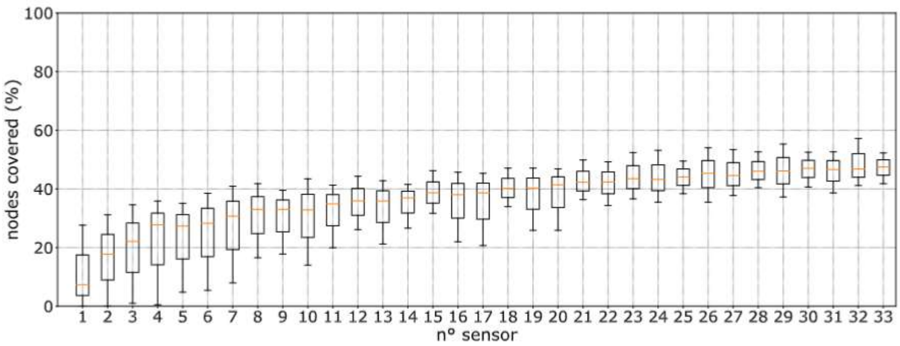
However, if the sensitivity of the sensors is better than 0.5m, better results are achieved (Fig. 3.13).

Figure 3.13 - Maximum, mean and minimum coverage rate by 0.5 meter - Sensors BattLeDIM). a) Leak demand: 0.5% (3.3 L/s). b) Leak demand: 1.0% (6.6 L/s). c) Leak demand: 1.5% (10L/s).

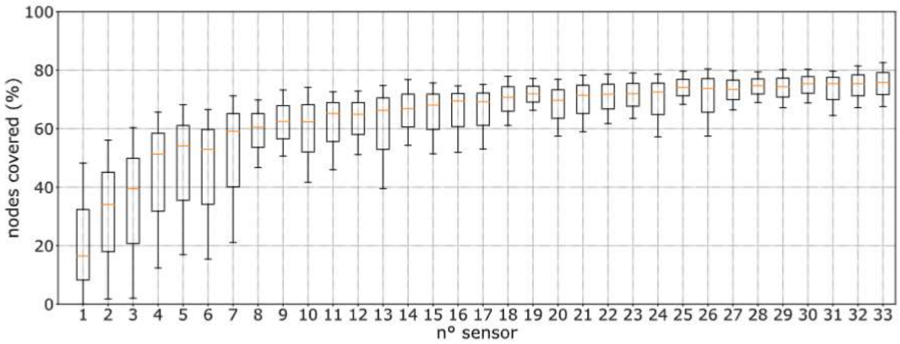
a)



b)



c)



The Figure 3.13a shows that the maximum, mean and minimum values almost do not change for leaks of 3.3 L/s, with an increase of 2% between the mean and maximum values and an increase of 3% between the minimums and averages. However, for leaks of 6.6 L/s (Fig.3.13b) a maximum coverage rate is achieved with 32 sensors placed (58%). For leaks of 7.8 L/s but analysing variations greater than or equal to

0.5m (Fig. 3.13c), a peak of 81% coverage is reached with 33 sensors placed. In this case, the data are more symmetrical since the coverage rate does not vary greatly when the sensor locations are changed. Since with smaller number of sensors, the maximum values are similar (mainly for the case exposed in (Fig. 3.13a), this corroborates once again about not needing a larger number of sensors. Thus, if a better observation of the number of sensors and locations were carried out, it would be possible to reduce the number of sensors.

This means that to ensure greater detection of leaks in this network, sensors would be needed to capture smaller changes in pressure. And that for sensors that are not so accurate, a smaller amount will already guarantee the maximum coverage rate. This shows that, for this network, many sensors placed would not improve the coverage rate for leaks of small intensities, because the coverage rate reaches its maximum value already with 6 sensors placed. Nonetheless, for leaks with 10% of the total network demand, there is an increase of 12% between the coverage rate of the minimum and mean values and of 8% between the mean and maximum values. This same percentage is observed for leaks with 10 L/s.

3.5 Discussion and partial conclusion

In this work, we focus on the placement of pressure sensors for leak detection to ensure an efficient distribution of sensors, considering network topological characteristics and centrality metrics in graphs. This process proved to be easy to implement, with low computational effort and achieved high leakage coverage rates. For this, clustering approach, graph theory and signal sampling on graph tools are used in the methodological process. The coverage rate in this work is presented in an innovative way, simulating leaks in all nodes of the studied networks, and considering it covered when the pressures in the sensors changed by more than 0.5 and 1 meter. The signal sampling metrics used for the selection of the monitored nodes proved to be effective and achieved higher coverage rates compared to sensors placed by another research. For the L-town network, the proposal proved to be more effective for smaller leaks.

This work also brought a focus on the number of sensors placed, which showed that with a smaller number of sensors it is possible to achieve results equal to those obtained when there are more sensors placed. This situation is also confirmed for the sensors placed in the L-town network by BattLeDIM, since an exhaustive process is carried out by varying the quantities of sensors and the maximum, mean and minimum coverage rates being exposed.

Another factor addressed in this research that deserves attention is the clustering process using the nodes coordinates as parameters and their sensitivity to leaks because this did not guarantee that these nodes are connected to each other by pipes, which makes it impossible to apply the SSG. Another limiting factor is that the change in pressure can occur due to pattern variability and therefore such technique is useful compared to monitoring of several days. Therefore, in future works it would be more feasible to use the connectivity of the nodes in the clustering process. Another point indicated for future work would be the use of new SSG metrics, also considering edge sampling, monitoring for example water quality, tank levels and flow rates.

4

Signal processing and pattern recognition for leak detection in water distribution network

This chapter is an adapted version of Barros, D., Pereira, T., Meirelles., G., Fernandes. W., Brentan, B. Signal processing and pattern recognition for leak detection in water distribution network. **Journal of Water Resources Planning and Management**, 2023.

Abstract

Leaks are a constant problem in water distribution systems, resulting in wasting resources, environmental impacts, and financial losses. Thus, it is crucial to develop effective and agile methods to detect network leaks. In this context, this study proposes a leak detection methodology using three different processes. The first consists of treating monitoring data through Independent Component Analysis, while the other two detection processes use the Interquartile Range (IQR) and Matrix Profile (MP) techniques, respectively. The methodology is evaluated based on a set of reference data provided by the Battle of Leakage Detection and Isolation Methods competition. The results indicate that the proposed approach is effective in detecting leaks, with some cases being detected in a few minutes after the beginning of the leak. It is worth mentioning that the IQR method presents better performance in detecting leaks with abrupt onset, while the MP method is more efficient in leaks with gradual increase in flow. In summary, the proposed methodology offers a robust and promising approach for fast and accurate leak detection in water distribution networks.

4.1 Introduction

Monitoring fluid transport represents a challenge for the sustainable and secure operation of any pressurised pipeline due to the imminent occurrence of leaks and ruptures, which can be attributed to the gradual deterioration of the intrinsic properties of the pipeline or to sudden unforeseen events (ZAMAN et al., 2020). Leaks and bursts in water distribution systems (WDS) in developing countries are the most important problem to be solved when thinking about the efficiency of the system. For example, only 61.5 Liters for every 100 Liters of treated water are in fact counted as used by consumers in Brazil, and this water loss can occur in two ways, namely: apparent or real (TRATA, 2020). The apparent loss is related to the use of water for unauthorised purposes or due to measurement errors, whereas the real loss consists of leaks in pipelines, branches, and reservoirs (TRATA, 2020). A study developed by the World Bank pointed out that 48.6 billion cubic meters of water are lost annually in the supply networks, representing about 14.6 billion dollars (RASHID et al., 2014). Thus, it is evident that water loss is a major problem that generates high negative impacts from environmental and economic perspectives.

Leaks can be classified as reported, unreported or background leaks (ADEDEJI et al., 2017). The first two have similar occurrences, however the identification in those reported is only visually possible by people who surround the occurrence region, allowing optimised repair operations. On the other hand, background leaks, eventually caused in joints, are incipient and have low flow magnitudes, presenting great detection complexity without the aid of more robust methods (ADEDEJI et al., 2017). The issue of leak detection in hydraulic systems has been extensively explored in the literature, with computational and statistical methods that can be applied to models for predicting anomalous network conditions (DARSANA and VARIJA, 2018). The known leak detection methodologies are generally classified as external and internal approaches (GOULET et al., 2013). Acoustic analysis has been extensively explored in external methods since the 1990s. Leakages can be identified by analysing noise patterns that arise from abnormal conditions. This detection can be achieved using portable equipment operated by personnel or through stationary noise measurement sensors placed at predetermined locations (HUNAIDI et al., 2004; CHEW et al., 2023). Nevertheless, the accuracy of noise measurement can be significantly impacted by external factors. Successful measurement in certain instances necessitates the operator's expertise (XU et al., 2019).

The continuous monitoring system of the Water Distribution Networks (WDN) is used in internal methods to take advantage of the interconnectivity of the hydraulic systems. With this, it is possible to optimise the sensors placement and improve the detection effectiveness of anomalous events (HUANG et al., 2015). Internal methods commence with the strategic placement of sensors within water supply infrastructures. There are two primary methodologies which are prominent in leak detection and localization. The model-based approach involves utilising a hydraulic model, implemented through simulation software to faithfully represent the network's hydraulic behaviour. This method entails comparing real-world hydraulic data with simulated information for accurate localization. On the other hand, data-driven approaches dismiss the need for a hydraulic model. Instead, they use measurements from in-network monitoring devices to extract insights for effective leak detection and localization. Both methodologies are commonly preferred and extensively utilised strategies in practical applications (RAJABI et al., 2023).

Algorithms can be modelled to collect and interpret data from WDS, which can significantly decrease instrumentation costs (HE et al., 2018). Network information, such as pressure and flow rate, can be monitored by sensors and later processed by computational methods, which enable identifying leak signal behaviours (CUGUERO-ESCOFET et al., 2017). Computational methods that treat and recognise leak patterns in time series can be exhaustive processes, and even so are widely used (SANTOS and PEREIRA, 2014; ZHANG et al., 2016). Thus, we can list the following machine learning algorithms to perform this processing work: Support Vector Machines (SVM) (MOUNCE et al., 2010); Singular Value Decomposition (SVD) (OLIVEIRA 2016); and Artificial Neural Networks (ANN) (MUHAREMI et al., 2019).

The pressure and flow rate parameters measured at a network point when analysing hydraulic networks can be influenced by several conditions, such as: the variable demand of consumption during the day; valve closing and opening operations; the roughness coefficient; the pipe diameters; and the occurrence of associated leaks (ORMSBEE and LINGIREDDY, 1997). When the value of a monitored parameter can be affected by several intrinsic elements of the network, the information can encompass a mixture of data with different attributes.

Blind Source Separation (BSS) has been explored since the 1980s as an important signal processing tool of a multivariate nature (COMON and JUTTEN, 2010). This technique aims to identify and filter out noise associated with a given set of observations, effectively separating the original signals into distinct sources. By doing so, the method enables more accurate analysis and interpretation of the underlying data. The use of Principal Component Analysis (PCA) makes it possible to convert a set of supposedly correlated variables into a data set with linearly uncorrelated variables. Another methodology widely used in the field of signal processing is Independent Component Analysis (ICA), which enables determining signal sources, also with the aid of mathematical operations (GAO et al., 2014).

The data obtained when modelling WDS and evaluating hydraulic parameters through a series of observations may exhibit amplitude variations within expected ranges. The Interquartile Range (IQR) is a useful tool for classifying a sample of data by dividing it into equal quartiles of a normal distribution, with the IQR value representing the difference between the third and first quartiles (JEONG et al., 2017). This tool can help visualise the range of the expected (or default) domain, aiding in decision-making. The IQR can also be used to identify points in a time series where deviations from the expected pattern occur. When no pattern is present, the points which deviate significantly from the expected results can represent system anomalies.

There are various approaches to detect anomalies in data, including methods that assess recurrent patterns, associate them with failures or trends, and perform sliding window analyses using specific evaluation parameters over the entire time series. One of these techniques is the matrix profile method, which annotates a time series and measures similarities between sub-sequences to identify differences (LI et al., 2022). The Matrix Profile (MP) is highly effective at extracting characteristic patterns of time series, such as motifs and discords, by evaluating the distances between all sub-sequences and their nearest neighbours (GUIDOTTI and D’ONOFRIO, 201). Motifs are sub-sequences of the time series which are like each other, while discords are those that differ significantly from the others. Researchers have used the MP method to identify discords in energy time series to highlight large loads between buildings (NICHIFOROV et al., 2020) and to label events in synchrophasor data (SHI et al., 2019).

Considering that the hydraulic state of water distribution networks is driven by the interaction of user consumption, leaks, and operation of control devices, in this work it is assumed that the information (i.e., flow rate and pressure data) collected by the water monitoring sensors can be understood as a mixture of vectors from different sources (HONGYU et al., 2016). Thus, the data available in the Battle of Leak Detection and Isolation Methods (BattLeDIM 2020) (VRACHIMIS et al., 2022) is used to develop a methodology to improve the detection of anomalous events in WDS. BattLeDIM is a significant benchmark for leakage detection and localization, which has been employed by researchers with model-based methods (LI et al., 2022; STEFFELBAUER et al., 2022), data-driven methods (DANIEL et al., 2022; WANG et al., 2022), and simulation-based approaches (MIN et al., 2022). Although recent research addresses leak detection, there are still techniques to be explored in terms of detection speed and accuracy. Certain methodologies require a calibrated network with established demand patterns, pump operation routines, and known tank levels in order to achieve a dependable detection process (STEFFELBAUER et al., 2022; DANIEL et al., 2022). However, as the water distribution network ages, the calibration precision might be compromised, and the inherent variability in demand could render some methodologies ineffective. There consequently exists a substantial opportunity to deploy robust and efficient approaches which directly focus on the monitored data, thereby circumventing any undue influence stemming from preprocessing of the network under examination.

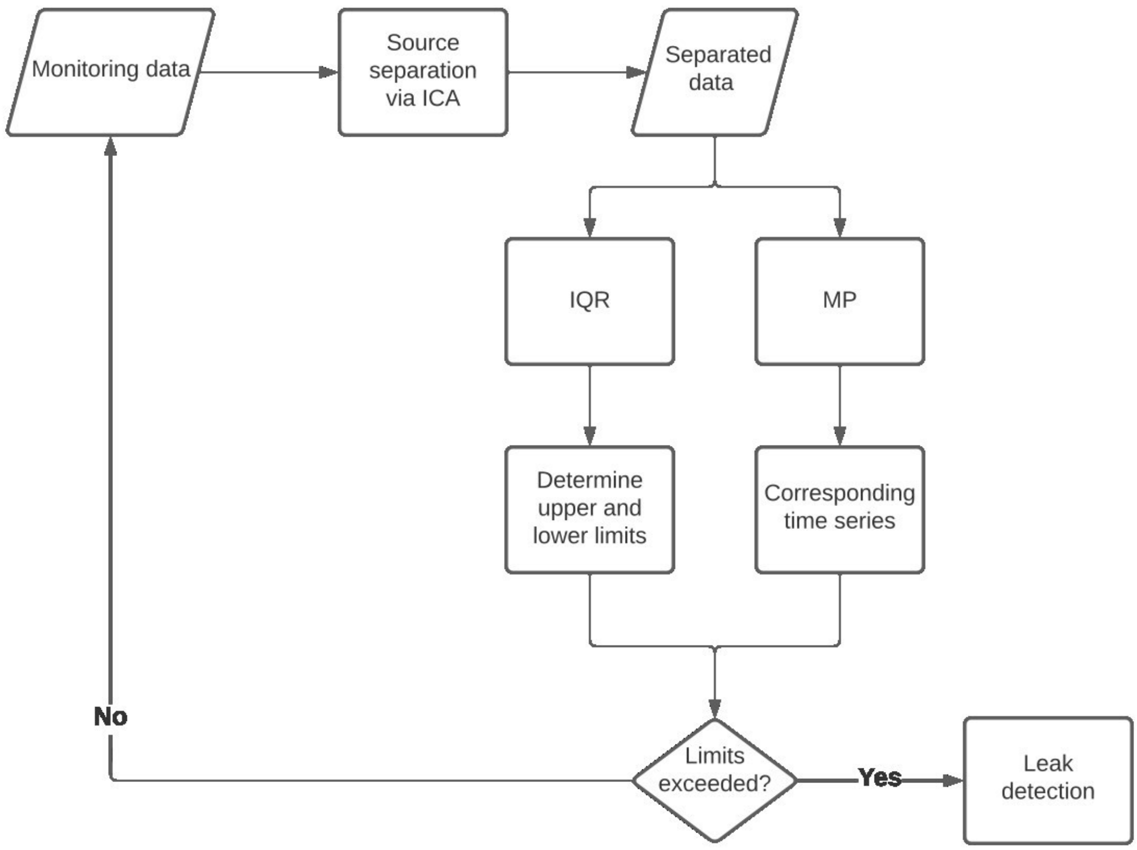
This article presents a data-based approach and BattLeDIM data is processed using the ICA algorithm, which separates it into sources associated or not with leak patterns. After this processing, the IQR and MP methods are applied in all data. Upper and lower limits are determined by the IQR method, and an anomaly is highlighted when the processed data exceed these limits. These limits are then updated throughout the data analysis when applying the method. A process seeking to identify the sensitivity of the methodology is also performed, comparing the results of the methods with the beginning of leaks. In turn, it is possible to point out the most accurate moment when the anomaly began. Finally, this methodology proved to be a tool which facilitates visualizing detected anomalous events.

4.2 Materials and Methods

This study focuses on the combination of a BSS method and statistical approaches for detecting leaks. Thus, the ICA algorithm, the IQR metric, and the MP method are used to achieve this goal. Pressure data of a monitoring system is initially processed and separated into two sources using the fast ICA algorithm. Both the IQR metric and the MP method are subsequently applied for the same purpose (i.e., to detect anomalies in the signal source separated by the ICA). However, in the case of the IQR metric, it is essential to determine upper and lower limits at each time step. Anomaly detection is then triggered whenever the data exceed these limits.

The methodological framework is described as a flowchart in Figure 4.1.

Figure 4.1 - Flowchart of the methodological process



The combination of the ICA algorithm, IQR metric and MP method seeks to ensure robust and reliable detection of anomalies in complex situations, where the application of only one technique may not guarantee the expected accuracy.

4.2.1 Independent Component Analysis (ICA)

Brown et al. (2001) presents a discussion named the Cocktail Party Problem, and this is a classic example of the signal separation methodology which refers to the process of isolating individual signals from mixed or composite signals. This problem occurs in various situations, including festive events where multiple sound sources are present, and distinguishing between them becomes a difficult task for the listener. The use of microphones strategically placed can help capture all the sounds present in the environment in such situations. The primary goal is to identify a single sound source, such as a person's voice, from the mixed signals using advanced BSS algorithms. By exploring the linear independence between the emitted signals by each source, these algorithms can extract the original, independent components from the mixed signals, enabling to identify the required voice within the festive environment (COMON, 2004). The ability to isolate individual sound sources has significant advantages in various fields, including speech processing, music analysis, and biomedical signal processing. Thus, the Cocktail Party Problem has become a crucial test scenario for evaluating and developing signal separation techniques, including ICA (COMON, 2004).

Making an analogy with the elements presented, we take a number N of vectors (recorded sounds) by a number M of samples collected (microphones) as input, and then processing the data with BSS makes it possible to identify components that are separated according to their previously unknown nature.

Equating the problem, a signal can be represented by a matrix X with transposed rows with dimension $N\ x\ M$:

$$X_{NxM} = \begin{bmatrix} X_{11} & \cdots & X_{1M} \\ \vdots & \ddots & \vdots \\ X_{N1} & \cdots & X_{NM} \end{bmatrix}$$

where

$$X_i = [X_{i,1}, X_{i,2}, \cdots, X_{i,M}]^T$$

It is assumed that X is a linear combination of the original vectors, corresponding to the multiplication between a matrix A (mixture matrix) and a matrix S with a structure like X:

$$X = AS \tag{4.1}$$

The X components correspond to the sounds recorded in the cocktail analogy and S to the original, individual sounds. The proposed algorithm has the role of finding matrix A. However, matrix S is also unknown. ICA does not guarantee that sources are separate or that all original sources are retrieved, but it seeks to find transformations that maximise statistical independence between components. A variation of ICA that aims to solve some of the limitations and computational challenges associated with the direct application of ICA is Fast-ICA. This variation is optimised to converge quickly to independent solutions and typically uses a negentropy approach to measure statistical independence.

Thus, to estimate both matrices, the fast-ICA algorithm is used for the separation process. The measured signals, X, are initially centred, removing the average value of each component. After centralization, the data are processed by a whitening technique, which is a linear transformation of the data based on decomposed eigenvectors of the correlation matrix from the centralised data (OLSHAUSEN and FIELD, 2004).

$$X = D^{0.5} V^T X \tag{4.2}$$

In which: V^T is the eigenvector transposition matrix calculated from the correlation matrix, D is the diagonal matrix of eigenvalues and X is the centred matrix of the measured signals.

Based on the lightening data, fast-ICA starts an iterative process to minimise the non-Gaussianity of a vector of definite weight w projected onto X Under the projection $u = wTX$, a non-quadratic function $f(u)$, the first term of the derivative $f'(u)$ and the second derivative term $f''(u)$ are computed to estimate a new value of w:

$$w = \frac{E\left(Xf'(u^t)\right) - E(X(f''(u)w))}{\left|E\left(Xf'(u^t)\right) - E(X(f''(u)w))\right|} \tag{4.3}$$

where E is the expectation operator over all argument values. By determining the value of w, the separate sources S are calculated as the final projection $u = wTX = S$.

The fast-ICA method is used to partition data into sources whose quantities are determined by the operator, but in smaller quantities than the original data. In the context of water supply networks, initial tests indicate that the flow and pressure signals ultimately resemble a demand signal and a noise signal (BRENTAN, et al., 2021. However, when dealing with the noise signal, additional methodologies are necessary to automatically identify anomalies. By identifying such anomalies, it then becomes possible to effectively carry out corrective measures.

4.2.2 Interquartile Range application

The interquartile range (IQR) method is commonly used to develop an effective anomaly detection methodology that can automatically identify anomalies in the output of the fast-ICA algorithm. The separation of the independent components obtained through fast-ICA can be evaluated using boxplots of the IQR method. The distance between the first ($Q1$) and third quartiles ($Q3$) is a variability measure of the data, and any values that fall beyond 1.5 times the IQR are considered as potential outliers. This approach enables identifying anomalies which significantly deviate from the expected values of the signal. By automating this process, the IQR-based anomaly detection methodology can be used in real-time data analysis without relying on manual inspection by operators, making it a valuable tool for a range of applications, including fault diagnosis and quality control. Overall, the IQR-based methodology provides a reliable and efficient way to detect anomalies in complex data sets, contributing to improve data analysis and decision-making processes.

The sample needs to be divided into quartiles and then the difference between the third and first quartiles is computed to calculate its IQR:

$$IQR = Q3 - Q1 \tag{4.4}$$

This method is also known as the “Rule of $1.5(IQR)$”, which means that data will be an outlier when it is “$1.5(IQR)$” higher than $Q3$, or lower than $Q1$. Or further, low outliers can be understood as those below “$Q1 - 1.5(IQR)$”, and high outliers when they are above “$Q3 + 1.5(IQR)$”. It is worth mentioning that an outlier is not necessarily an anomaly, it can represent a peak in demand or a measurement error of the sensors, for example. The coefficient 1.5 within the equation is subject to modification for performance optimization. To elaborate, the introduced variable, denoted as "V_{HL} (IQR)," has been conceived to facilitate adaptations to the V_{HL} value based on distinct datasets under analysis (WAN et al., 2014). This flexible nature of the V_{HL} parameter allows for temporal adjustments, and in this context, the present study uses dynamic calibration within defined time intervals. The initiation of the anomaly detection process aligns with application of the ascertained V_{HL}value following each stipulated time window. This strategic approach facilitates discerning anomalies within the dataset. In cases where persistent undetected anomalies persist, a re-calibration of values ensues, subsequently prompting a repetition of the detection procedure aimed at uncovering any additional anomalies.

4.2.3 Matrix profile concept

The Matrix Profile (MP) is a concept of time series analysis that consists of calculating the similarity between sub-sequences of a time series, developed by Yeh et al. (2016). The MP is a data structure that contains the minimum distance between each sub-sequence and the most similar sub-sequence in another part of the time series. The distance is calculated using a dissimilarity measure, such as Euclidean distance.

It is first necessary to define a window of length m to calculate the Matrix Profile, which is the size of the sub-sequences to be compared. Then, each sub-sequence of length m is compared with all other sub-sequences of length m in the time series. The distance between each pair of sub-sequences is calculated and the minimum distance value is stored in the MP. The result is a time series of length $n - m + 1$, where each value represents the minimum distance between a sub-sequence and its most similar sub-sequence (GHARGHABI et al., 2017; ZYMBLER and IVANOVA, 2021).

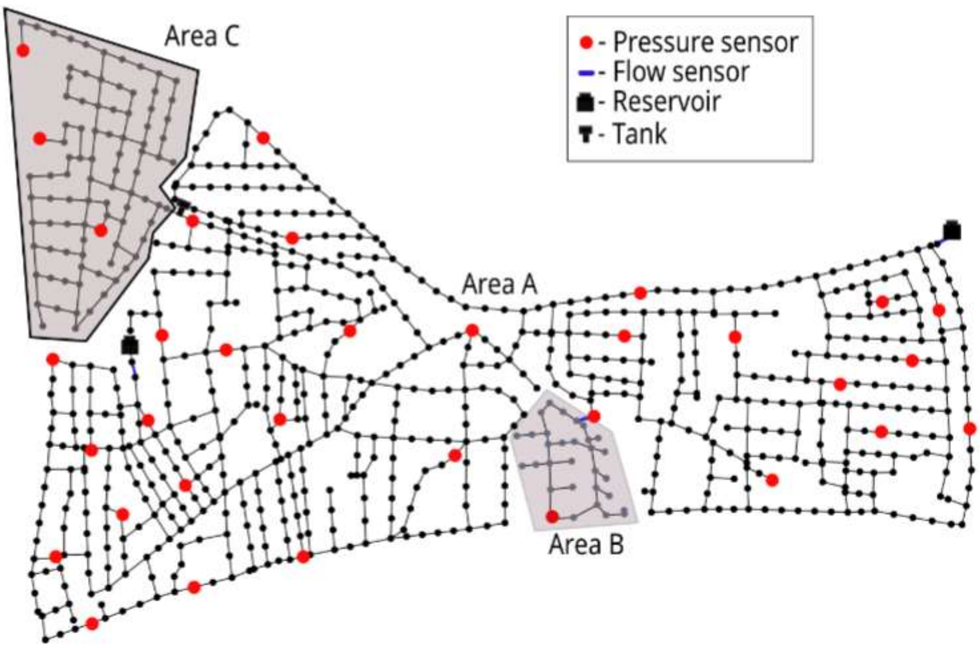
MP is used in various time series analysis applications such as anomaly detection (LAN et al., 2021), pattern classification (ASYALI et al., 2006) and time series clustering (LI et al., 2022). It enables efficiently and accurately identifying similar patterns and anomalies in a time series.

4.3 Case study - The Battle of Leakage Detection and Isolation Methods

The methodology presented herein was applied to the L-Town network to conduct simulations and verify the sensitivity of processing hydraulic data signals and generate statistical control charts which enables improved leak detection in water supply networks. This case-study network is inspired by a city on the island of Cyprus, and it was presented in the academic context of Battle of the Leakage Detection and Isolation Methods (BattLeDIM) (VRACHIMIS et al., 2022) as a study tool in leak detection.

To ensure conditions like reality, the L-Town network has 42.6 km of pipe network supplied by two reservoirs. The control system of this network is composed by one pressure reducing valve and one pump. The system meets a demand of 10,000 inhabitants in three sub-regions, where areas B and C are highlighted and the rest of the network is area A (Figure 4.2) There are 37 sensors installed along the network (1 water level sensor in the tank, 3 flow meters and 33 pressure sensors) which collect the average value data every 5 minutes by a Supervisory Control and Data Acquisition (SCADA) to analyse the pressure, reservoir level and flow variables.

Figure 4.2 - L-town network topology representing the sub-regions



One of BattLeDIM's goals is to identify leaks along L-Town pipelines. All data from L-Town sensors are grouped into two periods, one referring to the year 2018, used as a reference for the operation pattern, and another for 2019, which is the period for analysis of the events described in BattLeDIM.

It is important to highlight that leaks are observed simultaneously through the network databases, even if they did not start at the same time. This situation arises because many leaks are only addressed once they have reached a significant level of flow or have caused enough anomalous events that are easily detectable. As a result, each leak is individually analysed or examined within sets starting from the most recent periods. Figure 4.3 displays all the leaks that occurred during the data period analysed in this study.

Figure 4.3 - Leaks for the 2019 database.

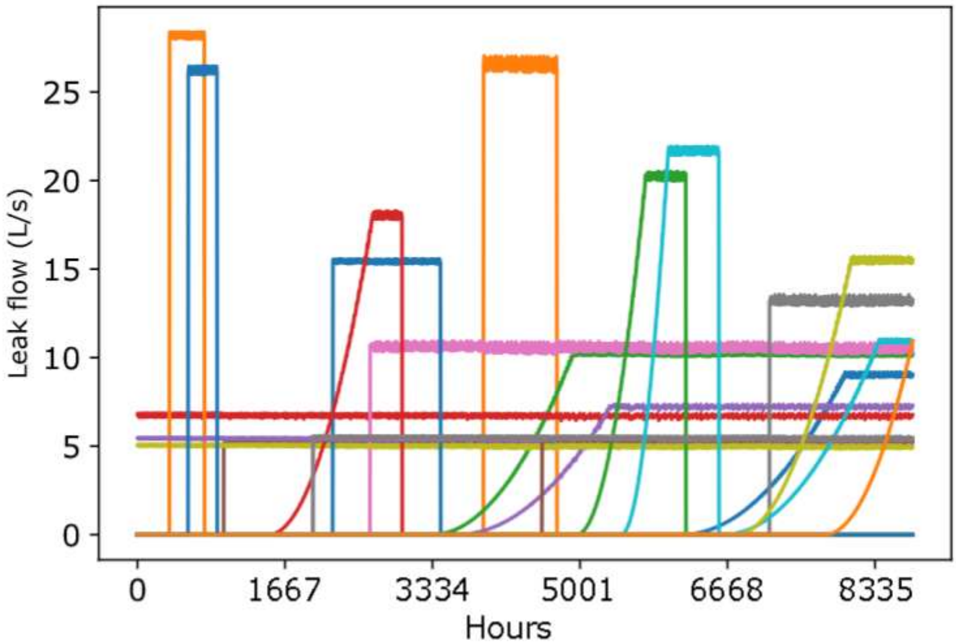


Figure 4.3 shows that 9 leaks have a gradual increase in flow until they reach a stable condition. A 7-day sliding window is applied to analyse these leaks (which in general are the most common and difficult to be detected quickly) before the leak starts and continues up to two days after the stabilised flow. The reason for selecting a 7-day duration is because the pressure/flow patterns exhibit a quasi-periodic nature with a frequency of one week, which can be attributed to the social behaviour of water usage (BRENTAN et al., 2017). This process has two objectives: first, it enables determining the detection sensitivity of the methods by indicating the flow rate at which the leak

was detected; and second, it enables detecting leaks with long gradual growth, since other leaks can start during this growth, and this affects the detection process.

It is important to note that three leaks remained constant during the entire period. Although these leaks were pre-existing, they provide a realistic scenario for detecting new leaks in water distribution systems, where the other methodologies may not be able to detect them.

A sliding window of 4 days prior to the event is utilised in leaks with gradual growth to analyse the temporal variations and determine the presence of other leakages. However, it is important to note that the duration of the analysed data greatly impacts the information that can be extracted from the MP technique. A larger dataset allows a larger sliding window, and therefore more detailed and accurate analyses. In turn, utilizing a 4-day sliding window for MP pre-processing is acceptable if it serves the specific analysis objective. It is essential to carefully select the window size to ensure a robust and precise analysis. The IQR calculation can also enhance the accuracy of the analysis with a larger sample size.

4.3.1 Evaluation metrics for leak detection algorithm

Data in which the beginning of the leaks are known will be used to evaluate the methodology presented. Thus, the detection time after the beginning of the anomalies will be used as an evaluation criterion. Another evaluation process will be related to the intensity value of the anomaly flow at the time of detection. Thus, it will be possible to observe the sensitivity of the method in relation to the size of the leaks. In the meantime, BattLeDIM presents evaluation methods which will also be used in the present study.

The assessments presented follow a purely economic approach with respect to the water profit saved over a year from successful leak detection (VRACHIMIS et al., 2022). The battle organisers also present methods for assessing the leak sites, but these will not be explored. One of the evaluation processes is called True Positive, where one considers true detection if the leak meets the following condition:

$$t_{st}^{l} \leq t_{d}^{h} \leq t_{end}^{l} \tag{4.5}$$

where t_d^h is the detection time, t_{st}^l and t_{end}^l is the start and end time of leakage l.

Another valuation approach presented is Profit from water saved related to profit p_w^h (euro) of water saved by detecting a leak and determined by:

$$p_w^h = \left(\sum_{k=t_d^h}^{t_{end}^l} q^l(k)\Delta t\right) c_w \tag{4.6}$$

where by detection l, $q^l(k)$ flow rate of leakage l at each discrete time step k. Δt is the duration of the discrete time step and c_w is the cost (euro) of water per cubic meter.

The total score Ts is determined by the detection set, which is determined by:

$$p_w^h = \sum_{h \in D} s_h + \sum_{h \in D}(p_w^h + c_h^r) \tag{4.7}$$

where s_h is the score per given detection k and c_h^r is the repair crew cost. These evaluation processes are presented by the BattLeDIM developers, who provide an algorithm for determining these and other evaluation processes (VRACHIMIS et al., 2022).

Three evaluation approaches in addition to these metrics are tested, called Recall, Precision and F_1. Recall identifies anomalous instances, Accuracy measures the ratio of correct anomaly detection to false alarms, and F_1 is the harmonic mean weighted between Recall and Accuracy, ranging from 0 (worst) to 1 (best) (DANIEL et al., 2022).

$$Recall = \frac{TP}{TP+FN} \tag{4.8}$$

$$Precision = \frac{TP}{TP+FP} \tag{4.9}$$

$$F_1 = \frac{2\ x\ Recall\ x\ Precision}{Recal \quad Precision} \tag{4.10}$$

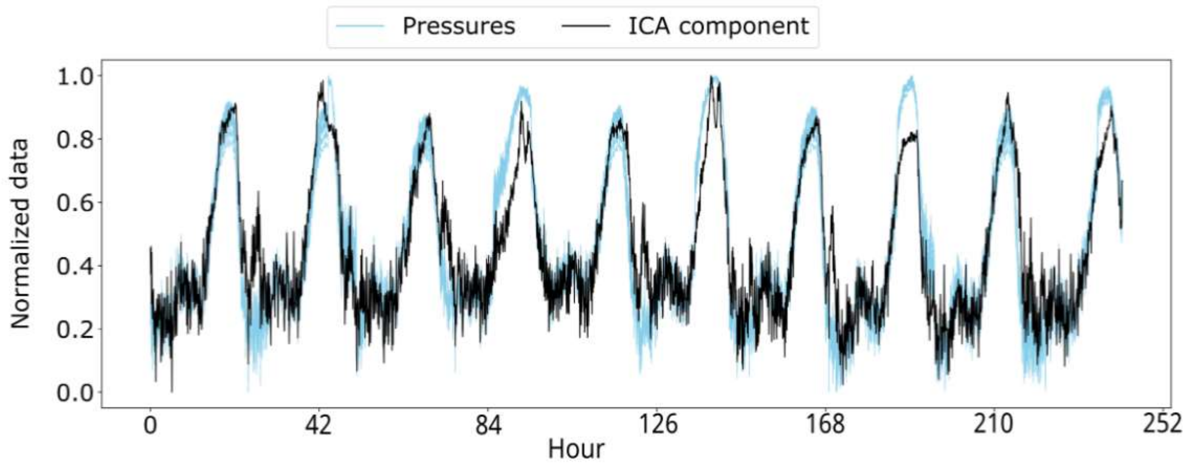
where TP is True Positive; FN is False Negative and FP is False Positive. These variables are related to the correct detection of leaks.

4.4 Results and Discussions

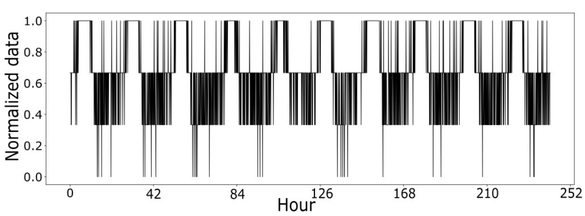
The results are divided into two parts: the first examines the constituent elements obtained through implementing the fast-ICA technique, while the other focuses on identifying leaks using IQR and MP. In this study, only the pressure data from the 33 sensors are used as input to the fast-ICA to partition the data into two sources. One of these sources exhibited patterns that were consistent with many sensors, specifically those positioned in area A. Figure 4.4 presents the normalised pressure from 29 sensors in area A in relation to the signal source separated using fast-ICA in two components. In addition, Figure 4.4 shows the normalised pressure of areas B and C, and the signal with noise separated by fast-ICA.

Figure 4.4 - Normalized pressures and Component 1 via fast-ICA. a) Component 1 and Pressure normalized – Area A. b) Pressure normalized - Area B. c) Normalized pressure - Area C.

a)



b)



c)

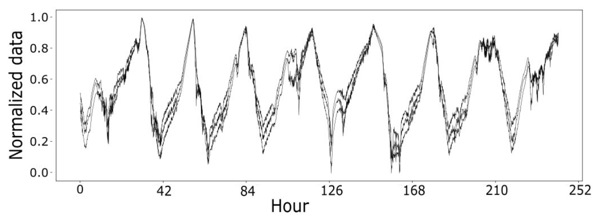
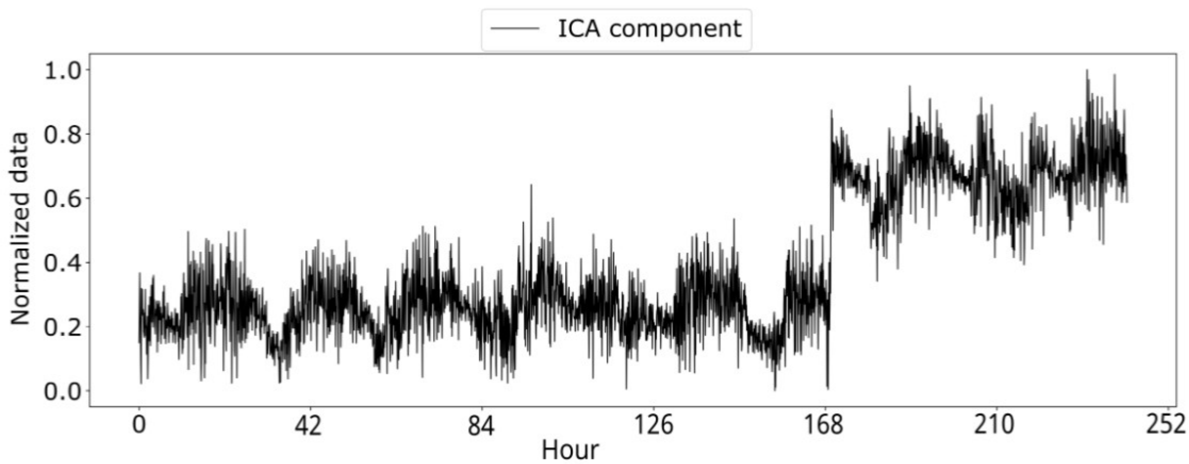


Figure 4.4a displays the data collected by sensors located in area A, along with the first component separated by fast-ICA. The pattern observed in this component

corresponds to changes in pressure in area A. Interestingly, the pressures measured in areas B and C (shown in Figure 4.4b and 4.4c, respectively), which were also used in the fast-ICA processing, have no effect on the behaviour of the data. The observations imply that alterations in pressure detected in region A are distinct from those occurring in regions B and C, and that the initial component isolated by fast-ICA is predominantly influenced by pressure fluctuations in region A. The contrast in pressure responses witnessed between regions B and C can be ascribed to the presence of a pressure reducing valve at the entry point of area B, as well as the fact that area C is being supplied by a tank.

These findings provide important insights into the dynamics of the system under study and highlight the utility of fast-ICA in identifying meaningful components in complex datasets. Since this signal source (Fig. 4.4a) does not have visible anomalies, it is not used in the leak detection process of this article. Nevertheless, the information can be utilised to fulfil the requirements of diverse research endeavours, for instance, forecasting demand patterns and scrutinizing extraneous signals present in the data. On the other hand, the component 2 (Fig. 4.5) follows a unique pattern using data from all areas, considered in this study as noise, and is used for the leak detection analysis.

Figure 4.5 - Component analyses 2 via fast-ICA.



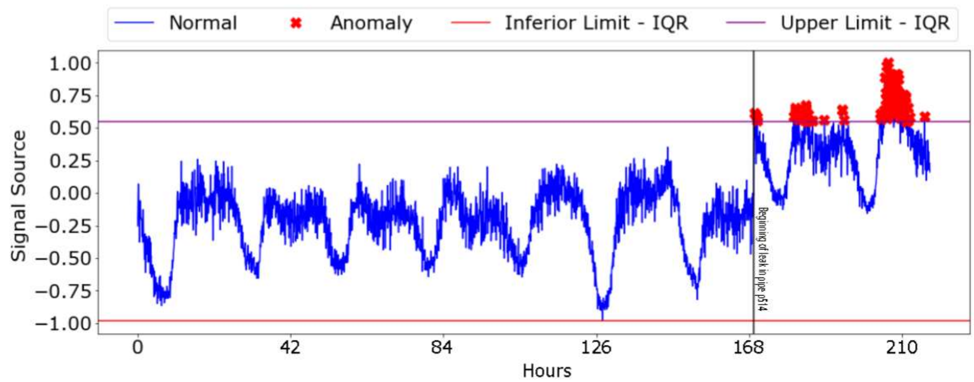
4.4.1 Leak detection via IQR and MP

The component 2 resulting from the ICA application is analysed by IQR and MP methods. For this, data from the separated source is analysed for 7 days before the abrupt leaks occur. However, a new process using a one-day sliding window is performed in leaks that grow gradually and have other leaks during the growth period, considering 4 days before the leak. Moreover, the analysis continues in all cases until 2 days after the beginning of the leaks. This ongoing analysis process allows us to assess whether leaks are detected during the growth of other leaks, how much time it takes to detect it and at what magnitude (flow rate) the detection was possible.

The analysis of two cases of leaks is presented in this work. The first refers to the leak in p514 pipe occurred after about 52 days and has an average leakage flow of 15.4 L/s (25% of total demand). By analysing this leak individually through the separate source by the ICA and IQR and MP methods it was possible to detect it with 10 minutes after its beginning. Figure 4.6 shows an example of the detection process for the leak in the p514 pipe.

Figure 4.6 - Leak detection - p514. a) Detection IQR. b) Detection MP

a)



b)

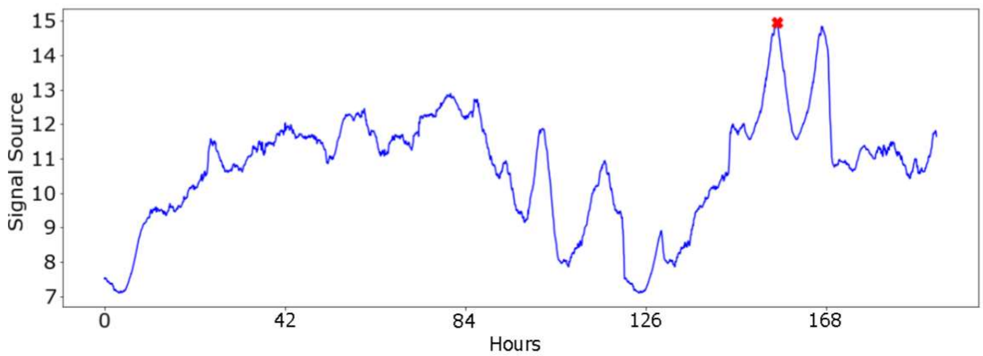


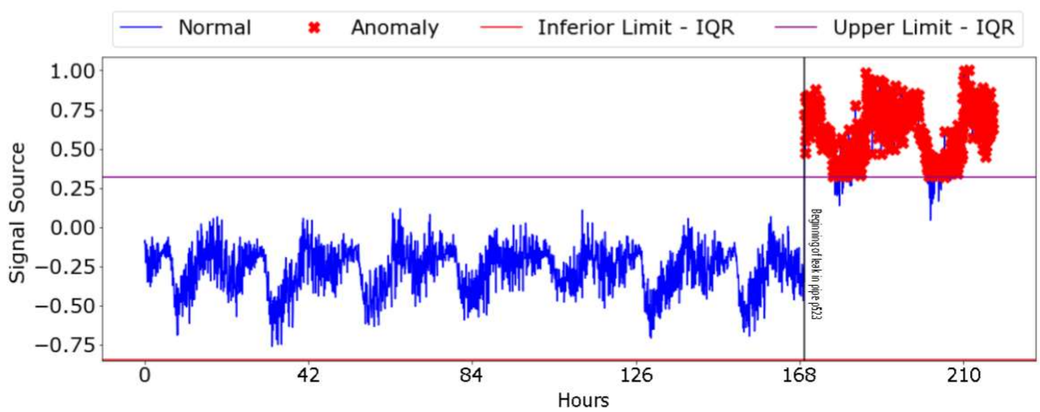
Fig. 4.6a shows the component 2 obtained by the ICA after applying the IQR, as well as the lower and upper limits, the moments when anomalies were detected and a vertical line indicating the beginning of the leak. The IQR method detected the anomaly

10 minutes after the leak appeared. On the other hand, the MP method (Fig. 4.6b) had a delay, detecting the anomaly four-time steps (20 minutes) after the leak started. The MP method requires a certain amount of data for training purposes, and as such, the time steps displayed in Figure 4.6b do not include the initial days used for the comparison process, as these time steps were used solely for training and are not represented in the graph.

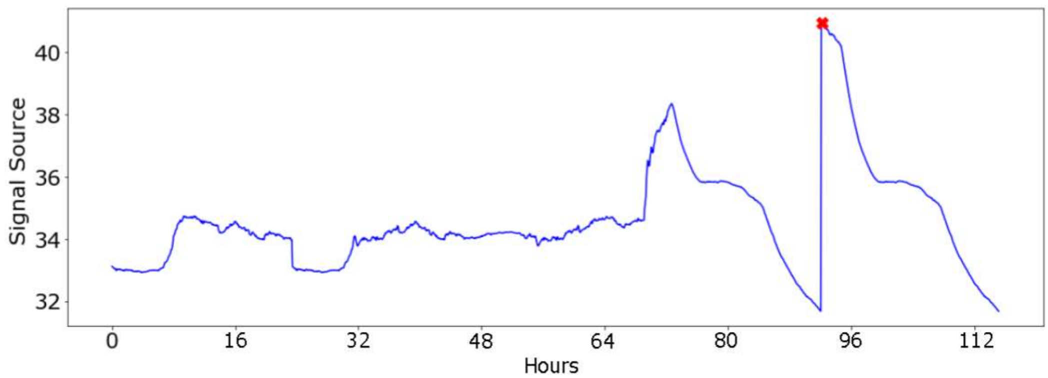
The second example of the detection process is presented in Figure 4.7, which shows the leak occurred in the p523 pipe starting after about 16 days and with an average flow of 28 L/s (48%).

Figure 4.7- Leak detection - p523. a) Detection IQR. b) Detection MP

a)



b)



The detection process presented in Figure 4.6 shows a large variation in the data from the source separated by the ICA. Thus, the detection processes are very accurate, being detected immediately after its start by the IQR method in the first 5 minutes (Fig.5.6a) and 10 minutes by the MP method (Fig. 4.6b). In contrast to the IQR approach, the MP method was utilised to solely detect the anomalous point in the time series, which stands out from the rest.

Two limitations of the proposed methodology should be highlighted: the same IQR limits may not be accurate in all cases and the number of days used by the MP method may affect the detection accuracy. The first limitation occurs due to the changes and number of days considered after the beginning of the leaks, since the interquartile interval is based on the maximum, average and minimum values of the data set. Outliers are easily exposed when an anomaly affects this data very intensely. However, in cases of minor changes the limits can consider this anomaly within the limits. Wan et al. (2014) seeks to solve this limitation developing equations that can determine the values of interquartiles. The second limitation is the number of days used for behaviour analysis in the MP application. In this study, the standard is to use the first 7 days prior to the leak, but sometimes more accurate results are obtained using only the initial 3 days. Table 4.1 presents the information with the best results for all leaks, including the maximum and minimum limits, days used by the MP method, detection delay and amount of water lost until leak detection.

Table 4.1 - Results and definitions - proposed methodology.

	IQR					MP		
Leaks	**Days before**	**Days after**	**Limits**	**Detection (h:m)**	**Flow in detection time (L/s)**	**Days base**	**Detection (h:m)**	**Flow in detection time (L/s)**
p123*	7	1	3.7	11:45	0.01	4	25:50:00	0.01
p142	7	1	3.3	13:15	26.29	4	29:10:00	26.74
p193*	7	5	3.3	71:40:00	0.05	4	126:45:00	0.11
p257					Undetected			
p277*	2	4	3.5	78:50:00	0.04	2	10:05	0.01
p280	2	1	1.9	18:40	5.23	1	17:45	5.24
p331	3	5	1.5	174:30:00	10.53	3	36:45:00	10.54
p426	3	5	1.5	00:55	13.10	3	01:55	13.15
p427					Undetected			
p455*	3	5	1.5	327:10:00	0.47	3	327:30:00	0.47
p514	7	2	1.8	00:10	15.39	7	24:05:00	15.38
p523	7	1	2.5	00:05	28.19	4	20:40	28.07
p586*	4	4	1.5	58:45:00	0.17	4	66:25:00	0.2
p653*	3	5	2.2	01:55	0.2	4	21:00	0.02
p654					Undetected			
p680	4	2	4.1	22:35	5.37	4	01:35	5.36
p710	3	5	2.2	17:15	5.48	3	09:50	5.53
p721*	3	5	2.2	61:00:00	0.3	3	33:55:00	0.3
p762*	3	5	2.2	75:25:00	0.2	3	31:40:00	0.3
p800*	2	4	2.0	09:45	0.2	2	66:50:00	0.47
p810					Undetected			
p827	7	1	2.7	00:05	26.11	4	10:25	26.44
p879*	3	5	2.2	00:45	0.2	3	17:05	0.2

* - Leaks with increasing start

Table 4.1 presents the results obtained by the proposed methodology for each leak, including the absence of detection in three cases (P257, p427 and p654). The processing period before and after the beginning of the leak is shown, and the values for leaks with a non-abrupt start are usually less than the 7 days initially proposed. This is because leaks with gradual onset affect the monitored data less intensively, starting with small flows. In addition, it is necessary to consider more days after the start of the event to identify the flow rate at which the method detected the leak. The processing window is adjusted in some cases of abruptly starting leaks due to the proximity to other events or the fact that they occur while other leaks have not yet reached stability, as occurs in cases p331, p420, p680 and p710.

Table 5.1 shows the V_{HL} values used to detect each leak. A value of 1.5 was initially considered, but values between 1 and 5 were tested to determine which would have better performance in terms of shorter detection time, meaning less time steps after the beginning of the leak. The IQR method generally proved to be more effective than the MP method during the detection process, but it is important to emphasise that the definition of the IQR limits is extremely sensitive to the available sample and quite subjective. The IQR method detected the leak faster than the MP method in 12 of the 19 detected leaks. However, the MP method was more effective in detecting leaks that occur during leaks with increasing onset before stabilizing. This behaviour is especially evident in the detection of p331, p680 and P710 leaks, in which the MP method was faster in the detection.

The days used as the basis for the MP method of each leak are also presented in Table 4.1. The standard value used was 4 days before the beginning of the leak, but the best results in some cases were obtained with different values, always lower or equal to the days of processing used by the IQR method. Finally, Table 4.1 also displays the flow rates of the leaks in the detection times. This process was performed to identify the detection method sensitivity to leaks. It was found that leaks are always identified after reaching 0.1 L/s in the case of leaks with increasing onset, and the highest detection rate was 0.47 L/s in the leakage of p455 pipeline. Flow rates for leaks with an abrupt start remain practically constant, only oscillating with varying pressures. However, it is important to note that it is necessary to make a value judgment regarding the detection time. The difference between the two methods in the first example presented was only 10 minutes, while it was 5 minutes in the second example.

The assessment considers the evaluation process presented by BattLeDIM, which provides an Economic score for the tested approaches. The IQR method shows an Economic score of €354,360, while the MP method shows €353,860. Both values represent water savings throughout the year, the two highest values achieved by battle participants are €264,873 and 260,562, and the perfect score value is €523,124 (VRACHIMIS et al., 2022). This highlights the effectiveness of the leak detection process, as the Economic Scores displayed exceed those of competitors in battle. Some other evaluation methods offered by BattLeDIM are not applicable in this study, as it also focuses on the location of leaks. However, the organisers of BattLeDIM present an open source that in addition to calculating the Economic score, also presents the True Positive and False Positive rates. The present methodology presents 23 True Positives through the code, which is the maximum possible value. This also results in maximum values for Recall, Precision and $F_1 = 1$, which are normally presented ratings. However, it should be noted that the code presents True Positives regarding the leak detection during the time it occurs, which means that even if the detection time is long, the leak is considered covered if it is detected during its occurrence. The detection times after the start of the leak are shown in Table 5.1.

Another aspect within the proposed methodology is evident in the undetected leakages (p257, p427, p654, and p810). These leaks originated before 2019, which is the data used, making the methodology unable to identify changes associated with these pre-existing leaks. This circumstance highlights the need for historical data for applying the methodology. This results in loss of information and consequent anomalies that occur during this data or prior to it are not detected. However, we highlight the possibility of applying the methodology to real-time data analysis or for detecting other anomalies, because there is no need for hydraulic simulations or calibrations. In addition, the methodology gains more and more robustness and accuracy over time due to the implementation of a sliding time window that encompasses the detection of new anomalies in cases where they persist without interruption.

4.5 Discussion and partial conclusions

This study describes a new methodological approach to detect leaks through pre-processing monitoring data using fast-ICA. This pre-processing generates separate signal sources, two of which are identified in the study: one follows the patterns of the monitoring data, and another contains noise. The source with noise is used as input for a detection process using the IQR and MP methods. However, the application of these methods requires some definitions, such as the IQR dimensions and the data used as the basis of behaviour for the MP method. Several values are tested for both techniques and those that presented better performance in terms of the leak detection speed are presented.

Finally, the benchmark data for leak detection methodologies presented in the Battle of Leak Detection and Isolation Methods (BattLeDIM) competition are used to validate the proposed methodology. The methodology can detect leaks with gradual start from 0.1 L/S. The detection for leaks with an abrupt start occurred in a few minutes after the beginning of the leak in some cases, but others were immediately detected, requiring only the data of the next monitoring time step. However, some steps still require adjustments, such as the definition of the interquartile interval values for the IQR method and a better definition for data used as a basis for applying the MP method.

5

Leak Detection in Water Distribution Networks Based on Graph Signal Processing of Pressure Data

This chapter is an adapted version of Barros, D., Souza, R., Meirelles., G., Brentan, B. Leak Detection in Water Distribution Networks Based on Graph Signal Processing of Pressure Data. **Journal of hydroinformatics**, 2023.

Abstract

Leakages in water distribution networks (WDNs) affect the hydraulic state of the entire or a large part of the network. Statistical correlation computed among pressure sensors monitoring network nodes aids the detection and localization of such leaks. This opens the possibility to work with water network databases, where graph signal processing (GSP) tools aid to understand changes in pressure signals due to leakages in the hydraulic system. This paper presents a methodology to time-varying pressure signals on graph structures. The core of this methodology is based on changing of pressure, due to leaks, that modifies the graph structure. Computing for each time step a new topology of the graph and applying centrality analysis based on PageRank, it’s possible to identify the presence of new leaks at water system. A confusion matrix evaluates the precision of the proposed methodology on defining where and when such leakages start and end. Seven leaks are used to validate the process, which presented 86% in accuracy terms. The results show the benefits of the method in terms of speed, computational efficiency, and precision in detecting leakages.

5.1 Introduction

The water distribution network (WDN) is an essential infrastructure responsible for supplying citizens with drinkable water. Therefore, damage to pipes has an important impact on the water distribution process and on water quality and network hydraulics (HU *et al*., 2021). In addition, a large portion of the treated water is lost in the WDN, due to leaks, theft, measurement failures and several other factors (SILVA *et al*., 2021). According to Fan *et al*. (2021) in UK around 3,281 Megalitres of water were lost between 2009 and 2011, and in US around 15% of the treated water supplied is lost annually. In Brazil, it’s estimated that in 2019 there were losses of about 38% in water distribution (OLIVEIRA *et al*., 2021). This shows that even with researches focused on loss reduction and leak detection, the methodologies proposed still require further development, mainly aimed at fast and effective leak detection and easy application in WDN.

Leaks are linked to losses in the water distribution process and are classified as reported, unreported or background leakage (ADEDEJI *et al*., 2017). Reported leaks are visible on the ground and are easily detected by the public or employees of network administrators (CHAN, CHENG and XIONGHU, 2018). On the other hand, unreported leaks are like those reported, but they do not emerge from the ground. And finally, background leakage is small and difficult to detect by normal methods, and often go unnoticed for a long time, resulting in significant losses (Abdulshaheed, Mustapha and GHAVAMIAN, 2017). Different methods are applied to detect leaks, and they can be classified as visual inspection, transient-based approach, model-based approach, and data-driven approach (CHAN, CHENG and XIONGHU, 2018).

Visual and sensor-based strategies require the use of mobile inspection equipment linked to optical, electromagnetic, or acoustic sensors. However, it is an expensive and time-consuming process, and often, especially the acoustic signals, are influenced by the soil type and pipe material (FAN *et al*., 2021). Transient based approaches analyse and evaluate the pressure transient wave caused by hydraulic chances in the system to detect leaks. This wave can quickly travel through the entire network and affect flows, pressures, contract or expand pipes and many other parameters to the system (AYATI *et al*., 2019). Since the wave speed usually are high, from $400 m/s$ for PVC pipes achieving 1200m/s for iron pipes, transient sensors are more expensive and transient responses decay rapidly, making it difficult for large applications (FAN *et al*., 2021). Model-based leak detection methods include, for example, the use of sensitivity matrices created in normal situations and with known leaks (SALGUERO *et al*., 2019; GENG *et al*., 2019) and calibration approaches (SOPHOCLEOUS, SAVIC and KAPELAN, 2019). These methods are capable to detect leaks; however, they require calibrated hydraulic models, with user demand data, pipe conditions and pressure BEHAVIOUR, leading to results very sensitive to modelling and measurement errors (FAN *et al*., 2021; HU *et al*., 2021). Finally, data-based methods include feature classification methods (Sun *et al*., 2019), prediction classification methods (LAUCELLI *et al*., 2016), mathematical-statistical methods (QUIÑONES-GRUEIRO *et al*., 2018) and unsupervised clustering (GEELEN *et al*., 2019). These methods use network monitoring data, such as pressure, flow, and reservoir levels. Hu *et al*. (2021) concludes that data-based detection methods do not require deep WDN comprehension, but require large amounts of data, being more suitable when there is a large amount of historical network data under analysis.

Leak detection process using mathematical-statistical tools has been successfully used in research related to WDN. Brentan *et al*. (2021) use time series of hydraulic data (pressure, flow, and reservoir levels) and apply the algorithm fast Independent Component Analysis (fastICA) to separate the hydraulic data into independent components. The authors analyse the independent components with a statistical control algorithm to detect abrupt changes to identify cyber-attacks in WDN. Gao *et al*. (2014) use a Blind Signal Separation (BSS) process on flow and pressure data to detect leaks and separate the leaks flow from the node total flow. Okeya *et al*. (2014) use historical demand data, a network hydraulic model and a modified Kalman Filter method to detect leaks in networks. The authors predict hourly demand through historical data, apply the predicted data in the hydraulic model to estimate flows and pressures, and apply the Kalman Filter to calculate corrected demands at the current time step driven by the difference between predicted and observed data (ZANFEI et al., 2022).

Computational mathematical methods successfully used in other research fields can be explored and applied to reduce losses in water distribution systems. In this sense, complex networks can be an effective tool for this analysis, since this theory models and identifies interactions through the relationship between objects and has been widely used in network analysis (TAKALA *et al*., 2020). Complex networks are linked to graph theory, but with a more irregular, complex structure that can evolve dynamically over time. The application of this theory began with the effort to define new concepts and measures to characterize the topology of graphs and thus united a series of principles and statistical properties (BOCCALETTI *et al*., 2006).

Systems modelled as a graph are represented by set of vertices and edges (STANKOVIC *et al*., 2019). Turning to the data acquired by the WDN monitoring sensors, it is possible to statistically correlate these data for the creation of graphs. As this monitoring occurs at each defined time steps, the graph is also updated with each time step. Thus, events in the physical WDN can affect the structure of the graph and consequently the metrics that evaluate such graph. In this perspective, studying the relationship and interactions of hydraulic data (e.g., flow, pressure, water quality parameters) can help in the leaks detection and location. Methods linked to graph theory have been successfully used to detect anomalies in radar images (PHAN, MARCIER and MICHEL, 2015), detect connectivity patterns in human brain networks (Farahani *et al*., 2019; Wright, Marco Venneri, 2021), detect anomalies in the integrity of structures (Kaveh, Rahmani and ESLAMLOU, 2022) etc. These researches create graphs from the relationship between the monitored data and as soon as new data are issued, the verification of some graph parameters happen, such as amount of information (PHAN, MARCIER and MICHEL, 2015) and information paths (FARAHANI *et al*., 2019; KAVEH, RAHMANI and ESLAMLOU, 2022).

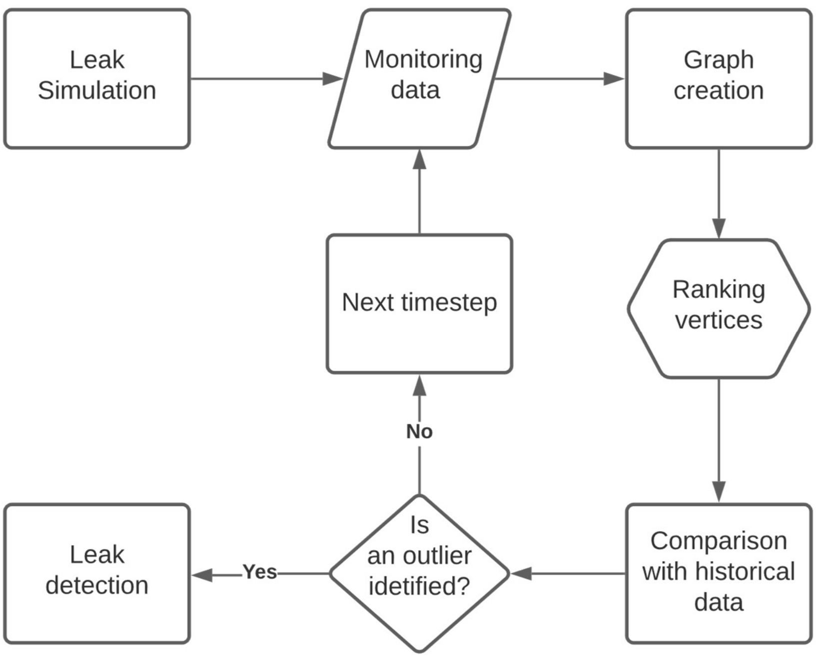
Since the application of data signal processing on water distribution systems is still incipient, this work presents a methodology for leak detection based on graph signal processing of pressure data. For this, the pressure data of each monitored nodes are used to create a graph structure and then, a vertex ranking metric is used. This process is performed with the historical data to recognize the pattern of vertex classifications. When the centrality metric deviates from the historical pattern behaviors, an anomaly is then detected and associated with leaks in the network. The proposed methodology proved to be fast, effective and with low computational effort for leak detection.

5.2 Methods

This section presents the methodology for leak detection based on graph signal processing. The process starts considering the historical pressure data and as they are made available for the graph's creation. For this, the data will be treated in a Python programming environment and the graph signal processing package (PyGSP) (DEFFERRARD *et al.,* 2017) is used to calculate the correlation between the data. The graphs will be created with the correlation calculated at each time step using the Network Analysis package in Python (NetworkX) (Hagberg and Conway, 2020). After creating the graph, PageRank node ranking metric is employed. This metric assigns a rank value to the vertices based on the edge structures. After classifying the nodes with the historical data, as more data is monitored, the new classification value of the vertices is determined, and, if this value does not follow the historical trend, a leak in the network is indicated.

The first step for the development of this methodology is the leak simulation. For this, the Water Network Tool for Resilience package is used (KLISE *et al*, 2017). After the simulations, the pressure data at the monitoring points of the network are correlated and used in the creation of graphs. Every process follows the flowchart shown in Figure 5.1.

Figure 5.1 - Flowchart of the methodological process.



To model a graph (G), in general, its objects are determined, called vertices (V) and the interaction between these vertices that is called edges (L), being then the graph representation as $G = (V, L)$. The graph and its interactions can be expressed as an adjacency matrix (A), which is a Boolean matrix with columns and rows indexed to vertices. Thus, when there is the interaction between the vertices, the edge is represented as value 1 in the matrix A. A weighted graph can be created by adding values to the edges, which is called edge weights. This results in a matrix of weights (W), now no longer Boolean, because it contains values other than 0 and 1 (STANKOVIC *et al.*, 2019).

The modelling of WDN as graphs is possible by modelling considering the network physical topology, with the vertices corresponding to the nodes (junction nodes, tanks, and reservoirs) and the edges representing the links (pipes, pumps, and valves) (DI NARDO and DI NATALE, 2011). Another way to model the WDN as a graph is by the correlation between the monitored data. In this case, the vertices represent the monitoring nodes, and the edges are the interactions between these nodes (BEZERRA *et al*, 2022).

5.2.1 Graph creation via pressure data correlation

The graphs are created using pressure data at each time step. For this, the research of Kalofolias (2016) is used. The authors present a process of creating graphs from temporal data, calculating distance between pairs of vertices and penalizing edges with high weights. The authors use a X data matrix, where the columns $(x_{v1}, x_{v2}, \cdots, x_{vi})$ represent the sensors and the rows the time signals. The method calculates the distances between data pairs, creating a Z distance matrix:

$$Z_{ij} = \|X_{vi} - X_{vj}\|^2 \tag{5.1}$$

denoting the distance between the temporal data of vertices v_i and v_j. This method assumes that the smoothness between Z and the edge weight matrix (W) is small:

$$\sum_{v_i} \sum_{v_j} W_{v_i v_j} Z_{v_i v_j} \tag{5.2}$$

and create a graph that minimizes very high weights on W, so that:

$$\min_{W \in \mathbf{W}_*} \|W \circ Z\|_{1,1} + f(W) \tag{5.3}$$

where $\boldsymbol{W}_*$ is a set of adjacency matrices (determined by the variation of weights over time). The matrix function $f(W)$ prevents the matrix W from having a value of zero, controlling its sparsity and imposing a data-dependent structure.

This methodology calculates the distance between data pairs to create the matrix W, being implemented by the package PyGSP. The data are separated into a square matrix (X) where the columns represent the pressure data acquired by a set of sensors $(N)_S$ in the network and the rows are the temporal data (t). A graph will be generated as monitoring data becomes available. However, the first graph will have the amount of data equal to the number of sensors, this to generate a square matrix of weights W. Once a new data is available, the oldest data is removed, always keeping the matrix W square. Finally, the matrix W is used in the graph creation and the vertices are ranked with each new graph.

5.2.2 Leak detection using ranking vertex

To measure the importance of web pages, Page (1999) proposed a method to rank each page based on the web structure. PageRank algorithm can be described as a way of evaluating the web page importance, based on the links quantity and quality that direct to it. In a classic discrete-time finite-state random walk model, Yao, Mark and Rabbat (2012) denote by P in $n\,x\,n$ transition matrix, where n is the number of states and P_{gh} is the probability of transition from state g to state h. A stationary distribution value s is defined as:

$$\boldsymbol{s}^T = \boldsymbol{s}^T P \tag{5.4}$$

$$s.t. \quad \boldsymbol{s}_g \geq 0 \quad and \quad \sum_{g=1}^{n} \boldsymbol{s}_1 = 1 \tag{5.5}$$

where s^T is the eigenvector of P that corresponds to the eigenvalue 1.

PageRank ranking modifies the random walk model to:

$$\boldsymbol{s}^T = \alpha\, \boldsymbol{s}^T P + (1 - \alpha)\boldsymbol{t}_v^{\boldsymbol{T}} \tag{5.6}$$

where $\boldsymbol{t_v}$ is a column vector named teleport vector, which satisfies $\sum_{g=1}^{n} t_{vg} = 1$ and α is a scalar called the damping factor. The teleport vector and the damping factor are considered so that there is communication between pages (vertices) in case of graph not connected. Thus, the equation 5.6 corresponds to the union of two random walks, one with the matrix P and the other that is the transition from state g to any other state with probability t_v, and has α as mixing parameter. In the initial iteration all vertices are considered equal with $t_{vg} = 1/n$ and $\alpha = 0.85$ (YAO, MARK, and RABBAT, 2012), but over the interactions the weights of vertices and edges are being updated. This makes the convergence rate limited and empirically mimics the behavior of web users.

The PageRank values of vertices are given in simplified form by:

$$PR(u) = \alpha \sum_{c \in \boldsymbol{B}(u)} \frac{PR(c)}{N_c} \tag{5.7}$$

where u represents the web page (or the vertex in a graph), $B(u)$ is the set of vertices that point to $uPR(u)$, $PR(c)$ are classification scores of vertices u and c; N_c is the number of edges leaving vertex c (XING *et al*., 2004).

To detect leaks, we used the maximum values determined by the PageRank metric. The maximum score does not refer to a specific vertex, but the highest value achieved by all vertices. The data of the first 7 days that are simulated without leaks, are used as maximum base values, and are compared with the data of the next days of simulation. If this comparison exceeds a threshold, it will be considered a data anomaly. The determination of this stipulated value will be determined by a sensitivity analysis and chosen the percentage that presents the best result. Finally, a confusion matrix will be generated between two sets of data, the first considering the simulated leaks, where the moments of beginning and end are known, and the second with the times with anomaly detections.

A confusion matrix is generally used in classification problems between two sets of data in four combinations: true positive (TP), true negative (TN), false positive (FP) and false negative (FN) (BERRY *et al*., 2021). Table 5.1 shows how the classifications are arranged in the confusion matrix.

Table 5.1 – Confusion matrix example.

		Observed	
		Positive	Negative
Estimated	Positive	TP	FP
	Negative	FN	TN

Between the two data sets in the confusion matrix, there is one with the real observations, in the case of this research they are the known leaks, and another with the observations predicted by the method. In other words, in this work, the observed data are the time that the known leaks start, and the estimated data are the time that PageRank values vary significantly from the normal trend, higher than a marginal error.

To evaluate the confusion matrix, the accuracy metric, which is the correct proportion of the method, is used. For this, the equation follows:

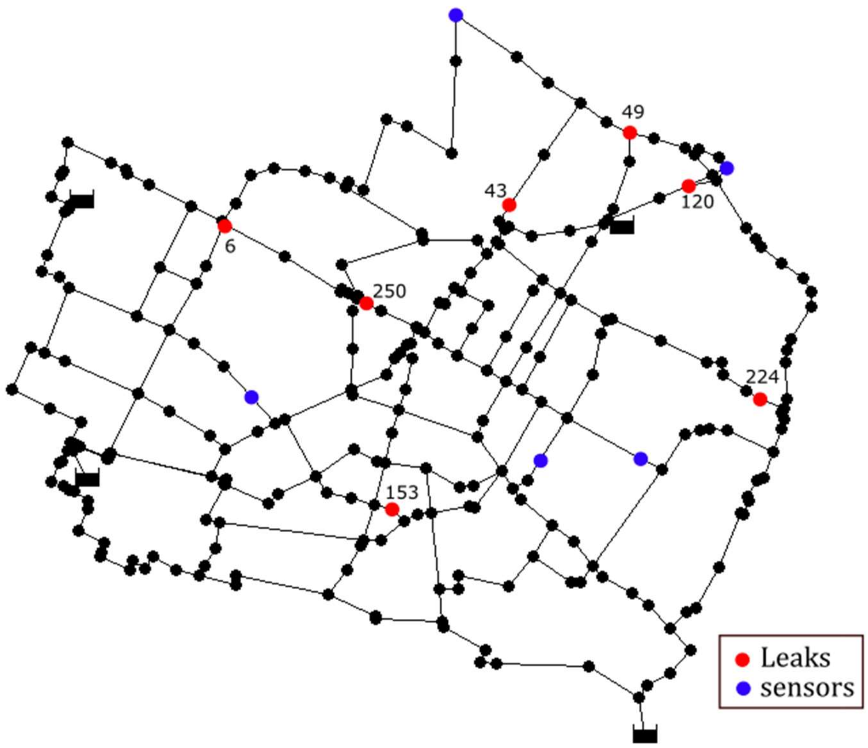
$$Ac = \frac{TP+TN}{TP + TN + FP+F} x\ 100 \tag{5.8}$$

The results are between 0 and 100, with 100 being a perfect prediction result (BERRY *et al*., 2021).

5.3 Case Study

The WDN presented by Bragalli *et al.* (2012) is adapted and a set of leaks are simulated for the application of the methodology presented in this work. This network is based on the distribution network of the Italian city Modena, in the Emilia-Romagna region, and it is composed of 268 nodes, 4 reservoirs, 317 pipes and does not have valves and pumps, as shown in Figure 5.2.

Figure 5.2– Modena network



The research of Mankad, Balasubramaniam and Babji (2021) uses this network to place pressure sensors to detect and locate leaks. The authors placed 5 pressure sensors (exposed in Fig. 5.2) and only the data from these sensors will be considered in the process of graphs generation. This means that the X data matrix (equation 4.1) will have 5 columns and will result in graphs with 5 vertices. A leak simulation process is also required due to lack of leakage scenarios in this network. Therefore, 7 nodes were selected with leak points, and these are shown in Figure 5.2.

5.3.1 Simulation process

Since there is no real monitoring data for this network, a leak simulation process will be performed using the Water Network Tool for Resilience (WNTR) package (KLISE *et al.,* 2017). Leaks will be considered as an additional demand flow at the network nodes, with the leakage flow rate (q) being determined by the orifice equation:

$$q = C_d A \sqrt{2gP} \tag{5.9}$$

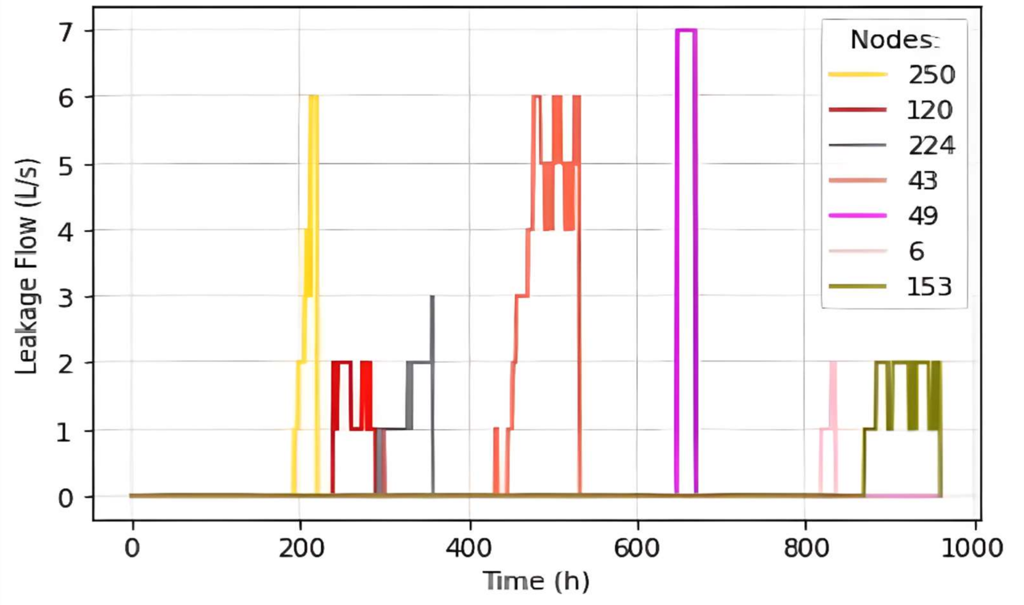
where C_d is the discharge coefficient, A is the orifice area, g is the gravity acceleration and P is the pressure at the leak point. As the leak flow affects the node pressure and the pressure determines the leak flow, an iterative process must be carried out until the pressure and the leak flow rate stabilize and then this flow is used in addition to the demand. The C_d utilized is equal to 0.6, following the research of Van Zyl (2014). The area values will be chosen randomly between 0.00012m^2 and 0.00050m^2, values used in the research of Van Zyl and Cassa (2014).

Another factor with random choice will be the leaks start times and their durations. The entire simulation process is done for 40 days, with the first 7 days without leaks to obtain a reference historical data. After that date, leaks are included. During the simulation process the base demand and the hourly multiplication factor are multiplied by a randomness factor, in this study considered between 0.9 to 1.1, making the process closer to reality. The sensors have an acquisition frequency of one hour, meaning that a graph will be created, and PageRank will be calculated every hour.

5.4 Results and discussion

Seven leak scenarios are performed at different locations on the network. The leaks are simulated with different magnitudes and behaviors: in some cases, there is an increase in flow over time and in others there was an abrupt appearance of leaks. Figure 5.3 shows the behavior of leak flows for each node where leaks are simulated. It can be observed that there were leaks occurring in a spaced way and a situation of simultaneous leaks in different locations (nodes 120 and 224).

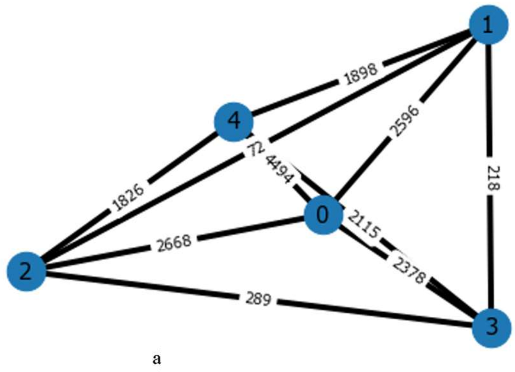
Figure 5.3 - Leakage flow behaviour



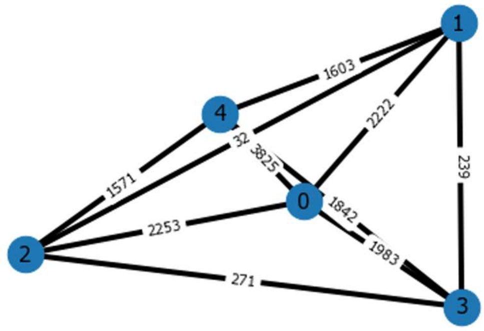
The entire simulation process considered that the sensors emitted monitoring signals every hour, resulting in a dataset with 960 measurements. In this sense, there are the same amount of graph creations and nodes rankings by the PageRank metric, since each measurement is used in the process. The graphs change as pressure data changes at monitoring points. This change can be observed in the edge weights in Figure 5.4.

Figure 5.4 - Graphs in anomalous situations.

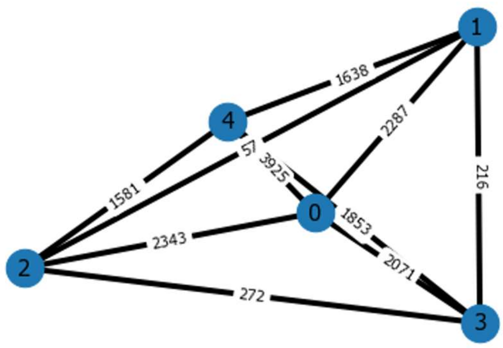
a) without leakage



b) node 250



c) node 120



d) node 49

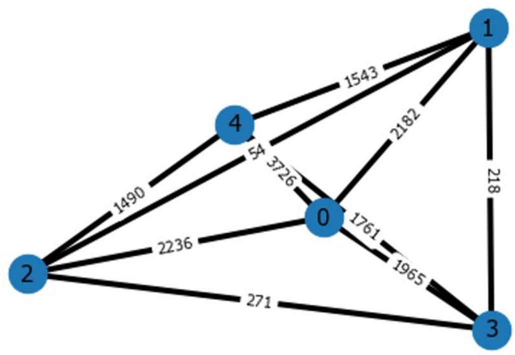
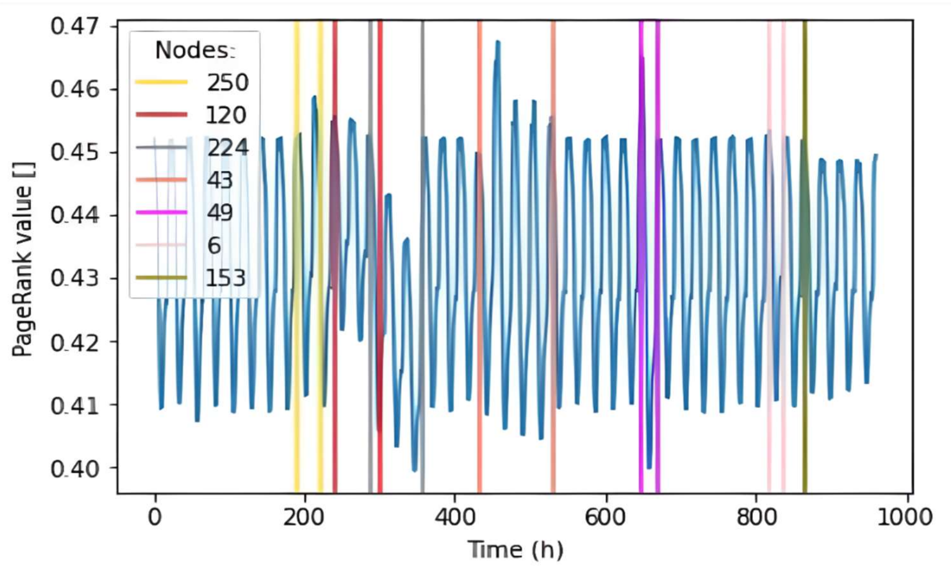


Figure 5.4 shows four distinct situations of the weights in the graph. The first (Fig. 5.4a) exposes the graph in a leak-free situation. And the other figures (Fig. 5.4 b, c, and d) show the graphs generated at the peak moments of the leaks at the 250, 120 and 49 nodes, respectively. Thus, Figure 5.5 shows these values throughout the simulation, including the moments when there are leaks.

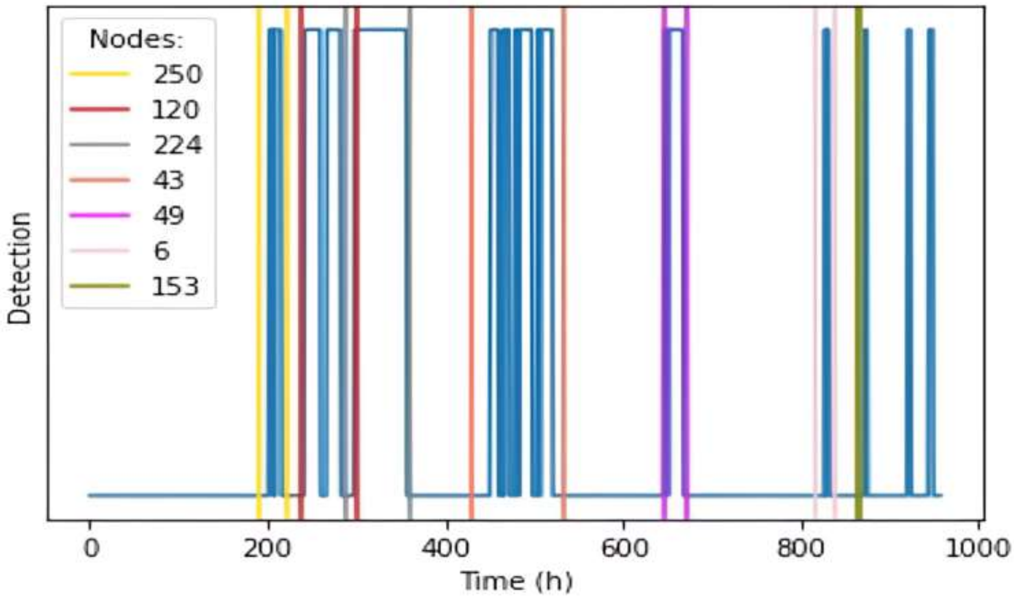
Figure 5.5 - PageRank maximum values



It can be observed in Figure 5.5 that the maximum PageRank values change significantly when there are leaks in the network, not having much relation with the intensity of the leaks. It can be seen, for example, that leakage at node 250, even with a high value of leakage, does not significantly impact the behavior of PageRank values, especially when compared to leakage at nodes 120 and 224, which are of lower intensity, but which cause considerable changes in PageRank values. On the other hand, the leak at node 6 does not cause a change in behavior, probably because it is a short time, low-flow leak, which the sensors may not have been able to monitor.

To define the marginal error, a sensitivity analysis is performed, in which the process varied between 1 and 10% the hourly PageRank values calculated during historical data (7 days of data) looking for the highest true positive values. This analysis showed that, considering a margin of error higher than 5%, the TP values are very low (134 - 13%, times detect and percentage of total). However, if the marginal error is 1%, the detection has a high value of false positives (202 - 21%). Thus, it indicates a margin of error of 2%, both in this application and in future applications, as it is the percentage that achieves the best TP values. The detection moments using the margin of error equal to 2% is exposed in Figure 5.6.

Figure 5.6- Anomaly detection.



It can be seen in Figure 5.6 that some leaks are detected throughout their duration (nodes 120, 224 and 49). As for the leaks at nodes 250, 43 and 6, the detection takes place in certain time steps, but during a large part of the leaks' duration. However, the leak at node 153, which lasts until the end of the simulation, is detected in just a few moments. For a better understanding of the results, a confusion matrix is created and is shown in Table 5.2.

Table 5.2 – Confusion matrix.

		Observed	
		Positive	Negative
Estimated	Positive	655(68%)	14(1.46%)
	Negative	113(11.7%)	176(1.3%)

Evaluating the confusion matrix in terms of accuracy, the detection method obtained an 86% score. This means, according to Chicco, Totsch and Jurman (2021), that there is high accuracy in leak detection. This shows that the leak detection process proposed in this research guarantees a good detection rate using monitoring data already performed based on the application. The proposed methodology is easy and quick to execute, proving to be an effective approach to leak detection.

5.5 Discussion and partial conclusions

This research presented a methodology for leak detection that makes use of the ranking of vertices in graphs by the PageRank metric. The proposed methodology creates graphs with the correlation between the pressure data emitted by monitoring sensors using graph creation and analysis packages in a Python programming environment. It was shown that the graphs undergo changes when leaks occur. Thus, there are also changes in the vertex ranking values, and these changes are used to detect leaks. A historical series of hourly pressure data is used to learn the vertices rank for the different days, and the leak is detected when the maximum PageRank value of the vertices is different from those learned from the historical data. The proposal achieved a score of 86% based on seven simulated leaks with the results evaluated in terms of the accuracy of a confusion matrix. The methodology proved to be easy and quick to apply.

Even so, there are still points to be improved. Some of these points can be addressed in future works, seeking, for example, to use new ways of creating graphs, such as methods that consider the topology of the network. New vertex classification metrics can also be used considering, for example, the centrality of the vertices or the number of degrees and using other sources of monitoring data, such as: flows, reservoir levels and water quality. Another approach that can be followed as plans is leak detection by methods linked to graph theory, as through the connection between the sensor nodes and the other network nodes.

6

Temporal Graph Analysis for Leak Detection: A Water Quality Perspective

This chapter is an adapted version of Barros, D., Brentan, B. Temporal Graph Analysis for Leak Detection: A Water Quality Perspective. **Journal of hydroinformatics**, 2024. (Under review)

Abstract

To reduce losses in water distribution networks, data-driven approaches have been explored, mainly with the aim of detecting leaks. Detection approaches have used monitored hydraulic data, such as pressure, flow, and tank levels, however monitored water quality data can present anomalies associated with leaks more intensely than hydraulic data. Therefore, this paper presents a methodology for detecting leaks through statistical analysis of pressure, water age and residual chlorine concentration data. By a preprocessing, the correlation between the monitored data is used as a weight matrix to create a temporal graph, and the z-score algorithm analyzes the graph structural values and points out anomalies. The detection method proved to be more effective when using the ranking values of the graph vertices created with water age and chlorine concentration data, especially in comparison to the use of pressure data.

6.1 Introduction

The reduction of losses in water distribution networks is continually studied to reduce, of course, the physical losses of the resource, but also reduce the price distributed to consumers. Losses in the water distribution stage are related to leaks, frauds and measurement failures and can represent a significant portion of the water collected and treated by companies (BOZTAŞ et al., 2019). Take as an example the Brazilian average of losses in 2021, which was approximately 40% of the water collected (TRATA, 2023). Furthermore, leaks can cause problems to urban infrastructure, such as structural damage, soil erosion and interruption of services (ADEDEJI et al., 2017). And repairing these leaks can influence other networks, such as traffic, internet, and energy. Therefore, WDN control approaches are studied and presented so that operators can take accurate forecasting and repair measures.

Leak detection approaches, especially for non-apparent leaks, are constantly evolving. An example of this are acoustic methods, widely studied in the 1990s (FUCHS and RIEHLE, 1991; HUNAIDI and CHU; 1999). This method is initially carried out manually using microphones and the operator's experience determined the occurrence and location of the leak (HUNAIDI et al., 2004). Due to the difficulty related to on-site inspection and the impossibility of analyzing the entire network in this way, noise analysis began to use automated methods to detect anomalies. Different approaches are presented for this, for example, microphones were installed at the beginning and end of the pipes and the noises from the microphones are compared and analyzed by statistical analysis algorithms in search of anomalies associated with leaks (SANTOS et al., 2013). However, acoustic inspection approaches require the installation of microphones in many pipes, which makes its application difficult. Furthermore, acoustic monitoring is influenced by noise from other networks, such as traffic, gas, and sewage, in addition to common everyday noise pollution (NAVARRO et al., 2020).

Hydraulic monitoring of the network at fixed points is also used to detect leaks in WDN. Pressure, flow, and reservoir level data are widely explored in anomaly prediction and detection (DARSANA and VARIJA, 2018). To achieve this, research addresses different techniques, for example, Di Nardo et al. (2015) which uses a genetic algorithm trained with historical pressure and flow data to predict the monitored values, if the new monitored value is not close to the predicted value, an anomaly is pointed out. Shao et al. (2019) also use monitored pressure and flow data to detect leaks by comparing monitored and simulated data. However, the method proposed by the authors is strongly influenced by noise in the data, since no pre-treatment of the data was not carried out, in addition, smaller leaks are not identified because they are ignored due to modeling uncertainties and the need for large memory. computational system to store data from simulated scenarios.

Model-based detection methods use WDN computational modeling, estimate hydraulic behavior through simulations and point out anomalies when monitored data do not coincide with simulated values. Nasirian et al. (2013) presents an approach that uses network calibration to detect and estimate the size of leaks, through comparison between monitored and simulated data. Moasheri and Ghazizadeh (2021) also present a model-based method for detecting anomalies by comparing simulated and monitored data. To do this, the authors calibrate the nodal demands and the roughness of the pipes and estimate the pressure values at different points in the network. However, model-based methods require a much information to calibrate the network, such as pipe roughness, nodal demands, pipe, and pump flows, as well as other information that may not be easily acquired or information that changes over time, such as the pipes roughness (JADHAO and GUPTA, 2018).

To avoid calibration processes, data-driven methods have also been explored to detect leaks in WDN. Detection approaches that use data-driven methods analyze monitored data to identify deviations from behavioral patterns, unusual data, associate data behavior with known anomalies, or outline points in the data (HU et al., 2021). Daniel et al. (2022) present a detection method data-driven through semi-supervised linear regression analysis of pressure data, using data from pairs of sensors. However, the method presented by the authors can be influenced by fluctuations in pressure values, as these change over time. Barros et al. (2023b) present a data-based detection method, but the authors use the correlation between the monitored data as a way of pre-treatment of the data, using the value of this correlation to create a temporal graph and analyze the topological information of the graph to detect leaks. The authors state that analyzing the structure of graphs can favor the application of statistical data analysis methods as they present marked anomalous behavior when leaks occur.

However, the approach presented by the authors can still be explored using all the information in the graph and analyzed using statistical methods.

Graph theory has favored the management and control of WDN, through structural and relationship analysis in WDN modeled as graphs (SITZENFREI, 2021). A graph has vertices, which represent objects, and edges representing the interaction between objects (BIGGS et al.,1986). To represent a WDN, research has represented demand nodes, tanks and reservoirs as vertices, the edges represent pipes, correlations between demand and pressure data. Torres et al. (2017) explore the performances of several WDNs using hydraulic, water quality and demand information as a basis for creating the graph. The authors show that the metrics related to graph analysis have strong links with network performance measures, such as maximum losses and average water age, and the authors also indicate the use of these metrics in approaches to detecting and locating leaks.

Using different data sources can favor the detection of anomalies using data-driven approaches. For example, as mentioned by Torres et al. (2017) which uses water quality information to create graphs and explore WDN performance. In this sense, Barros et al. (2023a) exposes the advantages of using water quality data to detect leaks. The others carry out leak simulations and monitor pressures, water age and chlorine concentration, in which more significant changes are observed in water age and chlorine concentration data in cases of leaks. The authors conclude and explain that to meet the flow of leaks, water may travel different paths or even require the supply of another tank or reservoir, a situation that changes water quality parameters.

In general, approaches that use PageRank values as a form of data pre-processing proved to be more efficient for detecting anomalies. Furthermore, water quality data are more affected in cases of leaks, which favors the detection process. Taking it into account, this research presents a data-driven leak detection approach that utilizes data on pressure, water age, and chlorine concentration. The correlation among the monitored data is used to create the temporal graph, in which all vertices in the graph are classified temporally using a graph analysis metric. The ranking values are then analyzed by a statistical analysis algorithm, which points out anomalies in the ranking data for each vertex and for each type of data used to create the graph. The study shows that the classification of data referring to graphs created with data on water age and chlorine concentration favors detection compared to the same approach that uses pressure data.

6.2 Material and methods

The present study is mainly based on the research of Barros et al. (2023a) and Barros et al. (2023b), in which the first research is used as a basis for the process of mathematical and computational modeling of leaks, and the second research used as a basis for the creation of a temporal graph based on the correlation between monitored data. This research is compound by computational simulations of leaks, creation of a temporal graph through the correlation among monitored data and the data anomaly detection via the Z-score statistical analysis algorithm.

6.2.1 Leakage modeling and simulation process

The leak simulation process is carried out in a Python programming environment using the Water Network Tool for Resilience (WNTR) library (KLISE at al., 2017). The WNTR library is based on the EPANET software (ROSSMAN et al. 2000), however, due to its use in a programming environment, it is possible to include different methods and algorithms both during the simulation process and in the analysis of the results obtained in the process. To simulate leaks with WNTR, it is possible to use the emitter equation, which is already implemented in the library and software. However, when using the emitter equation, it is not possible to select specific periods for the beginning and end of the leaks, since when including an emitter, it is considered throughout the entire simulation time.

However, using the standard orifice equation 6.1 it is possible to determine the leak flow rate, and add it as demand to the WDN nodes during the simulation. This determines the size of the leak, start and end times and whether leaks with an abrupt start or gradual growth in flow will be considered. For this, the orifice equation is described by:

$$q = C_d A \sqrt{2gP} \tag{6.1}$$

where q is the flow rate, P is the pressure head, C_d is the discharge coefficient, A is the orifice area and g is the gravity acceleration. The computational hydraulic simulation considers a total simulation time, and leaks are distributed across demand nodes in the WDN, with different dimensions and starting shape, all chosen randomly. The size of the leak is considered through the variation in area A in the equation 6.1 and the form of onset, such as abrupt or gradual growth, is determined by multiplying reduction factors until it reaches the maximum value of q.

The simulation process also considers monitoring sensors previously installed on the network and saves hydraulic and water quality data at the monitored points. Pressure data is used as the standard approach in the present research, since it is the data normally used for leak detection. However, water age data will also be used for leak detection. The water age information is related to the time that the water travels from the reservoir until it reaches the demand nodes (ROSSMAN et al. 2000). To consider this information in the simulation process, the 'AGE' function is used as a quality parameter in the WNTR library.

In addition to the water age, an additional simulation process is carried out, with the same characteristics as the simulated leaks, but considering the chlorine concentration. For this, the analyzed WDN reservoirs are considered as a source of continuous chlorination at a concentration of 3 mg/L of chlorine. Simulations with chlorine concentration also consider the flow reaction coefficient at -2.5 and pipe wall reaction coefficient equal to 0.15 (ROSSMAN et al. 2000; BARROS et al. 2023b). This information is added prior to the beginning of the simulation process, which includes information on chlorine concentration in the reservoirs individually and information on reaction coefficients globally. This approach during the simulation process uses the 'CHEMICAL' function as a quality parameter. The data obtained in this simulation process is then used in the leak detection process, but previously treated through the correlation between the monitored data to create a temporal graph.

6.2.2 Temporal graph creation

Barros et al. (2023b) presents an approach for creating a temporal graph based on the correlation of monitoring data. To do this, the authors consider a matrix of monitored data X, with each column of this matrix referring to a monitoring node (v_1, v_2, ..., v_i)

and the column lines refer to the data monitored at time t. Thus, the Euclidean distance of the data monitored at time t for each column v is assigned to the distance matrix Z and calculated by:

$$Z_{ij} = \|X_{vi} - X_{vj}\|^2 \tag{6.2}$$

A matrix Z is created every t in X and this matrix is used as a weight matrix in the graph creation (KALOFOLIAS, 2016). As the matrix Z is updated every t, this new information is assigned to the graph, thus creating a temporal graph.

A graph is normally represented by $G = (V, E)$ where V is the vertices of the graph and E is the set of edges (SITZENFREI, 2021). The representation of the graph is also possible through matrices, in which an adjacency matrix represents the relationship between the vertices, in which it is a Boolean matrix indexed by the vertices of the graph (BIGGS and = LLOYD, 1986). However, to represent graphs whose vertices have strong or weak relationships between them, a weight matrix (W) is used. The matrix W, also indexed by the vertices, has values assigned to the connection between the vertices, with values that can represent different types of information, for example, research related to WDN analysis normally uses hydraulic information such as flow, length, and roughness of pipes (TORRES et al., 2017; SITZENFREI, 2021).

The present research uses the matrix Z as the weight matrix W in creating the temporal graph (G_t). The vertices of G_t refer to the columns of X, that is, the graph G_t has vertices that represent the network monitoring nodes. The edges of G_t are the Euclidean distance values between the data monitored at each t. Barros et al. (2023b) observed that the graph structure changes temporally, and that these changes also reflect anomalous behavior in the WDN, for example in cases of leaks. Therefore, analyzing the temporal graph structure is treating of monitored data, since anomalies may not be strongly reflected in the monitored data. However, anomalies related to leaks can strongly impact the correlations between the data and the graph structure.

6.2.3 Detection approach via ranking vertex analysis

The structural values of G_t change over time, which is reflected in the behavioral analysis of the importance of vertices and edge weights. Therefore, analyzing the importance of vertices and edges in a temporal graph can shed light on issues not observed in other analyses. Therefore, the present research analyzes the structure of G_t based on the importance of the vertices. For this, the metric presented by Page (1999) is used, which ranks the importance of vertices based on the general structure of the graph, and the quantity and quality of edges directed to the vertices. The ranking metric presented by Page (1999) is called PageRank and is determined by:

$$PR(v_i) = (1 - d) \sum_{v_j \in V_{v_i}} \frac{PR(v_j)}{N_{v_j}} \tag{6.3}$$

where v_i is the vertex in analysis, V_{v_i} is the set of vertex that are connected to v_i. $PR(v_i)$ and $PR(v_i)$ are ranking score of vertex v_i and v_j. N_{v_j} denotes the number of outgoing edges of vertex v_j. d is a dampening factor that is usually set to 0.85.

Barros et al. (2023b) showed that PageRank values change over time and associated these changes with network leaks. However, the authors only use the maximum PageRank values, referring to a vertex of the graph, and detect changes based on direct comparison between the initial days of the simulated data, in which there are no leaks. The present research uses the PageRank values for all vertices of the graph G_t, that is, all the sensors in the network, and uses the z-score algorithm to statistically analyze and point out anomalies in the PageRank values.

6.2.3.1 z-score algorithm

The z-score algorithm developed by Altman (1968) evaluates the relationship of a value in terms of the mean and standard deviation in a set of values. The algorithm was initially developed to evaluate the prediction of company bankruptcy using financial and economic indices. However, the use of the z-score in other analyzes has proven effective, such as detecting anomalies in health, behavioral and investment data (ANUSHA et al.,2019). The z-score value is calculated in this research by:

$$Zs_{t,v_i} = \frac{PR_{(t,v_i)} - mean(PR_{(ws:t-1,v_i)})}{std(PR_{(ws:t-1,v_i)})} \tag{6.4}$$

where Zs_{t,v_i} is the z-score value for the data at vertex v_i at time t. $PR_{(t,v_i)}$ is the PageRank value of vertex v_i at time t. $mean(PR_{(ws:t-1,v_i)})$ is the mean of the PageRank values for v considering the temporal data $t-1$ up to a stipulated window size ws. The ws refers to a quantity of data prior to $PR_{(t,v_i)}$ under analysis, and consider a current evaluation, that is, a sliding window of data that runs through the data. At each iteration the window moves considering the next data and removing the oldest data. Thus, ws always has the same amount of data, but the behavior of the data throughout the analysis is considered when evaluating new data.

The analysis also considers a sliding window to allow evaluating data that changes over time, an example is related to the pressure behavior in the WDN, which changes over time due to changes in the roughness of the pipes and unrepaired background leaks. This research uses a sliding window of 7 days of data to obtain mean and standard deviation values, as it contains demand behavior on all days of the week. However, the process individually evaluates the data used in the first sliding window, that is, the data monitored during the initial 7 days, also evaluates the data from the first day. At the end of the evaluation of the first 7 days, each new monitored and analyzed data included in the data window slides and the oldest data is deleted.

To evaluate the performance of the proposed method, simulated leaks are used, since the start and end times are known, and the time elapsed after the start until detection is indicated. Leak simulations are used mainly because there is no monitoring of water quality in benchmark problems related to detecting and locating leaks. Detection is considered correct if it occurs during the leak, determined by:

$$t_{st}^{l} \leq t_{d}^{h} \leq t_{end}^{l} \tag{6.5}$$

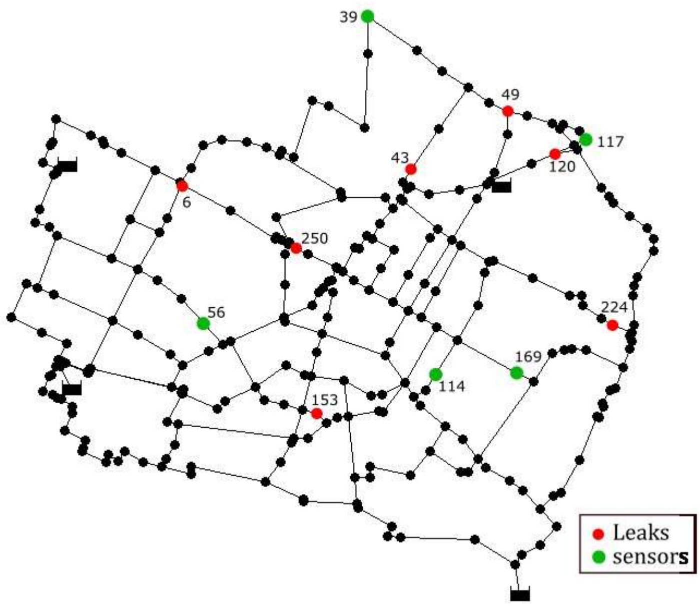
where t_d^h is the detection time, t_{st}^l and t_{end}^l is the start and end time of leakage l (VRACHIMIS et al., 2022). Thus, the detection time after the start of the leaks will be evaluated, and if detections occur that do not follow the equation 6.5 this is considered a False Positive, a failed detection.

6.3 Case study

As a case study, this research uses the Modena network presented by Bragalli et al. (2012). The Modena network (Fig. 6.1) is based in an Italian city and has 268 demand nodes, 4 reservoirs and 317 pipes. This network was also used by Barros et al. (2023a) and Barros et al. (2023b) in leak simulation and water quality analysis processes. Therefore, to provide a more robust analysis of the results, the same hydraulic parameters are used for leak simulations (flow, duration, and form of onset) and quality simulations (water age, chlorine concentration and reaction coefficients).

Figure 6.1 - Case study information

a) Modena Network



b) Leakages

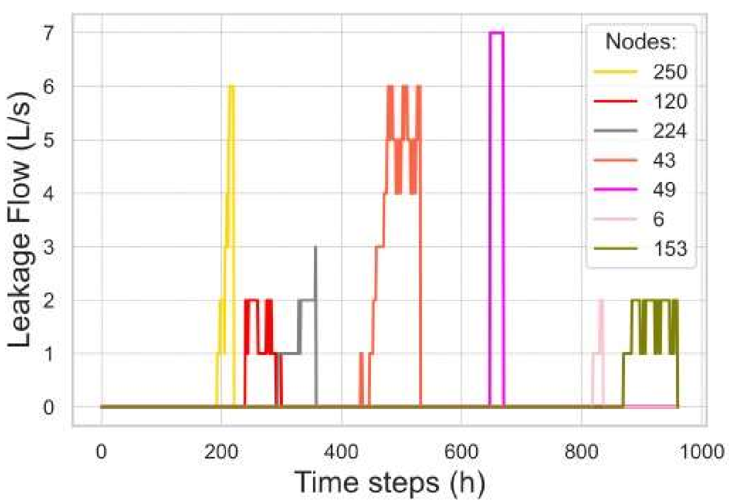


Figure 6.1a shows the network topology, in addition to the nodes used as leak sources and sensor nodes. The sensor nodes follow those presented by Mankad et al. (2022). Figure 6.1b shows the duration and flow rate of the simulated leaks. The simulation process follows, as mentioned, the standards presented by Barros et al. (2023b) in which it considers the duration of the simulation to be 40 days, data monitored every hour, and a sliding window of 7 days. However, the authors consider a random rate of the multiplication factor of the base demand of the consumption nodes. In this work, random values varying by 10% of the hourly multiplication factor are also considered.

At the end of the leak simulation process, three matrices of monitored data X are obtained, referring to data on pressure, water age and chlorine concentration. Each matrix is analyzed individually, in which the temporal graphs referring to each matrix are created, the vertices are ranked and anomalies in the ranking value of the vertices are detected using the z-score algorithm. Furthermore, the data is evaluated individually at each time t, which reflects the application in cases of real-time data analysis.

6.4 Results and discussion

The results presented corroborate the research by Barros et al. (2023a) and Barros et al. (2023b) in different aspects. Firstly, the use of quality data proved to be effective in detecting leaks. Furthermore, analyzing the structure of graphs can also be an efficient way of pre-processing data since its structure and PageRank values change in cases of anomalies in the monitored data. Therefore, to demonstrate the best efficiency in detecting leaks using the structural data of the time graph, the pressure and water age data resulting from the simulation process are shown in Figure 6.2.

Figure 6.2 - Monitoring data analysis

a) Pressure data

b) Water age data

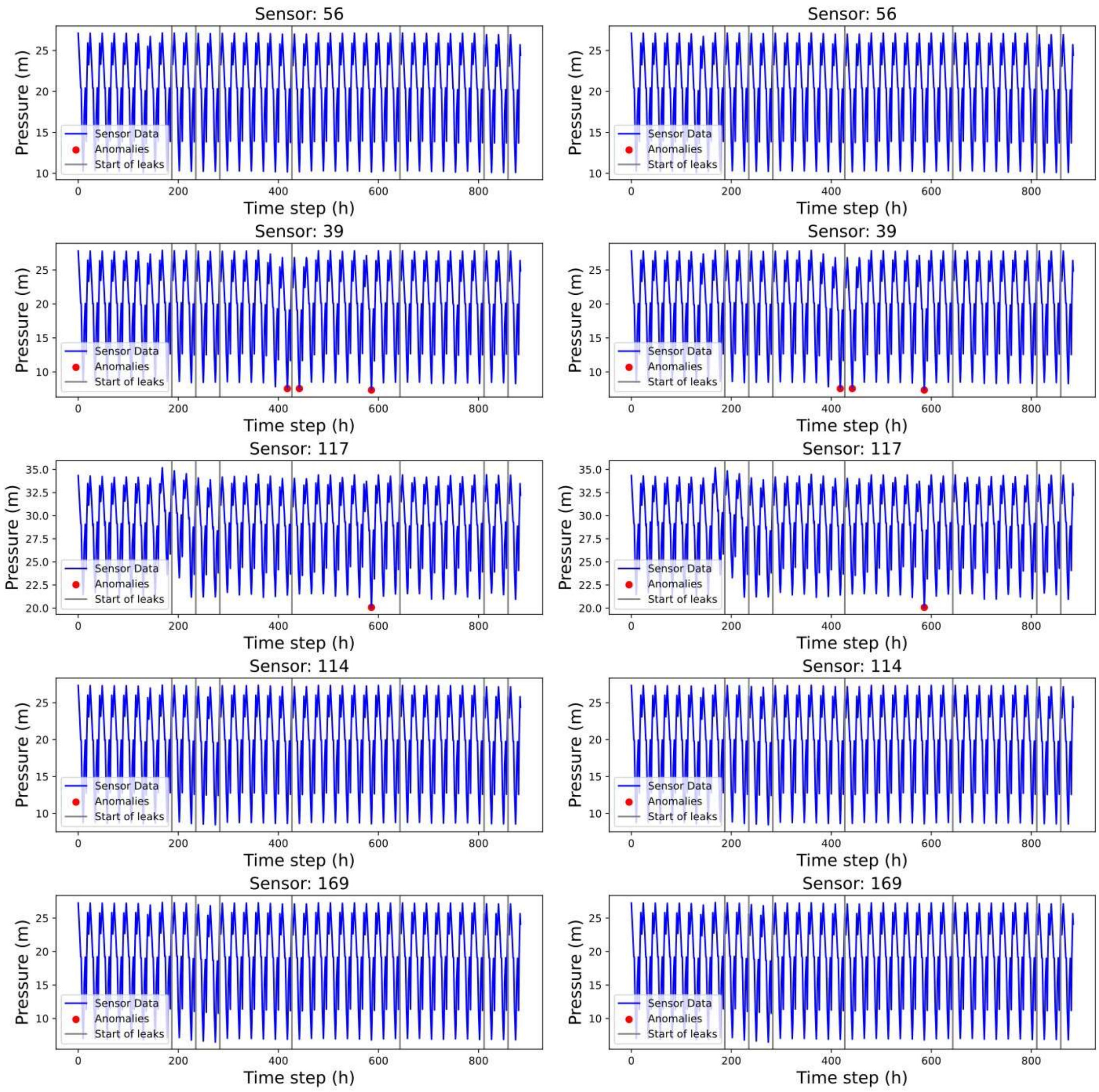
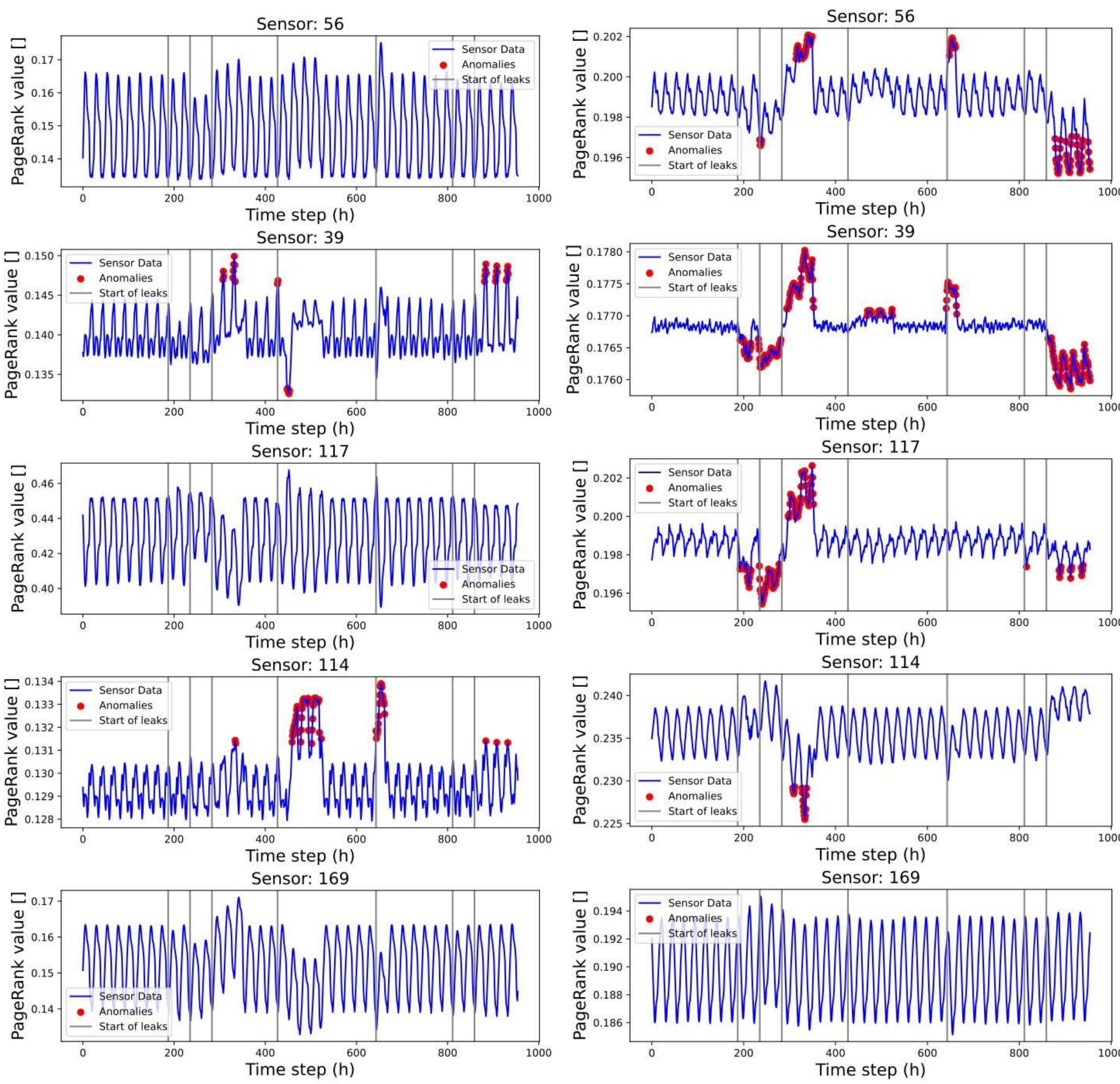


Figure 6.2a shows that some anomalies are detected in the sensor data at nodes 39 and 117. Even so, the sensor 39 detected 3 anomalies, but 2 cases are reported as false positives, since the detection is occurring at times when there were no leaks. A similar scenario is shown in Figure 6.2b as anomalies are also detected only in data from sensors 39 and 117. However, the z-score algorithm detected anomalies before the leaks began in the data from both sensors, this is because the data used in the sliding window is also analyzed in comparison to the initial data.

On the other hand, the use of the correlation between the monitored data obtained in the simulation in the creation of the temporal graph and the ranking of the vertices proved to be a practical and effective way of pre-processing the data. Figure 6.3 shows the PageRank values considering the pressure and water age data when creating the temporal graphs.

Figure 6.3 - PageRank value detection

a) Pressure data b) Water age data



It is possible to observe in Figure 6.3a the behavior of the PageRank values referring to the monitored pressure data change dramatically when leaks occur. The behavior of sensor data 117 (Fig. 6.2a) stands out as it presents behavior like that presented by Barros et al. (2023b), which once again corroborates the application of the proposed method. However, anomaly detection using the z-score algorithm is more effective using data from other sensors, for example data from sensors at nodes 114 and 39. In both sensors, detection occurs immediately after the start of the simulated leaks, and the number of false positives is also reduced. Anomalies are also detected in sensor data at nodes 56, 117 and 169, but with lower quantity and precision. Furthermore, it is noteworthy that in some cases leaks are not detected.

When using water age data in the process of creating the graph and ranking the vertices, the behavior of PageRank values is even more pronounced, as shown in Figure 6.3b the detection of anomalies in the PageRank values referring to data on nodes 39 and 117 also occurs immediately after the start of the leaks in some cases. The moments of detection always occur during the leak, and there are no false positive detections. Therefore, the approach that uses water age data to determine PageRank values is more accurate in detecting leaks, which can be seen numerically in Table 6.1, compared to the approach that uses PageRank data. related to monitored pressure and stands out even more than the direct use of pressure and water age data. Furthermore, it is noteworthy that this approach did not present false positives and that the data from the sensors are complementary for the effective detection of leaks. For example, the leak at node 6 is detected only by sensor 117 and the leak at node 49 was detected only by sensors 39 and 56. Even so, no anomalies were detected in the data from sensor 169.

However, the water age data is not actually monitored, therefore, the behavior of the chlorine concentration at the same monitoring points was considered, and the concentration data is also analyzed by the z-score algorithm and used in creating the temporal graph. In this sense, Figure 6.4 presents the behavior of chlorine concentration and PageRank values, and the moments with identified anomalies.

Figure 6.4 - Chlorine concentration

a) Monitoring data b) PageRank values

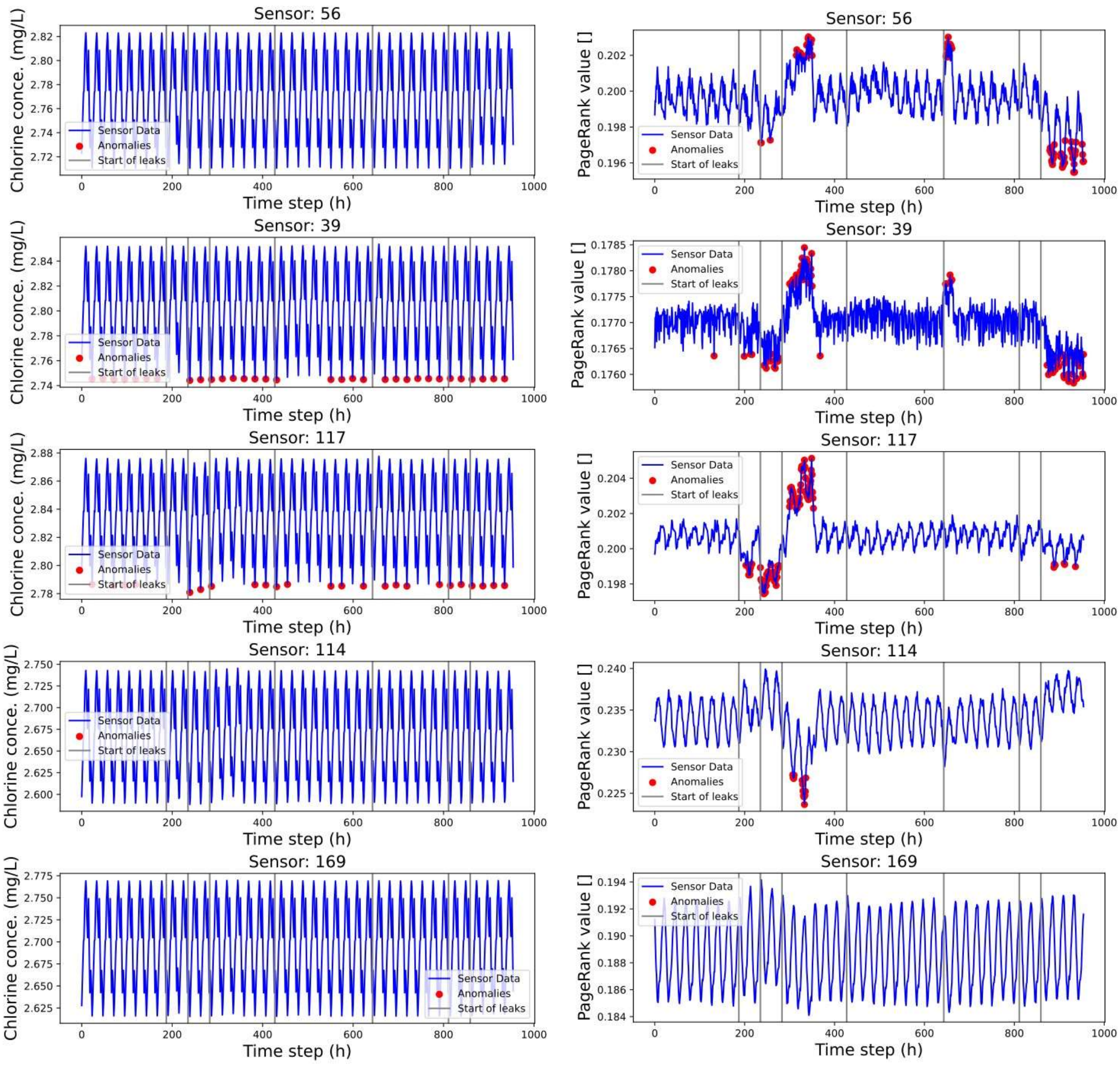


Figure 6.4a shows that more anomalies are highlighted in the chlorine concentration data for sensors 39 and 117, which follows the approach that uses water age (Fig. 6.2b). However, also detecting many false positives. Furthermore, once again changes in the data are not visibly noticeable. This influences the number of false positives identified, in which the detection algorithm using the chlorine concentration detects leaks a few hours after the beginning, which can be pointed out as a detection bias, since several anomalies are identified and many of them are not related to leaks.

On the other hand, the detection of anomalies in PageRank values is more significant compared to the direct use of chlorine concentration data. Still, the behavior of PageRank data is like the behavior of water age, which is expected due to the equations used by the WNTR software. However, it is observed that the PageRank values using the chlorine concentration (Fig. 6.4b) present greater noise than the PageRank values using the water age (Fig. 6.3b). An example of this can best be seen in the sensor 39 data in both cases.

The numerical results of the leak detection are highlighted in Table 6.1, which individually presents the simulated leaks, the leak onset time, and the detection time after the leak onset. Table 6.1 also shows the leak flow rates at the time of detection.

Table 6.1 - Detection process results

Node leaks	Start Time (h)	Max. leak flow (L/s)	Pressure Data		Water age data		chlorine concentration	
			Monitoring	PageRank	Monitoring	PageRank	Monitoring	PageRank
250	192	6	N.D.	N.D.	N.D.	1(1L/s)	N.D.	2(1L/s)
120	240	2	N.D.	N.D.	2(2L/s)	1(2L/s)	2(2L/s)	1(2L/s)
224	288	3	N.D.	14 (1L/s)	2(1L/s)	5(1L/s)	2(1L/s)	10(1L/s)
43	432	6	10(1L/s)	1(1L/s)	4(1L/s)	32(2L/s)	3(1L/s)	N.D.
49	648	7	N.D.	1(7L/s)	33(7L/s)	1(7L/s)	33(7L/s)	3(7L/s)
6	816	2	N.D.	N.D.	2(1L/s)	1(1L/s)	2(1L/s)	N.D.
153	864	2	N.D.	18(1L/s)	23(1L/s)	1(1L/s)	(1L/s)	7(1L/s)

N.D. - Not detected

The values presented in Table 1 refer to the hours after the start of the leaks, which considers the leak detected when detection occurs within the leak simulation period. If the detections are not associated with leaks, they are considered False Positives. It can be seen in Table 1 that when directly using pressure data, the leak detection method achieves the worst results. Only detecting the leak at node 43, 10 hours after the start. On the other hand, the PageRank values referring to the use of pressure data already show better results. It is observed that leaks at nodes 43 and 49 occur immediately after the start of the leak. However, even this approach did not detect the leaks at nodes 250, 120, and 6.

Although detections occur a few hours after the start of the leaks, directly using water age data may not mean the correct detection of leaks. The maximum peaks of the values are marked with anomalies, as seen in Fig. 6.2b, and some cases coincide with the beginning of the leak. Unlike this, PageRank data that uses water age data presents different behavior when in the presence of leaks and reflects in the detection times, where 5 of the 7 simulated leaks are detected immediately after the starting of the leak (nodes 250, 120, 49, 6 and 153).

Finally, the direct use of chlorine concentration presented, as expected, similar results to the approach that uses the water age. However, in this case several false positives were also reported. The approach that uses PageRank values referring to chlorine concentration data also presents results like the approach related to water age; however, two leaks were not detected, at nodes 43 and 6. It is also noteworthy that two false positives are pointed out in sensor data 39.

The individual analysis of data from each sensor proved to be complementary compared to using only data from one monitoring point and can favor approaches for leak localization. Taking as an example the leak at node 6 that was detected by sensors 56 and 39 in the approaches using PageRank values and water age and chlorine concentration data. In the example in question, both sensors are close to the leak, and the leak localization can be determined, for example, by the sensor coverage area and the intersections between the covered areas.

Furthermore, the approach proposed in this article can still be improved if optimally positioned quality monitoring sensors are used. To analyze the results in comparison with pressure and water quality data, this research uses the pressure monitoring sensors presented by Mankad et al. (2022). However, the authors use a methodology that exploits pressure data to estimate pressures at other nodes, detect leaks and estimate the state of the network. Therefore, the detection process presented here can have significant gains if it uses optimally placed water quality monitoring sensors, or even the joint use of hydraulic and water quality data.

6.5 Discussion and partial conclusions

This research addressed leak detection processes using pressure and water quality data. For this, approaches were tested, which use simulated data on pressure, water age and chlorine concentration. However, a form of data pre-treatment is used, in which the correlation between data is considered to create temporal graphs and the structure of these graphs is evaluated using the PageRank vertex ranking metric.

Although widely used for leak detection, pressure data proved to be less effective for leak detection compared to water age and chlorine concentration data. In general, directly using monitored data may not favor detection methods, therefore, the way to pre-treat this data is indicated using the ranking of the graph's vertices. This data pre-treatment made the application of the z-score algorithm for anomaly detection more efficient, as it is possible to visibly identify changes in the data. Even so, there are ways to implement the proposed method, mainly using other methods to determine the correlation of data to create the temporal graph. It is also possible to use different metrics to evaluate the structure of the temporal graph. Furthermore, the method can also be used to locate leaks.

7

Water distribution networks represented as Multilayer Graphs: leak detection and localization approach

This chapter is an adapted version of Barros, D., Meirelles., G., Brentan, B. Water distribution networks represented as Multilayer Graphs: leak detection and localization approach. **Water research**, 2023. (Under review).

Abstract

Water distribution networks (WDN) of increasing size and complexity pose significant management challenges and increase the risk of failures. Globally, it is estimated that 126 billion m³ of non-revenue water are lost annually, highlighting the urgency of measures to mitigate losses. This study proposes a methodology that uses graph-based correlation and multilayer graph analysis for leak detection and localization in WDNs. The detection process involves correlating monitored data to create a temporal graph and classify vertices. The classification values are then analysed by the z-score and IQR algorithms to detect anomalies. The localization process uses a multilayer graph approach that combines sensor data, the network topology, and an approach to determining the sensor coverage area. The DTW algorithm is also used to determine the similarity between monitored and simulated leak data, identifying likely leak locations. The results demonstrate the effectiveness of the methodology, detecting anomalies 15 minutes after the start of the leak and locating them 50 meters from the actual location of the leak. The research highlights the advantages of multi-layer graphs, offering insights into leak location, sensor coverage and reducing network sample space.

7.1 Introduction

Water distribution networks (WDNs), essential for supplying cities, are gradually becoming larger and more complex, making management challenging and increasing the likelihood of failures. In this context, it is worth noting that globally, an estimated 126 billion m³ of unbilled water are lost annually, representing significant losses in terms of both financial and water resource aspects (LIEMBERGER and WYATT, 2019). A survey conducted by the Brazilian National Sanitation Information System showed that approximately 40% of captured and treated water is lost due to leaks, measurement errors, and theft (TRATA, 2023). These estimates highlight the need for the implementation and development of measures to mitigate these losses, focusing on agile and effective approaches.

Strategies to reduce losses, especially those related to leaks, have been subject to constant development (VAIRAVAMOORTHY and LUMBERS, 1998; DE VRIES, et al., 2015; RAJABI et al., 2023). Conventional field approaches often involve the use of acoustic devices to detect the noises associated with leaks (HUNAIDI et al., 2004). However, the effectiveness of these methods has been increasingly compromised due to the presence of underground utilities, such as gas, electricity, and internet cables, as well as the need for technical teams to physically inspect the networks, making the process more time-consuming and costly (SAGNARD et al., 2016). Therefore, the automation of leak detection and localization processes has become a frequent research topic. In this context, hydraulic monitoring data collected by sensors are analysed using mathematical and statistical approaches that identify anomalies and associate them with potential leak locations (CHOUDHARY et al., 2021).

Monitored pressure and flow data are applied in different leak detection and localization approaches. Perez et al. (2009) use pressure data, a calibrated network model, and a genetic algorithm to identify leaks based on the discrepancy between the monitored and simulated data. Meanwhile, Romano et al. (2010) use pressure and flow data to detect leaks in real-time, employing artificial neural network (ANN) techniques to predict the values of the monitored data and subsequently analyse the difference between the observed and predicted data. In recent years, methods involving artificial intelligence, such as artificial neural networks and machine learning, have been widely explored (LIU et al., 2019; MASHHADI et al., 2021).

However, it is important to note that the quality of results obtained through these methods depends on the quality of the monitored data and the calibration of the analysed network, which becomes a challenge due to the constant modifications and expansions of WDNs. In response to this challenge, there is a growing trend towards the application of methods that focus exclusively on the monitored data, aiming to identify failures and represent the network in a more mathematical way (KIRSTEIN et al., 2019). Kaghazchi et al. (2021), for example, model a WDN for irrigation using Hybrid Bayesian Networks for hydraulic simulations and operational performance evaluation. Wu et al. (2021) represent WDN through game-theoretic to consider the water network's operating characteristics and seek to minimize the worst-case disruption impacts. Yu et al. (2023) model a WDN as a graph to evaluate the resilience of networks through the individual importance of nodes and the proportion of indispensable nodes.

Graph theory, a branch of mathematics that explores connections between objects, has played an important role in network analysis in various research domains (BEELER, 2015). A graph contains a set of vertices that represent objects, and these vertices are connected by edges indicating the relationship between the objects. The relationship between vertices can consider different approaches, such as physical (MUNIKOTI et al, 2021), temporal (SHINKUMA et al., 2019), similarity relationships (XU et al., 2019), among others. Thus, the values of these relationships are considered as weights of these edges and represent stronger or weaker connections between the vertices. This theory has been applied in WDN as a representation and analysis method, so that the vertices represent the nodes (demand, tanks, and reservoirs) of the networks and the edges represent the pipes, pumps, and valves. The weight of the relationships between the vertices follows different approaches, such as using the flow rate, diameters, and length of the pipes (TZATCHKOV et al., 2008; SITZENFREI, 2021).

The application of graph theory allows the evaluation of WDN through approaches linked to complex network analysis, such as, for example, identifying critical vertices through centrality metrics (AGATHOKLEOUS et al., 2017) or evaluating the relationship between vertices to detect leaks (BARROS et al., 2023). Furthermore, this theory allows the representation of correlations in monitored data (KALOFOLIAS,

2016) and in the networks themselves as graphs (SITZENFREI, 2021; GIUDICIANNI et al., 2021).

Recently, methodologies have been developed that represent WDNs as graphs for leak detection. For example, SHEKOFTEH and JALILI (2020) employed the Girvan-Newman algorithm to partition a graph based on edge weights, adopting an approach that uses ANN to classify the monitored data. However, this method is sensitive to fluctuations in demand and uncertainty in pressure data, which can lead to inaccurate results. To mitigate errors resulting from uncertainties in monitoring data, Barros et al. (2023) present an approach based on the maximum classification values of graph vertices using the PageRank metric. To detect anomalies in the monitored data, the authors constructed graphs based on the correlation between pressure data monitored by sensors, which are susceptible to variations due to hydraulic behavior or network faults, affecting the graph structure and consequently impacting the classification of graph vertices. However, the authors exclusively employ maximum vertex classification, where classification values are assigned to all vertices, but the analysis is limited to only the maximum values. Nevertheless, the PageRank metric assigns ratings to all vertices, with these vertices representing the monitored nodes. Thus, notable fluctuations in PageRank values can serve as indicators of anomalies within the sensor's coverage area, thus providing information that favors the detection and localization of anomalies.

In a sensor monitored WDN, the representation of the network can consider both the network elements and the vertices and edges of a graph, to represent the network topology itself, as well as to analyse the correlation between the monitored data. This implies the existence of two distinct graphs, although they share the same vertices. This situation opens possibilities for the application of approaches related to multilayer graphs, which represent graphs that share vertices, considering multiple types of relationships, with each relationship between vertices resulting in a new perspective of the graph (LIU et al., 2022). In the context of this approach, Herrera et al. (2023) explored performance indicators of a communication network as layers in a multilayer graph, allowing for the analysis and evaluation of network performance in different aspects. Additionally, the authors highlighted how interlayer information can be helpful in monitoring, control, and problem classification within the network.

Multilayer graphs are explored for different purposes, such as the one presented by Stahl et al. (2019) to plan the trajectory of racing vehicles based on a multilayer graph that has layers related to the actions to be taken, the costs for traveling different paths, and the relationship of speed between competitor in the race. This approach can also favour the detection of subgraphs through mapping between the relationships between layers (BREDERECK et al., 2019). Although this approach has been widely explored in other areas of research, its use in relation to WDN is still scarce. For example, one can explore, as layers in a Multilayer graph, the relationship between monitored data and topological relationships. Also explore subgraphs through the relationship of sensors with their coverage areas. Thus, it may be possible to apply this approach to the process of detecting and locating leaks.

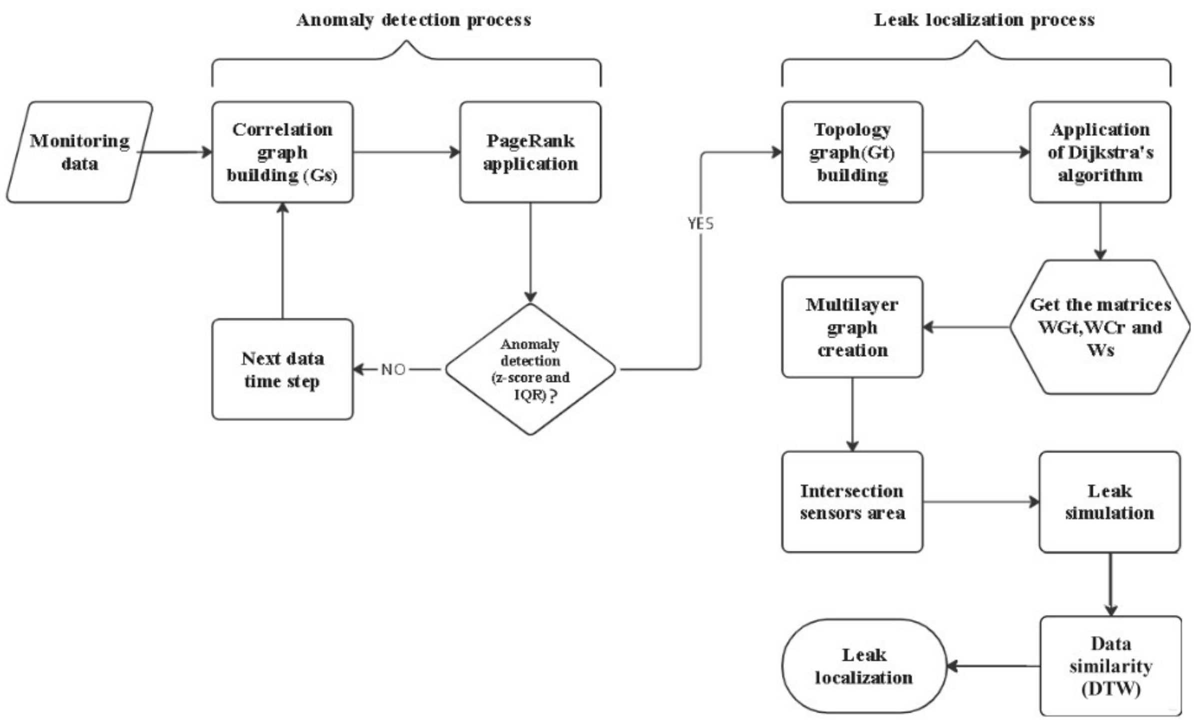
Based on the considerations presented, the main objective of this research is to develop a methodology for leak detection and localization that utilizes multiple graphs within a multilayer structure. To achieve this purpose, the methodology employs three distinct approaches in graph creation: one based on network elements, another that creates a graph with temporal variations through the correlation between monitored data, and finally, the integration of these two graphs into a multilayer structure, where the interaction between the layers is represented by the areas considered covered by the sensors. The detection process begins with the use of the correlation graph to rank the vertices, employing statistical methods to identify anomalies and classify the vertices with the highest discrepancies. Then, the vertices with the highest anomalies are integrated into a graph based on the network topology, forming the layers of a multilayer graph that is used in the leak localization process. Finally, considering the nodes covered by the sensors in the multilayer structure, leak simulations are performed, and a percentage of similarity between the monitored and simulated data indicates the likely leak locations.

7.2 Methodology

The methodology developed and applied in this study is based on two main processes. The first creates a graph through the correlation between monitored data, classifies the vertices of this graph, and uses the classification information to detect leaks. The second application addresses a multilayer graph in which a topological graph of the entire network is considered a layer, the correlation graph of monitored data is another layer and the interlayer relationship is given by the sensors' coverage area.

The vertex classification process uses methods such as z-score (ALTMAN, 2017) and IQR (WAN et al., 2014) to identify vertices that are likely to leak. The edge multilayer analysis process identifies edges that are shared by multiple vertices, which can be used to locate the leaks. In this approach, two graphs are built. The first one represents the correlation among monitored data, where the vertices correspond to the monitored nodes. The second graph is an abstraction the WDN, where network nodes are represented by vertices, and the pipes are represented by edges. By combining these two graphs, a multilayer graph is built, interconnecting the sensor graph and the WDN graph through edges that represent the node coverage. The leak detection process employs the z-score and IQR methods to identify and classify sensor vertices that exhibit anomalies in the data. Finally, leakage isolation is accomplished through the analysis of the edges in the multilayer graph. Figure 7.1 illustrates the steps involved in applying this methodology.

Figure 7.1 - Flowchart methodology



7.2.1 Graph theory application

Graph theory is explored here in three different approaches to detecting and locating leaks. The first approach uses the relationship between monitored data to create a temporal graph. Then, the WDN topology is modelled as a graph weighted by the pipe flows. And finally, a multilayer graph is created with the correlation graph and topological graph, being interconnected by a sensor coverage relationship determined by a metric that determines shortest paths and proximity between vertices.

7.2.1.1 Hydraulic data-based graph (G_s)

Given a monitored water network, each monitored point is represented as a vertex in a graph, and the interaction between these vertices can be established through the Pearson correlation between the information at each vertex. The method proposed by Kalofolias (2016) addresses the creation and learning of graphs through soft signals. The author presents the concepts for creating a sparse graph, that is, a graph with a reduced number of vertices but which maintains structural properties of the complete graph. In this work, this method is applied, but without reducing the number of vertices in the graph. That is, a graph is constructed from a matrix of monitoring data X, where the columns represent different signal sources (nodes monitored in the WDN), and the rows correspond to temporal data t. Each column of X is treated as a vertex in the graph (x_{v1}, x_{v1}, ..., x_{v1}), and edges are defined based on the distances between data pairs. each column.

A graph can be expressed through a square matrix that denotes the interconnections between its vertices, being called an adjacency matrix (A) when the edges do not have weights, and a weight matrix (W) when specific information is associated to these edges. The method proposed by Kalofolias (2016) calculates a pairwise distance matrix (Z) for each t of the data and between all vertices, which also results in a square matrix, and it is used to create the graph. This matrix Z is determined by:

$$Z_{ij} = \|X_{vi} - X_{vj}\|^2 \tag{7.1}$$

However, the process of creating the matrix Z results in correlation values between all vertices, thus, this amount of information makes it difficult to understand the behaviour of the graph and the underlying tasks (YAN et al., 2006). Barros et al. (2023) presents a similar process for the matrix Z creation but does not prioritize strong connections, which we observed influence the achievement of better results. For this, an extra analysis is conducted to prioritize only strong edges between vertices. Thus, Matrix W_s is created:

$$\text{if } W_{ij} \leq mean\ (Z) \text{ then } W_{ij} = 0 \tag{7.2}$$

$$\text{if } W_{ij} > mean\ (Z) \text{ then } W_{ij} = Z_{ij}$$

The W_s matrix is used to create a G_s graph which is applied in the anomaly detection process on the monitored data. For this, the process of creating the Z and W_s matrices, in addition to the creation of the G_s graph, occurs every t. This results in an application that can be used in a database and can also be applied to real-time data.

7.2.1.2 Topological based graph (G_T)

To represent the WDN as a graph (G_T). In this representation, the set of vertices (N), correspond to the junctions (demand nodes, tanks, and reservoirs), while the set of edges (E) represents the pipes, pumps and valves that connect these vertices. These connections are mathematically represented by an adjacency matrix (A_{NxN}), where the elements a_{ij} describe the network topology (TZATCHKOV et al., 2008; SITZENFREI, 2021).

Each edge E of the graph G_T can have different weights, according to the analysis to be performed. In the case of leak localization, the edge weights are determined by the maximum pipe flows in the WDN. In this work, to obtain the edge weights, a simulation of the network is performed using the library Water Network Tool for Resilience (WNTR) (KLISE et al., 2017). Through this simulation, information about the amount of water flowing through each pipe of the network is obtained, allowing for appropriate weights to be assigned to the G_T edges.

The approach makes the representation more realistic by incorporating information about the flow capacity of each pipe. When considering the edge weights in the construction of the W_{GT} matrix, the specific transport capacity along the connections is reflected, making the modelling more faithful to real network conditions. This inclusion of details about the flow capacity of the pipes enhances the analysis by considering the physical constraints of the system, which in turn contributes to more accurate and applicable results in identifying potential leak locations. In essence, when considering the practical aspect of flow capacity, the graph representation comes closer to the effective operational dynamics of the water network, providing more realistic and practically useful results (ANCHIETA et al., 2023).

7.2.1.3 Multilayer graph creation (G_{ML})

A multilayer graph is a structure in which each layer represents different types of interactions or relationships between system elements. These layers can be interconnected through edges, which can be weighted, directed or undirected, depending on the nature of the interactions (KIVELÄ et al., 2014). This approach allows the individual analysis of each layer, understanding the specific dynamics between the elements in each context. Furthermore, the joint analysis of the multilayer graph provides a holistic view of the system, revealing patterns of interaction between the layers (LIU et al., 2022).

In this work, the first layer involves the graph G_T based on the topological features of the water network model, and he second layer is the graph G_S based on hydraulic data correlation. A process is used to evaluate the correlation between layers, choosing to analyze the sensor coverage area with an emphasis on the proximity between nodes and sensors. To achieve this, we adopted a proposal based on network partitioning approaches. In the method presented by Goa et al. (2017) the Dijkstra algorithm is employed to calculate the shortest paths between nodes, resulting in the division of WDN. Within the scope of this work, sensors play a central role, and a coverage region is delimited because of this process.

This process involves the use of the graph G_T and the application of Dijkstra's algorithm (DIJKSTRA, 2022). Dijkstra's algorithm is used to find the shortest path between nodes in the network, considering the edge weights. In this paper, the maximum pipe flowrates are used because this information can indicate the importance of the pope's' contribution to contribute the total water demand (SITZENFREI et al., 2020). Using Dijkstra's algorithm, one can accurately estimate the areas around sensor nodes based on the edge weights. This algorithm plays a crucial role in determining which network nodes are in the operational range of each sensor. By identifying this information, it is used in the process of locating possible leaks. However, considering all the vertices covered by the sensors is not a realistic option, because usually one sensor reaches values between 8% (BARROS et al., 2023) to 25% coverage (ZHAO et al., 2020). Therefore, this considers the sensors coverage to 8% of the total network nodes. In case of specific areas in a WDN, the operator can choose the most applicable quantity for the problem. Thus, this application is mathematical formulation:

$$\text{if } d[i] + \omega(i,j) > d[j] \tag{7.3}$$

$$\text{then } d[j] = d[i] + \omega(i,j)$$

where $d[j]$ represents the estimated distance between vertex i and the source vertex j, while $\omega(i,j)$ is the weight of the edge connecting vertices (i,j). The application of this algorithm results in the construction of matrix W_{Cr} with dimensions N x X_s where N corresponds to the number of vertices in the graph G_T. The W_{Cr} matrix presents the distances between the sensor and all network nodes and this information is used as a method of determining the sensor coverage rate.

These matrices W_{GT}, W_{Cr} and W_s are used for building the extended matrix (W_{ex}) which will be used for the creation of the multilayer graph (G_{ML}) (GARG et al., 2021). Figure 7.2 presents the ordering of the matrices for the matrix W_{ex} creation.

Figure 7.2 - Extended matrix (W_{ex}) composition

$$\begin{bmatrix} [\,W_{GT}\,] & [\,W_{Cr}\,] \\ [\,W_{Cr}^{t}\,] & [\,W_{S}\,] \end{bmatrix}$$

Figure 7.2 shows the matrix W_{GT} that represents the graph of the first layer, W_s the graph in the second layer and W_{Cr} are the inter-layer connections. These inter-layer connections are the search spaces for the anomaly location, so it will be considering only the coverage area of the most affected sensors.

7.2.2 Leak detection approach

The methodology for detecting leaks follows three consecutive steps. Initially, the temporal graph $G_s$is constructed, followed by the classification of the vertices using the PageRank metric (PAGE, 1999). The classification values assigned to the vertices are then analyzed by the z-score and IQR algorithms, which look for anomalous moments in the data. During this procedure, the algorithms not only detect anomalies, but also point out which vertices present the most striking changes in the behavior of PageRank values. Finally, the vertices identified with the greatest changes are then used in the second stage of this study, which consists of locating the leaks.

7.2.2.1 Vertex ranking process

The vertex ranking process is conducted sequentially, using the PageRank metric, which is derived from the PageRank matrix. The PageRank algorithm, originally developed by Page (1999), plays a key role in this methodology for assessing the importance and influence of each vertex within the network. By employing the G_s graph, which can change with each step t due to the calculation of correction between the monitored data, the importance of vertices can also be changed. Therefore, this dynamic application of the PageRank metric can quantify the importance of vertices as the data is employed, ensuring a continuous and adaptive evaluation of the structural importance of the graph.

The PageRank method was originally developed by Page (1999) with the goal of ranking web pages, considering the endorsements from other important pages. This method can be applied to graphs by considering the connections between vertices. The PageRank metric is based on a random walk model on the graph, assigning a "PageRank" value to each vertex, represented by a probability vector. PageRank $PR(v_i)$ is calculated by using a recursive equation expressed by:

$$PR(v_i) = \frac{(1-d)}{i} + d \times \left(\frac{PR(v_1)}{L(v_1)} + \frac{PR(v_2)}{L(v_2)} + \cdots\right) \quad (7.4)$$

where $L(v_1)$ represents the number of edges outgoing from vertex v_i, and d is a damping factor (usually set to 0.85) (YAN and DING, 2011). Initially, each vertex is assigned an equal PageRank value. Through an iterative equation, the relative relevance of the vertices is successively updated, considering the importance of their neighbors. This process is repeated until the scores converge, indicating that the vertex relevance has stabilized. The steady state reflects the final rankings of vertices based on their PageRank scores, evidencing their importance and influence on the structure of the graph. Vertices with higher PageRank scores are considered more important or influential in the network (Gu et al., 2022).

Using this process helps reduce noise in the data and reveals more significant changes when anomalies occur, making the application of the detection algorithm more effective due to the data noise reduction.

7.2.2.2 Anomaly detection and affected sensors

Despite the data noise reduction obtained through the determination of PageRank values, an automated process is necessary to reliably point out the anomalies and which sensors have the greatest anomalies. Therefore, the PageRank values for each sensor over time are analyzed using two distinct approaches, which are jointly applied to detect anomalies and identify sensors most affected by anomalies. This process identifies anomalies in the data, which suggests potential anomalies in the system and pinpoints the sensors with the most significant deviations. These identified sensors play a crucial role in the leak localization phase during the second stage of the research.

The first stage uses a statistical measure known as the z-score (Z_s), which quantifies how deviant a data point is from the mean in terms of standard deviations. The z-score is determined by the following equation:

$$Z_s = \frac{PR(v_{i,j}) - \mu}{\sigma} \tag{7.5}$$

where $PR(v_{i,j})$ is PageRank value for vertex v_i in time t, μ is the average value of the variable, and σ is the standard deviation of the variable (KIZILÖZ et al., 2022).

The second stage of this process uses the Interquartile Range (IQR) (WAN et al., 2014) method to evaluate the PageRank value of each sensor. The IQR is a statistical measure that indicates the dispersion of values around the median. It can be calculated by subtracting the first quartile (Q1) from the third quartile (Q3) and represents a measure of data variability. Values exceeding 1.5 times the IQR are considered possible outliers. This can be expressed as an equation:

$$IQR = Q3 - Q1 \tag{7.6}$$

and the Outliers are expressed:

$$Outliers = Q3 + 1.5\, x\, IQR \tag{7.7}$$

$$Outliers = Q1 - 1.5\, x\, IQR \tag{6.8}$$

This exceed value (1.5) is usually an empirical choice that aims to balance the detection of significant outliers without removing many values that can be normal variations. This approach allows to identify anomalies that deviate significantly from the expected signal values. By automating this process, the anomaly detection methodology based on IQR can be applied to real-time data analysis, without relying on manual inspection by operators.

The z-score helps identify data points that deviate significantly from the mean, while the IQR method assists in detecting outliers in the dataset (CHIKODILI et al., 2020). The z-score considers all values in the calculation, making it sensitive to outliers, while the IQR focuses on quartiles and data variability, making it more robust against the influence of such values (CHIKODILI et al., 2020). By using the z-score and IQR in combination, we leverage the sensitivity of the z-score in identifying deviations in average and standard deviation, while benefiting from the robustness of the IQR in handling extreme values.

The detection process takes place when the data exhibits significant changes based on the defined thresholds for z-score and IQR. In the detention process, the time-varying PageRank data for each vertex is evaluated, and if a data point exceeds the criteria set by the z-score and IQR, it is considered a deviation, and the vertex is identified as a potential anomaly. Once all the vertices with deviations are identified, a ranking process is initiated to rank the vertices, using the maximum z-score value and the IQR value as the primary criteria. The higher the values, the higher the ranking of the vertex in relation to the set of vertices indicated with anomalies. In this way, a set of vertices X_s is generated, representing the potential nodes with detected anomalies.

The vertices in X_s are also vertices in the graphs G_T and G_s, with the sensors' coverage area represented as vertices in the graph G_{ML}. The sensors' coverage area in X_s are then used in a leak simulation process, carried out with the WNTR package. Then, the behavior of the simulated data is compared with the monitored data and a similarity method is applied to indicate the location of the leak.

7.2.3 Leak isolation approach

After the preliminary identification of the potential leakage region, i.e., sensors with the most significant alterations, a refinement of the localization process is initiated. This process has two main objectives: locating the leak by comparing monitored and simulated data and narrowing down the search space, as simulating and comparing all possible points demands substantial computational effort.

In the leakage simulation stage, a flow of approximately 3% of the network total demand is individually added to each demand node within the coverage area of sensors X_s. This percentage of total demand is based on that proposed by Quinones et al. (2019) for large leaks. This flow is chosen to strongly impact the sensors and exhibit characteristics in monitored data that may be associated with leaks. During the simulation, only data from sensors X_s are stored and compared with monitored values (presented in the monitored data matrix X to measure the similarity between the temporal data sequences, both monitored and simulated.

To carry out the comparison process, the Dynamic Time Warping (DTW) algorithm (SAKOE and CHIBA, 1978) is employed. DTW is an algorithm used for comparing time series with different lengths or temporal changes (BURSTYN et al., 2021). It enables the comparison of data sequences even when they are not temporally aligned. The DTW algorithm is used to calculate the similarity between time series and identify anomaly points based on discrepancies in behavior patterns between simulated and monitored data. The application of this algorithm involves several steps, with the first being the creation of the cost matrix. Initially, the algorithm creates a cost matrix (or distance matrix) of size gxh, where h is the size of the simulated data sequence, and g is the size of the monitored data sequence. The matrix is filled iteratively, with each value determined by:

$$D(x,z) = |s(x) - m(z)|^2 \tag{7.9}$$

where $D(x,z)$ represents the cost of aligning point x in the simulated sequence with point z in the monitored sequence. $s(x)$ is the value at position x in the simulated sequence, and $m(z)$ is the value at position z in the monitored sequence.

After this step, the values in the matrix D are updated, also iteratively, based on the costs of adjacent positions, aiming to find the minimum cost from the starting position $(0x0)$ to the ending position (hxg) (KEOGH and RATANAMAHATANA, 2005). For this purpose, the recurrent equation 7.10 is apply:

$$D(x,z) = |s(x) - m(z)|^2 + min\,\{D(x-1,z), D(x,z-1), D(x-1,z-1)\} \tag{7.10}$$

Thus, the total alignment cost between the data sequences is determined by the value in the last cell of the cost matrix, i.e., $D(hxg)$. The subsequent step in the algorithm's application reconstructs the path of the minimum cost traversed in the matrix by tracking the positions that minimized the cost in the previous step. This step indicates the alignment between simulated and monitored data. Finally, the algorithm assigns a similarity measure to each simulated node, utilizing the average similarity measure calculated by:

$$Sim(n) = 1 - \frac{D(n,g)}{g} \tag{7.11}$$

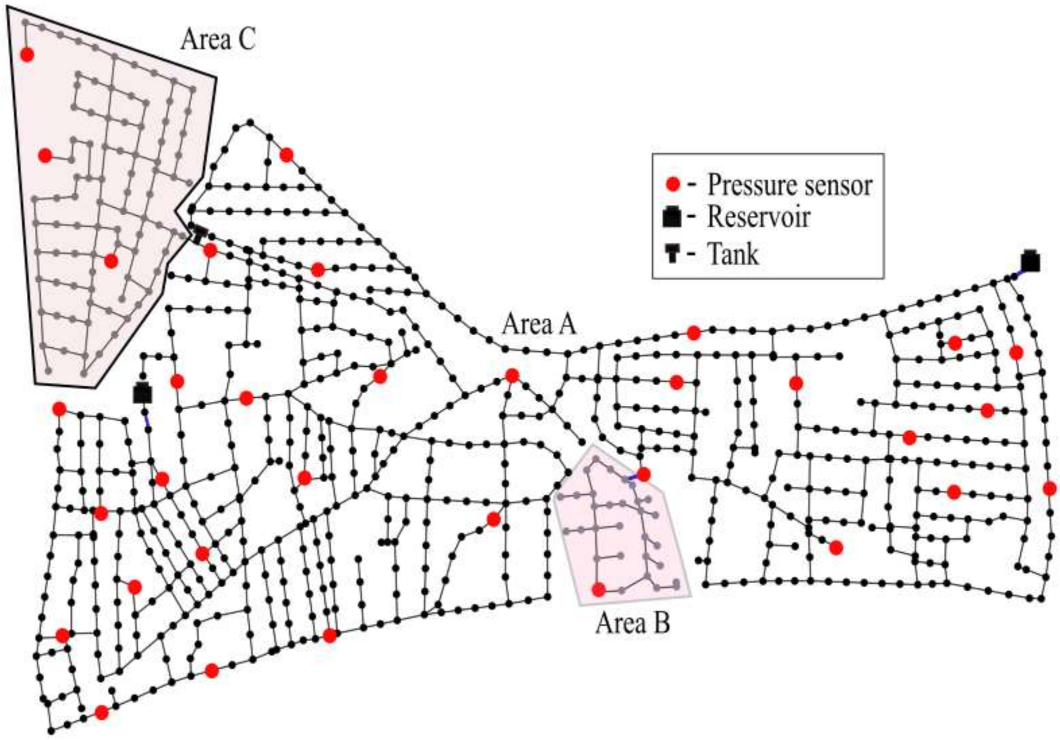
Where $Sim(n)$ is the measure of similarity between the nth simulated node, $D(n,g)$ is the alignment cost of the nth simulated node with the last position of the monitored sequence (g), and g is the size of the monitored data sequence. Thus, nodes with higher percentage values of similarity are identified as potential leak locations. The effectiveness and accuracy of the proposed methodology are determined by comparing the identified leakage locations with the actual leakage site. Reference data and networks are employed to validate and assess the methodology, as outlined in the subsequent sections of the study.

7.3 Aplication: Battle of Leakage Detection and Isolation Methods (BattLeDIM)

The proposed methodology is evaluated using the benchmark problem presented in the Battle of Leakage Detection and Isolation Methods (BattLeDIM), described in detail by Vrachimis et al. (2022). The benchmark utilizes the L-Town WDN, which is designed to resemble a real network. The L-Town network is characterized by variations in flows and pressures over time, including changes in demand, pump, and valve operations, as well as the occurrence of multiple leaks introduced at different moments, intensities, and duration. The L-Town network consists of 905 pipes, totalling approximately 42 km in length, 782 junctions, 3 pressure reducing valves, 2 reservoirs, and 1 pump. Additionally, the network has 33 pressure monitoring points distributed at different locations. The data provided for evaluation spans a period of two years, with accurate measurements taken every 5 minutes, without delays.

Figure 7.3 presents the topology of the L-Town network, including its structural elements such as pipes, junctions, reservoirs, and the tank. This visual representation provides an overview of the complexity and size of the network, which serves as a challenging scenario for leak detection and localization. The L-Town benchmark scenario provides a realistic and comprehensive testbed for evaluating the effectiveness and performance of leakage detection and localization methods. It allows researchers and practitioners to compare different approaches and algorithms under consistent conditions, fostering advancements in the field of water distribution network management.

Figure 7.3 - L-town network



The use of the L-Town network as a benchmark allows for the evaluation of the efficiency of the proposed methodology in a realistic and complex context. Based on the data provided by the network, simulations can be conducted, leakage detection and localization algorithms can be applied, and the obtained results can be compared with the actual locations of leaks present in the benchmark. This analysis validates the methodology's ability to correctly identify the leak locations and assesses its precision in relation to the expected results. Furthermore, the data is processed considering District Metered Areas (DMA), as in general, the behavior of sectors follows similar patterns. Similarly, since the areas in the L-Town network are separated by a pressure control valve (Area B) and a tank (Area C), the pressures in these areas exhibit distinct characteristics. Therefore, the methodology is applied individually for each area.

7.3.1 BattLeDIM evaluation metrics

To evaluate the methodological performance proposals for detecting and locating BattLeDIM leaks, the organizers present a database in which the start time, duration and size of leaks are known. Thus, four initial criteria are considered in the evaluation: detection time after the start of the leak; the flow rate of the leak at the time of detection; the maximum leak flow; and the location distance to the actual leak location. In addition to these criteria, the assessment approaches presented by BattLeDIM are also used in the present study. These assessments address the economic impact related to profit saved throughout the year due to successful detection and localization (VRACHIMIS et al., 2022).

The first evaluation metric is denominated True Positive and considers leak detection if the following condition is met:

$$t_{st}^{l} \leq t_{d}^{h} \leq t_{end}^{l} \tag{7.12}$$

where t_d^h is the detection time, t_{st}^l and t_{end}^l is the start and end time of leakage l. The organizers also present the Profit from Water Saved related to value p_w^h (euro) of water saved due to correct leak detection, described by:

$$p_w^h = \left(\sum_{k=t_d^h}^{t_{end}^l} q^l\,(k)\Delta t\right) c_w \tag{7.13}$$

where by detection l, q^l(k) flow rate of leakage l at each discrete time step k. Δt is the duration of the discrete time step and c_w is the cost (euro) of water per cubic meter.

Finally, a Total Score (Ts) is determined to evaluate the entire detection and localization set. For this, the total score is determined by:

$$Ts = \sum_{h \in D} S_h = \sum_{h \in D} (p_w^h + c_h^r) \tag{7.14}$$

where S_h is the score per given detection l and c_h^r is the repair crew cost.

7.4 Results and discussions

The proposed methodology allows real-time analysis of monitoring data, and this approach is used on data provided by BattLeDIM, which is evaluated at monitoring intervals immediately after issuance. The matrix X represents the monitoring data and the first step consists of creating the graph G_s that explores the correlation between the monitored pressure data and the determination of PageRank values. The entire database is analyzed sequentially, that is, at each time step, but the application of the methodology for 3 periods that refer to a leak in each area of the L-town network is exposed in detail. These graphs are generated considering the pressure sensors in the network area under analysis.

7.4.1 Multilayer graph creation applied to L-Town

The first analysis presented refers to the application of the methodology using the sensors in area A and highlights the first 24 days of 2019 of the monitored data. The other applications are performed with the 3 sensors in Area C and the only sensor in Area B. However, related to Area B, it does not involve the first stage of the methodology because there is only one sensor which makes it impossible to create the Z matrix for this area.

The temporal graph G_s is monitored at each time step and, also at each time step, the PageRank value is assigned to each vertex of the graph G_s. Figure 7.4 shows the normalized PageRank values for the first 24 days of data and referring to the 29 sensors in Area A.

Figure 7.4 - PageRank values - Area A

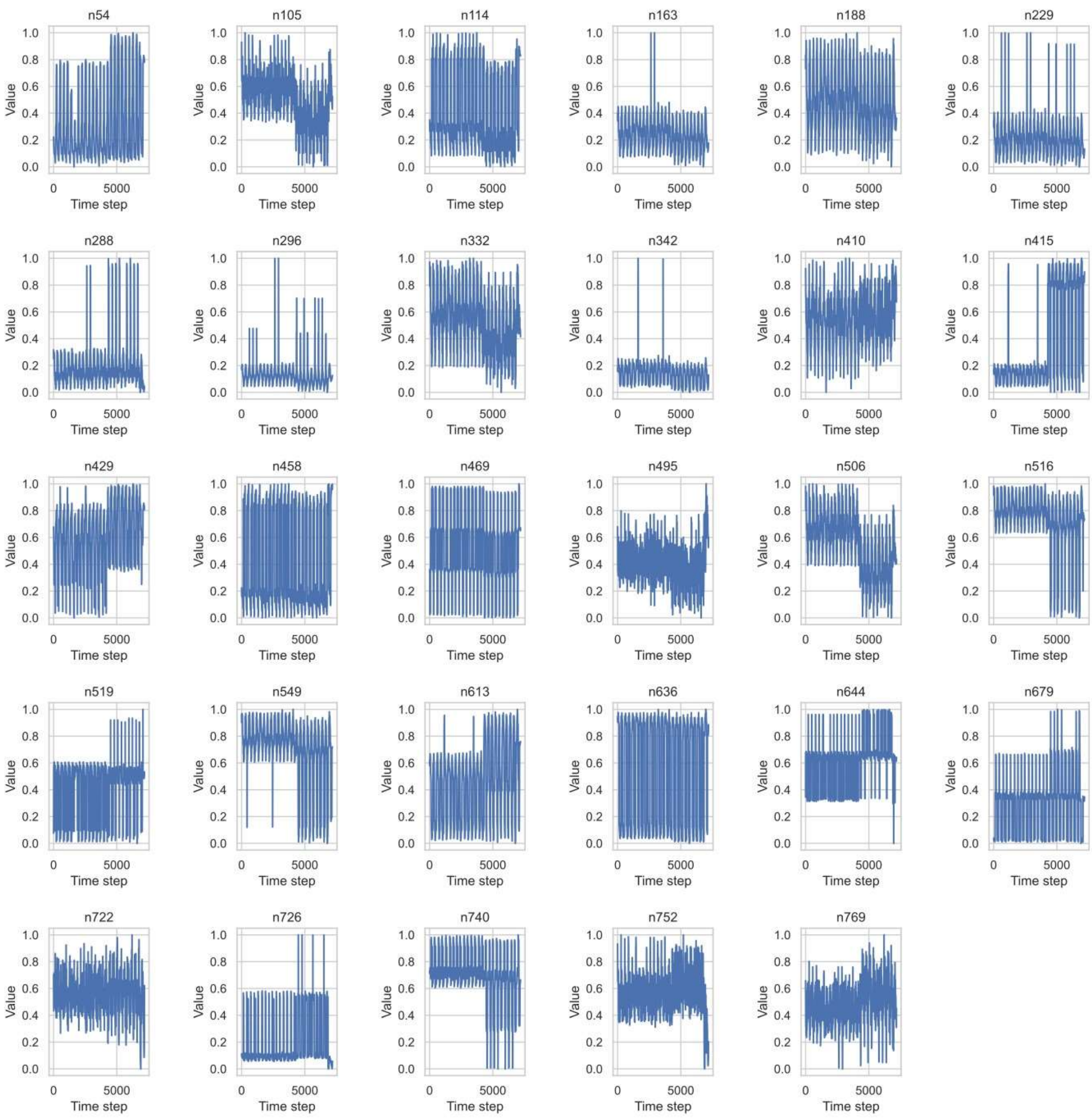
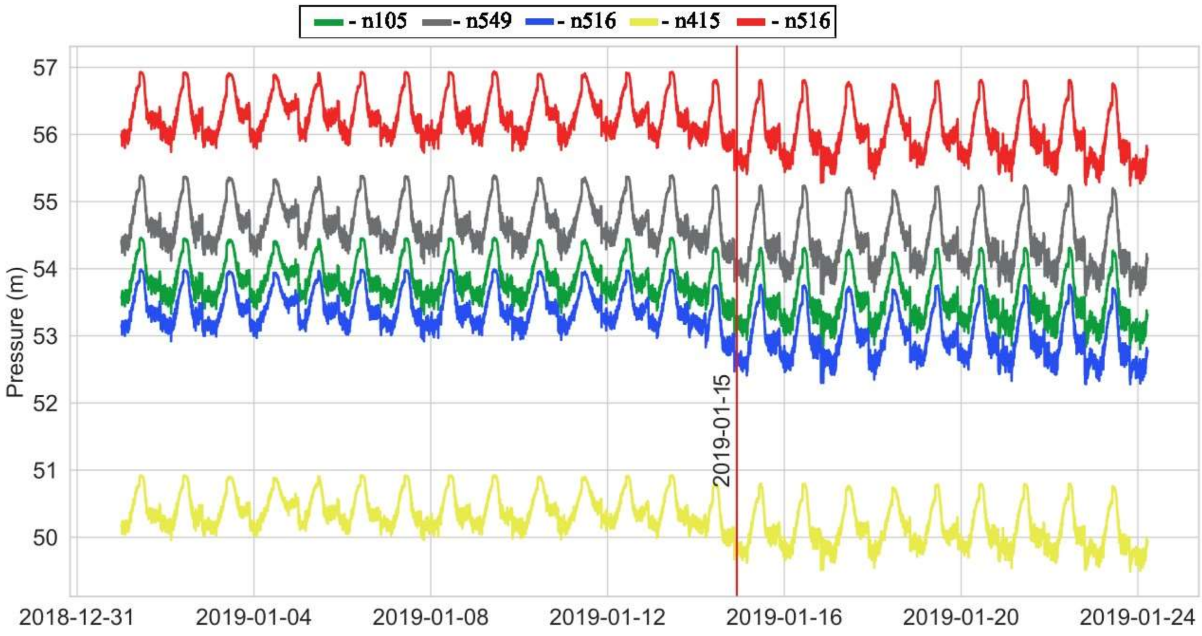


Figure 7.4 shows that PageRank values on some sensors change after a few days. This change is more noticeable in some cases, and the anomaly detection method using the z-score and IQR algorithm pointed out anomaly points and which sensors show greater changes in behavior in the data. Figure 6.5 shows the 5 sensors with the biggest anomalies identified and the pressures that were monitored by these same sensors.

Figure 7.5 - Data behavior

a) Pressure



b) PageRank values

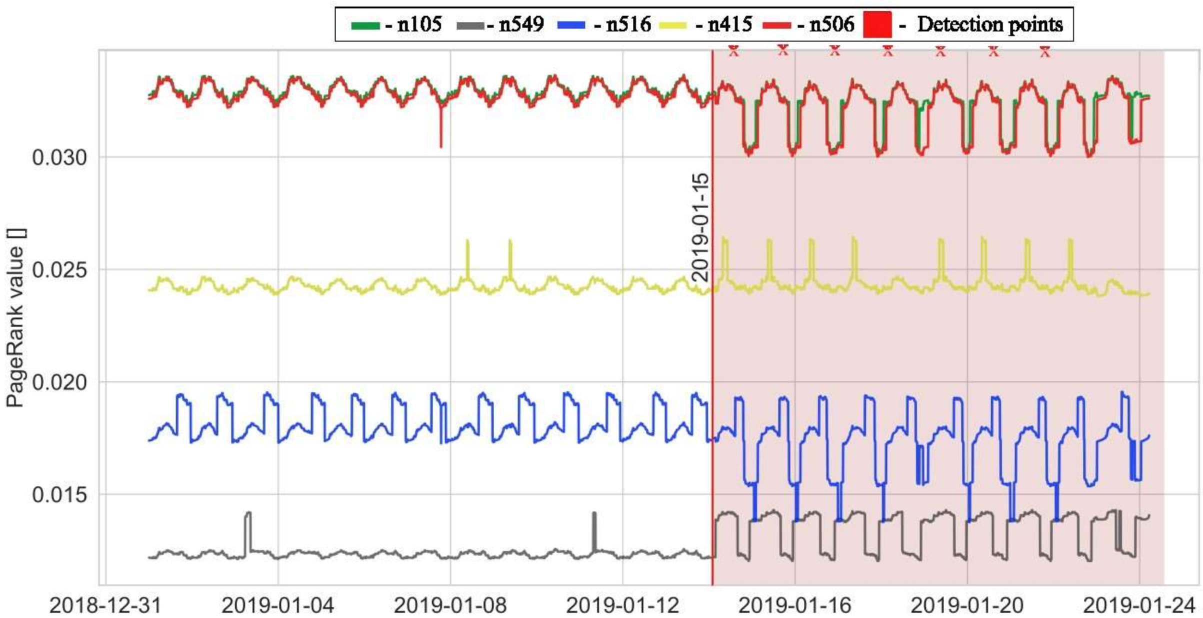


Figure 7.5a presents a portion of the pressure monitoring data provided by BattLeDIM for Area A. By visually inspection of the data, one can notice the behavior and similarity among them. However, starting from 2019-01-15, a leakage in pipe p523 appears at the network. The changes on pressure domain are not as easily apparent by just analyzing the monitored data. On the other hand, PageRank values also show similar

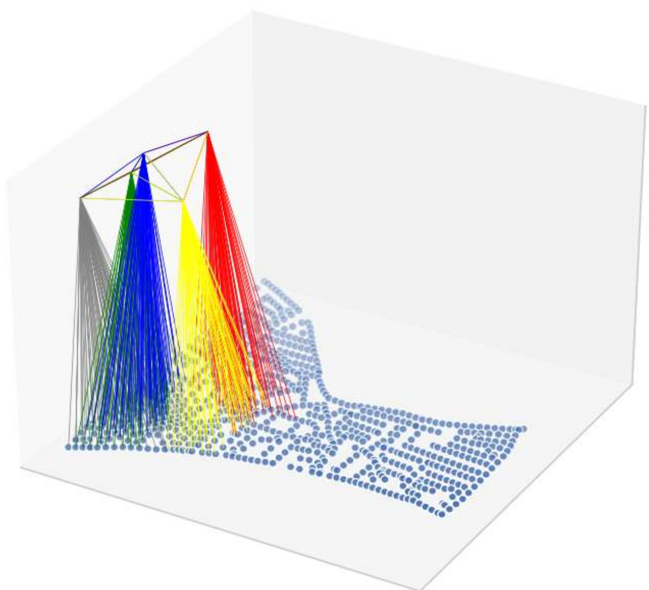
and repetitive behavior, but the behavior changes visibly when the leak begins (Fig.7.5b). The detection method is applied and was able to identify moments of anomalies, as well as point out, for each leakage case. All detection moments pointed out by the detection process are exposed in Table 6.1. In addition to this process, the z-score and IQR methods pointed to each case the sensors with the highest values of changes and thus begins the process of leak localization.

For this, Area A is considered with the graph G_T, and the 5 sensors with the highest anomalies are selected to create the graph G_s. This number of sensors is chosen because they present significantly higher discrepancies than other quantities. The Dijkstra algorithm is applied and considers around 60 nodes as covered by sensors; this represents around 8% of the vertices of the graph G_T as the sensor coverage area.

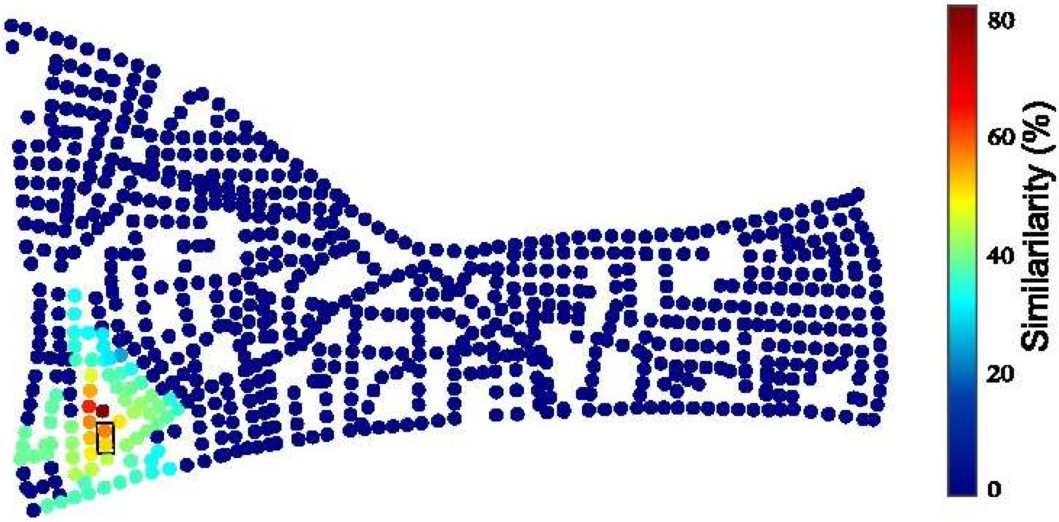
The nodes present in the sensors' coverage area are used as sources of leaks in a simulation process. This occurs mainly to reduce the search space and processing time of these simulations. The flow rate used in the simulations was approximately 4 L/s, the simulated data is only saved from the monitoring points. Thus, the DTW algorithm determines the similarity value between the monitored data and the simulated data. The results of these processes, considering the leakage in pipe p523, can be seen in Figure 7.6.

Figure 7.6 - Methodology application - Area A

a) Multilayer Graph - Area A



b) Localization Leak - p523

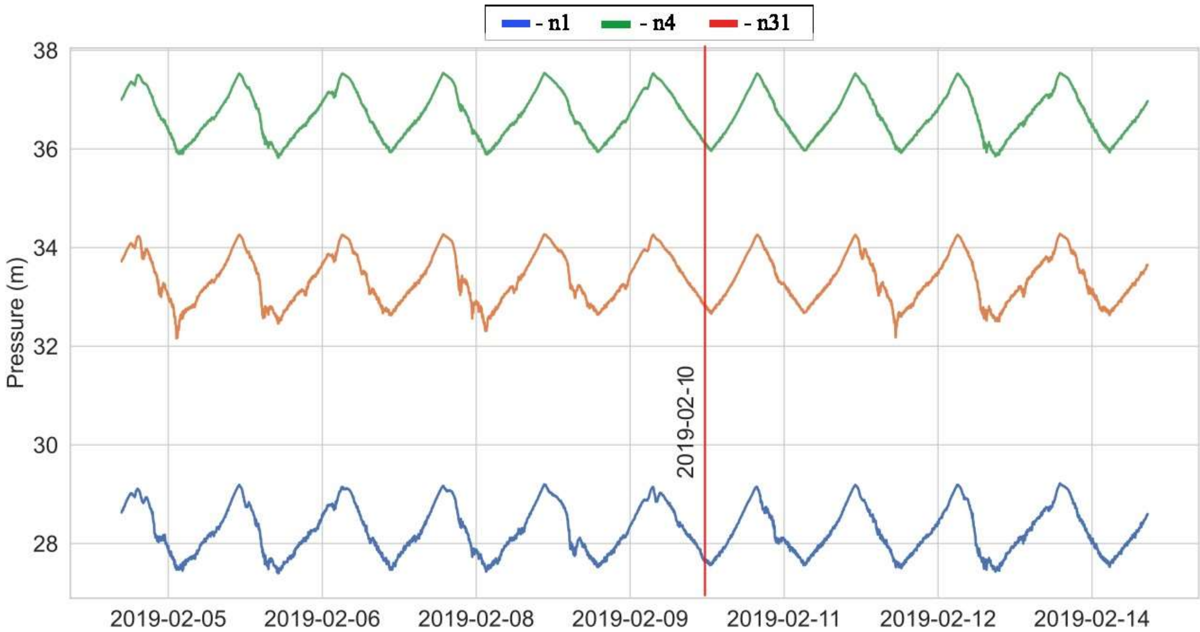


The Figure 7.6a shows the resulting multilayer graph from the proposed methodology. It can be observed that there is a dispersion among the sensors with the highest detected anomalies. However, the three most indicated sensors (green, blue, and grey edges) show higher alterations compared to the other sensor. Therefore, the vertices that have edges with both sensors are used to simulate leaks, and the similarity between the simulated and monitored data is determined. The Figure 6.6b shows the vertices with the highest similarity values determined by the DTW method. It is worth noting that the vertices with the highest similarities surround the actual leak location (black box) at about 50 meters. This distance is calculated using the coordinates of the node identified as the leak source and the average of the coordinates of the initial and final nodes of the leaking pipe.

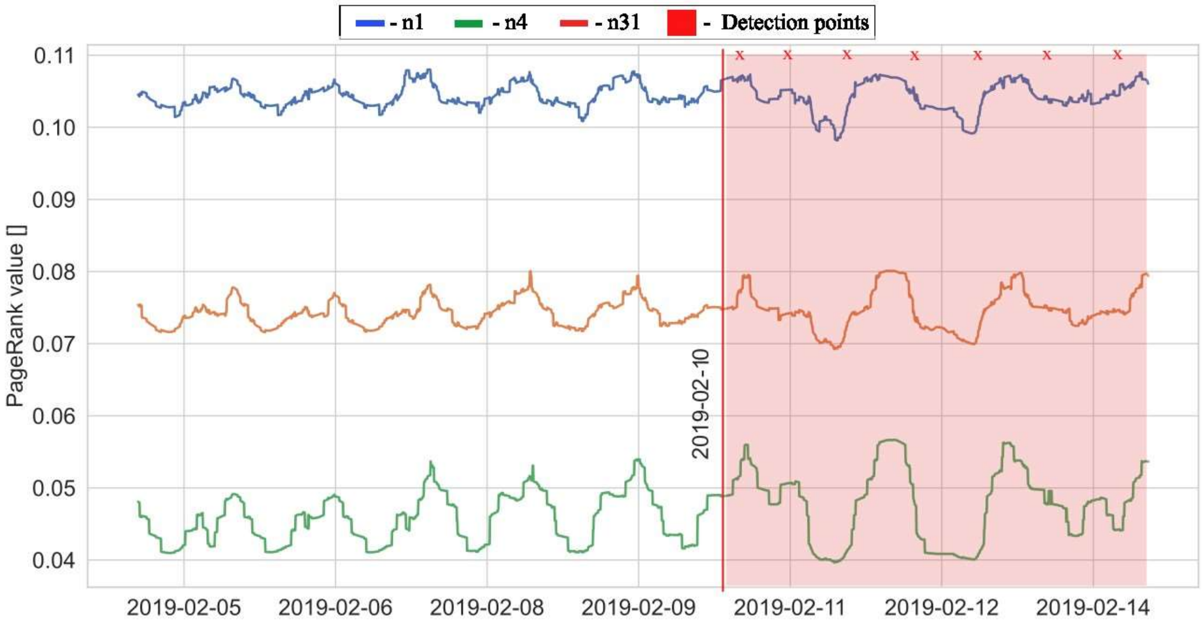
The application of the methodology related to Area C exposed here, presents the detection related to the leak in node n280. The leak in this node began on 02/10/2019, and the monitored data and PageRank values for the sensors in this area are shown in Figure 7.7.

Figure 7.7 - Data behavior – Area C

a) Pressures



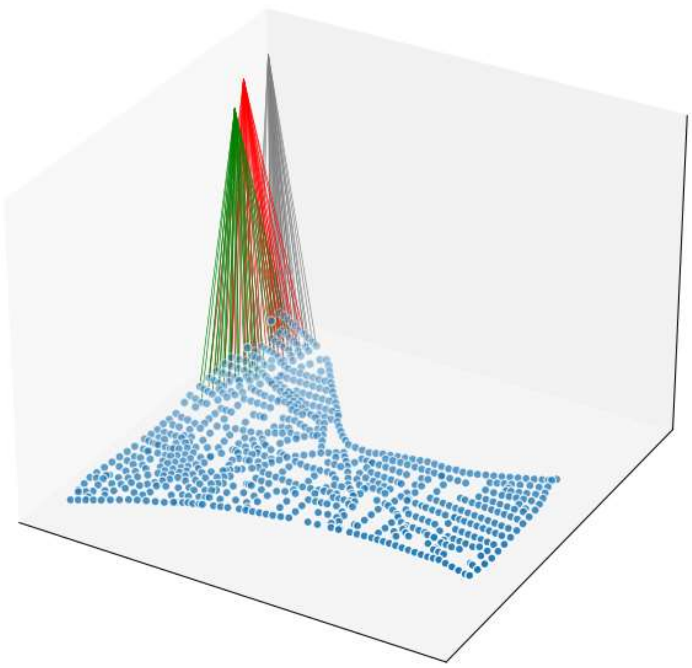
b) PageRank value



In Figure 7.7a it is possible to observe that the behavior of the monitored pressures hardly changes after the start of the leak. Although less intense, PageRank values change over time and this change was detected by the proposed algorithms (Fig. 7.7b). Thus, the leak localization process began, using the 3 sensors in the area to create the G_s graph, mainly due to the small number of sensors in this area. Thus, the coverage area is determined using the Dijkstra algorithm and the nodes in this area are used in the leak simulation process. The multilayer graph and the locations with the highest probability of leaks are shown in Figure 7.8.

Figure 7.8 - Methodology application - Area C

a) Multilayer Graph - Area C



b) Localization Leak - p280

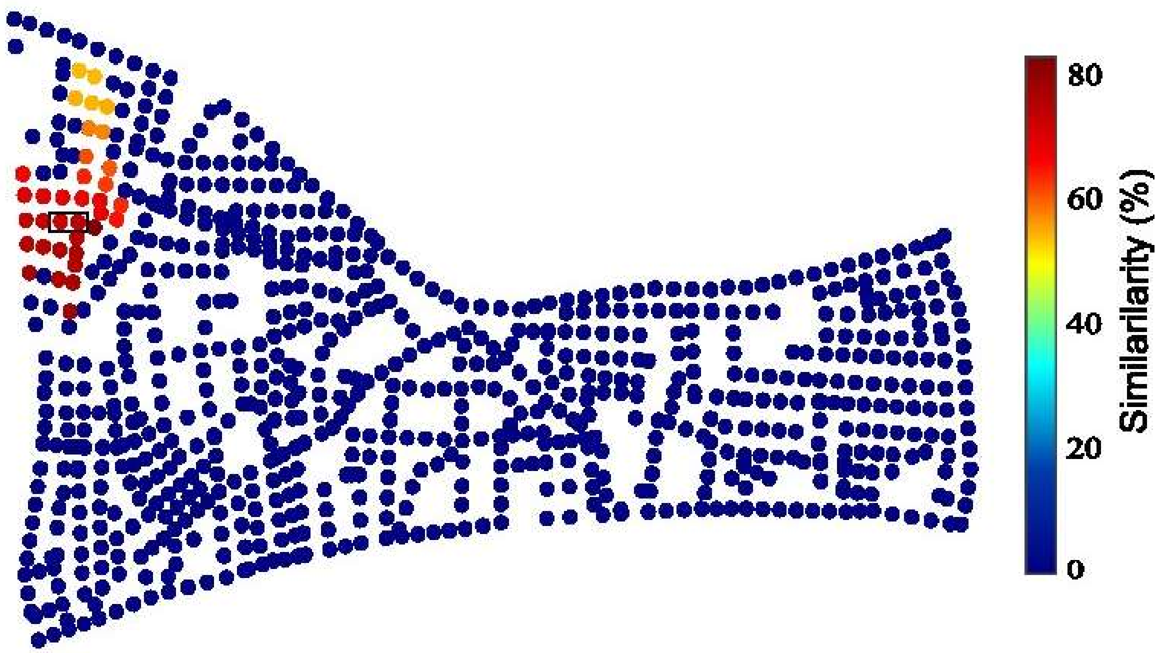
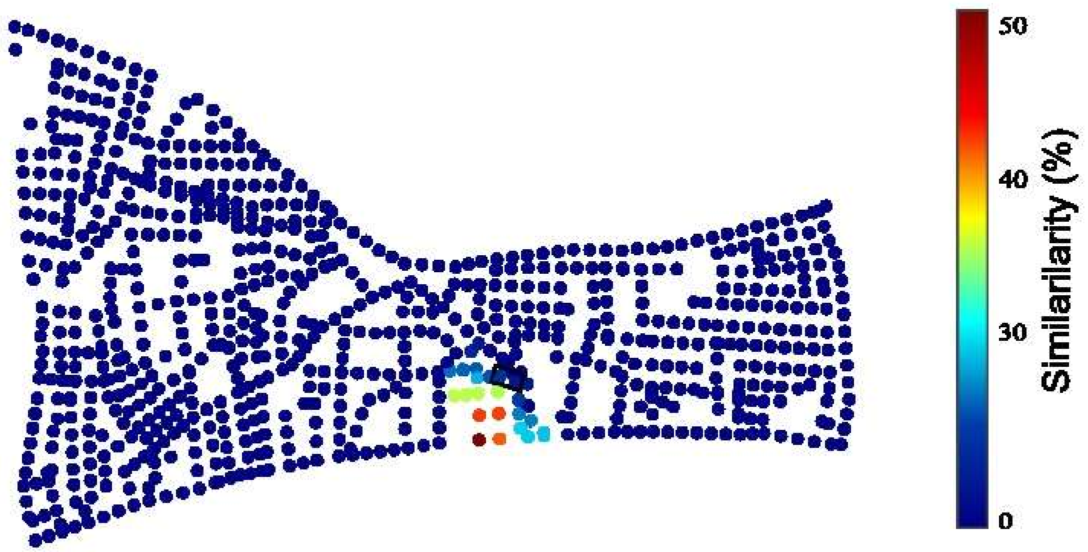


Figure 7.8c shows the multilayer graph related to Area C. An anomaly is detected, and the sensors at nodes n31 and n4 (red and green, respectively) are pointed out as having the highest changes. Thus, the nodes that showed the highest behavior similarity values with the real data (Fig. 7.8b) are approximately 60 meters away from the actual leakage location. It can be observed in Figure 7.8b that all nodes present high similarity values, which may be because nodes in this network have similar pressures controlled by the tank. Therefore, a leak can affect the water level in the tank, and this level change can affect the pressure in all nodes of the DMA.

The third application of the methodology is performed in Area B (Fig. 7.9), but the localization process occurs based on the determination of similarity values.

Figure 7.9 - Leak Localization - Area B p680



The accuracy in locating the leak in the p680 pipe, shown in Figure 7.9, is not as high compared to other leaks. However, the approach can identify anomalies in the data after 45 minutes of the start of the leak. The methodology is applied to all leaks identified by BattLeDIM, and Table 7.1 shows the time elapsed from the actual leakage start time to detection, the leak flow at the time of detection, the maximum leak flow rate, the distance between the location indicated and the report time detection. until leak detection.

Table 7.1 - Detection and localization results.

	Detec. time (h:m)	Flow leak detec. (L/s)	Max flow leak (L/s)	Local. dist. (m)	Report time
p123*	1016:20	3.2	919	237	2019-10-20 12:25
p142	04:40	26.88	2704	132	2019-06-13 05:45
p193*	869:30	3.31	1036	378	2019-07-08 20:10
p257			Undetected		
p277*	1023:00	2.74	736	290	2019-07-20 21:55
p280	00:20	5.16	526	46	2019-02-10 13:25
p331	00:35	10.65	1093	327	2019-04-20 10:45
p426	08:40	13.25	1356	224	2019-10-26 22:05
p427			Undetected		
p455*	885:00	3.1	1105	75	2019-11-17 05:00
p514	00:25	15:38	1558	192	2019-04-02 21:05
p523	00:15	28.6	2839	43	2019-01-15 23:15
p586*	296:35	3.13	2052	127	2019-08-22 07:15
p653*	48:30	3.29	1828	164	2019-03-05 13:40
p654			Undetected		
p680	00:45	5.36	537	204	2019-07-10 09:30
p710	02:55	5.56	558	42	2019-03-24 17:15
p721*	763:15	5.12	1318	147	2019-09-08 22:15
p762*	314:50	1.03	1571	238	2019-11-14 01:05
p800*	193:20	3.11	2195	48	2019-08-25 03:20
p810			Undetected		
p827	01:35	26.05	2646	152	2019-01-24 20:05
p879*	501:30	3.02	1093	286	2019-12-17 19:25

* - Leaks with increasing start.

The Table 7.1 displays the results of the application, generating pertinent discussions. Initially, the effectiveness of the detection process stands out, revealing the identification of anomalies in the data just 15 minutes after the start of the leak (leak p523). Leaks were detected at nodes p280, p514, p331 and p680 up to 45 minutes after the beginning, all these cases are leaks that started abruptly. However, it is worth noting that in some cases, detection occurred several hours after the leak began. These cases present leaks characterized by a gradual increase in flow, such as those observed in leaks in p123, p277, p455 and p721. Leak detection in these pipes occurs more than 700 hours after the start of the leak. However, as these leaks have a gradual increase in flow rates, this also gradually changes the monitored data. Such behavior introduces complexity into data analysis, as temporal changes are minimal and impact the data uniformly. This complexity is even more evident in the localization process, where establishing similarities between data points produces less pronounced results. Consequently, there are increasing disparities between the location suggested by the method and the actual location of the leak.

Table 7.1 also presents the flow rate of the leak at the time of detection due to the difficulty in quickly detecting leaks with gradual flow growth. It is observed that in most cases of leaks with gradual growth, detection occurs when the flow reaches about 3 L/s (p123, p193, p277, p455, p586, p653, P800, p879). Except for two cases, the leak in p721, where detection occurs when the flow reaches 5.12 L/s, and the leak in p762, where it is detected when the flow is at about 1 L/s. However, it is noteworthy that these detection times are lower than the maximum values, such as the flow rate at which the leak was detected in p762 (1.03 L/s), while its maximum flow value reached 15.71 L/s.

As mentioned, the process of locating these leaks is not as accurate, but in some cases, the distance is less than 100 meters (p455, P800). Noteworthy are the leaks in Pipes p523 and P710, with distances to the exact location of the leak of 43 and 42 meters, respectively. On the other hand, the leak in the p193 pipe results in the greatest distance (378 meters), but this leak presents gradual growth of the flow. Additionally, the leak in the p331 pipe stands out, which has an abrupt start and was indicated 327 meters from the actual location of the leak. This situation occurs in a region where the pressures show little variation and do not significantly affect the data during the simulation of leaks.

These greater distances are also reflected in the evaluation process presented by the BattLeDIM organizers. The approach identifies 17 leaks as True Positive and 2 as False Positive. This means that of the 23 leaks, 17 are detected correctly and two others are detected but in the wrong locations. These values corroborate the results presented in Table 6.1 in two main factors: regarding the 4 leaks that are not detected (p257, p427, p654 and p810); and the leaks in tubes p193 and p331 where the distance indicated between the actual location of the leak is greater than 300 meters. Locations greater than 300 meters are considered failures by BattLeDIM organizers. Even so, the Total score presented by the methodology proposed in this work is equal to €243491, a higher value close to the highest achieved by BattLeDIM participants (€264873 and €260562). The perfect Total Score, if all leaks were detected and located immediately, would be €523154. The Report file indicating the locations and time of the report is presented as an appendix; this file can be used in the BattLeDIM evaluation software.

Some relevant points of this research are the approaches that can be considered when working with Multilayer graphs, as they offer several advantages, such as determining the coverage area of the sensors by different methods and reducing the graph using subgraphs. Another important point is the determination of similarity through the application of DTW. Although it is a fast process, it becomes impractical when the sampling space is too large. This problem was solved by reducing the sample space to only the nodes covered by the sensors indicated in the simulation process. However, even with good results, there is still room for improvement, especially because this process indicates which sensors are most impacted, which influences the search space for the leak.

7.5 Discussion and partial conclusions

The proposed methodology for leak detection and localization in WDN demonstrated promising results when applied to the benchmark problem presented in BattLeDIM. This approach, based on graph analysis and monitoring data, was able to detect anomalies in the network pressure data, indicating the presence of leaks within 15 minutes of their onset. However, it is important to highlight that the effectiveness of leak detection and localization may vary depending on the nature of the leaks. Leaks with a gradual flow increase presented additional challenges, as changes in data over time were smaller, resulting in delayed detections and less precise localizations. Despite these limitations, the overall score achieved by the proposed methodology was competitive compared to other participants in BattLeDIM, demonstrating its ability to compete with established approaches.

The presented methodology represents a significant advancement in leak detection and localization in water distribution networks, especially in situations where changes in data are more subtle. Nevertheless, there is still room for improvement, particularly in addressing a WDN as a graph since the interaction between vertices can be represented in various ways. Studying and analyzing these representations could contribute to refining the methodology and expanding its potential in leak detection and localization in water distribution networks.

8

Final considerations

8.1 Final considerations

The present thesis addresses methodologies aimed at mitigating leaks in the water distribution phase, presenting specific approaches for the detection and localization of these leaks. The research also examines the impacts of leaks on water quality. To achieve these purposes, equations employed in the computational modelling of leaks were tested to enhance the realism of computational hydraulic simulations, striving to strike a balance between reducing computational effort and obtaining results that are more faithful to reality. To carry out effective detection and localization, the strategic placement of monitoring sensors in the network is essential, ensuring coverage in cases of leaks and providing monitored data that contribute to the comprehensive control and management of the system.

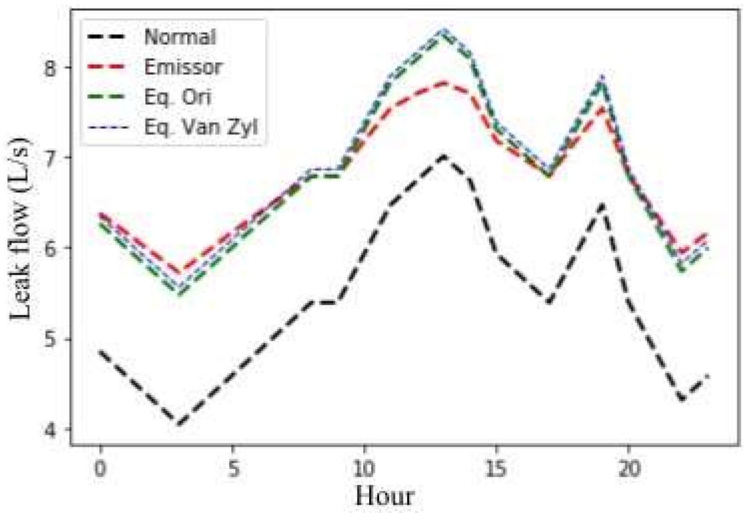
The efficiency of the detection and localization processes is ensured through the efficient placement of sensors, especially concerning detection time and accuracy in localization. The thesis presents various approaches to detection, directly utilizing monitored pressure data in statistical analysis algorithms and exploring the correlations between these data to optimize results. Additionally, it proposes a methodology for leak localization that incorporates topological information of the network, as well as the correlation between monitored data in the creation of distinct graphs, using these graphs as layers of a multilayer graph. The proposals have proven effective, and advancements in each stage are detailed in the subsequent subsections.

8.1.1 Computational Modeling of Leaks

To simulate leaks in water distribution network (WDN) models, the emitter equation (Eq. 2.1) and the standard orifice equation (Eq. 2.2) are widely employed. While these equations provide representative results, there are formulations that consider the material behavior and shape of the orifices. Van Zyl and Cassa (2014) propose a modified orifice equation to calculate leakage flows in viscoelastic polyethylene pipes, addressing different hydrodynamic simulation processes and orifice shapes. The authors conclude that the orifice area increases linearly with pressure, except for circular orifices. Van Zyl et al. (2017) also introduce modifications to the orifice equation, considering elastic, viscoelastic, plastic deformations, and fractures in the pipes.

Although modifications to the orifice equation provide more accurate results regarding the real behaviour of the material during leakage simulations, their application increases computational effort and the time required to obtain results. During the development of this thesis, the equations proposed by Van Zyl and Cassa (2014) and Van Zyl et al. (2017) were tested and compared with conventional emitter and orifice equations. It was observed that the flow rates calculated by the modified equations exhibited similar behaviours to the standard orifice equation, but due to high sensitivity to pressure, flow peaks occurred during periods of elevated pressure, as illustrated in Figure 8.1. The figure depicts flow rates under normal network conditions and during leakage simulations using the standard orifice equation, the emitter equation, and the modified orifice equation proposed by Van Zyl and Cassa (2014).

Figure 8.1 – Flow behavior of leaks - equations



While modelling leaks with the modified orifice equation is more realistic, the standard orifice equation yields similar results. However, using the modified equation requires additional information about the pipes, such as material, elastic modulus, and wall thickness, making the simulation process more computationally intensive. For this reason, the studies in this thesis exclusively utilized the orifice and emitter equations in leakage simulations, primarily because the presented scenarios considered all nodes in the network as potential sources of leaks.

8.1.2 Effects of Leaks on Water Pressure and Quality

Computational leak simulations were employed at different stages of the present study. Initially, all nodes in the network were designated as potential sources of leaks, with the flow from these leaks added to the demand of consumer nodes. The purpose of these simulations was to analyse two primary behaviours: the sensitivity of all nodes in the network to leaks in other nodes and the effects of leaks at a specific node on the pressure and water quality in other nodes. The second chapter of this thesis explores these two situations, addressing both the sensitivity of nodes and the influence of simulated leaks on pressure and water quality in other network nodes. It is observed that the pressure sensitivity of nodes in leak scenarios yields similar results for neighbouring nodes, particularly in regions where node pressures and elevations are comparable.

However, water quality sensitivity does not entirely follow this pattern of similarity among neighbouring nodes. Some network nodes exhibit significantly higher sensitivity values for water quality than other nodes, including their neighbours. It was noted that, due to the demand of certain nodes and the water trajectory to meet this demand, water quality can cause substantial changes in the water reaching the nodes. This directly affects sensitivity to water quality in some nodes, making them more susceptible to leaks in other parts of the network. Therefore, it was found that modifications in water flows to meet leak demands cause water to travel new paths or receive water from additional sources. Thus, the research demonstrated that water quality can be an additional source of information for more effective leak detection, considering that, in some cases, variations in water quality are more pronounced than changes in node pressures. However, it is important to emphasize that water quality monitoring is less common, especially in Brazilian networks, and the corresponding sensors are more expensive compared to pressure and flow sensors.

8.1.3 Leak Detection Approach

The leak detection proposals presented in this study incorporate different data preprocessing approaches. In Chapter 4, an approach employs Independent Component Analysis (ICA) for data processing, using the Interquartile Range (IQR) and Matrix Profile (MP) statistical methods to automatically identify anomaly points in the data. A second approach, presented in Chapters 5, 6 and 7, uses the correlation between monitored data to create temporal graphs and analyzes the structural relationship using Z-score algorithms, in chapters 6 and 7 and IQR, in chapter 7. In the approaches exposed in chapters 6 and 7, in addition to pointing out the moments of anomalies in the monitored data, the algorithm also indicates which sensors have undergone changes and which have had the greatest changes in values, which are then used in the leak location stage, also presented in chapter 7.

Both detection methods, especially those discussed in Chapters 4 and 7, utilized the Battle of the Leakage Detection and Isolation Methods (BattLeDIM) reference database. This makes it possible to compare and analyze results between the proposed methodologies. Leaks in the p514 and p523 pipelines were quickly detected by both approaches. However, the method using ICA and the IQR algorithm detected the leaks 10 and 5 minutes after onset, respectively, while the MP approach detected the same leaks at 24 and 20 hours, respectively. The method that uses graphical structural relationships and Z and IQR scoring algorithms detected the leak in the p514 pipe within 25 minutes and the leak in the p523 pipe within 15 minutes after the start.

However, both methods faced challenges in detecting leaks with increasing flow, as observed in the p455 pipeline leak. The IQR and MP methods, applying the ICA algorithm, detected the leak 327 hours after the beginning when it reached a flow rate of 0.45 L/s. The same leak was detected by Z-score and IQR methods using graphical structural data just 885 hours after onset, when the flow reached 3 L/s. However, this was not the last moment of detection of this second approach; this method detected only 1,016 hours after the start of the leak in p123 and 1,023 hours after the start of the leak in the p277 pipe when they reached 3.2 and 2.7 L/s, respectively.

Despite late detections, the method presented in Chapter 7 demonstrates consistency in flow at the time of detection, except for the leak in the p762 pipeline. All leaks with a gradual increase in flow rate were detected when approaching 3 L/s. In contrast, the method presented in Chapter 4 detected leaks with increasing flow more quickly, although in some rupture cases detection occurred many hours after onset. The leak in the p331 pipeline, which occurs abruptly, reaching flow rates close to 10 L/s, was detected by the ICA and IQR method 174 hours after the beginning. This same leak is detected by the Z-score and IQR approach just 35 minutes after initiation.

These results highlight that both methods have advantages and limitations, making them options that meet the needs of systems managers. However, even considering late detections, the method that uses the correlation between monitored data to create the temporal graph seems, from my perspective, more advantageous, mainly because this approach can facilitate the application of other techniques, such as the one presented in Chapter 7 for leak location and represent the relationship between data and WDN in a more comprehensive way.

Although not using the BattLeDIM benchmark problem, the approaches covered in chapters 5 and 6 use the Modena network (BRAGALLI et al. 2012) and a sequence of simulated leaks to generate data and address techniques for detecting anomalies in data. However, although the simulation process follows the same parameters, the studies focus on leak detection through the analysis of pressure and water quality data, in this case evaluating the age of the water and chlorine concentration. Focusing on what is presented in chapter 3 on the behavior of quality data in leak situations, the research presented in chapter 6 explores the advantages of using quality data to detect leaks and compares these results with an approach that uses pressure data.

Even though the anomaly detection process presented in chapter 5 has achieved satisfactory results, there is only a comparison between the values from the initial days of the leak simulation with the pressure data for the next simulated days. This approach, although functional for validating the model, does not have guaranteed applicability due to the behavior pattern of the monitored parameters changing over time, mainly due to seasonality and changes in tube roughness. Therefore, the method proposed in chapter 6 applies a current data analysis window, so that changes in the standard behavior of the parameters are included throughout the analysis. Using the mean and standard deviation of the data in this sliding window, the z-score algorithm points out moments of anomalies in 6 databases: pressure data, water age and chlorine concentration, and vertex ranking data in created temporal graphs with data on pressure, water age and chlorine concentration.

Using vertex ranking data from a graph created through the correlation between data monitored by sensors proved to be an effective way of pre-treating data. Detecting anomalies in vertex ranking data generally occurs faster than directly using monitoring data. These results are explored in chapter 6, for example the leak at node 43 which was detected 10 hours after the start if pressure data was used and was detected 1 hour after the start if the vertex ranking data of a created graph was used. with the correction between pressure data.

Also noteworthy is the use of water quality data to detect leaks. The advantages of using data on water age and chlorine concentration compared to pressure data, normally used for detection, proved to be more assertive for detecting leaks, both using the data directly and using correlation and ranking of vertices. The simulation process carried out 7 leak simulations at different nodes of the network, however directly using the pressure data in search of anomalies resulted in the detection of only one leak, as mentioned, 10 hours after its beginning. When directly using water age, 6 leaks are detected, with 5 detected less than 5 hours after the start. On the other hand, when using the correlation between the water age data and the ranking process of the vertices of the temporal graph, the detection process occurs for all simulated leaks, with 5 of them being detected 1 hour after the beginning.

All the above shows that exploring water quality to detect leaks can represent a significant increase in reducing losses in water distribution networks. Even though the water quality parameters are slowly changed, related to the water path time to the monitoring points, the change in the parameters is more pronounced than the pressure data and this is reflected in the detection process. Furthermore, the detection process presented in Chapter 6 can still be explored using data from optimally positioned quality sensors. The research uses as monitoring points the locations determined by the approach of Mankad et al. (2022), however, the authors use a methodology focused on pressure monitoring. Therefore, by using an approach that considers optimization of water quality coverage, the detection process can become more comprehensive and accurate. Even so, using the same locations for positioning pressure and quality sensors can be advantageous in terms of operational efficiency, by centralizing the sensors in a single location; in reducing costs due to infrastructure, cables, and energy; in ease of maintenance, repairs, and replacement.

Furthermore, the different changes in monitored water quality data can favor leak location methods, since the changes are more pronounced and this behavior is different for each sensor analyzed, as explained in chapter 6. Thus, the methods locations can point out the sensor coverage area with the greatest changes as leak locations and apply other approaches for more precise location of leaks.

8.1.4 Leak Localization Approach

The proposed approach for leak localization is based on the use of sensors that exhibited significant variations during the detection process. Initially, the sensor coverage region is delimited, considering the probable leakage area. The coverage extension of these sensors is determined through an algorithm that considers the proximity and weights of edges in a topological graph weighted by pipe flow. Subsequently, only the nodes within this area are used as leakage sources in a simulation process, aiming to minimize computational efforts and reduce processing time. The simulation results, including monitored pressure information, are then compared with monitored pressure data. From this comparison, a similarity value is determined between the data from each simulated node and the monitored data, resulting in the presentation of a probability distribution for the location, indicating the likely leakage region.

The methodology employed for localization demonstrated effectiveness in identifying leaks according to the BattLeDIM benchmark, locating them within less than 50 meters from the actual leak location. However, the difficulty in detecting leaks with gradual flow growth was also evident in the localization. For instance, the leak in pipe p193 was located at 378 meters from the actual leak location. This leak was identified by the method proposed in Chapter 6 after 869 hours from the onset of the leak when the flow rate reached approximately 3.3 L/s. In contrast, using the IQR algorithm, the leak was detected 71 hours after the onset, while the MP algorithm detected the leak in 126 hours, both presented in Chapter 4.

Additionally, notable are the leaks in pipes p455 and p800, characterized by gradual flow growth, located at 75 and 48 meters, respectively, from the actual leak location. Despite the detection time being 885 hours for the leak in p455 and 193 hours for the leak in pipe p800, both were also identified when reaching approximately 3 L/s. The effectiveness in localization in these cases can be attributed to the proximity of leaks to sensors and reservoirs, resulting in consistent pressures, especially due to proximity to reservoirs, influencing pressures at monitored points. Another noteworthy case is related to the leak in pipe p331, detected 35 minutes after the onset, but localized 327 meters from the actual leak location. It was observed that, due to the pipe in question being in a sequential line of pipes, the method identified the pipe at the end of this sequence as the probable leak location.

Overall, the proposal for leak localization presented satisfactory results compared to other methods proposed in the literature. This is primarily due to the use of an approach that capitalizes on the interconnection of network topology, correlation between monitored data, and computational reduction during hydraulic simulation processes. Furthermore, the approach provides an alternative to determining the area monitored by sensors, using an algorithm for the determination of shorter paths between two vertices in graphs, considering the weights of connections between vertices and the distance to the monitored vertex.

8.1.5 Advances and Next Steps

Despite the lack of frequent and continuous monitoring in parts of water distribution networks, this thesis encompasses phases related to leak detection and localization in these systems. The research scope includes modelling leaks, sensor implementation, and culminates in the detection and localization of leaks. Overall, the adopted approach prioritizes the use of freely accessible tools, easily applicable algorithms, and data analysis methods previously validated in other research areas. The purpose is to facilitate the adoption of these methods by system operators and managers, aiming at reducing financial and resource losses. Notably, a 1% reduction in system losses can translate into substantial savings over the year, also impacting users' economies.

The proposal begins with leak modelling and sensor placement, relying exclusively on a computational network model that network management companies may possess. The proposed leak modelling covers pressure and water quality behaviours using available network models. Given the significant investment required for sensor installation, the proposed approaches allow operators to seek a balance by considering the installation of pressure or quality sensors, thus favouring leak detection and localization methods through network monitoring.

The approach for leak detection using Independent Component Analysis demonstrates advantages when applied to sets of multiple monitored data. This technique allows the separation of independent components in complex and multivariate data sets, as evidenced in Chapter 4, where it proved effective in leak detection. Conversely, if only one type of data is monitored, creating a temporal graph through the correlation between data from different monitored points can facilitate leak detection through graph structural analysis.

The above highlights opportunities to improve leak detection and localization methods. An example would be exploring the correlation between diverse monitored data in the creation of temporal graphs. Although the correlation between pressure data has been used in building the temporal graph, the use of multiple data for this purpose, such as flow, pressure, water quality, and reservoir levels, has not been fully explored. Thus, analysing the structures and relationships of this graph can result in significant advancements in network management, especially using complex network analysis methods widely employed in sectors like energy networks, the internet, and interpersonal relationships.

Furthermore, there is room to explore different ways of representing water distribution networks as graphs, attributing hydraulic, physical, and relational information between elements to vertices and edges. This approach can favour not only the reduction of physical losses but also pressure control, the determination of measurement districts, and the optimized operation of pumps and valves.

References

sensor data analysis inside RDBMS. **Mathematics**, v. 9, n. 17, p. 2146, 2021. ABDULSHAHEED, A., MUSTAPHA, F., & GHAVAMIAN, A. A pressure-based method for monitoring leaks in a pipe distribution system: A Review. **Renewable and Sustainable Energy Reviews**, 69, 902-911. 2017.

ABOWD, Gregory D.; MYNATT, Elizabeth D. Charting past, present, and future research in ubiquitous computing. ACM **Transactions on Computer-Human Interaction** (TOCHI), v. 7, n. 1, p. 29-58, 2000.

ADAMIC, Lada A. et al. Search in power-law networks. **Physical review** E, v. 64, n. 4, p. 046135, 2001.

ADEDEJI, K. B., HAMAM, Y., ABE, B. T., & ABU-MAHFOUZ, A. M. Towards achieving a reliable leakage detection and localization algorithm for application in water piping networks: An overview. **IEEE Access**, 5, 20272-20285. 2017.

AGATHOKLEOUS, A. et al. Robustness and vulnerability assessment of water networks by use of centrality metrics. **European Water Resources Association**, v. 58, p. 489-495, 2017.

ALMAZYAD, Abdulaziz S. et al. A proposed scalable design and simulation of wireless sensor network-based long-distance water pipeline leakage monitoring system. **Sensors**, v. 14, n. 2, p. 3557-3577, 2014.

ALTMAN, Edward I. et al. Financial distress prediction in an international context: A review and empirical analysis of Altman's Z-score model. **Journal of International Financial Management & Accounting**, v. 28, n. 2, p. 131-171, 2017.

ALTMAN, Edward I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. **The journal of finance**, v. 23, n. 4, p. 589-609, 1968.

ANCHIETA, Thomaz et al. Water distribution network expansion: an evaluation from the perspective of complex networks and hydraulic criteria. **Journal of Hydroinformatics**, v. 25, n. 3, p. 628-644, 2023.

ANUSHA, Peruri Venkata et al. Detecting outliers in high dimensional data sets using Z-score methodology. **International Journal of Innovative Technology and Exploring Engineering**, v. 9, n. 1, p. 48-53, 2019.

ARAD, Jonathan et al. A dynamic thresholds scheme for contaminant event detection in water distribution systems. **Water research**, v. 47, n. 5, p. 1899-1908, 2013.

ASLAM, H. et al. Detection of leaks in water distribution system using non-destructive techniques. In: **IOP Conference Series: Earth and Environmental Science**. IOP Publishing, 2018. p. 012004.

ASYALI, Musa H. et al. Gene expression profile classification: a review. **Current Bioinformatics**, v. 1, n. 1, p. 55-73, 2006.

AYATI, A. H., HAGHIGHI, A., & LEE, P. Statistical review of major standpoints in hydraulic transient-based leak detection**. Journal of Hydraulic Structures**, 5(1), 1-26. 2019.

AYATI, Amir Houshang; HAGHIGHI, Ali; GHAFOURI, Hamid Reza. Machine Learning–Assisted Model for Leak Detection in Water Distribution Networks Using Hydraulic Transient Flows. **Journal of Water Resources Planning and Management**, v. 148, n. 2, p. 04021104, 2022.

BARROS, D., ALMEIDA, I., ZANFEI, A., MEIRELLES, G., LUVIZOTTO, E. & BRENTAN, B. An Investigation on the Effect of Leakages on the Water Quality Parameters in Distribution Networks. **Water**, v. 15, n. 2, p. 324, 2023a.

BARROS, D., SOUZA, R., MEIRELLES, G., & BRENTAN, B. Leak detection in water distribution networks based on graph signal processing of pressure data. **journal of hydroinformatics**. 2023b.

BARROS, D.; Giaducianni, C.; Herrera, M.; Di Nardo, A.; Brentan, B. Pressure Sensors Placement for Leakage Detection Using Graph Signal Processing and Sampling Theory. **39th IAHR World Congress** 2022.

BEELER, Robert A. Application: Graph Theory. How to Count: An Introduction to Combinatorics and Its Applications, p. 309-343, 2015.

BERAHMAND, Kamal et al. A novel method of spectral clustering in attributed networks by constructing parameter-free affinity matrix. **Cluster Computing**, p. 1-20, 2022.

BERRY, M. W., ZAIN, J. M., YAP, B. W., & MOHAMED, A. Soft computing in data science. **Springer Singapore**, 2021.

BEZERRA, D., SOUZA, R., MEIRELLES, G., & BRENTAN, B. Leak detection in water distribution networks based on graph signal processing of pressure data. **journal of hydroinformatics**. 2023.

BIGGS, Norman; LLOYD, E. Keith; WILSON, Robin J. Graph Theory, 1736-1936. **Oxford University Press**, 1986.

BOCCALETTI, S., LATORA, V., MORENO, Y., CHAVEZ, M., & HWANG, D. U. Complex networks: Structure and dynamics. **Physics reports**, *424*(4-5), 175-308. 2006.

BONILLA, Carlos A. et al. A digital twin of a water distribution system by using graph convolutional networks for pump speed-based state estimation. **Water**, v. 14, n. 4, p. 514, 2022.

BOZTAŞ, F. et al. Analyzing the effect of the unreported leakages in service connections of water distribution networks on non-revenue water**. International Journal of Environmental Science and Technology**, v. 16, p. 4393-4406, 2019.

BRAGALLI, C., D'AMBROSIO, C., LEE, J., LODI, A., & TOTH, P. On the optimal design of water distribution networks: a practical MINLP approach. **Optimization and Engineering,** 13(2), 219-246. 2012.

BRANDES, Ulrik. A faster algorithm for betweenness centrality. **Journal of mathematical sociology**, v. 25, n. 2, p. 163-177, 2001.

Brasil, g. O Brasil e os ODM — ODM Brasil. http://www.odmbrasil.gov.br/o-brasil-e-os-odm, 2013. (Accessed on 10/26/2022).

BRASIL, T. Desafios para Disponibilidade Hídrica e Avanço da Eficiência do Saneamento Básico, 2019. Acesso Em 2020, 22,341–345.

BRASIL, TRATA. Desafios para Disponibilidade Hídrica e Avanço da Eficiência do Saneamento Básico, 2023.

BREDERECK, Robert et al. Assessing the computational complexity of multilayer subgraph detection**. Network Science**, v. 7, n. 2, p. 215-241, 2019.

BRENTAN, B., REZENDE, P., BARROS, D., MEIRELLES, G., LUVIZOTTO Jr, E., & IZQUIERDO, J. Cyber-attack detection in water distribution systems based on blind sources separation technique. **Water**, 13(6), 795. 2021.

BRENTAN, Bruno et al. Water quality sensor placement: A multi-objective and multi-criteria approach. **Water Resources Management**, v. 35, p. 225-241, 2021.

BROWN, Glen D.; YAMADA, Satoshi; SEJNOWSKI, Terrence J. Independent component analysis at the neural cocktail party. **Trends in neurosciences**, v. 24, n. 1, p. 54-63, 2001.

BUCHBERGER, Steven G.; NADIMPALLI, Gayatri. Leak estimation in water distribution systems by statistical analysis of flow readings. **Journal of water resources planning and management**, v. 130, n. 4, p. 321-329, 2004.3

BURSTYN, Yuval; GAZIT, Asaf; DVIR, Omri. Hierarchical Dynamic Time Warping methodology for aggregating multiple geological time series. **Computers & Geosciences**, v. 150, p. 104704, 2021.

CAI, Lei; GAO, Jingyang; ZHAO, Di. A review of the application of deep learning in medical image classification and segmentation. **Annals of translational medicine**, v. 8, n. 11, 2020.

CAPPONI, Caterina et al. Leak detection in a branched system by inverse transient analysis with the admittance matrix method. **Water Resources Management**, v. 31, p. 4075-4089, 2017.

CARDOSO, Sandra Maria et al. Optimal sensor placement for contamination detection: A multi-objective and probabilistic approach. **Environmental Modelling & Software**, v. 135, p. 104896, 2021.

CARRICK, Christina; MACLEOD, Kevin. An Evaluation of Genetic Algorithm Solutions in Optimization and Machine Learning. Proceedings of the **Annual Conference of CAIS / Actes du congrès annuel de l'ACSI**, [S. l.], 2013.

CASILLAS, Myrna V. et al. Optimal sensor placement for leak location in water distribution networks using genetic algorithms. **Sensors**, v. 13, n. 11, p. 14984-15005, 2013.

CASILLAS, Myrna V.; GARZA-CASTAÑÓN, Luis E.; PUIG, Vicenç. Optimal sensor placement for leak location in water distribution networks using evolutionary algorithms. **Water**, v. 7, n. 11, p. 6496-6515, 2015.

CHAN, T. K., CHIN, C. S., & ZHONG, X. Review of current technologies and proposed intelligent methodologies for water distributed network leakage detection. **IEEE Access**, 6, 78846-78867. 2018.

CHANG, Hong; YEUNG, Dit-Yan. Robust path-based spectral clustering. **Pattern Recognition**, v. 41, n. 1, p. 191-203, 2008.

CHEN, Siheng et al. Signal recovery on graphs: Fundamental limits of sampling strategies. **IEEE Transactions on Signal and Information Processing over Networks**, v. 2, n. 4, p. 539-554, 2016.

CHENG, Wei-ping et al. A novel statistical model for water age estimation in water distribution networks. **Mathematical Problems in Engineering**, v. 2015, 2015.

CHEW, Alvin Wei Ze et al. Generalized Acoustic Data Analysis Framework for Leakage Detection and Localization in Field Operational Water Distribution Networks. **Journal of Water Resources Planning and Management,** v. 149, n. 11, p. 04023056, 2023.

CHICCO, D., TÖTSCH, N., & JURMAN, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. **BioData mining**, 14(1), 1-22. 2021.

CHIKODILI, Nwodo Benita et al. Outlier detection in multivariate time series data using a fusion of K-medoid, standardized euclidean distance and Z-score. In: International Conference on Information and Communication Technology and Applications. **Cham: Springer International Publishing**, 2020. p. 259-271.

CHRISTODOULOU, Symeon E. Smarting up water distribution networks with an entropy-based optimal sensor placement stra. **Journal of Smart Cities**, v. 1, n. 1, p. 47-58, 2019.

COLLINS, Richard; BOXALL, Joby. Influence of ground conditions on intrusion flows through apertures in distribution pipes. **Journal of Hydraulic Engineering**, v. 139, n. 10, p. 1052-1061, 2013.

COMON, Pierre. Blind identification and source separation in 2/spl times/3 under-determined mixtures. **IEEE Transactions on Signal Processing**, v. 52, n. 1, p. 11-22, 2004.

COMON, Pierre; JUTTEN, Christian (Ed.). Handbook of Blind Source Separation: Independent component analysis and applications. **Academic press**, 2010.

CUGUERO-ESCOFET, Miquel A.; PUIG, Vicenç; QUEVEDO, Joseba. Optimal pressure sensor placement and assessment for leak location using a relaxed isolation index: Application to the Barcelona water network. **Control Engineering Practice**, v. 63, p. 1-12, 2017.

DANIEL, Ivo et al. A sequential pressure-based algorithm for data-driven leakage identification and model-based localization in water distribution networks**. Journal of Water Resources Planning and Management,** v. 148, n. 6, p. 04022025, 2022.

DARSANA, P.; VARIJA, K. Leakage detection studies for water supply systems—A review. **Water Resources Management: Select Proceedings of ICWEES-2016**, p. 141-150, 2018.

DEFFERRARD, M., MARTIN, L., PENA, R., & PERRAUDIN, N. Pygsp: Graph signal processing in python. URL https://github. com/epfl-lts2/pygsp. 2017.

DEUERLEIN, Jochen et al. Sensitivity analysis of topological subgraphs within water distribution systems. **Procedia Engineering**, v. 186, p. 252-260, 2017.

DI NARDO, A. et al. Simplified approach to water distribution system management via identification of a primary network. **Journal of Water Resources Planning and Management**, v. 144, n. 2, p. 04017089, 2018b.

DI NARDO, A., & DI NATALE, M. A heuristic design support methodology based on graph theory for district metering of water supply networks. **Engineering Optimization**, *43*(2), 193-211. 2011.

DI NARDO, Armando et al. A genetic algorithm for demand pattern and leakage estimation in a water distribution network. **Journal of Water Supply: Research and Technology**—AQUA, v. 64, n. 1, p. 35-46, 2015.

DI NARDO, Armando et al. Applications of graph spectral techniques to water distribution network management. **Water**, v. 10, n. 1, p. 45, 2018.

DIAO, Xu et al. Leak detection and location in liquid pipelines by analyzing the first transient pressure wave with unsteady friction. **Journal of Loss Prevention in the Process Industries**, v. 60, p. 303-310, 2019.

DIJKSTRA, Edsger W. A note on two problems in connexion with graphs. In: Edsger Wybe Dijkstra: His Life, Work, and Legacy. 2022. p. 287-290.

EMMANUEL, Tlamelo et al. A survey on missing data in machine learning. **Journal of Big Data**, v. 8, n. 1, p. 1-37, 2021.

ESTRADA, Ernesto; RODRIGUEZ-VELAZQUEZ, Juan A. Subgraph centrality in complex networks**. Physical Review** E, v. 71, n. 5, p. 056103, 2005.

FAN, X., ZHANG, X., & YU, X. B. Machine learning model and strategy for fast and accurate detection of leaks in water supply network. **Journal of Infrastructure Preservation and Resilience**, **2(1)**, 1-21. 2021.

FAN, Xudong; YU, Xiong. An innovative machine learning based framework for water distribution network leakage detection and localization. **Structural Health Monitoring**, v. 21, n. 4, p. 1626-1644, 2022.

FARAHANI, F. V., KARWOWSKI, W., & LIGHTHALL, N. R. Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review. **frontiers in Neuroscience**, 13, 585. 2019.

FARES, Ali et al. Leak detection in real water distribution networks based on acoustic emission and machine learning. **Environmental Technology**, v. 44, n. 25, p. 3850-3866, 2023.

FENG, Yelai; WANG, Huaixi. A Faster Algorithm for Betweenness Centrality Based on Adjacency Matrices. **arXiv** preprint arXiv:2205.00162, 2022.

FERREIRA, Bruno et al. Multi-objective optimization of pressure sensor location for burst detection and network calibration. **Computers & Chemical Engineering**, v. 162, p. 107826, 2022.

FISHER, Ian; KASTL, George; SATHASIVAN, Arumugam. New model of chlorine-wall reaction for simulating chlorine concentration in drinking water distribution systems. **Water research**, v. 125, p. 427-437, 2017.

FOX, Sam et al. Experimental quantification of contaminant ingress into a buried leaking pipe during transient events. **Journal of Hydraulic Engineering**, v. 142, n. 1, p. 04015036, 2016.

FUCHS, Helmut V.; RIEHLE, Rainer. Ten years of experience with leak detection by acoustic signal analysis. **Applied acoustics**, v. 33, n. 1, p. 1-19, 1991.

GALUPPINI, Giacomo; CREACO, Enrico; MAGNI, Lalo. A gain scheduling approach to improve pressure control in water distribution networks. **Control Engineering Practice**, v. 103, p. 104612, 2020.

GAO, J., Qi, S., WU, W., Li, D., RUAN, T., CHEN, L., ... & ZHUANG, Y. Study on leakage rate in water distribution network using fast independent component analysis. **Procedia Engineering**, 89, 934-941. 2014.

GAO, Jinliang et al. PMA Partition Method of Water Distribution Network Combined with Gragh Theory. **Procedia Engineering**, v. 186, p. 278-285, 2017.

GARG, Deepak et al. Dynamizing Dijkstra: A solution to dynamic shortest path problem through retroactive priority queue. **Journal of King Saud University-Computer and Information Sciences**, v. 33, n. 3, p. 364-373, 2021.

GEELEN, C. V., YNTEMA, D. R., MOLENAAR, J., & KEESMAN, K. J. Monitoring support for water distribution systems based on pressure sensor data. **Water Resources Management**, 33(10), 3339-3353. 2019.

GENG, Z., HU, X., HAN, Y., & ZHONG, Y. A novel leakage-detection method based on sensitivity matrix of pipe flow: case study of water distribution systems. **Journal of Water Resources Planning and Management,** 145(2), 04018094. 2019.

GHARGHABI, Shaghayegh et al. Matrix profile VIII: domain agnostic online semantic segmentation at superhuman performance levels. In: 2017 **IEEE international conference on data mining (ICDM)**. IEEE, 2017. p. 117-126.

GIUDICIANNI, C. et al. Multi-criteria method for the realistic placement of water quality sensors on pipes of water distribution systems. **Environmental Modelling & Software**, v. 152, p. 105405, 2022.

GIUDICIANNI, C. et al. Topological placement of quality sensors in water-distribution networks without the recourse to hydraulic modeling. **Journal of Water Resources Planning and Management**, v. 146, n. 6, p. 04020030, 2020.

GIUDICIANNI, C., Herrera, M., Di Nardo, A., Oliva, G., & Scala, A. The faster the better: On the shortest paths role for near real-time decision making of water utilities. **Reliability Engineering & System Safety**, v. 212, p. 107589, 2021.

GIUDICIANNI, Carlo et al. Zero-net energy management for the monitoring and control of dynamically-partitioned smart water systems. **Journal of Cleaner Production**, v. 252, p. 119745, 2020.

GOBET, J. et al. Development of an on-line chlorine sensor for water quality monitoring in public distribution networks. **Water Science and Technology: Water Supply**, v. 1, n. 2, p. 211-216, 2001.

GOULET, James-A.; COUTU, Sylvain; SMITH, Ian FC. Model falsification diagnosis and sensor placement for leak detection in pressurized pipe networks. **Advanced Engineering Informatics**, v. 27, n. 2, p. 261-269, 2013.

GRBČIĆ, Luka; KRANJČEVIĆ, Lado; DRUŽETA, Siniša. Machine learning and simulation-optimization coupling for water distribution network contamination source detection. **Sensors**, v. 21, n. 4, p. 1157, 2021.

GU, Xian-Ming et al. A Hessenberg-type algorithm for computing PageRank Problems. **Numerical Algorithms**, v. 89, n. 4, p. 1845-1863, 2022.

GUIDOTTI, Riccardo; D'ONOFRIO, Matteo. Matrix profile-based interpretable time series classifier**. Frontiers in Artificial Intelligence**, v. 4, p. 699448, 2021.

GUIDOTTI, Roberto et al. Modeling the resilience of critical infrastructure: The role of network dependencies. S**ustainable and resilient infrastructure**, v. 1, n. 3-4, p. 153-168, 2016.

GUTIÉRREZ-PÉREZ, Joanna A. et al. Application of graph-spectral methods in the vulnerability assessment of water supply networks. **Mathematical and Computer Modelling**, v. 57, n. 7-8, p. 1853-1859, 2013.

HAGBERG, Aric, and Drew CONWAY. "NetworkX: Network Analysis with Python." 2020.

HAGBERG, Aric; SWART, Pieter; S CHULT, Daniel. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining concepts and techniques, Morgan Kaufmann Publishers. San Francisco, CA, p. 335-391, 2001.

HANH, Nguyen Thi et al. An efficient genetic algorithm for maximizing area coverage in wireless sensor networks. **Information Sciences**, v. 488, p. 58-75, 2019.

HE, Guilin et al. An efficient multi-objective optimization method for water quality sensor placement within water distribution systems considering contamination probability variations. **Water research**, v. 143, p. 165-175, 2018.

HERRERA, Manuel et al. Performance assessment of a communication infrastructure with redundant topology: A complex network approach. **Computer Networks**, v. 228, p. 109747, 2023.

HONGYU, Kuang; SANDANIELO, Vera Lúcia Martins; DE OLIVEIRA JUNIOR, Gilmar Jorge. Análise de componentes principais: resumo teórico, aplicação e interpretação. **E&S Engineering and science**, v. 5, n. 1, p. 83-90, 2016.

HOWARD, Guy et al. Domestic water quantity, service level and health. 2003.

HU, Chengyu et al. A Spark-based genetic algorithm for sensor placement in large scale drinking water distribution systems. **Cluster Computing**, v. 20, p. 1089-1099, 2017.

HU, X., HAN, Y., YU, B., GENG, Z., & FAN, J. Novel leakage detection and water loss management of urban water supply network using multiscale neural networks. **Journal of Cleaner Production**, 278, 123611. 2021.

HU, Xuguang et al. Small leak location for intelligent pipeline system via action-dependent heuristic dynamic programming. IEEE **Transactions on Industrial Electronics**, v. 69, n. 11, p. 11723-11732, 2021.

HU, Z., CHEN, B., CHEN, W., TAN, D., & SHEN, D. Review of model-based and data-driven approaches for leak detection and location in water distribution systems. **Water Supply**, 21(7), 3282-3306. 2021.

HU, Zukang et al. Multi-objective and risk-based optimal sensor placement for leak detection in a water distribution system. **Environmental Technology & Innovation**, v. 28, p. 102565, 2022.

HUANG, Yen-Chen; LIN, Chao-Chin; YEH, Hund-Der. An optimization approach to leak detection in pipe networks using simulated annealing. **Water Resources Management**, v. 29, p. 4185-4201, 2015.

HUNAIDI, O., WANG, A., BRACKEN, M., GAMBINO, T., & FRICKE, C Acoustic methods for locating leaks in municipal water pipe networks. In: **International conference on water demand management**. Jordan: Dead Sea, 2004. p. 1-14.

HUNAIDI, Osama et al. Acoustic methods for locating leaks in municipal water pipe networks. In: **International conference on water demand management**. Jordan: Dead Sea, 2004. p. 1-14.

HUNAIDI, Osama; CHU, Wing T. Acoustical characteristics of leak signals in plastic water distribution pipes. **Applied Acoustics**, v. 58, n. 3, p. 235-254, 1999.

HUNTER, Paul R.; MACDONALD, Alan M.; CARTER, Richard C. Water supply and health. **PLoS medicine**, v. 7, n. 11, p. e1000361, 2010.

JADHAO, Ramrao D.; GUPTA, Rajesh. Calibration of water distribution network of the Ramnagar zone in Nagpur City using online pressure and flow data. **Applied Water Science**, v. 8, p. 1-10, 2018.

JEONG, Jina et al. Identifying outliers of non-Gaussian groundwater state data based on ensemble estimation for long-term trends. **Journal of Hydrology**, v. 548, p. 135-144, 2017.

JUN, Sanghyun; KWON, Hyuk Jae. The optimum monitoring location of pressure in water distribution system. **Water**, v. 11, n. 2, p. 307, 2019.

JUNG, Donghwi; LANSEY, Kevin. Water distribution system burst detection using a nonlinear Kalman filter. **Journal of Water Resources Planning and Management**, v. 141, n. 5, p. 04014070, 2015.

KAGHAZCHI, Afsaneh; SHAHDANY, S. Mehdy Hashemy; ROOZBAHANI, Abbas. Simulation and evaluation of agricultural water distribution and delivery systems with a Hybrid Bayesian network model. **Agricultural Water Management**, v. 245, p. 106578, 2021.

KALOFOLIAS, V. How to learn a graph from smooth signals. In Artificial Intelligence and Statistics (pp. 920-929). **PMLR**. 2016.

KAVEH, A., RAHMANI, P., & DADRAS ESLAMLOU, A. A multistage damage detection approach using graph theory and water strider algorithm. **Iranian Journal of Science and Technology, Transactions of Civil Engineering**, 46(1), 33-54. 2022.

KEOGH, Eamonn; RATANAMAHATANA, Chotirat Ann. Exact indexing of dynamic time warping. **Knowledge and information systems**, v. 7, p. 358-386, 2005.

KHORSHIDI, Mohammad Sadegh et al. Pressure sensor placement in water distribution networks for leak detection using a hybrid information-entropy approach. **Information Sciences**, v. 516, p. 56-71, 2020.

KHULIEF, Y. A. et al. Acoustic detection of leaks in water pipelines using measurements inside pipe. **Journal of Pipeline Systems Engineering and Practice**, v. 3, n. 2, p. 47-54, 2012.

KIRSTEIN, Jonas Kjeld et al. A semi-automated approach to validation and error diagnostics of water network data. **Urban water journal**, v. 16, n. 1, p. 1-10, 2019.

KIVELÄ, Mikko et al. Multilayer networks. **Journal of complex networks**, v. 2, n. 3, p. 203-271, 2014.

KIZILÖZ, Burak; ŞIŞMAN, Eyüp; ORUÇ, Halil Nurullah. Predicting a water infrastructure leakage index via machine learning. **Utilities Policy**, v. 75, p. 101357, 2022.

KLISE, K. A., HART, D., MORIARTY, D. M., BYNUM, M. L., MURRAY, R., BURKHARDT, J., & HAXTON, T. Water network tool for resilience (WNTR) user manual (No. SAND2017-8883R). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States). 2017.

KLISE, Katherine A. et al. A software framework for assessing the resilience of drinking water systems to disasters with an example earthquake case study. **Environmental modelling & software**, v. 95, p. 420-431, 2017.

KLISE, Katherine A.; MURRAY, Regan; HAXTON, Terra. An Overview of the Water Network Tool for Resilience (WNTR). 2018.

KUMAR, Jitendra et al. Detection of leaks in water distribution system using routine water quality measurements. In: **World Environmental and Water Resources Congress 2010**: Challenges of Change. 2010. p. 4185-4192.

LAN, Shi-Ying; CHEN, Run-Qing; ZHAO, Wan-Lei. Anomaly detection on it operation series via online matrix profile. **arXiv preprint** arXiv:2108.12093, 2021.

LAUCELLI, D., ROMANO, M., SAVIĆ, D., & GIUSTOLISI, O. Detecting anomalies in water distribution networks using EPR modelling paradigm. **Journal of Hydroinformatics**, 18(3), 409-427. 2016.

LENZI, C. et al. From energy balance to energy efficiency indicators including water losses. **Water science and technology: water supply**, v. 13, n. 4, p. 889-895, 2013.

LESKOVEC, Jure; FALOUTSOS, Christos. Sampling from large graphs. In: **Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining**. 2006. p. 631-636.

LI, Hailin et al. Time series clustering via matrix profile and community detection. **Advanced Engineering Informatics**, v. 54, p. 101771, 2022.

LI, Juan et al. Optimal sensor placement for leak localization in water distribution networks based on a novel semi-supervised strategy. **Journal of Process Control**, v. 82, p. 13-21, 2019.

LIANG, Haoran et al. Consistent anomaly detection and localization of multivariate time series via cross-correlation graph-based encoder–decoder GAN. IEEE **Transactions on Instrumentation and Measurement**, v. 71, p. 1-10, 2021.

LIEMBERGER, Roland; WYATT, Alan. Quantifying the global non-revenue water problem. **Water Supply**, v. 19, n. 3, p. 831-837, 2019.

LIU, Liang et al. Multilayer graph contrastive clustering network. **Information Sciences**, v. 613, p. 256-267, 2022.

LIU, Shuming; AUCKENTHALER, Pierre. Optimal sensor placement for event detection and source identification in water distribution networks. **Journal of Water Supply: Research and Technology—AQUA**, v. 63, n. 1, p. 51-57, 2014.

LIU, Yang et al. Water pipeline leakage detection based on machine learning and wireless sensor networks. **Sensors**, v. 19, n. 23, p. 5086, 2019.

LORENZO, PaoloDi; BARBAROSSA, Sergio; BANELLI, Paolo. Sampling and recovery of graph signals. In: **Cooperative and Graph Signal Processing**. Academic Press, 2018. p. 261-282.

MANKAD, Jaivik; NATARAJAN, Balasubramaniam; SRINIVASAN, Babji. Integrated approach for optimal sensor placement and state estimation: A case study on water distribution networks**. ISA transactions**, v. 123, p. 272-285, 2022.

MAO, Guoyong; ZHANG, Ning. Analysis of average shortest-path length of scale-free network. **Journal of Applied Mathematics**, v. 2013, p. 1-5, 2013.

MARINHO, Cristiane da Silva Ramos et al. Objetivos de Desenvolvimento do Milênio: impacto de ações assistenciais e mudanças socioeconômicas e sanitárias na mortalidade de crianças. **Cadernos de Saúde Pública**, v. 36, 2020.

MARTINI, Alberto; TRONCOSSI, Marco; RIVOLA, Alessandro. Leak detection in water-filled small-diameter polyethylene pipes by means of acoustic emission measurements. **Applied Sciences**, v. 7, n. 1, p. 2, 2016.

MASHHADI, Neda et al. Use of machine learning for leak detection and localization in water distribution systems. **Smart Cities**, v. 4, n. 4, p. 1293-1315, 2021.

MIN, Kyoung Won et al. Detecting and Localizing Leakages in Water Distribution Systems Using a Two-Phase Model. **Journal of Water Resources Planning and Management,** v. 148, n. 10, p. 04022051, 2022.

MOASHERI, Reza; GHAZIZADEH, Mohammadreza Jalili; TASHAYOEI, Mohammadreza. Leakage detection in water networks by a calibration method. **Flow Measurement and Instrumentation**, v. 80, p. 101995, 2021.

MOUNCE, S. R.; BOXALL, J. B.; MACHELL, J. Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows. **Journal of Water Resources Planning and Management,** v. 136, n. 3, p. 309-318, 2010.

MOUNCE, Stephen R.; MACHELL, John. Burst detection using hydraulic data from water distribution systems with artificial neural networks. **Urban Water Journal**, v. 3, n. 1, p. 21-31, 2006.

MUHAREMI, Fitore; LOGOFĂTU, Doina; LEON, Florin. Machine learning approaches for anomaly detection of water quality on a real-world data set. **Journal of Information and Telecommunication**, v. 3, n. 3, p. 294-307, 2019.

MUNIKOTI, Sai; LAI, Kexing; NATARAJAN, Balasubramaniam. Robustness assessment of hetero-functional graph theory based model of interdependent urban utility networks. **Reliability Engineering & System Safety**, v. 212, p. 107627, 2021.

NASIRIAN, Ali; MAGHREBI, Mahmoud F.; YAZDANI, Siavash. Leakage detection in water distribution network based on a new heuristic genetic algorithm model. 2013.

NAVARRO, Juan M. et al. Sound levels forecasting in an acoustic sensor network using a deep neural network. **Sensors**, v. 20, n. 3, p. 903, 2020.

NICHIFOROV, Cristina et al. Information extraction approach for energy time series modelling. In: 2020 **24th International Conference on System Theory, Control and Computing (ICSTCC)**. IEEE, 2020. p. 886-891.

OKEYA, I., KAPELAN, Z., HUTTON, C., & NAGA, D. Online burst detection in a water distribution system using the Kalman filter and hydraulic modelling. **Procedia Engineering**, 89, 418-427. 2014.

OLIVEIRA Gesner; SCAZUFCA, Pedro; SAYON, Pedro; OLIVEIRA, Rafael. Perdas de água 2021 (SNIS 2019): Desafios para disponibilidade hídrica e avanço da eficiência do saneamento básico. São Paulo, 2021.

OLIVEIRA, Juliano Vieira de et al. Estudo da decomposição em valores singulares e análise dos componentes principais. 2016.

OLSHAUSEN, Bruno A.; FIELD, David J. Sparse coding of sensory inputs**. Current opinion in neurobiology**, v. 14, n. 4, p. 481-487, 2004.

ORMSBEE, Lindell E.; LINGIREDDY, Srinivasa. Calibrating hydraulic network models. **Journal-American Water Works Association**, v. 89, n. 2, p. 42-50, 1997.

PAGE, Lawrence. The pagerank citation ranking: Bringing order to the web. Technical report. **Stanford Digital Library Technologies Project**, 1998, 1998.

PENG, Sen et al. Pressure sensor placement in water supply network based on graph neural network clustering method. **Water**, v. 14, n. 2, p. 150, 2022.

PEREIRA, T.; BARROS, D.; MEIRELLES, G.; Brentan, B. Real-time leak detection from signal filtering via independent component analysis. **Xxiv SimpóSio Bras. Recur. Hídrico,** 2021.

PERELMAN, Lina et al. Event detection in water distribution systems from multivariate water quality time series. **Environmental science & technology**, v. 46, n. 15, p. 8212-8219, 2012.

PERELMAN, Lina Sela et al. Sensor placement for fault location identification in water networks: A minimum test cover approach. **Automatica**, v. 72, p. 166-176, 2016.

PÉREZ, Ramon et al. Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks. **Control Engineering Practice**, v. 19, n. 10, p. 1157-1167, 2011.

PÉREZ-PADILLO, José et al. Design and implementation of a pressure monitoring system based on iot for water supply networks. **Sensors**, v. 20, n. 15, p. 4247, 2020.

PESENSON, Isaac. Sampling in Paley-Wiener spaces on combinatorial graphs. **Transactions of the American Mathematical Society**, v. 360, n. 10, p. 5603-5627, 2008.

PHAM, M. T., MERCIER, G., & MICHEL, J. Change detection between SAR images using a pointwise approach and graph theory. **IEEE Transactions on Geoscience and Remote Sensing**, **54**(4), 2020-2032. 2015.

PILLER, Olivier et al. Local sensitivity of pressure-driven modeling and demand-driven modeling steady-state solutions to variations in parameters. **Journal of Water Resources Planning and Management**, v. 143, n. 2, p. 04016074, 2017.

QI, Zhexian et al. Better understanding of the capacity of pressure sensor systems to detect pipe burst within water distribution networks. **Journal of Water Resources Planning and Management**, v. 144, n. 7, p. 04018035, 2018.

QUIÑONES-GRUEIRO, M., VERDE, C., PRIETO-MORENO, A., & LLANES-SANTIAGO, O. An unsupervised approach to leak detection and location in water distribution networks**. International Journal of Applied Mathematics and Computer Science**, **28**(2). 2018.

QUINONES-GRUEIRO, Marcos et al. Robust leak localization in water distribution networks using computational intelligence. **Neurocomputing**, v. 438, p. 195-208, 2021.

QUINTILIANI, Claudia et al. Optimal pressure sensor locations for leak detection in a Dutch water distribution network**. Environmental Sciences Proceedings**, v. 2, n. 1, p. 40, 2020.

RAEI, Ehsan et al. Placing an ensemble of pressure sensors for leak detection in water distribution networks under measurement uncertainty. **Journal of Hydroinformatics**, v. 21, n. 2, p. 223-239, 2019.

RAJABI, Mohammad Mahdi et al. Leak detection and localization in water distribution networks using conditional deep convolutional generative adversarial networks. **Water Research**, v. 238, p. 120012, 2023.

RAMOS, Helena M. et al. Smart water management towards future water sustainable networks. **Water**, v. 12, n. 1, p. 58, 2020.

RASHID, Sidra et al. A method for distributed pipeline burst and leakage detection in wireless sensor networks using transform analysis. **International Journal of Distributed Sensor Networks**, v. 10, n. 7, p. 939657, 2014.

REBOUÇAS, Aldo C. Água no Brasil: abundância, desperdício e escassez. Bahia análise & dados, v. 13, n. esp., p. 341-345, 2003.

ROMANO, M.; KAPELAN, Z.; SAVIĆ, D. A. Burst detection and location in water distribution systems. In: **World Environmental and Water Resources Congress 2011**: Bearing Knowledge for Sustainability. 2011. p. 1-10.

ROMANO, Michele; KAPELAN, Zoran; SAVIĆ, Dragan A. Automated detection of pipe bursts and other events in water distribution systems. **Journal of Water Resources Planning and Management**, v. 140, n. 4, p. 457-467, 2014.

ROMERO-BEN, Luis et al. Model-free sensor placement for water distribution networks using genetic algorithms and clustering. **IFAC**-PapersOnLine, v. 55, n. 33, p. 54-59, 2022.

ROSSMAN, Lewis A. et al. EPANET 2: users manual. 2000.

SADEGHIOON, Ali M. et al. Water pipeline failure detection using distributed relative pressure and temperature measurements and anomaly detection algorithms. **Urban Water Journal**, v. 15, n. 4, p. 287-295, 2018.

SAGNARD, F., NORGEOT, C., DEROBERT, X., BALTAZART, V., MERLIOT, E., DERKX, F., & LEBENTAL, B. FLORENCE et al. Utility detection and positioning on the urban site Sense-City using Ground-Penetrating Radar systems. **Measurement**, v. 88, p. 318-330, 2016.

SAKIYAMA, Akie et al. Efficient sensor position selection using graph signal sampling theory. In: 2016 **IEEE International Conference on Acoustics, Speech and Signal Processing** (ICASSP). IEEE, 2016. p. 6225-6229.

SAKOE, Hiroaki; CHIBA, Seibi. Dynamic programming algorithm optimization for spoken word recognition**. IEEE transactions on acoustics, speech, and signal processing**, v. 26, n. 1, p. 43-49, 1978.

SALGUERO, F. J., COBACHO, R., & PARDO, M. A. Unreported leaks location using pressure and flow sensitivity in water distribution networks. **Water Supply**, **19**(1), 11-18. 2019.

SÁNCHEZ-GARCÍA, Rubén J. et al. Hierarchical spectral clustering of power grids. **IEEE Transactions on Power Systems**, v. 29, n. 5, p. 2229-2237, 2014.

SANTOS, Cláudia Cristina; PEREIRA, Augusto José. Water demand forecasting model for the metropolitan area of São Paulo, Brazil. **Water resources management**, v. 28, p. 4401-4414, 2014.

SANTOS, R. et al. Spectral analysis for detection of leaks in pipes carrying compressed air. **Chemical Engineering Transactions**, v. 32, p. 1363-1368, 2013.

SANTOS-RUIZ, Ildeberto et al. Pressure sensor placement for leak localization in water distribution networks using information theory. **Sensors**, v. 22, n. 2, p. 443, 2022.

SARRATE, Ramon et al. Sensor placement for leak detection and location in water distribution networks. **Water Science and Technology: Water Supply**, v. 14, n. 5, p. 795-803, 2014.

SESHADHRI, C.; PINAR, Ali; KOLDA, Tamara G. Wedge sampling for computing clustering coefficients and triangle counts on large graphs. **Statistical Analysis and Data Mining: The ASA Data Science Journal**, v. 7, n. 4, p. 294-307, 2014.

SEYOUM, Alemtsehay G.; TANYIMBOH, Tiku T. Integration of hydraulic and water quality modelling in distribution networks: EPANET-PMX. **Water Resources Management**, v. 31, p. 4485-4503, 2017.

SHAHRA, Essa Q.; WU, Wenyan. Water contaminants detection using sensor placement approach in smart water networks. **Journal of Ambient Intelligence and Humanized Computing**, p. 1-16, 2020.

SHAO, Yu et al. Time-series-based leakage detection using multiple pressure sensors in water distribution systems. **Sensors**, v. 19, n. 14, p. 3070, 2019.

SHEKOFTEH, Mohammadreza; JALILI GHAZIZADEH, Mohammadreza; YAZDI, Jafar. A methodology for leak detection in water distribution networks using graph theory and artificial neural network. **Urban Water Journal**, v. 17, n. 6, p. 525-533, 2020.

SHI, Jie et al. Discovering and labeling power system events in synchrophasor data with matrix profile. In: **2019 IEEE Sustainable Power and Energy Conference (iSPEC)**. IEEE, 2019. p. 1827-1832.

SHINKUMA, Ryoichi; SUGIMOTO, Yuki; INAGAKI, Yuichi. Weighted network graph for interpersonal communication with temporal regularity. **Soft Computing**, v. 23, p. 3037-3051, 2019.

SILVA, M. A., LOUREIRO, D., AMADO, C., & ARREGUI, F. Análise de dados de telemedição de consumos domésticos para apoio na avaliação de perdas aparentes e reais em sistemas de distribuição de água. **Ingeniería del agua**, **25**(1), 37-47. 2021.

SITZENFREI, Robert et al. Using complex network analysis for optimization of water distribution networks. **Water resources research**, v. 56, n. 8, p. e2020WR027929, 2020.

SITZENFREI, Robert. Using complex network analysis for water quality assessment in large water distribution systems. **Water Research**, v. 201, p. 117359, 2021.

SMITH, Allan H.; LINGAS, Elena O.; RAHMAN, Mahfuzar. Contamination of drinking-water by arsenic in Bangladesh: a public health emergency. **Bulletin of the world health organizatio**n, v. 78, n. 9, p. 1093-1103, 2000.

SOLDEVILA, Adrià et al. Leak detection and localization in water distribution networks by combining expert knowledge and data-driven models. **Neural Computing and Applications,** p. 1-21, 2022.

SOPHOCLEOUS, S., SAVIĆ, D., & KAPELAN, Z. Leak localization in a real water distribution network based on search-space reduction. **Journal of Water Resources Planning and Management.** 2019.

SOROUSH, Fattah; ABEDINI, Mohammad J. Optimal selection of number and location of pressure sensors in water distribution systems using geostatistical tools coupled with genetic algorithm. **Journal of Hydroinformatics**, v. 21, n. 6, p. 1030-1047, 2019.

STAHL, Tim et al. Multilayer graph-based trajectory planning for race vehicles in dynamic scenarios. In: 2019 **IEEE Intelligent Transportation Systems Conference (ITSC)**. IEEE, 2019. p. 3149-3154.

STANKOVIĆ, L., DAKOVIĆ, M., & SEJDIĆ, E. Introduction to graph signal processing. In **Vertex-Frequency Analysis of Graph Signals** (pp. 3-108). Springer, Cham. 2019.

STEFFELBAUER, David B. et al. Pressure-leak duality for leak detection and localization in water distribution systems. **Journal of Water Resources Planning and Management**, v. 148, n. 3, p. 04021106, 2022.

STRAMARI, Marcos Roberto; KALBUSCH, Andreza; HENNING, Elisa. Random forest for the detection of unauthorized consumption in water supply systems: a case study in Southern Brazil. **Urban Water Journal**, v. 20, n. 3, p. 394-404, 2023.

SUN, C., PARELLADA, B., PUIG, V., & CEMBRANO, G. . Leak localization in water distribution networks using pressure and data-driven classifier approach. **Water**, **12**(1), 54. 2019.

TANAKA, Y., ELDAR, Y. C., ORTEGA, A., & CHEUNG, G. Sampling signals on graphs: From theory to applications**. IEEE Signal Processing Magazine**, **37**(6), 14-30. 2020.

TANYIMBOH, Tiku T.; CZAJKOWSKA, Anna M. Joint entropy based multi-objective evolutionary optimization of water distribution networks. **Water resources management**, v. 32, p. 2569-2584, 2018.

TORRES, Jacob M. et al. Exploring topological effects on water distribution system performance using graph theory and statistical models. **Journal of Water Resources Planning and Management**, v. 143, n. 1, p. 04016068, 2017.

TRATA, B. Desafios para Disponibilidade Hídrica e Avanço da Eficiência do Saneamento Básico. **Instituto Trata Brasil**, p. 21, 2023.

TSITSVERO, Mikhail; BARBAROSSA, Sergio; DI LORENZO, Paolo. Signals on graphs: Uncertainty principle and sampling**. IEEE Transactions on Signal Processing**, v. 64, n. 18, p. 4845-4860, 2016.

TZATCHKOV, Velitchko G.; ALCOCER-YAMANAKA, Victor H.; BOURGUETT ORTÍZ, Víctor. Graph theory based algorithms for water distribution network sectorization projects. In: **Water Distribution Systems Analysis Symposium 2006**. 2008. p. 1-15.

VAN ZYL, J. E. Theoretical modeling of pressure and leakage in water distribution systems. **Procedia Engineering**, **89**, 273-277. 2014.

VAN ZYL, J. E., & CASSA, A. M. Modeling elastically deforming leaks in water distribution pipes. **J. Hydraul**. Eng, **140**(2), 182-189. 2014.

VAN ZYL, J. E.; LAMBERT, A. O.; COLLINS, R. Realistic modeling of leakage and intrusion flows through leak openings in pipes. **Journal of Hydraulic Engineering**, v. 143, n. 9, p. 04017030, 2017.

VAN ZYL, Jakobus E.; CASSA, A. M. Modeling elastically deforming leaks in water distribution pipes. **Journal of Hydraulic Engineering**, v. 140, n. 2, p. 182-189, 2014.

VON LUXBURG, Ulrike. A tutorial on spectral clustering. **Statistics and computing**, v. 17, p. 395-416, 2007.

VRACHIMIS, Stelios G. et al. Battle of the leakage detection and isolation methods. **Journal of Water Resources Planning and Management**, v. 148, n. 12, p. 04022068, 2022.

WAN, Xiang et al. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. **BMC medical research methodology**, v. 14, p. 1-13, 2014.

WANG, Xiaoting et al. Multiple leakage detection and isolation in district metering areas using a multistage approach. **Journal of Water Resources Planning and Management**, v. 148, n. 6, p. 04022021, 2022.

WEI, Zhuangkun et al. Optimal sampling of water distribution network dynamics using graph fourier transform**. IEEE Transactions on Network Science and Engineering**, v. 7, n. 3, p. 1570-1582, 2019.

WORLD HEALTH ORGANIZATION et al. Joint monitoring programme for water supply and sanitation: report on intercountry workshop, Jakarta, Indonesia, 27-29 April 2009. WHO Regional Office for South-East Asia, 2010.

WRIGHT, L., De MARCO, M., & VENNERI, A. A graph theory approach to detect age and Alzheimer’s disease related changes to structural neural networks. **Alzheimer's & Dementia**, **17**, e050583. 2021.

WU, Yipeng et al. Defender–attacker–operator: Tri-level game-theoretic interdiction analysis of urban water distribution networks. **Reliability Engineering & System Safety**, v. 214, p. 107703, 2021.

WU, Yipeng; LIU, Shuming. A review of data-driven approaches for burst detection in water distribution systems. **Urban Water Journal**, v. 14, n. 9, p. 972-983, 2017

XIE, Lingqiang et al. Graph neural network approach for anomaly detection. **Measurement**, v. 180, p. 109546, 2021.

XIN, Kunlun et al. Apparent losses analysis in district metered areas of water distribution systems. **Water resources managemen**t, v. 28, p. 683-696, 2014.

XING, W., & GHORBANI, A. Weighted pagerank algorithm. In Proceedings. **Second Annual Conference on Communication Networks and Services Research**, 2004. (pp. 305-314). IEEE. 2004

XU, Jianhua et al. Identifying sets of key nodes for placing sensors in dynamic water distribution networks. **Journal of Water Resources Planning and Management**, v. 134, n. 4, p. 378-385, 2008.

XU, Tianshu et al. A small leakage detection approach for oil pipeline using an inner spherical ball. **Process Safety and Environmental Protection**, v. 124, p. 279-289, 2019.

XU, Tianshu et al. Pipeline leak detection based on variational mode decomposition and support vector machine using an interior spherical detector. **Process Safety and Environmental Protection**, v. 153, p. 167-177, 2021.

XU, Yifan et al. Small-world network analysis on fault propagation characteristics of water networks in eco-industrial parks. **Resources, Conservation and Recycling**, v. 149, p. 343-351, 2019.

XU, Yue; YU, Tao; YANG, Bo. Reliability assessment of distribution networks through graph theory, topology similarity and statistical analysis. **IET Generation, Transmission & Distribution**, v. 13, n. 1, p. 37-45, 2019.

YAN, Erjia; DING, Ying. Discovering author impact: A PageRank perspective. **Information processing & management**, v. 47, n. 1, p. 125-134, 2011.

YAN, Shuicheng et al. Graph embedding and extensions: A general framework for dimensionality reduction**. IEEE transactions on pattern analysis and machine intelligence**, v. 29, n. 1, p. 40-51, 2006.

YAO, Z., MARK, P., & RABBAT, M. Anomaly detection using proximity graph and pagerank algorithm. **IEEE Transactions on Information Forensics and Security**, **7**(4), 1288-1300. 2012.

YEH, Chin-Chia Michael et al. Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In: **2016 IEEE 16th international conference on data mining (ICDM)**. Ieee, 2016. p. 1317-1322.

YU, Xipeng et al. Resilience evaluation for water distribution system based on partial nodes' hydraulic information. **Water Research**, p. 120148, 2023.

YUNJUN, Zhang; FATTAHI, Heresh; AMELUNG, Falk. Small baseline InSAR time series analysis: Unwrapping error correction and noise reduction. **Computers & Geosciences**, v. 133, p. 104331, 2019.

ZAMAN, Dina et al. A review of leakage detection strategies for pressurised pipeline in steady-state. **Engineering Failure Analysis**, v. 109, p. 104264, 2020.

ZANFEI, A., MENAPACE, A., BRENTAN, B. M., RIGHETTI, M., & HERRERA, M. Novel approach for burst detection in water distribution systems based on graph neural networks. **Sustainable Cities and Society**, *86*, 104090. 2022.

ZHANG, Hong; LI, Zhanming. Anomaly detection approach for urban sensing based on credibility and time-series analysis optimization model. IEEE **Access**, v. 7, p. 49102-49110, 2019.

ZHANG, Qingzhou et al. Leakage zone identification in large-scale water distribution systems using multiclass support vector machines. **Journal of Water Resources Planning and Management,** v. 142, n. 11, p. 04016042, 2016.

ZHANG, Zhiyu; ZHANG, Kounan. Identification of critical nodes in power system based on graph theory and pagerank algorithm. In: 2019 IEEE **8th International Conference on Advanced Power System Automation and Protection** (APAP). IEEE, 2019. p. 213-217.

ZHAO, Mengke et al. Optimal sensor placement for pipe burst detection in water distribution systems using cost–benefit analysis. **Journal of Hydroinformatics**, v. 22, n. 3, p. 606-618, 2020.

ZHENG, Wenbo et al. An improved pagerank algorithm based on fuzzy C-means clustering and information entropy. In**: 2017 3rd IEEE International Conference on Control Science and Systems Engineering** (ICCSSE). IEEE, 2017. p. 615-618.

ZHOU, Xiao et al. Bridging hydraulics and graph signal processing: A new perspective to estimate water distribution network pressures**. Water Research**, v. 217, p. 118416, 2022.

ZHOU, Xiao et al. Deep learning identifies accurate burst locations in water distribution networks. **Water research**, v. 166, p. 115058, 2019.

ZYMBLER, Mikhail; IVANOVA, Elena. Matrix profile-based approach to industrial

REPÚBLICA FEDERATIVA DO BRASIL

Estado do Tocantins - Comarca de Paraiso do Tocantins - To

Cartório Samuel Miranda Acácio

Registro Civil

EMÍLIA ACÁCIO LUZ - Titular Vitalícia

DIVA LUZ ACÁCIO - Sub-oficial

CERTIDÃO DE NASCIMENTO

Livro A N.º 23 Fls. 251-V. N.º 18737

Sob nºs acima, foi registrado: Daniel Bezerra Barros —" "— Nascidos aos 28/11/92, às 02:55 hs. em Paraiso do Tocantins - To.

de sexo masculino, cor branca

Filho de João Luis Barros de Abreu

e de Aucirene Bezerra de Natividade Barros

Com as Profissões de: Motorista e professora

Naturais de: Cristalandia - To e Porto Nacional - To.

Casados no Cartório

Residente em nesta cidade

Avós paternos Raimundo Abreu Barros e Maria Leal Barros

Avós Maternos Francisco Bezerra Lima e Vitória Maria de Natividade

E testemunhas as constantes do termo

Declarante o pai

Obs. Feito o registro em: 30-11-92

O referido é verdade e dou fé

CARTÓRIO SAMUEL MIRANDA ACÁCIO
REGISTRO CIVIL
Drª Emília Acácio Luz
- Oficial Vitalícia -
Diva Luz Acácio
- Sub - Oficial -
PARAÍSO DO TOCANTINS - TO

P. do Tocantins 30 / 11 / 92

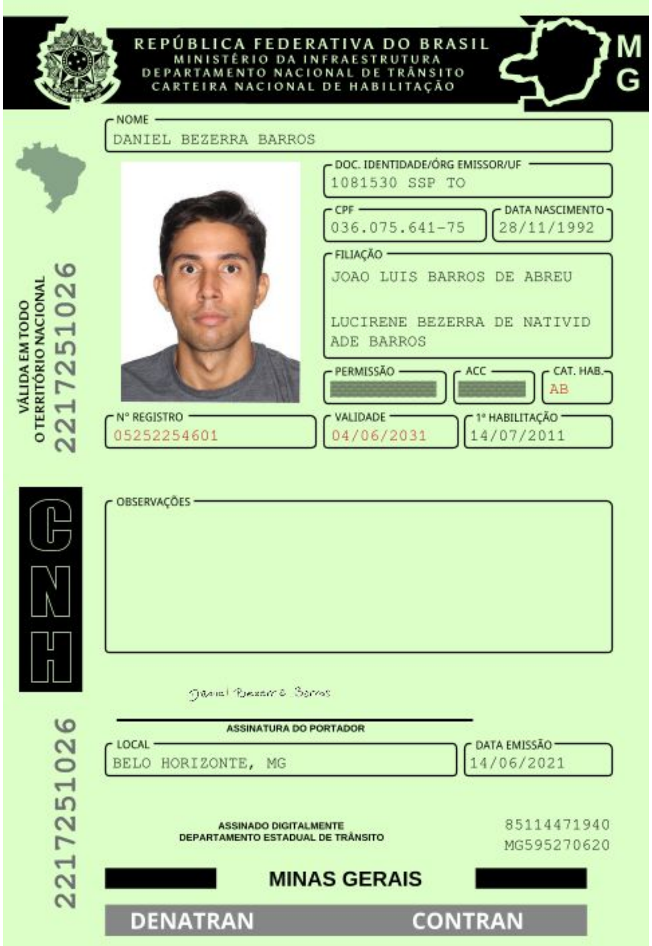
Emília Acácio Luz
— Oficial —

Emília Acácio

Oficial

CNH Digital

Departamento Nacional de Trânsito



REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DA INFRAESTRUTURA
DEPARTAMENTO NACIONAL DE TRÂNSITO
CARTEIRA NACIONAL DE HABILITAÇÃO

MG

VÁLIDA EM TODO O TERRITÓRIO NACIONAL

2217251026

NOME: DANIEL BEZERRA BARROS

DOC. IDENTIDADE/ÓRG EMISSOR/UF: 1081530 SSP TO

CPF: 036.075.641-75

DATA NASCIMENTO: 28/11/1992

FILIAÇÃO: JOAO LUIS BARROS DE ABREU

LUCIRENE BEZERRA DE NATIVIDADE BARROS

PERMISSÃO:

ACC:

CAT. HAB.: AB

Nº REGISTRO: 05252254601

VALIDADE: 04/06/2031

1ª HABILITAÇÃO: 14/07/2011

OBSERVAÇÕES:

CNH

Daniel Bezerra Barros

ASSINATURA DO PORTADOR

2217251026

LOCAL: BELO HORIZONTE, MG

DATA EMISSÃO: 14/06/2021

ASSINADO DIGITALMENTE
DEPARTAMENTO ESTADUAL DE TRÂNSITO

85114471940
MG595270620

MINAS GERAIS

DENATRAN CONTRAN

QR-CODE

Documento assinado com certificado digital em conformidade com a Medida Provisória nº 2200-2/2001. Sua validade poderá ser confirmada por meio do programa Assinador Serpro.

As orientações para instalar o Assinador Serpro e realizar a validação do documento digital estão disponíveis em: < http://www.serpro.gov.br/assinador-digital >, opção Validar Assinatura.

SERPRO / DENATRAN

CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS

ASSOCIAÇÃO EDUCACIONAL LUTERANA DO BRASIL - AELBRA
Credenciado pelo Decreto nº 130 – DOU de 07/07/2000
Recredenciado pela Portaria nº 3.607 – DOU de 20/10/2005

*O Reitor do Centro Universitário Luterano de Palmas, no uso de suas atribuições e tendo em vista a conclusão do curso de **ENGENHARIA CIVIL** em 30 de junho de 2016 e respectiva colação de grau no dia 28 de julho de 2016, confere o título de*

BACHAREL EM ENGENHARIA CIVIL

a

DANIEL BEZERRA BARROS

Brasileiro, natural do estado do Tocantins, nascido no dia 28 de novembro de 1992, portador da Cédula de Identidade nº 1.081.530 2ª via SSP/TO e outorga-lhe o presente Diploma, a fim de que possa gozar de todos os direitos e prerrogativas legais.

Palmas - TO, 18 de agosto de 2016.

Araujo
Drieli Drívela Cabral Araújo Varão
Secretária Geral

Daniel Bezerra Barros
Daniel Bezerra Barros
Diplomado

Adriano Chiarani da Silva
Reitor

CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS

ULBRA DIREÇÃO DE LEGISLAÇÃO E REGISTRO

Curso com renovação reconhecimento pela Portaria nº 286 publicada no D.O.U. de 27/12/2012.
Diploma registrado sob nº 7418, livro nº 98, folha nº 355, nos termos do art. 2º, § 4º do Decreto Nº 5.786 de 24 de [illegible] de 2006, DOU de 25 de maio de 2006.

Palmas - TO, 29 de agosto de 2014

TATIANE CARVALHO
REGISTRO DE DIPLOMAS

SELO DE AUTENTICIDADE

CREA-TO

Diploma Apresentado no CREA-TO para registro.

Palmas, 06/09/16

66-300

Assinatura do Funcionário e Matrícula

Termo de Autorização para Submissão de Recurso Educacional no RI-UFMG

1. Dados do autor responsável pela submissão

Nome completo: Daniel Bezerra Barros	
Matrícula UFMG: 209792	**Login MinhaUFMG:** danielbezerrab
Email institucional: danielbezerrab@eng-smarh.dout.ufmg.br	
Unidade (Faculdade/Instituto/Escola): Faculdade de engenharia	
Curso/Programa: Programa de pós graduação em Saneamento, Meio Ambiente e Recursos Hidricos	**Email do Curso/Programa:** danielbezerrab@gmail.com
Projeto:	**Email do Projeto:**
Setor:	**Email do Setor:**

2. Checklist para depósito do recurso educacional

Confira, abaixo, os requisitos e os dados obrigatórios para constarem no recurso educational:

1. Título completo
2. Nome completo e a indicação de responsabilidade intelectual de todos os envolvidos.
3. Data de publicação/criação
4. Nome da Unidade (Faculdade, Instituto, Escola, etc.) ao(s) qual(is) o(s) autor(es) está(ão) vinculado(s).
5. Palavras-chave (termos que resumem os temas principais do recurso)
6. Resumo
7. Objetivos pedagógicos
8. Tipo de Licença Creative Commons
9. Arquivo digital ou endereço externo de publicação isento de erros ou falhas técnicas

3. Dados do recurso educacional

O recurso já foi publicado e/ou revisado por pares e/ou comissão avaliadora?
( ) Sim ( ) Não

Título completo: **LEAK DETECTION AND LOCATION BASED ON HYDRAULIC AND QUALITY MONITORED DATA**

Ano de publicação/criação: **2024**

Selecione 01 (uma) coleção conforme a atividade acadêmica que motivou a produção do recurso:
( ) Ensino (**X**) Pesquisa ( ) Extensão

Responsabilidade pela revisão da língua portuguesa
Declaro, para fins de comprovação, ter realizado a revisão da língua portuguesa do recurso apresentado, consistindo em correção gramatical, adequação do vocabulário e inteligibilidade do texto.

Cessão e concessão de direitos autorais
Asseguro que o recurso educacional entregue é original, não infringe direitos de qualquer outra pessoa e contendo qualquer material do qual não detenho direitos de autor, inclusive o uso de imagem, som da voz e/ou nome exposto de terceiros e obtive a autorização prévia do detentor dos referidos direitos para conceder à UFMG. Estou ciente de que o depósito no Repositório Institucional preserva os direitos do autor e, dessa forma, não implica em transferência dos meus direitos sobre o recurso educacional para a Universidade.

Licenciamento de uso
Na qualidade de titular dos direitos de autor e com a anuência dos coautor(es) e colaborador(es) do recurso supracitado, de acordo com a Lei nº 9610/98, cedo e transfiro para a UFMG o direito de disponibilizá-lo gratuitamente, e de forma não-exclusiva, sem ressarcimento dos direitos autorais, para fins de leitura, impressão e/ou download pela internet a título da produção intelectual gerada pela Universidade, em qualquer idioma e em todos os países, a partir desta data, conforme as permissões da Licença Creative Commons, indicada no recurso educacional depositado.

Submissão do recurso no Repositório Institucional da UFMG
Afirmo ter ciência de que a disponibilização do recurso no RI-UFMG está condicionada a aprovação da Comissão Avaliadora da Unidade de Origem, conforme a análise dos aspectos de conteúdo, didático-pedagógicos, design de apresentação, acessibilidade e inclusão.