Leak Supervision in Water Distribution Networks based on model-based and data-driven approaches

Débora Cristina Costa da Silva Alves

THESE

présentée en vue
d'obtenir le grade de

DOCTEUR

En
Informatique, Automatique

Par

Débora Cristina Costa da Silva Alves

DOCTORAT DE L'UNIVERSITE DE LILLE
DELIVRE PAR IMT LILLE DOUAI

Titre de la thèse :

Leak Supervision in Water Distribution Networks based on model-based and data-driven approaches

Soutenue le 18 novembre 2022 devant le jury d'examen :

Rapporteur	Olivier PILLER	Directeur de recherche / INRAE Bordeaux, France
Rapporteur	Zoran KAPELAN	Professor / TU DELFT, Pays-Bas
Examinatrice	Pascale CHIRON	Maître de conférences, / Ecole Nationale d'Ingénieurs de Tarbes | INP ENIT,France
Examinatrice	Lizeth TORRES ORTIZ	Chargée de recherche, / Instituto de Ingeniería - UNAM, Mexico
Examinateur	Isam SHAHROUR	Professor / Université de Lille, France
Encadrant	Lala RAJAOARISOA	Maître de conférences / IMT Nord Europe, France
Directeur de thèse	Eric DUVIELLA	Professor / IMT Nord Europe, France
Directeur de thèse	Joaquim BLESA	Professor / UPC Universitat Politècnica de Catalunya, Espagne

Laboratoire d'accueil : Centre d'Enseignement, de Recherche et d'Innovation (CERI) Systèmes Numériques
Ecole Doctorale MADIS

To my parents: Divino Antonio Alves and

Helena M. Costa da Silva Alves,

To my husband: Carlos Scarpini da Silva

Dedication is a talent all on its own.

— Alphonse Elric

This book is dedicated to the voices in my head, the most remarkable of my friends.

— Fredrik Backman

ACKNOWLEDGEMENTS

These last three years of doctoral studies were full of learning and emotions. During these years, I met several people who helped me grow professionally and as a person. Therefore, I would like to express my sincere gratitude to all those who supported and helped me during this journey.

First, I would like to thank my supervisors, who have always supported me and been by my side to teach in the best possible way. I want to thank Dr. Joaquim Blesa for his patience and willingness to teach me. Dr. Eric Duviella for presenting me with the opportunity for a doctorate and always solving problems with lightness and good humor. And Dr. Lala Rajaoarisoa, who always helped me in my moments of doubt and always presented me with new challenges so that I could grow. You supported and helped me in these three years; a part of me is sad because the period you were my supervisor is ending, and another part of me is happy because I know you will be great friends for life.

Secondly, I want to thank my colleagues and friends who made these three years unforgettable. You were my family in Terrassa and Douai. A special thanks to Marjan Savadkooh, who has always been by my side to support, help me, and relieve pressure encountered from difficulties. Thanks to Boutrous Khoury and Sergio Samada, who made the pizza evenings more lively with conversations about Matlab and control theory; in these three years, you have become important friends, and I hope that all our dreams come true.

Third, I want to thank the research projects that funded my research. I sincerely thank the projects UTILITIES 4.0–P1 from AGAUR ACCIO RIS3CAT and L-BEST from the Spanish Ministry of Science and Innovation.

Lastly, I owe my gratitude to my parents, Divino Alves and Helena Maria, for always supporting me and never doubting me. I also express my most profound thankfulness to my husband, Carlos Scarpini, for his unlimited love. My life is happiest with you.

Débora Alves
2022

ABSTRACT

Water distribution networks are complex systems that are difficult to manage and monitor with extreme importance nowadays. The detection and location of leaks have become crucial for water distribution because when there are bursts or leaks, this can generate economic losses and an environmental issue and represents a potential risk to public health with contaminated water. However, with all these risks, currently, this infrastructure does not perform satisfactorily in practice.

The infrastructure in a medium-sized city can have pipes that span hundreds of kilometers connected to hundreds of nodes (pipe junctions or customers that relate to the network). Therefore, several factors can generate water loss during transport between the treatment plants and the reservoir for consumers, usually attributed to several causes, including leaks, measurement errors, and theft. Adequate leak supervision is vital for all of the factors mentioned above to save financial resources and water.

In this context, this dissertation intends to provide a methodology for leak supervision in the water distribution network (WDN), focusing on detecting, estimating, and locating leaks in the system. Additionally, sensor placement and validation approaches are presented to bring more robustness to the methods presented earlier. Finally, the entire methodology is introduced and tested in different water distribution networks based on simulated and real case studies.

The methods developed in this manuscript can be divided into model-based and data-driven. Model-based methods use hydraulic models that must be well-calibrated to obtain reliable results. This type of method can get more accurate results in comparison with the data-driven; however, not all WDNs have hydraulic models, or they are not updated to the current network since, over the years, the city tends to expand. Because of that, the hydraulic model must be constantly updated to be more faithful. Therefore, data-driven methods that use only sensor measurements and system topological information can be applied when hydraulic models are inaccurate or unavailable.

The leak detection methods presented in this thesis are fully data-driven that only require historical data from the sensors installed in the WDN. The first method uses only information from the flow sensors installed in the inlet of the network, detecting and estimating the leak's magnitude. The second is a complementary technique to focus on the multi-leaks problem, i.e., when more than one leak co-occurs in the WDN. This method uses inlet flow and pressure sensors installed around the WDN to group them by clustering and studying the leak's impact in all of the zone.

On the other hand, the leak localization strategies developed in this thesis are divided into model-based and data-driven. All the leak localization techniques are based on the residual pressure study generated using the comparison between leak-free pressure estimations and leak pressure measurements. The first method is a model-based methodology that focuses on the problem of multi-leaks in the WDN ideal when it is available a well-calibrated hydraulic model of the WDN. Following, three data-driven approaches are presented: (i) starting with a technique that aims to reduce the area of the leak localization in the WDN, to proceed that a distance clustering with pre-defined centroids is generated, (ii) following an approach based on analyzing the extra flow effect that appears in a system with leakage with the goal of pass from the cluster analysis to a node study. (iii) The last method focuses on the multi-leak problem by applying the fusion of the pressure residues applying the radial base function (RBF) interpolation to obtain the network zone with the highest leak probability.

To complement the leak supervision, this thesis developed a method for sensor placement ideal for the leak localization techniques based on the residual pressure analysis. In addition, a

sensor validation method was designed to make the leak detection and localization more robust since the technique allows the analysis to know if pressure sensors are working correctly. If not, the measurement data from this sensor can be removed for other leak supervision techniques not to be affected by wrong measurements.

Keywords: Water distribution networks, Leak detection, Leak localization, Real applications, Sensor placement, Sensor validation.

RÉSUMÉ

Les réseaux de distribution d'eau sont des systèmes complexes dont la gestion et la surveillance revêtent une extrême importance. La détection et la localisation des fuites sont devenues cruciales pour ce type de réseau, car la présence de fissures ou de fuites peut entraîner des pertes économiques, des problèmes environnementaux, et un risque potentiel de contamination de l'eau qui pourrait nuire à la santé publique. Malgré tous les risques encourus, les infrastructures de distribution d'eau potable sont encore aujourd'hui grandement impactés par les fuites.

Dans une ville de taille moyenne les infrastructures peuvent comporter des tuyaux de plusieurs centaines de kilomètres reliés à des centaines de nœuds (jonctions de tuyaux ou clients liés au réseau). Plusieurs facteurs, notamment les fuites, les erreurs de mesure et le vol, peuvent générer des pertes durant l'acheminement de l'eau entre les stations de traitement, les réservoirs et les consommateurs. Une surveillance appropriée des fuites est donc essentielle pour tous les facteurs mentionnés ci-dessus afin de préserver l'eau et d'économiser des ressources financières.

Dans ce contexte, cette thèse vise à développer une méthodologie pour la détection, l'estimation et la localisation de fuites dans les réseaux de distribution d'eau potable. Elle présente également une approche permettant de bien positionner les capteurs, en validant leur position, afin de rendre la méthodologie plus robuste. Enfin, l'ensemble des techniques sont présentées et testées dans différents réseaux de distribution d'eau sur la base d'études de cas simulées et réelles.

Les méthodes développées dans cette thèse peuvent être divisées en deux catégories : les méthodes basées sur des modèles et les méthodes basées sur des données. Les méthodes basées sur des modèles utilisent des modèles hydrauliques des réseaux de distribution qui doivent être bien calibrés pour obtenir des résultats fiables. Ce type de méthodes permet d'obtenir des résultats plus précis que les méthodes basées sur des données. Cependant, tous les réseaux de distribution d'eau ne disposent pas de modèles hydrauliques, ou ceux-ci ne sont pas réactualisés assez souvent pour tenir compte de l'expansion des villes et des modifications que peuvent subir les réseaux. Par conséquent, lorsque les modèles hydrauliques ne sont pas disponibles, les approches basées sur les données offrent une très bonne alternative puisqu'elles n'utilisent uniquement que les mesures issues des capteurs et les informations topologiques des réseaux.

Deux méthodes de détection de fuites entièrement guidées par les données sont présentées dans cette thèse. La première méthode utilise uniquement les informations des capteurs de débit installés, détectant et estimant l'ampleur d'une fuite unique. La seconde est une technique complémentaire qui se concentre sur la détection de fuites multiples, c'est-à-dire lorsque plusieurs fuites se produisent simultanément dans le réseau de distribution d'eau. Cette dernière méthode utilise les capteurs de pression déjà installés et consiste à les regrouper par zone (*i.e.* par classe), puis à étudier l'impact des fuites dans toutes ces zones.

Les stratégies de localisation des fuites sont basées sur l'étude de la pression résiduelle générée par la comparaison entre les estimations de pression et les mesures de pression des fuites. Les méthodologies développées peuvent être divisées en deux parties : (1) une approche basée sur un modèle qui se concentre sur le problème de fuites multiples dans le réseau de distribution d'eau. Cette méthode est particulièrement adaptée aux réseaux équipés d'un système de lecture automatique des compteurs garant de modèles hydrauliques précis. Cependant, cette méthodologie nécessite que les modèles hydrauliques soient bien calibrés pour être efficiente. (2) Les approches basées sur les données qui étudient l'effet des fuites sur les capteurs de pression installés dans les réseaux afin de sélectionner les zone les plus susceptibles d'être impactées par une fuite. Ce type d'approches utilise principalement la théorie des graphes pour représenter le réseau. Dans cette thèse, trois approches basées sur les données sont proposées : (i) La première technique consiste à réduire la zone où l'occurence d'une fuite est la plus plausible. Cette

technique repose sur une méthode de classification à base de distances et de noyaux qui sont définis à partir des informations de pression des capteurs, et de certains nœuds sélectionnés. (ii) La deuxième méthode proposée est basée sur l'analyse des effets des fuites avec pour objectif de réduire la zone de recherche en passant de l'analyse de clusters de nœuds à l'analyse de nœuds. (iii) La dernière méthode utilise la fusion des résidus de pression en interpolant une fonction de base radiale (RBF) pour obtenir la zone du réseau ayant la plus forte probabilité d'être impactée par une fuite.

Par ailleurs, le positionnement des capteurs dans les réseaux de distribution d'eau est fondamental car il peut grandement affecter les performances de localisation des fuites. Dans cette thèse, une méthode de positionnement des capteurs de pression, idéale pour les techniques de localisation de fuites qui analysent les résidus de pression, est proposée. Afin de robustifier cette méthode, une approche de validation du positionnement des capteurs est développée, en analysant le processus de détection des fuites en fonction de la position des capteurs. Si le positionnement d'un capteur n'entraine pas d'amélioration lors de la détection des fuites, les données issues de ce capteur peuvent être retirées afin que les autres techniques de détection de fuites n'en soient pas affectées.

Mots clés: Réseau de distribution d'eau, Détection de fuites, Localisation des fuites, Positionnement de capteurs, Validation du positionnement des capteurs, Cas d'étude réels.

RESUMEN

Las redes de distribución de agua son sistemas complejos, difíciles de gestionar y monitorizar con extrema importancia en la actualidad. La detección y localización de fugas se ha vuelto crucial para la distribución de agua, ya que cuando se presentan reventones o fugas de agua, esto puede generar pérdidas económicas, un problema ambiental y representa un riesgo potencial para la salud pública con agua contaminada. Sin embargo, con todos estos riesgos, actualmente, estas infraestructuras no funcionan satisfactoriamente en la práctica. La infraestructura en una ciudad mediana puede tener tuberías que se extienden por cientos de kilómetros conectadas a cientos de nodos (uniones de tuberías o clientes que se relacionan con la red). Por lo tanto, varios factores pueden generar pérdidas de agua durante el transporte entre las plantas de tratamiento y el embalse para los consumidores, generalmente atribuidos a varias causas, incluidas fugas, errores de medición y fraudes. La supervisión adecuada de fugas es vital para todos los factores mencionados anteriormente para ahorrar recursos financieros y agua.

En este contexto, esta tesis doctoral presenta una nueva metodología para la supervisión de fugas en las redes de distribución de agua (RDA), enfocándose en la detección, estimación y localización de fugas en el sistema. Además, se presentan enfoques de validación y ubicación de sensores para brindar más solidez a los métodos presentados anteriormente. Finalmente, toda la metodología se introduce y se prueba en diferentes redes de distribución de agua en base a estudios de casos reales y simulados. Los métodos desarrollados en este manuscrito se

pueden dividir en basados en modelos y basados en datos. Los métodos basados en modelos usan modelos hidráulicos que deben estar bien calibrados para obtener resultados confiables. Este tipo de métodos puede obtener resultados más precisos en comparación con el basado en datos; sin embargo, no todas las RDA tienen modelos hidráulicos o no están actualizados a la red actual ya que, con el paso de los años, la ciudad tiende a expandirse. Por esta razón, el modelo hidráulico debe ser actualizado constantemente para ser más fiel. Por lo tanto, los métodos basados en datos que usan solo las mediciones de los sensores las y la información topológica del sistema se pueden aplicar cuando los modelos hidráulicos son inexactos o no están disponibles. Los métodos de detección de fugas presentados en esta tesis están totalmente basados en datos que solo requieren datos históricos de los sensores instalados en la RDA. El primer método utiliza sólo información de los sensores de caudal instalados en la entrada de la red, detectando y estimando la magnitud de la fuga. La segunda es una técnica complementaria para centrarse en el problema de las fugas múltiples, es decir, cuando ocurre más de una fuga en la RDA. Este método utiliza el flujo de entrada y los sensores de presión instalados en la RDA para agruparlos y estudiar el impacto de la fuga en todas las zonas. Por otro lado, las estrategias de localización de fugas desarrolladas en esta tesis se dividen en basadas en modelos y basadas en datos. Todas las técnicas de localización de fugas se basan en el estudio de presión generado a partir de la comparación entre estimaciones de presión sin fugas y mediciones de presión con la fuga. El primer método es una metodología basada en modelos que se centra en el problema de las fugas múltiples en la RDA ideal cuando se dispone de un modelo hidráulico bien calibrado de la red. A continuación, se presentan tres enfoques basados en datos: (i) comenzando con una técnica que apunta a reducir el área de localización de la fuga en la RDA, para proceder a generar un agrupamiento basada en la distancia con centroides predefinidos, (ii) siguiendo una enfoque basado en analizar el efecto de flujo extra que aparece en un sistema con fugas con el objetivo de pasar del análisis de fuga a nivel de nodo. (iii) El último método se centra en el problema de las fugas múltiples aplicando la fusión de los residuos de presión mediante la interpolación dada por una función de base radial para obtener la zona de la red con la mayor probabilidad de fuga. Para complementar la supervisión de fugas, esta tesis desarrolla un método de colocación de sensores ideal para las técnicas de localización de fugas basadas en el análisis del residuo de presión. Además, se diseña un método de validación de sensores para hacer más

robusta la detección y localización de fugas, ya que permite el análisis para saber si los sensores de presión están funcionando correctamente. De lo contrario, los datos de medición de este sensor se pueden eliminar para que las técnicas de supervisión de fugas no se vean afectadas por mediciones incorrectas.

Palabras clave: Redes de distribución de agua, Detección de fugas, Localización de fugas, Aplicaciones reales, Colocación de sensores, Validación de sensores.

RESUM

Les xarxes de distribució d'aigua són sistemes complexos, difícils de gestionar i monitoritzar amb molta importància actualment. La detecció i la localització de fuites s'ha tornat crucial per a la distribució d'aigua, ja que quan es presenten rebentades o fuites d'aigua, això pot generar pèrdues econòmiques, un problema ambiental i representa un risc potencial per a la salut pública amb aigua contaminada. No obstant això, amb tots aquests riscos, actualment, aquestes infraestructures no funcionen satisfactòriament a la pràctica. La infraestructura en una ciutat mitjana pot tenir canonades que s'estenen per centenars de quilòmetres connectades a centenars de nodes (unions de canonades o clients que es relacionen amb la xarxa). Per tant, diversos factors poden generar pèrdues d'aigua durant el transport entre les plantes de tractament i l'embassament per als consumidors, generalment atribuïts a diverses causes, incloses les fuites, els errors de mesura i els fraus. La supervisió adequada de fuites és vital per a tots els factors esmentats anteriorment per estalviar recursos financers i aigua.

En aquest context, aquesta tesi doctoral presenta una nova metodologia per a la supervisió de fuites a les xarxes de distribució d'aigua (XDA), enfocant-se en la detecció, estimació i localització de fuites al sistema. A més, es presenten enfocaments de validació i ubicació de sensors per brindar més solidesa als mètodes presentats anteriorment. Finalment, tota la metodologia s'introdueix i es prova a diferents xarxes de distribució d'aigua segons estudis de casos reals i simulats. Els mètodes desenvolupats en aquest manuscrit es poden dividir en basats en models i

basats en dades. Els mètodes basats en models usen models hidràulics que han d'estar ben calibrats per obtenir resultats fiables. Aquest tipus de mètodes pot obtenir resultats més precisos en comparació del basat en dades; no obstant, no totes les XDA tenen models hidràulics o no estan actualitzats a la xarxa actual ja que, amb el pas dels anys, la ciutat tendeix a expandir-se. Per això, el model hidràulic ha de ser actualitzat constantment per ser més fidel. Per tant, els mètodes basats en dades que fan servir només les mesures dels sensors i la informació topològica del sistema es poden aplicar quan els models hidràulics són inexactes o no estan disponibles. Els mètodes de detecció de fuites presentats en aquesta tesi estan totalment basats en dades que només requereixen dades històriques dels sensors instal·lats a la XDA. El primer mètode utilitza només informació dels sensors de cabal instal·lats a l'entrada de la xarxa, detectant i estimant la magnitud de la fuita. La segona és una tècnica complementària per centrar-se en el problema de les fuites múltiples, és a dir, quan hi ha més d'una fuita a la XDA. Aquest mètode utilitza el cabal d'entrada a la xarxa i els sensors de pressió instal·lats a la XDA per agrupar-los i estudiar l'impacte de la fugida a totes les zones. D'altra banda, les estratègies de localització de fuites desenvolupades en aquesta tesi es divideixen en basades en models i basades en dades. Totes les tècniques de localització de fuites es basen en l'estudi de pressió generat a partir de la comparació entre estimacions de pressió sense fugides i mesures de pressió amb la fuita. El primer mètode és una metodologia basada en models que se centra en el problema de les fuites múltiples a la XDA ideal quan es disposa d'un model hidràulic ben calibrat de la xarxa. A continuació, es presenten tres enfocaments basats en dades: (i) començant amb una tècnica que apunta a reduir l'àrea de localització de la fuita a la XDA, mitjançant la generació d'un agrupament basat en la distància amb centroides predefinits, (ii) seguint un enfocament basat en analitzar l'efecte de flux extra que apareix en un sistema amb fuites amb l'objectiu de passar de l'anàlisi de fuita a nivell de node. (iii) L'últim mètode se centra en el problema de les fuites múltiples aplicant la fusió dels residus de pressió mitjançant la interpolació donada per una funció de base radial per obtenir la zona de la xarxa amb la probabilitat de fuita més gran. Per complementar la supervisió de fuites, aquesta tesi desenvolupa un mètode de col·locació de sensors ideal per a les tècniques de localització de fuites basades en l'anàlisi del residu de pressió. A més, es dissenya un mètode de validació de sensors per fer més robusta la detecció i la localització de fuites, ja que permet l'anàlisi per saber si els sensors de pressió estan funcionant correctament. Si no, les dades de

mesures d'aquest sensor es poden eliminar perquè les tècniques de supervisió de fuites no es vegin afectades per les mesures incorrectes.

Paraules clau: Xarxes de distribució d'aigua, Detecció de fuites, Localització de fuites, Aplicacions reals, Col·locació de sensors, Validació de sensors.

RESUMO

As redes de distribuição de água são sistemas complexos de difícil gerenciamento e monitoramento com extrema importância nos dias atuais. A detecção e localização de vazamentos tornaram-se cruciais para o monitoramento de distribuição de água, pois quando ocorrem estouros ou vazamentos, isso gera perdas econômicas e problemas ambientais e representa um risco potencial à saúde com água contaminada. No entanto, com todos esses riscos, atualmente, essa infraestrutura na prática não funciona de forma satisfatória.

A infraestrutura em uma cidade de médio porte pode ter tubulações que se estendem por centenas de quilômetros conectadas a centenas de nós (junções de tubulação ou clientes que se reportam à rede). Portanto, diversos fatores podem gerar perdas de água durante o transporte entre as estações de tratamento e para os consumidores, geralmente atribuídas às causas do reservatório, incluindo vazamentos, erros de medição e furtos. A supervisão adequada de vazamentos é vital para todos os fatores mencionados acima para economizar recursos financeiros e água.

Neste contexto, esta dissertação pretende fornecer uma metodologia para supervisão de vazamentos na rede de distribuição de água (RDA), com foco na detecção, estimativa e localização de vazamentos no sistema. Além disso, abordagens de posicionamento e validação de sensores são apresentadas para trazer mais robustez aos métodos apresentados anteriormente. Por fim, toda a metodologia é introduzida e testada em diferentes redes de distribuição de água com base

em estudos de caso simulados e reais.

Os métodos desenvolvidos neste manuscrito podem ser divididos em baseados em modelos e orientados por dados. Os métodos baseados em modelos usam modelos hidráulicos que devem ser bem calibrados para obter resultados confiáveis. Esse tipo de método pode obter resultados mais precisos em comparação com o orientado a dados; no entanto, nem todas as WDNs possuem modelos hidráulicos, ou não são atualizadas para a rede atual, pois, com o passar dos anos, a cidade tende a se expandir. Por isso, o modelo hidráulico deve ser constantemente atualizado para ser mais fiel. Portanto, métodos baseados em dados que usam apenas medições de sensores e informações topológicas do sistema podem ser aplicados quando os modelos hidráulicos são imprecisos ou indisponíveis.

Os métodos de detecção de vazamentos apresentados nesta tese são totalmente orientados a dados que requerem apenas dados históricos dos sensores instalados na WDN. O primeiro método utiliza apenas informações dos sensores de vazão instalados na entrada da rede, detectando e estimando a magnitude do vazamento. A segunda é uma técnica complementar para focar no problema de multivazamento, ou seja, quando mais de um vazamento co-ocorre na WDN. Este método usa sensores de vazão e pressão de entrada instalados ao redor do WDN para agrupá-los agrupando e estudando o impacto do vazamento em toda a zona.

Por outro lado, as estratégias de localização de vazamentos desenvolvidas nesta tese são divididas em baseadas em modelos e baseadas em dados. Todas as técnicas de localização de vazamentos são baseadas no estudo de pressão residual gerado usando a comparação entre estimativas de pressão livre de vazamento e medições de pressão de vazamento. O primeiro método é uma metodologia baseada em modelo que foca no problema de multi-vazamento no WDN ideal quando está disponível um modelo hidráulico bem calibrado do WDN. A seguir, três abordagens orientadas a dados são apresentadas: (i) começando com uma técnica que visa reduzir a área de localização do vazamento no WDN, para proceder que um agrupamento de distância com centroides pré-definidos seja gerado , (ii) seguindo uma abordagem baseada na análise do efeito de fluxo extra que aparece em um sistema com vazamento com o objetivo de passar da análise de cluster para um estudo de nós. (iii) O último método foca no problema de multivazamento aplicando a fusão dos resíduos de pressão aplicando a interpolação da função

de base radial (RBF) para obter a zona da rede com maior probabilidade de vazamento.

Para complementar a supervisão de vazamentos, esta tese desenvolveu um método de posicionamento de sensores ideal para as técnicas de localização de vazamentos baseado na análise de pressão residual. Além disso, o método de validação de sensores foi projetado para tornar a detecção e localização de vazamentos mais robusta, pois a técnica permite que a análise saiba se os sensores de pressão estão funcionando corretamente. Caso contrário, os dados de medição deste sensor podem ser removidos para que outras técnicas de supervisão de vazamento não sejam afetadas por medições erradas.

Palabras chaves: Redes de distribuição de água, Detecção de vazamentos, Localização de vazamentos, Aplicações reais, Posicionamento de sensores, Validação de sensores.

NOMENCLATURE

α^u	Amplitude of the demand uncertainty [%]
α^{l_j}	New demand pattern distribution with a simulated leak with magnitude l in the node j
α_i	Normalized proportional outflow in node i
$\bar{\alpha}_i$	New demand with uncertainties in node i
$\boldsymbol{\Gamma}$	Confusion matrix
$\boldsymbol{\nu^e}$	Signature vector of events e
$\boldsymbol{\Omega}$	Leak sensitivity matrix
$\boldsymbol{A}$	Matrix containing the minimum topological distance
$\boldsymbol{d}$	Vector of nodal demands
$\boldsymbol{H}$	Incidence matrix
$\boldsymbol{p}, \boldsymbol{p}^{(in)}, \boldsymbol{p}^{(r)}$	Vector of absolute pressures at the nodes, vector of inlet pressure, vector of pressure in the reservoir
$\boldsymbol{q}$	Vector of flows in the edges
$\boldsymbol{z}$	Vector of geodesic levels at each vertex
$\Delta\boldsymbol{p}$	Vector of differential pressures across the pipes
$\Delta\overline{r_{Ci}}$	The group index of leak detection of the Group $\mathcal{C}$ of $i = 1, \ldots, \alpha$
δ_h	Threshold considering hourly measurements
Δ_{C_i}	Threshold in the leak detection method in the group C_i
Δ_W	Threshold in the leak detection method
ℓ	The number of elements of $\mathcal{S}$

γ^i Leak localization index of node i

$\hat{\bar{l}}(k)$ Virtual fused to leak estimate measurement at the instant k

$\hat{d}_i$ Estimated demand at node i,

$\hat{p}_i, p_i$ Leak-free pressure estimation, and pressure measurement at inner node i.

$\hat{y}(k)$ Demand forecasting at the instant k

$\hat{y}_h$ Hour demand forecast

λ_{max} Largest eigenvalue

$\mho(x)$ Function of the interpolation technique

$\mathbf{q}_{AMR}(k)$ Vector of AMR measurements

$\mathcal{E}$ The set of edges

$\mathcal{G}$ The directed graph

$\mathcal{N}(0, \sigma^2)$ Normal distribution with μ the mean and σ^2 the variance

$\mathcal{P}_{ij}^{min}$ The minimum resistance of a path connect the node i and j

$\mathcal{S}$ The set of nodes in the sensor placement approach

$\mathcal{V}$ The set of vertices

$\nabla P^{max}, \nabla P^{min}$ Maximum and minimum value of the comparing between matrix $\boldsymbol{P}$

$\nabla_1 \overline{r_i}$ The finite difference of the residuals in of sensor i

∂_i The likelihood index of the leak localization method

ϕ Sensor faulty signals

$\Phi_{i,j}$ Spatial faulty signals

ρ Number of node selected to be a cluster centroid

σ_W^2 New variance with time window W at the instant k

σ_{st} Total number of shortest paths from node s to node t

$\tilde{d}_{WDN}$ Total inflow of water in the WDN

$\underline{\varepsilon}_{i,j}, \bar{\varepsilon}_{i,j}$ Upper and lower spatial residual bounds in the nodes i and j

$\underline{\varpi}, \bar{\varpi}$ Lower and Upper residuals bounds

A_s Adjacency matrix of the subgraph G

$C, \hat{C}$ Cluster, index cluster

$C, \hat{C}$ Cluster, index cluster

C_c Closeness centrality

C_{st} Betweenness centrality

$d(i,j)$ Distance between the nodes i and j

d_i Demand at node i

$D_{i,j}$ Diameter of the pipe

$e(k)$ Error at the instant k

$e_W(k)$ Fused error using the time window W at the instant k

$f(x)$ ATD function

f_i Sum of the of the flows that pass through the node i

G Directed graph

h_i, h_j Head of the nodes i or in node j

$l(k), \hat{l}(k)$ Leak size magnitude and leak estimation at the instant k

$L_{i,j}$ Length of the pipe

m Links in the network(pipes, valves, and pumps)

n Nodes in the network

N_d Number of historical inlet flow free-leak data

n_d Following leak estimations bigger than the threshold that is necessary to trigger the leak detection

n_I Number of the inlets

n_o Number of inner nodes in the network

n_s Number of sensors

$q_{i,j}$ Flow in the pipe between the nodes i and j

r Residuals

s Number of set of cluster C

T_s Sample time

W Time window

X Groups of nodes that contain in a cluster

x^u ,x^l Upper and lower bound of the integer optimization problem

$y(k)$ Input flow at the instant k

ACRONYMS

AMR Automatic Meter Reading. 22, 85

ATD Difference Time Detection. 32

BattLeDIM Battle of the Leakage Detection and Isolation Methods. 74

CNN Convolutional Neural Networks. 12

DMA District Metered Areas. 21

DTD Difference Time Detection. 32

FPR False Positive Rate. 13, 32

GA Genetic Algorithm. 45

GN Girvan-Newman. 42

NRW Non-Revenue Water. 1

PRVs Pressure Reducing Valves. 22

RBF Radial Basis Function interpolation. 31, 71, 117

TPR True Positive Rat. 32

WDN Water Distribution Networks. vi, xviii, xxviii, 1, 2, 15, 16, 18, 19, 21, 39–41, 45, 47, 48, 54, 56, 123, 124

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Water Distribution Networks (WDN) are essential infrastructures in modern cities for several socioeconomic reasons. They are complex networks due to their size (thousands of pipes) and hydraulic behavior due to their nonlinearity. One of the concerns to be managed in these systems is water leakage, which may account for up to 30% of the total amount of distributed water [74], being significant because water is a limited resource. Another concern is the scarcity of water that can occur in 2025, which may affect half the world's population that will not have access to safe and accessible water for their basic needs [32]. However, with all these risks, currently, this infrastructure does not perform satisfactorily in practice. According to Forum, [97] of Networks of Intelligent Water, it can have an estimated loss of water called Non-Revenue Water (NRW) of up to 70% in some cities.

Water loss can be divided into "real losses" and "apparent losses." Apparent losses are constituted by badly read measurements, data handling errors, and illegal water tapping. In contrast, the real losses comprise leakage from all system parts and overflow at storage tanks. Finally, real losses are composed of "background leakage" made up of small undetectable and detectable leaks relevant for detection as they represent significant losses for the water distribution com-

pany. The efficient and effective management of water distribution networks, which are in charge of conveying water to the final consumption points, is crucial for the development and sustainability of urban areas in modern society. The leak supervision process is a part of leak management that aims to monitor a physical system. It takes actions to keep it functioning in case of failures, which comprises leak detection, localization, and the study of sensor placement and validation to get optimal results.

The critical nature of the problems generated when a leak occurs implies the need for management capable of taking into account the detection of the leak as quickly as possible and their localization in the WDN. In addition, a study of the positioning of pressure sensors is essential, as the zone can affect the result for the location of the leak [92] and a constant sensor validation to guarantee that the sensors' measures are proper. The methods of leak supervision in WDN can be classified into Hardware-based and Software-based methods.

Hardware-based methods utilize hardware sensors to detect a leak directly and help the localization of the leak. As there are various types of sensors and instruments available, they can be further sub-classified as acoustic and non-acoustic detection methods.

Software-based methods generally rely on an algorithm or model for detecting leaks. Unlike Hardware-based methods, these methods do not seek to locate the leak point accurately but minimize possible leakage areas. Since these methods are based on information, such as the pressure of the pipe network, flow data, etc., they work well on any type of pipe. These methods can be divided into physical modeling methods and data-driven methods. The physical model-based methods identify the leak using a numerical model and compare the results with the field data. For designing a model-based, an appropriate mathematical model is required to represent the most relevant system dynamics.

On the other hand, data-driven methods analyze the monitoring data, combining tools such as artificial intelligence or artificial neural networks. Thus it is possible to identify potential areas of the leak based on specific rules or principles without resorting to the simulation of the physical model results. However, these methods need, in general, an actual number of non-leak and leak data scenarios in the training process to obtain good results. A hydraulic simulator can

be used to generate leak data as a specific amount of leak scenarios are not available in general.

The research in this thesis is motivated by the UTILITIES 4.0–P1 and L-BEST Project from the Spanish Ministry of Science and Innovation and AGAUR ACCIO RIS3CAT. This doctoral thesis is dedicated to investigating methods for leak supervision at WDN to identify and isolate "background leakage". To this end, the development of techniques for leak detection and location applying model-based and data-driven approaches are investigated. In addition, the study of sensor positioning and sensor validation to achieve the optimal results in leak supervision techniques.

1.2 State of the art

Various techniques for detecting, estimating, and isolating leakage in WDNs are already applied or under development. The following subsections introduce an overview of the leak detection and location methods. In the literature, leak detection and localization are summarized and classified from different perspectives. As we can see, the [74] delivered a comprehensive review of leakage management methods divided them into three categories: leakage assessment quantifying the amount of water lost, leakage detection that notices leakage hotspots, and control methods effective control of current and future leakage levels. Other reviews such as [20] give a comprehensive review of transient-based methods; the [57] classifies the methods as hardware-based and software-based; the [22] review displays that acoustic reflectometry is most suitable for the blockage detection techniques. Some reviews are focused just on leak detection techniques as of [1, 105, 17, 113].

This manuscript proposes a classification for leak detection and isolation classification into two sets: Hardware-based and Software-based. Figure 1.1 attempts to summarize the different leakage management and detection methods. The hardware-based technique is the approach that utilizes hardware sensors to detect the occurrence of a leak directly and help the localization of the leak. As various types of sensors and instruments are available, they can be further subclassified as acoustic and non-acoustic detection methods. Software-based systems use different software packages to monitor pressure, temperature, flow, or other pipeline parameters. They can be divided into the model base using numerical modeling methods that identify possible

Table 1.1: Summary of current reviews about leakage detection and location methods

References	Focus on	Classification
[20]	Transient-based leak detection methods	Transient leak detection Inverse-transient analysis Frequency domain techniques Direct transient analysis
[74]	Leakage management methods	Leakage assessment Leakage detection Leakage control
[57]	Burst/leakage detection and location	Hardware-based methods Software-based methods
[22]	Pipeline fault detection methods	Blockage detection techniques Leakage detection techniques
[1]	Pressure-based leakage detection method	Inverse-transient analysis method Transient steady state method Transient damping method Inverse resonance method Pressure-flow deviation method Negative pressure wave method Pressure residual vector method
[2]	Leakage detection and localization	Externally-based methods Internally-based methods Leak localization Utilization of wireless sensor networks
[105]	Data-driven approaches for burst detection	Classification method Prediction-classification method Statistical method
[17]	Leakage detection	Current technologies Intelligent methodologies
[113]	Steady-state leakage detection strategies	Hardware-based methods Software-based methods Hybrid leak detection techniques
[39]	Leak detection and location strategies	Data-driven Model-based

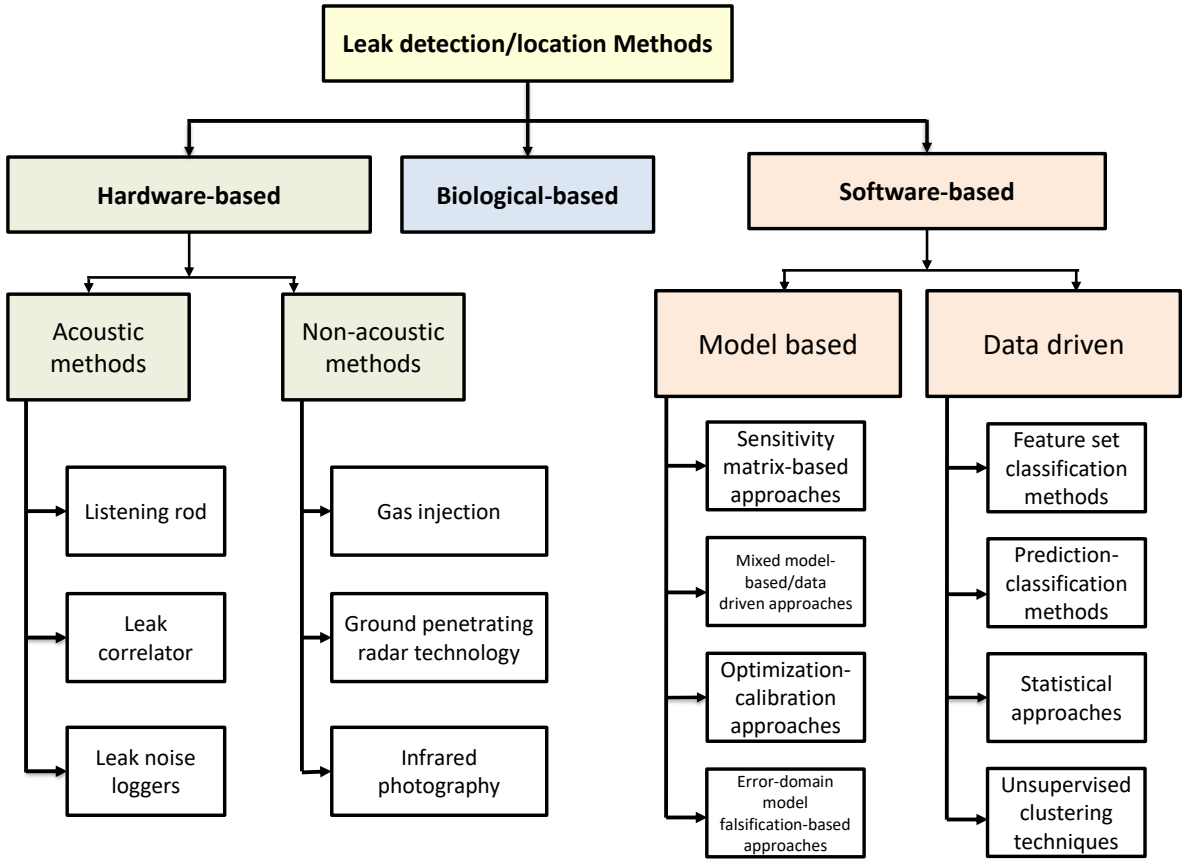


Figure 1.1: Categorization of leak detection and localization methods

leak areas, establishing a relevant numerical model to obtain relevant data to compare with simulation results. Moreover, the data-driven analysis monitoring data combined with tools such as data mining or artificial intelligence algorithms and then identifies possible leak areas based on specific rules or principles without resorting to the model simulation results.

1.2.1 Hardware-based methods

- **Acoustic detection methods**

These methods are based on the characteristics or principles of sound for detection. In general, acoustic detection methods work well to detect leaks from medium and large metal pipes. The current widely used methods are leak noise recorders, leak correlators, and listening bars.

– Listening rod

The listening rods or the listening sticks are basic acoustic instruments [37, 12, 35]. They have been used to detect and locate leaks since the mid-1960s and were called ground microphones. This device is placed on the floor, which amplifies the sound produced by a leak to facilitate detection, with the advantage that it is a cheap and useful tool. In contact with the surface of the detection equipment, the listening rod can detect a variety of faint sounds that human ears cannot hear due to air isolation or external interference when the equipment is in operation. This technique depends on the engineer's ability to hear the leak.

– Leak correlator

Leak noise recorders [86, 63] use acoustic receivers installed at different locations in the pipeline network, such as hydrants, valves, and other exposed tubes. Usually, they are switched on at predetermined times. Then, a particular software automatically registers all the recordings, and the leaks are identified with the difference of the recording with some noise recorders already predetermined. Each leak point produces a continuous leak sound, which determines the presence of leaks close to the recorder according to the degree of intensity and frequency. The computer software will automatically identify this and make a two-dimensional or three-dimensional map of the leak points.

– Leak noise loggers

Leak noise loggers [86, 63] use acoustic receivers installed in different locations in the pipeline network, such as hydrants, valves and other exposed pipes. Usually, they are switched on at predetermined times. Then, special software automatically registers all recordings. The leaks are differentiated as each leak point will produce a continuous leak sound, which determines the presence of leaks close to the recorder according to the degree of intensity and frequency that the recorder has noise records. The computer software will automatically identify this and make a two-dimensional or three-dimensional map of the leak points.

- **Non-acoustic detection methods**

The non-acoustic methods used for leak detection and location are able to detect small leaks and not suffer a loss of accuracy for the pipe material. Here the authors discuss mainly the gas injection method, ground penetration radar detection technology, and the method of thermal infrared imaging.

– Gas injection

This method uses a gas detector to detect the presence of a leak. A gas with lighter properties than air must be used. It is possible to use helium, but the most common is hydrogen due to a lower cost and high performance. The gas is placed inside the pipes, and in the presence of a leak they will escape to the outside. Hence the location of the leak in the pipe can be determined according to the location of the detected gas. Since the gas detector is very sensitive to a minimal amount of this gas, a tiny leak can be detected using this method [42]. However, this method is costly if it is necessary to consider filling the entire WDN and consuming much time to transport the detector over the network until it finds the leak. In addition, the technology has a significant disadvantage, as only the leak above the pipe can be detected using this method, as the gas cannot overflow from the bottom of the pipe.

– Ground penetrating radar

Ground-penetrating radar inspection (GPR) [41] is a non-destructive and non-invasive geophysical method that produces a continuous section or record of continuous underground resources. This technique uses an emitter of electromagnetic waves that are propagated through the ground and spread back to an electromagnetic detector, to infer the spatial location, structure, shape, and depth of the underground burial according to the electromagnetic wave received, amplitude and changes in intensity and time [26]. Although this technique is cheap, it needs specialized personnel and, even so, it is easy to obtain false positives (e.g. joints or valves that alter the image), and it cannot be applied in cold climates or saturated soil.

– **Thermal infrared imaging**

The principle of a thermal infrared (IR) camera [25] is based on thermal images; therefore, when a leak occurs, the temperature of the soil around the leak can change and consequently can be located. This method has several positive points such as high efficiency, accurate judgment, intuitive, safe, non-contact detection, immunity to electromagnetic interference, long-range detection, high speed, and is independent of the type and size of the pipe material. However, it also has some disadvantages, for example, it is affected by many factors, such as climatic conditions, soil conditions, and pavement surface.

1.2.2 Software-based

Software-based methods generally rely on an algorithm or model for detecting leaks. Unlike hardware-based methods, these methods do not seek to locate the leak point accurately but minimize possible leakage areas. Since these methods are based on information, such as the pressure of the pipe network, flow data, etc., they work well on any pipe. Software-based methods can be divided into model-based and data-driven.

- **Model-based**

A hydraulic model of the WDS is used to simulate its operating state in the model-based leak detection and location techniques. The hydraulic model, once built, needs to be calibrated to ensure it provides an accurate reflection of current operating conditions. Model-based approaches involve four key steps: (i) Construction of a hydraulic model, (ii) Calibration of the hydraulic model, (iii) Leak detection, (iv) Leak localization.

The two first steps are similar in all model-based approaches, while the other depends on the type of data used and the method development. Table 1.2 compares and summarizes several model-based leak detection techniques, with the method category, the reference the method adopted, the data request, and the technique type: LD to leak detection and LL to leak localization.

Table 1.2: Summary and comparison among model-based approaches

Category	References	Technique adopted	Data request	Type
Sensitivity matrix-based approaches	[72, 70, 71] [14, 16]	Pressure sensitivity matrix	Pressure	LD-LL
	[49, 31]	Flow sensitivity matrix	Flow	LD-LL
	[82]	Pressure/Flow sensitivity matrix	Pressure/Flow	LL
	[68]	Pressure/Flow sensitivity	Pressure/Flow	LD
Mixed model-based/data driven approaches	[114, 67]	Multiclass support vector machine	Pressure/flow	LL
	[90, 91]	K-nearest neighbors; Bayesian classifiers	Pressure	LL
	[109]	K-means clustering, linear classifier	Pressure	LL
	[116]	Fully linear DenseNet	Pressure	LL
	[38]	Multiscale neural networks	Pressure/Flow	LD
	[27]	Decision Tree, KNN, random forest, and Bayesian network	Pressure/flow	LD
Optimization-calibration approaches	[108, 65, 93]	Genetic algorithm	Pressure	LD-LL
	[83]	Least squares (LS), geographically allocated demand parameters	Demand	LD-LL
	[36]	Multi-objective and colony optimization	Pressure/demand	LD
Error-domain model falsification-based approaches	[33]	Error-domain model falsification	Demand	LD
	[85]	Time-series-based	Pressure/flow	LD
	[47]	Bayesian model updating approach; multilevel Markov chain Monte Carlo algorithm	Demand	LD

– (i) Construction of a hydraulic model

In this step, the hydraulic model of the WDN is generated. For this, software such as EPANET [81], LOOP [117], CADRE flow [13], Pipe flow expert [73], Synergi pipeline simulator [98], InfoWorks WS [44], WaterGEMS [104], and NextGen Simulation Suite[66] are used.

– (ii) Calibration of the hydraulic model

In this step, the hydraulic parameters of the model are set. Examples are the nodal head, water consumption, pipe length, pipe diameter, and pipe roughness coefficients. The most uncertain input variables in the simulation model are the pipe roughness and water consumption at a demand node because they are not typically directly measurable. Therefore, they require calibration.

– (iii) Leak detection and (iv) Leak localization

After the hydraulic model calibration, the supervision for leak monitoring in the WDN starts with detecting leaks in the system, and then the location technique is applied. Model-based

techniques may be classified as sensitivity matrix-based approaches, mixed model-based/data-driven approaches, optimization-calibration approaches, and error-domain model falsification methods.

The sensitivity matrix-based approaches are based on pressure and flow measurement. They use the interdependence of all the operating parameters to generate a sensitivity matrix to analyze the distribution of the network. The method is very efficient under ideal conditions using such a sensitivity matrix; nevertheless, uncertainties in nodal demands and measurement noise negatively affect the performance of these methods—the first time used was for leak location in the distribution network of Barcelona by [72]. Subsequently, similar studies have been conducted [70, 71, 16, 49, 82]. Finally, in [14, 31], the sensitivity matrix-based methodology is applied to leak detection.

Mixed model-based/data-driven approaches usually are applied in the leak localization problem. In [114, 90], the model is used to generate a sensitivity matrix, and the leak location is formulated like a classification problem. Researchers have used a variety of classification methods, such as Bayesian classification [91], linear classification [109] and deep learning identifies [116]. In [27] the leak detection was performed applying the hybrid method using AI algorithms and hydraulic relations developed and in [38], used the Multi-scale neural networks.

The optimization/calibration uses the observational values at the monitoring point to derive parameters for unknown leaks, such as the location and size of the leak. In [108] developed a method to leak detection using optimization problems to optimize the leakage node locations and their associated emitter coefficients such that the differences between the model predicted and the field observed values for pressure and flow are minimized and used the Genetic algorithm. In [93] developed an optimization problem to leak localization using the Genetic algorithm to solve it. In [83] developed an approach to compare the calibrated parameters with their historical values to assess if changes in these parameters are caused by a leak using the least-squares and geographically allocated demand parameters.

The error-domain model falsification-based approach uses the model parameters (leakage location and intensity) to calibrate, attempting to minimize the difference between predicted and

measured values by falsifying the model in the error domain. This approach does not require the error structure of the model predictions to be fully defined, as it is sufficient to falsify the model instances where the difference between the predicted and measured values exceeds the maximum plausible error. In [33], developed a technique to leak detection by falsification approach locates candidate leaks by falsifying improbable leak scenarios. Model falsification has become widely adopted as a leak detection method. Nonetheless, these methods also have a few weaknesses; for example, they are insensitive to small leaks [85], and they perform poorly if the WDN contains a relatively small number of sensors [47].

Recent work has created model-based methods combining multiple categories to improve the limitations of each type. For example, the work [95] developed a method of leak detection and localization that uses the combination of time series, cluster analysis, and least squares to calibrate the hydraulic model and study pressure derivations. This method detects multi-leak problems with high true-positive rates for the leak isolation index.

- **Data-driven**

Data-driven approaches are techniques that do not use simulated results from network models. They analyze the monitoring data combined with tools like data mining or artificial intelligence algorithms, using some rules to identify and locate the leak. These methods have a significant feature of not needing to establish a hydraulic model. Therefore, detailed piping and equipment information and parameters are not needed; they need to analyze large amounts of data and find the rule itself to identify system failures. These techniques typically utilize flow or pressure data but may also use end-user water demands for leak analysis. Table 1.3 compares and summarizes some of the data-driven leak detection and localization techniques, with the method category, the reference, the method adopted, the data request, and the technique type: LD to leak detection and LL to leak localization.

In most cases, a data-driven technique is performed in two steps: first, data acquisition, preprocessing, and transformation; the last two are selected according to leak detection method and data properties, and the primary purpose is to eliminate erroneous or missing data from time-series data to facilitate subsequent analysis. And the second step is a leak detection strategy

that can be divided into feature set classification methods, prediction-classification methods, statistical methods, and unsupervised clustering methods.

Feature set classification methods use the data information to distinguish the leaks from regular and fault operations. This approach needs considerable data to train the technique, which often becomes complicated. Several studies have been conducted to solve problems caused by a lack of training samples, e.g., [89] used Kriging spatial interpolation to obtain the pressure data of all nodes. Feature extraction is key to the success of feature set classification methods. To improve the accuracy of leak detection and localization, various methods have been used to extract leak features, including Convolutional Neural Networks (CNN) in [19, 52], linear prediction in [18, 96], discrete Fourier transform in [110] and Principal Component Analysis [115].

Prediction-classification methods consist of two stages: first, to predict the system's normal operation using historical data of normal operations to produce predictions, and second, to classify the data by comparing predictions and measurements. The prediction stage estimates the ideal WDS data under normal conditions. The use techniques such as artificial neural network [77], Kalman filter [111, 50, 53], support vector regression [62], evolutionary polynomial regression [56]), and long short-term memory [103]. Some researchers have also explored other approaches for leak detection; for example, prediction and classification using dynamic time warping [40], polynomial function based on weighted least squares [112].

Statistical approaches apply statistical analyses to the acquired data. The Statistical Process Controls is an essential technique for this method, which uses graphical analysis to set control limits to identify outliers measurements caused by a leak. The statistical approaches can predict leaks simply by applying statistical analyses on the acquired data [69, 51, 64]. Although statistical methods have shown promise for leak detection, they are deficient in certain aspects. For example, the method used by [59] fails to account for unexpected water demands or the demands of large users, whereas the method used by [76] quite time-consuming for leak detection.

Unsupervised clustering techniques rely on historical data of normal operations to cluster the WDN in groups with similar behavior. This method is called unsupervised because cluster

approaches do not need a priori information about the leakage scenarios. In this class of methods, the flow [106, 107] or pressure [30, 59] data are compared in terms of similarities via the clustering analysis.

1.2.3 Discussion

As seen in the section above, the hardware-based methods are great methods to use in the WDN to obtain accurate leak detection and location result. However, they are best suited for small DMA or directly in pipe studies. On the other hand, model-based and data-driven are more suitable for large-scale networks. When the system has a hydraulic model of the WDN, it is recommended to use methods based on these techniques. However, it is necessary to ensure that these hydraulic models are accurately calibrated to ensure that their pressure and flow predictions are realistic reflections of the operating states. When models are unavailable or inaccurate, data-driven methods are the most indicated. These techniques show promising applicability for leak detection in pipeline systems with ample monitoring sensors. Nevertheless, these techniques exhibit high False Positive Rate (FPR) due to uncertainties in the monitored data and their intrinsic limitations, which pose difficulties for effective decision-making.

The analysis of night-time flows for leak detection methods are applied because the flow during the night usually does not have a significant variation during the year, as seen in [17]; however, the leak can be detected just when the data is updated during the night. In addition, the leak detection methods focus on data-driven that use only the information of the flow sensor, usually detecting bursts or outline events with the burst magnitude of more than 5% $10\%$of the inflow average in the WDN as seen in the work [103, 69, 59]. Another concern that has been studied recently is the study of leak detection when a multi-leak is happening in the network [101]. This type of problem is complicated because multi-leaks affect the sensors installed in the WDN in different magnitudes.

On the other hand, the leak localization methods still have a large field of research focusing on the problem of multi-leak and data-driven approaches. For example, in model-based techniques, the work of [91] has a satisfactory result applying pressure residuals; however, it is a single leak method analysis. On the other hand, in the methods based on data-driven, the [89,

Table 1.3: Summary and comparison among data-driven approaches

Category	References	Technique adopted	Data request	Type
Feature set classification methods	[52]	One-dimensional convolutional neural network; supportvector machine	Wave velocity	LD-LL
	[115]	Kernel principal component analysis; cascade support vectordata description	Pressure	LD
	[89]	Kriging spatial interpolation; Bayesian reasoning	Pressure	LL
	[96]	Linear discriminant analysis; neural networks	Pressure	LL
	[18]	Linear prediction	Acoustic signals	LD-LL
	[110]	Discrete Fourier transform; isolation forest techniques	Pressure	LD
	[19]	Convolutional neural network; variational autoencoder	Acoustic signals	LD
Prediction-classification methods	[77]	Artificial neural network	Pressure/flow	LD-LL
	[61]	Artificial neural network and density network	Pressure/flow	LD
	[111]	Linear Kalman filter	Pressure/flow	LD
	[62]	Support vector regression	Pressure/flow	LD
	[8]	Adaptive forecasting model and deviation analysis	Demand/pressure	LD
	[112]	Polynomial function based on weighted least squareswith EM algorithm	Flow	LD
	[78]	Bayesian inference system	Pressure/flow	LD
	[43]	Probabilistic demand forecasting model	Demand	LD
	[50]	Nonlinear Kalman filter	Demand	LD
	[53]	Predictive Kalman filter	Wave velocity	LD-LL
	[56]	Evolutionary polynomial regression	Pressure/flow	LD
	[40]	Dynamic time warping; supervised learning	Demand/flow	LD
	[103]	Deep learning recurrent neural networks	Flow	LD
Statistical approaches	[69]	Principal component analysis; statistical process control	Flow	LD
	[51]	Statistical process control	Pressure/flow	LD
	[59]	Modified statistical process control	Flow	LD
	[76]	Principal component analysis; periodic transformation; vector extension	Flow	LD-LL
	[64]	Principal component analysis; standardized exponential weighted moving average (EWMA)	Pressure/flow	LD-LL
Unsupervised clustering techniques	[106]	Clustering	Flow	LD-LL
	[107]	Clustering; cosine distance	Flow	LD
	[30]	Feature-based clustering	Pressure	LD
	[59]	Time series data mining	Pressure transients	LD

96] pointed out the need for more studies on the interpolation methods.

Furthermore, the [91, 96] highlights the importance of a previous study on the sensor position and the sensor anomalies, communication issues, and noise often results in poverty data . Consequently, the data/sensors must be validated by different techniques.

1.3 Objectives

After discussing the techniques for monitoring water distribution networks and the gaps observed in the literature, three critical questions arise. First, how the leak detection can be improved to detect a leak magnitude of 1% to 5% of the average system inflow, and how to detect multi-leaks in the WDN? Second, how can the leak localization methods be improved to a multi-leak scenario by applying only the flow and pressure sensor operating in a network of more than one inlet? And last, how to improve the leak localization methods based on residual analyses and assure the sensors' good function?

To answer these questions, the main objective of this thesis is to develop methods focused on model-based and data-driven leak detection and localization using the information of flow, pressure sensors, and ATD equipment aiming to study incipient leaks. In addition, develop methods in sensor validation and optimal sensor placement to assist the leak localization methodology based on the residual analysis in achieving the optimal result. The specific thesis objectives are summarized as follows:

- **Objective 1** Propose leak detection and estimation methods able to detect incipient leaks and distinguish multi-leak scenarios in WDNs.
- **Objective 2** Develop a leak localization scheme that combines hydraulic models and leak detection analysis to tackle the multi-leak problem.
- **Objective 3** Propose new data-driven approaches that tackle the problem of leak localization at cluster and node level in WDN.
- **Objective 4** Develop a new data-driven approach that tackles the multi-leak problem.

- **Objective 5** Tackle practical problems of using pressure sensors: Sensor validation and optimal sensor placement.

1.4 Thesis Outline

Figure 1.2 shows the leakage supervision proposed in this thesis. It can be divided into two steps, (i) off-line, Figure 1.2 .(a), and (ii) online phase, Figure 1.2 .(b). The main purpose of the offline phase is to recreate the WDN through a hydraulic model or network representation, the alternative of installing pressure sensors or not in the WDN is usually a decision of the water company, and the calibration of the leak detection and localization methods. The online phase is focused on the supervision techniques to keep the system working properly; it is first necessary to check if the WDN leaks; if it does, it is necessary to locate it; if the network does not leak, it is possible to validate the pressure sensors to make the method of leak detection and localization more robust.

The techniques used in this doctorate are demonstrated in Chapter 2. Subsequently, a study of the positioning of pressure sensors can be carried out; this step is not mandatory for all WDNs, as some networks are already equipped with the sensors, and the water company does not want to make any changes to the network. Nevertheless, if the WDNs install or modify the positioning of the sensors, two models of sensor positioning is shown in Chapter 3, concerning Objective 5. In the online phase of the leak supervision (Figure 1.2.(b)), it is first arranged in the detection of failures in the system, as shown in Chapter 4 (Objective 1); if the system is operating under normal conditions, the sensor validation method can be applied to ensure that all sensors in the network still have a good performance index, a method presented in Chapter 3 for sensor validation (Objectives 2, 3 and 4). Finally, if the leak detection method detects a leak in the WDN, the leak localization technique will be applied; different leak localization methods in Chapter 5 are depicted, focusing on model-based and data-driven.

In addition to the abovementioned chapters, Chapter 6 concludes this document. Chapters 2-6 are summarized as follows:

Chapter 2: Background

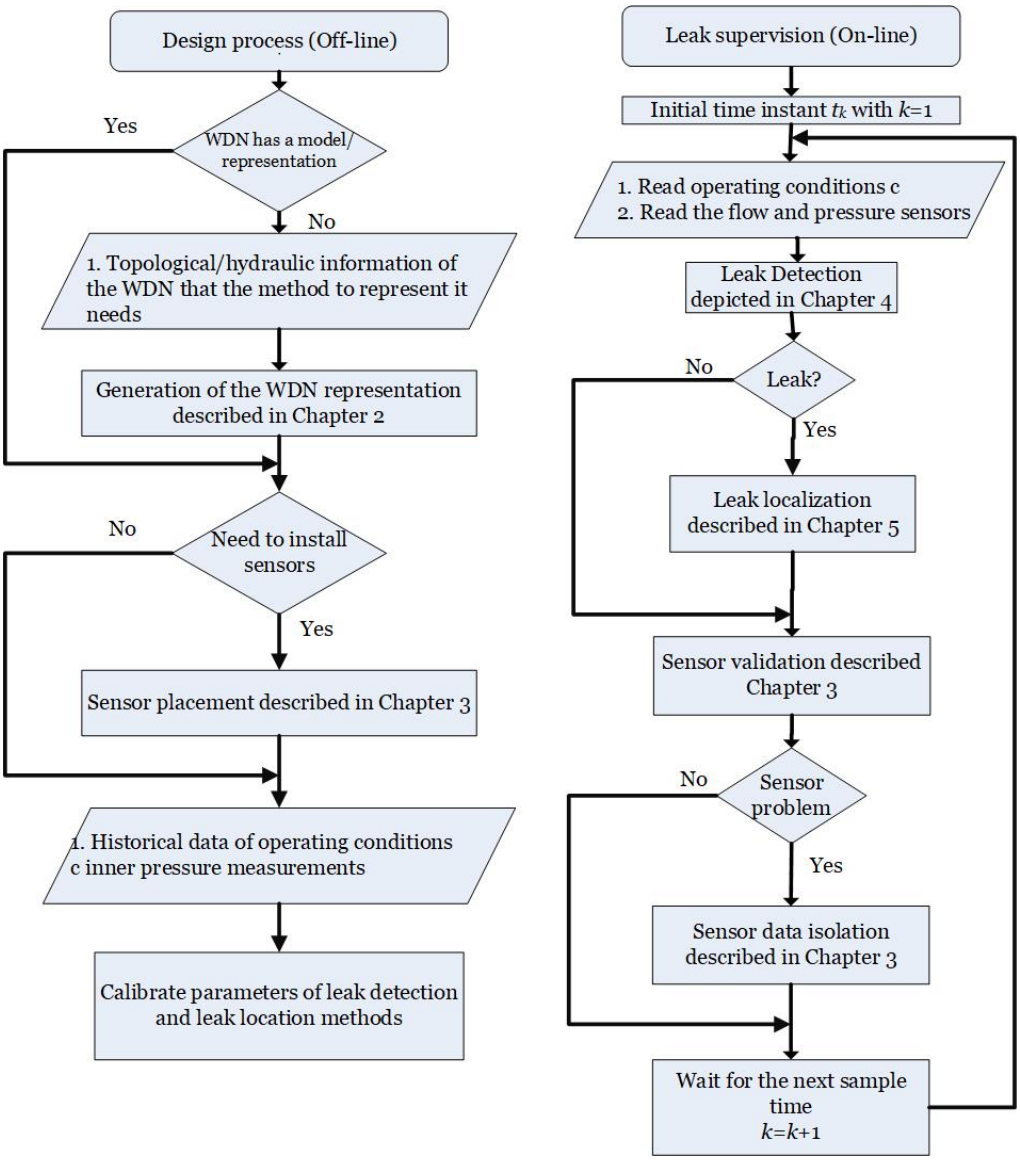


Figure 1.2: Flow chart of the leak supervision developed in this thesis

This chapter recalls a general definition of the water distribution network mode with the principal components of the WDN and the three forms to represent the network in this thesis. Following the introduction of the residual study to the leak localization technique. Then some explanations of strategies applied in this manuscript are introduced as the data fusion and interpolation techniques. Finally, the validation and evaluation indicators are shown.

Also, this chapter systematically presents the different case studies where the proposed methods have been applied. The network illustrated is the real data from the Barcelona DMA, the Hanoi network in Vietnam, Modena city in Italy, and the hypothetical city L-Town presented

for the first time in the global challenge BattLeDIM.

Chapter 3: Sensor Placement and Sensor Validation

This chapter proposes two methodologies to provide an optimum sensor deployment layout, one based on a model-based approach and the other entirely data-driven. The first method is formulated as an integer optimization problem, an optimization criterion that minimizes the average topological distance. The second method is a new methodology to provide an optimum sensor placement regarding how many sensors to install without using hydraulic information but exploiting the knowledge of the topology of the Water Distribution Networks. In addition, a pressure sensor validation method based on pressure residuals that allows the detection of sensor faults is proposed.

Chapter 4: Leak Detection

This chapter deals with the leak detection problem in WDNs. The first technique is a leak detection method based on the water demand analysis of District Metered Areas (DMAs). Then, historical leak-free data of water demand flow is used to extract minimum, maximum values, and statistical distributions of differences (errors) between demand flow and predicted values at different time hours of the day. The concept of sensor fusion is applied to reduce measurement uncertainties. For this, a virtual measurement is generated that considers each hour of the day as a feature and, combined, develops a more accurate error analysis capable of detecting leaks and estimating the leak size magnitude.

The second technique is a complementary study focused on the case with multi-leaks in the system. The leak detection approach uses the fusion data of the flow and pressure measurements, thus obtaining the instant where the leak starts and if there is more than one simultaneous leak (multi-leak) occurring in the network.

Chapter 5: Leak Localization

This chapter presents the leak localization approaches developed in this thesis. The methodology can be divided into two types: a model-based approach that uses the hydraulic model of

the network and pressure and demand information; data-driven approaches only require measurements from the network operation. The model-based model investigates the problem of multi-leaks using the automated meter reading that offers a better understanding of the consumption modes and can generate more accurate models.

In sequence, three data-driven strategies are introduced; the first develops a technique to reduce the area of the leak localization in the WDN, using Graph theory to represent the network, a distance clustering with pre-defined centroids that are the sensor pressure information, and some selected nodes. Furthermore, extra pressure information of leak events in the selected centroids is studied to develop a correlation between the pressure measurement and the event. The second proposed method approach is based on the use of inlet pressure and flow measurements, other pressure measurements available at some selected inner nodes of the WDN, and the topological information of the network. A reduced-order model structure calculates non-leak pressure estimations at sensed inner nodes. Residuals are generated using the comparison between these estimations and leak pressure measurements. In a leak scenario, it is possible to determine the relative incidence of a leak in a node by using the network topology, what it means to correlate the probable leaking nodes with the available residual information. Topological information and residual information can be integrated into a likelihood index used to determine the most probable leak node in the WDN at a given instant k or, through applying the Bayes rule, in a time horizon. The last method uses the fusion of the pressure residues by applying the radial base function (RBF) interpolation to obtain the network zone with the highest leak probability.

Chapter 6: Conclusion

This chapter addresses the conclusions of this PhD thesis, and highlights the main contributions made during its elaboration. Finally, some future works to be done in line of this PhD thesis is proposed.

1.4.1 Related Publications

Much of the work and results presented in this thesis have been published in scientific journals and conferences. A complete list of the publications is given in the following:

Journals papers

- [6] Alves, D., Blesa, J., Duviella, E.,and Rajaoarisoa, L. (2021). Robust data-driven leak localization in water distribution networks using pressure measurements and topological information. Sensors, 21(22), 7551.
- [80] Romero-Ben, L., Alves, D., Blesa, J., Cembrano, G., Puig, V., and Duviella, E. (2022) "Leak Localization in Water Distribution Networks Using Data-Driven and Model-Based Approaches". In: Journal of Water Resources Planning and Management 148.5 (2022), p. 04022016.
- Romero-Ben, L., Alves, D., Blesa, J., Cembrano, G., Puig, V., and Duviella, E. - Leak detection and localization in water distribution networks: review and perspective. (submitted)

Conference papers

- [4] Alves, D., Blesa, J., Duviella, E.,and Rajaoarisoa, L. (2022). Data-driven leak localization in WDN using pressure sensor and hydraulic information. Integrated Assessment Modelling for Environmental Systems - 2nd IAMES 2022.
- [5] Alves, D., Blesa, J., Duviella, E.,and Rajaoarisoa, L. (2022) Leak Detection in Water Distribution Networks Based on Water Demand Analysis . 11th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes - SAFEPROCESS.
- [3] Alves, D., Blesa, J., and Duviella, E(2020) "Detecçao de vazamento de distribuiçao de água". In: Conferência de Estudos em Engenharia Elétrica (CEEL) . DOI: 10.5281/2596-2221.xviiiceel.2020.548.
- Alves, D., Blesa, J., Duviella, E.,and Rajaoarisoa, L. (2022) Multi-leak detection and isolation in water distribution networks.- 2nd WDSA/CCWI Joint Conference. (to appear)
- Alves, D., Blesa, J., Duviella, E., Rajaoarisoa,and L. (2022). Topological analysis of water distribution networks for optimal leak localization. International Conference on Hydroinformatics HIC 2022 − Water INFLUENCE. (to appear)

CHAPTER 2

BACKGROUND

2.1 Introduction

This section introduces some necessary definitions, mathematical tools, and properties, which will be used in this thesis. First, the principal components and the techniques to represent the WDNs are presented, followed by the introduction of the residual pressure study used in Chapter 3, 4,5. Next, some explanation of the concept of the data fusion technique applied in Chapter 4 subsequent two techniques of interpolation that is used in Chapter 5 . Following the evaluation indicators used to evaluate the methods, finalizing with the WDN network used in the case studies used in Chapter 3, 4,5 .

2.2 Water Distribution Networks

WDNs are large systems present in cities around the world. This system is usually organized in District Metered Areas (DMAs) to facilitate the management of the water companies. The principal components of the network are:

- Inlets, also called reservoirs: feed the network with water. The networks can have one or several inlets.

- Pressure Reducing Valves (PRVs): ensure that an appropriate and comfortable water usage pressure is available at all times.
- Pipes: distribute the water across the network
- Node: can be junctions between pipes or points where the consumer users are connected with the network.
- Sensors: different sensor types can be installed in the WDN; in this thesis, two types were used (I) Flow sensors are usually installed in the reservoir for control and billing purposes (II) pressure sensors are installed around the network to measure the pressure information.
- Automatic meter reading (AMR): is a remote reading device that attaches to a existing water meter. This device allows the hydraulic models to be more accurate as it can inform the consumption demand of the node in which it is installed.

Throughout this Ph.D., the methodology developed is divided into model-based and data-driven, and the way the network is represented is different in each approach. The following subsection will introduce the WDN representation, starting with the hydraulic model used in the model-based methods; the following will present the Graph theory used in the leak localization methods based on data-driven, and the reduced-order model applied to calculate the pressure estimation in the leak detection method.

2.2.1 Hydraulic model

A WDN can be modeled using the non-differential Hazen-Williams equations as a static system, considering that changes in demands and flows are slow enough to consider the system operating in a steady-state. The relation of the flow in the pipes can be described as follow:

$$q_{i,j} = (h_i - h_j)|h_i - h_j|^{\frac{1}{a}-1} \left(\frac{10.7 L_{i,j}}{C_{i,j}^{a} D_{i,j}^{4.87}} \right)^{1/a} \tag{2.1}$$

where $q_{i,j}$ is the flow in the pipe between the nodes i and j $[m^3/h]$, the h_i and h_j are the head of the nodes i and j respectively in $[m]$, $L_{i,j}$ is the length of the pipe and $D_{i,j}$ is the diameter

of the pipe, both in meters $[m]$,C_j is the pipe roughness coefficient and a is the flow exponent coefficient, which is equal to 1.852. In addition, a flow balance can be establish in the node in the conservation law:

$$f_i - d_i = 0 \tag{2.2}$$

where d_i is the demand at node i and f_i is the sum of the flows that pass through the node i, both in $[m^3/s]$.Typically, demands on each node are estimated using billing records. These records are used to calculate the average daily consumption of each node concerning global consumption. Then, the demands on the nodes are considered that has the same pattern distribution, i.e., each node always has the same proportion of the total inflow water in the WDN, so the demand on each node is calculated as:

$$\hat{d}_i = \alpha_i \tilde{d}_{WDN} \tag{2.3}$$

where the $\hat{d}_i$ is the estimated demand at node i, $\tilde{d}_{WDN}$ is the total inflow of water in the WDN, and α_i is the normalized proportional outflow in node i with the sum of all α equal to one.

With the knowledge of the demand pattern distribution and the hydraulic characteristics (i.e., pipe shape, length, etc.), and the measurement of the boundary conditions (Pressure Reducing Valves (PRVs), total water consumption, etc.), the WDN hydraulic system can be simulated using a hydraulic simulator that provides a numerical solution.

A leak in the WDN can be modeled considering an extra demand in the network, and in this Ph.D. thesis, for simplicity without loss of generality, it is considered that the leaks can occur only in the nodes of the WDN. To consider the extra demand of a leak in node j (l_j), a new pattern distribution will be generated as

$$\alpha_i^{l_j} = \begin{cases} \alpha_i - \alpha_i \frac{l_j}{\tilde{d}_{WDN}} & i \neq j \\ \alpha_i - \alpha_i \frac{l_j}{\tilde{d}_{WDN}} + \frac{l_j}{\tilde{d}_{WDN}} & i = j \end{cases} \tag{2.4}$$

where l_j is the leak (with leak size magnitude l in $[l/s]$) at the node j, and α^{l_j} is the new demand

pattern distribution with a simulated leak, with the condition to the sum of all α equal to one.

For the research to be more robust, uncertainties on demand, α, and leak magnitude,l, can be added. These uncertainties are artificially generated using the methodology presented in [21]. To calculate the demand uncertainties, the pattern is modified as:

$$\bar{\alpha_i} = \alpha_i + \frac{\alpha_i \alpha^u rand}{100} \tag{2.5}$$

where the $\bar{\alpha_i}$ is the new demand with uncertainties in node i, α^u is the amplitude of the demand uncertainty [%], and $rand$ is a random value uniformly distributed in a range $[-0.5, 0.5]$. To satisfy the condition of all sums of alpha equal to one, the following normalization is done:

$$\bar{\alpha_i} = \frac{\bar{\alpha_i}}{\sum_{j=1}^{n} \bar{\alpha_j}} \tag{2.6}$$

Then, the new generated demands with uncertainty can be computed as

$$\bar{d_i} = \bar{\alpha_i} \tilde{d}_{WDN} \tag{2.7}$$

where $\bar{d_i}$ is the demand at node i considering the uncertainty. Finally, to generate the uncertainties in the leak magnitude, the following equation is used

$$\bar{l_i} = l_i + \frac{l_i l^u rand}{100} \tag{2.8}$$

where the $\bar{l_i}$ is the new leak magnitude with uncertainties in node i, l^u is the amplitude of the leak magnitude uncertainty [%], and $rand$ is a random value uniformly distributed in a range $[-0.5, 0.5]$. As seen in the Chapter 1.2, there exist different software programs to simulate the hydraulic model; in this thesis, the hydraulic simulator EPANET [81] is used.

2.2.2 Graph theory

When using data-driven approaches, most often Graph theory is used to represent the network. In this case, WDNs is composed of m pipes, n internal consumer nodes and can be described by a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, [23], with $\mathcal{V} = \{v_1, \ldots, v_n\}$ is the set of vertices that represents connections between the components of the network, additionally the last $\{v_{n-n_I+1}, ..., v_n\}$, represent the vertices of the system's input, being n_I the number of the inlets, with $n_I \geq 1$. The elements of the set $\mathcal{E} = \{e_1, \ldots, e_m\}$ are the edges, which represent the m pipes in the network.

The graph $\mathcal{G}$, can be represented by the incidence matrix $\mathbf{H} = [h_{ij}]$, in which the elements h_{ij} with $i = 1, .., n$ and $j = 1, .., m$ are defined as:

$$h_{ij} = \begin{cases} -1 & \text{if the } j^{th} \text{ edge is entering } i^{th} \text{ vertex.} \\ 0 & \text{if the } j^{th} \text{ edge is not connected to} \\ & \text{the } i^{th} \text{ vertex.} \\ 1 & \text{if the } j^{th} \text{ edge is leaving } i^{th} \text{ vertex.} \end{cases}$$

The direction of the edge represents a reference direction for the flow in the corresponding pipe. The H is the incidence matrix of a connected and directed graph [46]. The incidence matrix is composed of $H \in \{-1, 0, 1\}^{n \times m}$ with each row corresponding to a node and column corresponding to a pipe.

The WDN must fulfill mass conservation law, which expresses the conservation of mass in each vertex, described by:

$$\mathbf{H} \cdot \mathbf{q} = \mathbf{d} \tag{2.9}$$

where $\mathbf{d} \in \mathbb{R}^n$ is the vector of nodal demands, with $d_i > 0$ when the flow is into the node i, and $\mathbf{q} \in \mathbb{R}^m$ is the vector of flows in the edges. By virtue of the mass conservation, it is possible to have only $n - 1$ independent nodal demand, $\sum_{i=1}^{n} d_i = 0$, therefore the supply flow must equal the end-user demands as there is no storage in the network.

Let $\mathbf{p}$ be the vector of absolute pressures at the nodes and $\Delta\mathbf{p}$ be the vector of differential

pressures across the pipes, both in meters of water column [mwc], then the energy law for water networks gives:

$$\Delta \mathbf{p} = \mathbf{H}^T \mathbf{p} = f(\mathbf{q}) - \mathbf{H}^T \mathbf{z} \tag{2.10}$$

where $\mathbf{p} \in \mathbb{R}^n$, and $f : \mathbb{R}^m \to \mathbb{R}^m$, $f(\mathbf{q}) = (f_1(q_1), \ldots, f_m(q_m))$. The function $f_j(\cdot)$ describes the flow dependent pressure drop due to the hydraulic resistance in the j^{th} edge. The relationship between pipe flow and energy loss caused by friction in individual pipes can be computed using Hazen-Williams formula to expression $f_j(\cdot)$:

$$f_j(q_j) = \frac{10.7 \cdot L_j}{C_j^a \cdot D_j^{4.87}} \cdot q_j |q_j|^{a-1} \tag{2.11}$$

where L_j is the length of the pipe and D_j is the diameter of the pipe, both in meters [m], q_j is the pipe flow in m^3/s and ρ_j is the pipe roughness coefficient.

The term $\mathbf{H}^T \mathbf{z}$ is the pressure drop across the pipes due to difference in geodesic level (i.e. elevation) in meters [m] between the ends of the pipes with $\mathbf{z} \in \mathbb{R}^n$ the vector of geodesic levels at each vertex.

2.2.3 Reduced-order network model

In the leak diagnosis techniques, the representation of the node position is not required; however, a good representation of the nominal pressure at all the sensors is necessary. The reduced-order network model is used in this thesis to calculate the nominal pressure at the measured internal nodes in the Chapter 4. The model uses the pressure dependence of the network's internal nodes with the pressure and flow measurements of the inlets. The details of the model derivation can be found in [48] and [46].

A network can be divided into nodes connected with reservoirs (the inlets nodes) and internal nodes that compose the system. To facilitate the explanation in this work, the information re-

garding inlet nodes will be represented by (r) superscript and those of the internal nodes, which will be expressed by (i_n) superscript. In particular, vector $\mathbf{p}^{(in)}$ will contain pressure node values $p_1, ..., p_{n-n_I}$ and $\mathbf{p}^{(r)}$ inlet pressure values $p_{n+1-n_I}, ..., p_n$. The network requires to fulfill some conditions for using the reduced model proposed:

Condition 1: corresponds to the demands of the internal nodes of the system, where equation (2.9) can be defined as:

$$\mathbf{d}(k) = -\alpha(k)\sigma(k) \tag{2.12}$$

where $\sigma(k)$ denotes the total inlet flow into the network at time instant k, the vector $\alpha(k)$ defines the distribution of the total demand in the internal nodes at every time k, with the property $\sum_i^n \alpha_i(k) = 1$. Notice that if all consumers are residential, and all nodes demand have the same consumption profile, in consequence, the $\alpha(k)$ will be constant $\alpha(k) = \alpha$.

Condition 2: is a particularly case when the vector $\mathbf{p^{(r)}}$ of control inputs fulfill the following case:

$$\mathbf{p}^{(r)}(k) + \mathbf{z}^{(r)} = \kappa(k)\mathbb{1} \tag{2.13}$$

for some $\kappa \in \mathbb{R}$ which is the total head at the inlets in [mwc] and where $\mathbb{1}$ denote the vector consisting of ones. In [46], there is a discussion on this definition's feasibility where the controllers should satisfy this premise at least in networks with the low total consumption.

If these two conditions are fulfilled, the pressure at the i^{th} internal node can be expressed by:

$$p_i^{(i_n)}(k) = \alpha_i\sigma^2(k) + \sum_{j=1}^{n_I} \beta_{ij}(k)p_j^{(r)}(k) \tag{2.14}$$

where α_i is parameter dependent on the network topology and the distribution of demands in the

network, and β_{ij} is dependent on the network topology with $j = 1, ..., n_I$. The total inlet flows σ is typically well-known since inlet flows are measured.

Some methods of identifying parameters can be used to identify parameters α_i and β_{ij} since model (2.14) of $p_i^{(i_n)}$ is linear [87], using the measures of σ, $\mathbf{p}^{(r)}$, and $p_i^{(i_n)}$ with nodes that contain pressure sensors that will be denoted as $p_i \, \forall i = 1, ..., n_s$ in the following where n_s is the number of sensors installed in inner nodes.

If the conditions are not fulfilled, it is possible to generate an estimation of the pressure in inner nodes $\hat{p}_i(c)$ using the historical data directly as a lookup table as was proposed in [89]. That means, given the particular operating conditions, c provides the inner pressures from historical data that had the closest operating conditions $\hat{c}$ to c.

2.3 Residual study

A study widely used for the location of leaks is the study of residuals, i.e., differences between estimations provided by the model and measurements provided by sensors installed indicative of leaks, some works are [72, 75, 15, 11]. An important factor in the system in the presence of a leak will be that it will affect the measurement of all pressure sensors, but the sensors closer to the failure will be affected more [48, 79].

This section presents a basic leak location methodology based on residuals of pressure measurements to isolate leaks in a water distribution network. Usually, the methods are triggered when a leak is detected. Leak detection is usually done using inlet flow analysis. Inlet flow and pressure sensors in inner nodes are installed in the WDN. Leak localization can be carried out by employing the analysis of pressure residuals generated by the comparison of inner pressure measurements and leak-free pressure estimations as

$$r_i = \hat{p}_i(c) - p_i(c), \quad \forall i = 1, \ldots, n_s, \tag{2.15}$$

where $r_i, \hat{p}_i$ and p_i are the residual, leak-free pressure estimation, and pressure measurement at

inner node i. c is the operating condition defined by inlet measurements, and n_s is the number of inner sensors installed in the WDN. Physical models or historical data can compute leak-free pressure estimations. If a physical model of the WDN is available, a leak sensitivity matrix $\mathbf{\Omega}$ can be computed as

$$\mathbf{\Omega} = \begin{bmatrix} \frac{\mathrm{d}r_1}{\mathrm{d}f_1} & \cdots & \frac{\mathrm{d}r_1}{\mathrm{d}f_n} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}r_{n_s}}{\mathrm{d}f_1} & \cdots & \frac{\mathrm{d}r_{n_s}}{\mathrm{d}f_n} \end{bmatrix} \tag{2.16}$$

where $\frac{\mathrm{d}r_1}{\mathrm{d}f_1} = \hat{p}_i - \hat{p}_i^j$, $\hat{p}_i^j$ with is the pressure in node i considering a leak in node j denoted as f_i. Then, leak localization can be formulated as the maximum correlation between the observed residuals and the different leak hypothesis

$$\arg \max_{j \in 1,\ldots,n_s} \frac{\omega_j r}{||\omega_j|| ||r||} \tag{2.17}$$

where ω_j is the j^{th} column of a sensitivity matrix (2.16) and r is the residual whose components are computed in equation (2.15). Alternatively, if it is not available any hydraulic model of the WDN the leak localization method can be formulated as the maximum residual component

$$\arg \max_{j \in 1,\ldots,n_s} r_i \tag{2.18}$$

The main disadvantage in the use of leak localization in equation (2.18) compared with leak localization in equation (2.17) is that the result of the leak localization is not a node, but a cluster related to one of the n_s inner pressure sensors. But as it is very simple, and it does not require any physical model, it is a good reference point to develop new data-driven leak localization methods.

2.4 Data fusion technique

The methods presented in the leak detection in Chapter 4 use the base of the data fusion. Integrating data and knowledge from several sources is known as data fusion. Briefly, data fusion can be defined as combining multiple sources to obtain improved information; in this context, improved information means less expensive, higher quality, or more relevant information.

The equations methodology used in this thesis is found in the work [99]. Figure 2.1 shows the impact of the fusion technique on the sensor's variance, highlighting that the fused estimate's variance is smaller than the variances of the individual sensor measurement.

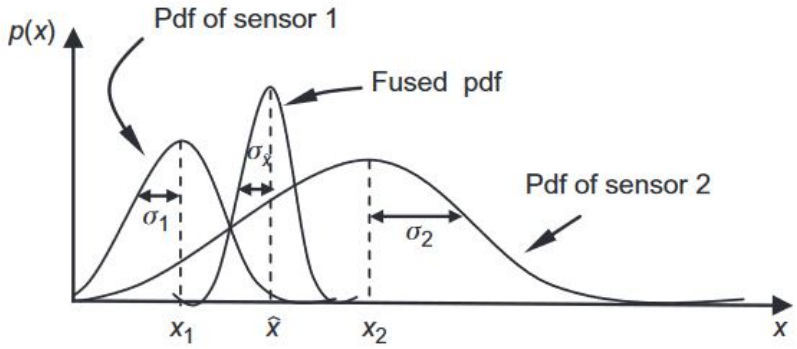


Figure 2.1: The variance $\sigma_{\hat{x}}$ of the fused estimate $\hat{x}$ is smaller than both σ_1 and σ_2

2.5 Interpolation techniques

Interpolation is the process of using known data values to estimate unknown data values. This thesis uses interpolation in leak localization based on data-driven approaches. The interpolation technique predicts the value of a function, $\mho$, at a given point, x by computing a relation to the known values of the function, $\underline{x_j}$ where j is the node that has sensors, in the neighborhood of the point. The following section will introduce the Kriging and radial basis function interpolations.

Kriging

Kriging interpolation method is a multivariate regression approach well-known in the area of geostatistics [54]. Universal kriging predictor [58] is given by

$$\mho(x) = w(x)\boldsymbol{\beta} + g(x)\boldsymbol{\psi} \tag{2.19}$$

where $w(x)$ defines the regression model as a function of spatial location x, $g(x)$ considers spatial correlations between sampling locations $\underline{x_j}$ and spatial location x. And $\boldsymbol{\beta}, \boldsymbol{\psi}$ are vectors calibrated with the sampling observations.

Radial basis function interpolation

Radial basis function interpolation (RBF) provides a very general and flexible way of interpolation in multidimensional spaces, even for unstructured data, where it is often impossible to apply polynomial or spline interpolation. Due to its good approximation properties, it was chosen in this work.

The method usually works in s dimensional Euclidean space which is $\mathbb{R}^s$ fitted with the Euclidean norm $||\cdot||$. The interpolation space consists of all functions of the form:

$$\mho(x) = \sum_{j=1}^{n_s} \lambda_j \phi(||x - \underline{x_j}||) \tag{2.20}$$

where x is a point in $\mathbb{R}^s$ which is the value to be obtained with the interpolation, $\underline{x_j}$ are the center points for the RBFs, λ_j are coefficients to determine, n_s are points in this space at which the function to be approximated is known, in this thesis is the value in the node that has sensors, and $\phi(r)$ is a radial basis function, set as a multiquadric problem:

$$\phi(r) = \sqrt{1 + \epsilon^2 r^2} \tag{2.21}$$

2.6 Validation and Evaluation Indicators

In this section, different indicators will be introduced to evaluate the performance of the methods presented, focused on the problem of leak detection and localization and the sensor placement and validation.

2.6.1 Leak detection indicators

To assess whether a fault detection method is working correctly, different scenarios should be generated with different magnitude values of leaks comprised of a few days with the system in normal operating mode and then a few days in a leak operating mode. The first performance index to be evaluated is the True Positive Rate (TPR), which tells us the percentage of correctly identified leaks as such. The second index is the False Positive Rate (FPR), which is the percentage of leak-free data that triggered the leak detection method, i.e., no leak is present in the network when the detection is raised of the total number of sequences analyzed. Finally, the Difference Time Detection (DTD) index measures the time in hours from the leak appearance to the leak detection.

2.6.2 Leak localization indicators

The Average Topological Distance (ATD) index is applied in evaluating leak localization methodologies. The index represents the node's distance between the node predicted as leaking and the actual node with the leak. To calculate the ATD is first necessary to create a matrix containing the minimum topological distance (in nodes or meters), $\mathbf{A} \in \mathbb{R}^{n \times n}$.

Second is necessary to calculate the confusion matrix, confusion that is a tabular way of visualizing the performance of a prediction model. Each entry in a confusion matrix denotes the number of predictions made by the model where it classified the classes correctly or incorrectly. For our study, only the confusion matrix analysis would be inclusive since the objective is the analysis of the proximity that the method points to the leakage and not its precision to the node with the leakage. The confusion matrix $\Gamma_{i,j}(n - n_I \times n - n_I)$ depicted in Table 2.1 that is used to assess the performance of leak localization approaches. The rows of this matrix correspond

to the leak scenario and the columns to where the leak is located ($\hat{l}$) by the leak localization method. Considering confusion matrix $\mathbf{\Gamma}$ the ATD can be computed as follows:

Table 2.1: Confusion matrix $\mathbf{\Gamma}$.

	$\hat{l}_1$	$\cdots$	$\hat{l}_i$	$\cdots$	$\hat{l}_{n-n_I}$
l_1	$\Gamma_{1,1}$	$\cdots$	$\Gamma_{1,i}$	$\cdots$	$\Gamma_{1,n-n_I}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
l_i	$\Gamma_{i,1}$	$\cdots$	$\Gamma_{i,i}$	$\cdots$	$\Gamma_{i,n-n_I}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
l_{n-n_I}	$\Gamma_{n-n_I,1}$	$\cdots$	$\Gamma_{n-n_I,i}$	$\cdots$	$\Gamma_{n-n_I,n-n_I}$

$$ATD = \frac{\sum_{i=1}^{n-n_I}\sum_{j=1}^{n-n_I}\Gamma_{i,j}A_{i,j}}{\sum_{i=1}^{n-n_I}\sum_{j=1}^{n-n_I}\Gamma_{i,j}} \tag{2.22}$$

2.6.3 Sensor placement indicators

The indicators applied to evaluate the optimal sensor placement techniques are:

- F1-score is the weighted average of precision and recall, where precision is the analysis of all positive predictions, how many are positive, and recall is the study of real positive cases, how many are predicted positives. The F1 score is a good indicator of imbalanced data [55].
- Cohen's Kappa: represents the degree of accuracy and reliability, which is the difference between the observed overall accuracy of the model and the overall accuracy obtained by chance. It is a more practical measure to use on problems with an imbalance in the classes. The kappa has a range from −1 to +1. Values superior to 0.6 are considered good [79].

2.7 Case Studies

This section presents the case studies used along with the Ph.D. thesis. In particular, the WDN and DMA networks are detailed.

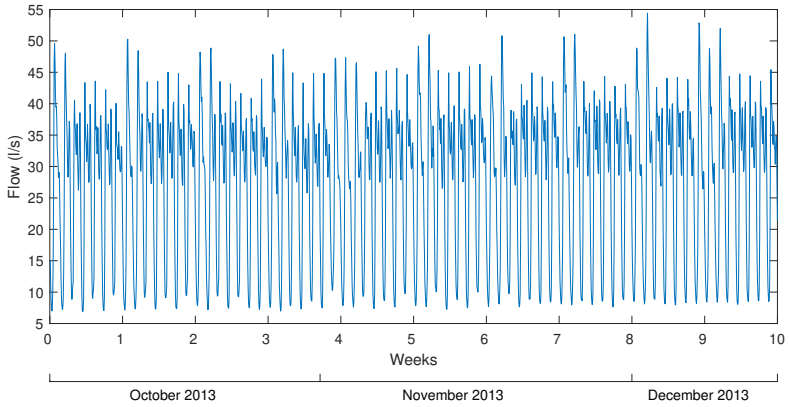


Figure 2.2: Historical leak-free inlet flow in Barcelona DMA1.

First, a DMA of the Barcelona WDN with only flow measurement at the inlet is presented. The data was used to test the leak detection method proposed in Chapter 4. Then, the reduced Hanoi benchmark WDN is presented, and the case is used for leak localization in Chapters 5. Next, the Modena network is introduced and applied in the leak localization, sensor validation, and sensor placement in Chapters 3 and 5. Finally, the hypothetical L-Town WDN is raised and employed to evaluate the proposed leak detection and localization techniques presented in Chapters 4 and 5.

2.7.1 DMA Barcelona WDN

Barcelona WDN supply water in 23 municipipalities that correspond to around three million consumers. This network convers and area of 424 km^2 and it is divided in more than 200 DMAs. In this Thesis a DMA of the Barcelona WDN has been used as a case study in the framework of the project ACCIO RIS3CAT UTILITIES 4.0–P1 ACTIV 4.0. ref. COMRDI-16-1-0054-03. This DMA, that will be refered as BCN DMA1 , is an extensive network formed by two reservoirs without PRVs, 3373 consumer nodes, and 3482 pipes. The data available is the total water consumption from October to December 2013. And the water consumption varies from approximately 8 to $50[l/s]$ shown in Figure 2.2.

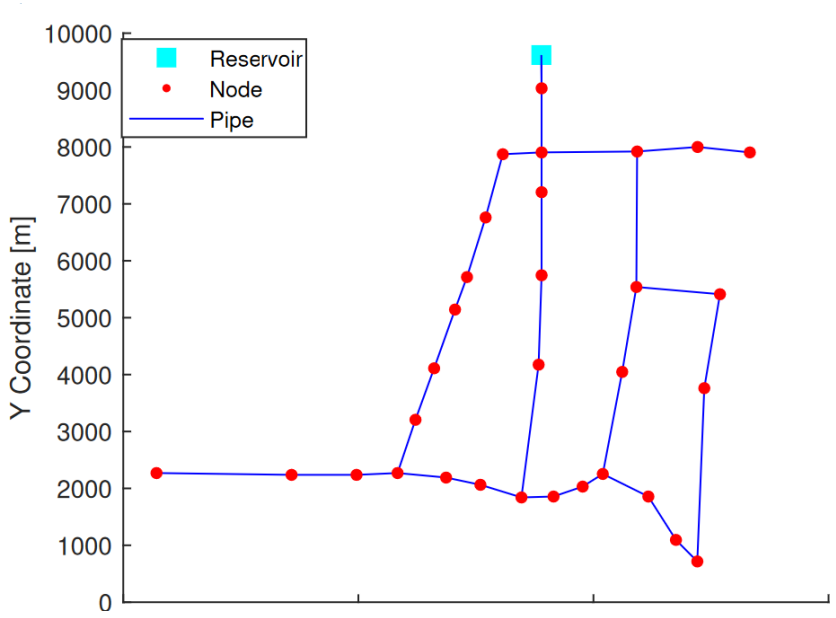


Figure 2.3: Simplified Hanoi WDN.

2.7.2 Hanoi

Hanoi's WDN is a reduced city's network model, from (Vietnam). The network is composed of one inlet (reservoir), 34 pipes, 31 nodes, no pumping facilities are considered since only a single fixed head source at elevation of $100m$ is available. The WDN is represented by figure 2.3. The Hazen Williams coefficient for all the links is 130. The diameter of the pipes in this network are between $305mm$ and $1,016mm$. The minimum flow requirement in this network is $5m^3/h$. The network was introduced by [29] and widely used to test optimization approaches.

2.7.3 Modena

Modena WDN is a reduced model of the real water distribution network of the Italian city Modena. The gravity-fed large-scale network shown in Figure 2.4. This large-scale network comprises 268 junctions (nodes) connected through 317 pipes and served by four reservoirs.The network is completely gravity-fed [102], therefore, it has no pumps.

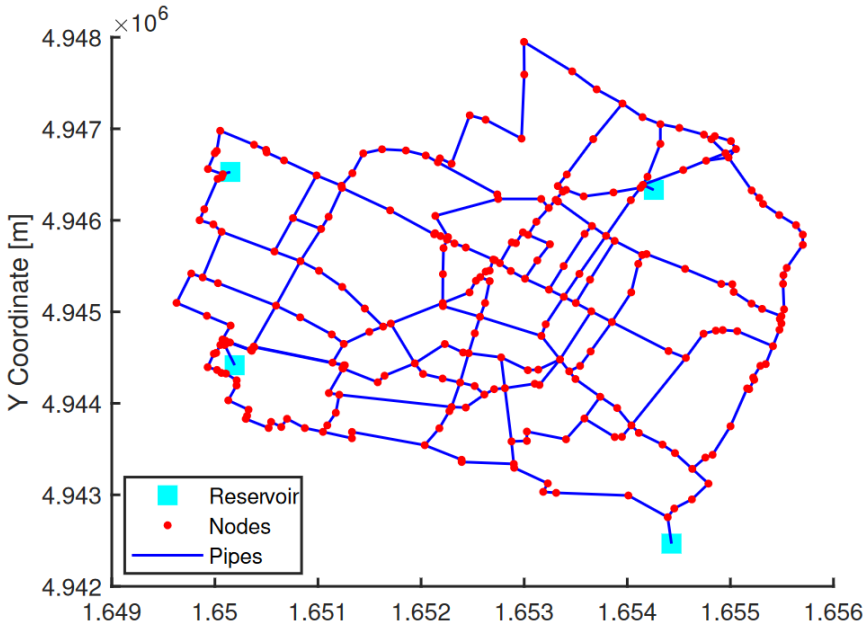


Figure 2.4: Modena WDN

2.7.4 L-Town

The Battle of the Leakage Detection and Isolation Methods is a challenge provided by the organizers of the BattLeDIM [101]. The aim is to detect and locate several leaks in a hypothetical city created with this intent, as depicted in Figure 2.5. The city is located in the Northern hemisphere and regroups a population of about 10,000 people. Thus, higher water usage is expected around July/August and lower in December/January. The network is divided into three distinct areas: Area A is supplied by two reservoirs, each containing flow sensors; Area B was installed with a pressure reduction valve (PRV) to help reduce background leakages; and Area C was installed with a pump and a water tank, with a flow sensor in this pump to control the flow that enters in the tank. The network has 33 pressure sensors (see Figure 2.5), all transmitting their measurements every 5 minutes to the utility's Supervisory Control and Data Acquisition (SCADA) System. In addition, it has been installed in Area C 82 Automated Metered Readings (AMRs), a technology used in utility meters for collecting data that does not require physical access or visual inspection (see Figure 2.6 . In this area, only ten regular sensors were distributed, and because Area C has a significant quantity of AMRS installed in the zone, a model-based approach is a good option to solve the leak localization problem in this area.

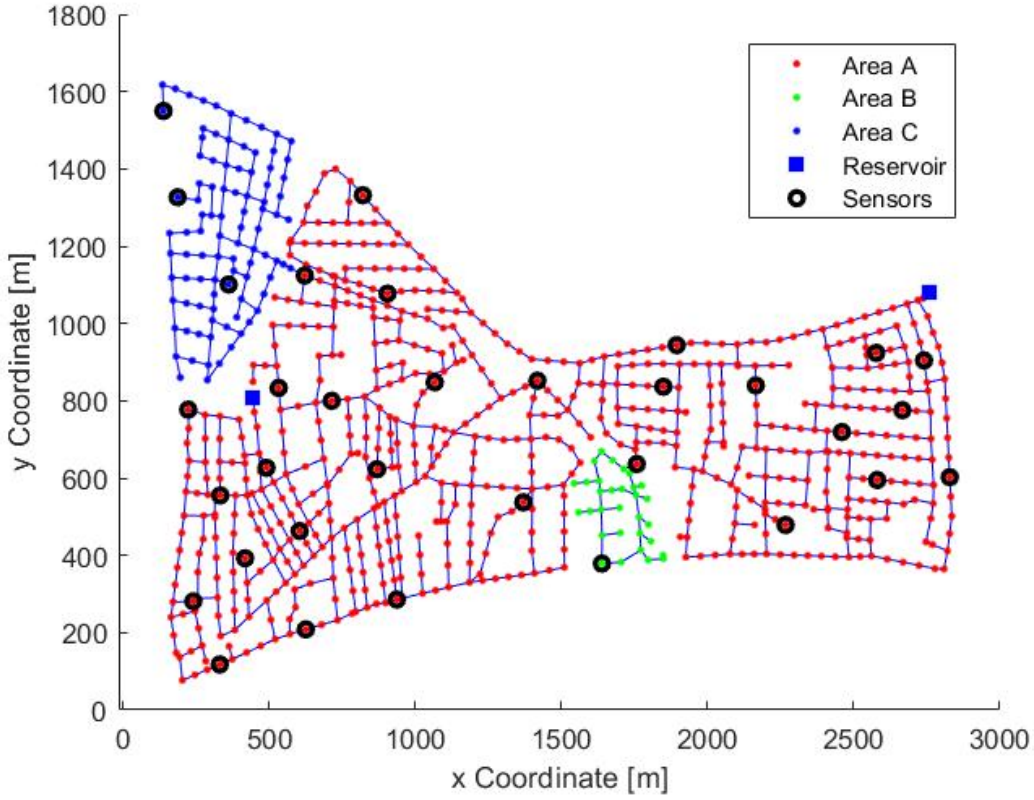


Figure 2.5: L-town WDN

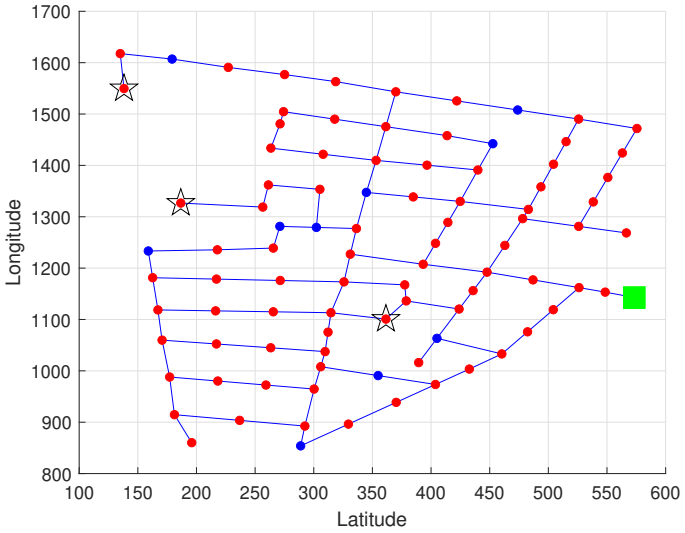


Figure 2.6: The 82 AMR locations (nodes with red color), the 10 nodes without AMRs (node with blue color), and the location of the pressure sensor (node with a star) installed in Area C.

2.8 Summary

This chapter summarizes the general knowledge necessary to understand the methods of detection and location of leaks and the technique of sensor placement and validation developed throughout this manuscript. First, an introduction of the hydraulic model equations, the representation of the WDN in graphs, and a reduced network model were presented. Additionally, the residual studies' explanation and the data-fusion concept were presented. Finally, the evaluation indices and the networks used in the case study were presented.

CHAPTER 3

SENSOR PLACEMENT AND SENSOR VALIDATION

3.1 Introduction

This chapter will focus on developing sensor placement and sensor validation methods that are an essential stage in WDN supervision. The sensor placement is in the offline phase of the flowchart Figure 1.2, and the sensor validation is applied in the online step when the network operates in normal conditions defined in the leak detection phase.

Leak localization in WDN is very sensitive to the positions of the sensors installed along with the system, and the wrong positioning of sensors affects its performance. This sensitivity in the sensor position can be noticed in Chapter 5, where different leak localization methods are analyzed. The most straightforward optimal sensor placement algorithm using a model is to test each possible sensor configuration subset, finding the one that minimizes the error in the leak localization considering all possible leaks (simulated with the model). However, this exhaustive search of the space is computationally intractable except for small sensor configuration sets, and not all the WDN has models to simulate all possible leaks.

In this chapter, two methods developed for the positioning of pressure sensors will be presented, the first being a model-based approach, ideal for when the WDN has a reliable hydraulic

model. The second will be data-driven, only needing the topological information of the network, being used for systems that do not have hydraulic models or that are not reliable. The two methods will be compared using the Moderna Network in the case study presented in Chapter 2.7.

Subsequently, the approach for validating pressure sensors is applied while the detection leak methods are operating. When supervising the WDN, it must be ensured that the sensors are working correctly since the leak localization methods will have satisfactory results if the information is reliable. Therefore, the method will check if the pressure sensors installed on the WDN are operating correctly. The information from the system's pressure measurements will be used, and the method will be validated using the Hanoi and Modena Network.

The contributions of this chapter have been published in:

- [6] Alves, D., Blesa, J., Duviella, E., and Rajaoarisoa, L. (2021). Robust data-driven leak localization in water distribution networks using pressure measurements and topological information. Sensors, 21(22), 7551.
- Alves, D., Blesa, J., Duviella, E., and Rajaoarisoa, L. (2022). Topological analysis of water distribution networks for optimal leak localization. International Conference on Hydroinformatics HIC 2022 − Water INFLUENCE. (to appear)

3.2 Sensor placement

Several works on leak localization were released using model-based approaches and commonly used demand-driven hydraulic simulators [90, 45]. The results based on hydraulic models are excellent. Nevertheless, the main difficulties characterize the model-based approaches are the calibration of accurate models and data availability for all possible complex scenarios.

An element that has the potential to significantly improve the localization of leaks is the sensor pressure configuration. Several strategies have been proposed in the last years that tackle the optimal sensor placement problem in WDNs for leak localization. As a branch and bound

searches [84], Genetic Algorithms [94], feature selection techniques [92], and game theory approaches [7]. This section aims to present the development of the approach to placing a given number of sensors, n_s, in a WDN to obtain a sensor configuration with a maximized leak localization performance.

Two new methodologies for sensor placement in the WDN. The first is formulated as an integer optimization problem bounded within upper and lower limits, the method uses the hydraulic model to simulate leak scenarios in the network to calculate the objective function. Moreover, the other uses only the topological network information to improve the leak localization methods. The second approach aspires to simplify the problem of sensor placement by eliminating the need to calibrate the hydraulic water models and reducing the computational burden. The methodology is based on the complex network theory applying the graph approach with the topological information to represent the WDN.

In both methodologies, the goal is to improve the leak localization methods that use residual analysis, first is necessary to set a simple leak localization method based on the pressure residual analysis. The leak localization method in Chapter 2.3 is used in this approach to improve the result of leak localization taking into account the sensor position.

In order to cope with the combinatory complexity of the sensor placement problem, following the ideas of [9], a two-step suboptimal search algorithm is proposed:

- STEP 1: Divide the set of nodes $\mathcal{S}$ of the WDN where a sensor can be installed into s clusters $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_{n_s}\}$, i.e., a cluster for every sensor to be installed in the WDN.
- STEP 2: Choose a node among all nodes of a cluster as the optimal place to install a sensor.

To carry out the STEP 1, the WDN can be represented as a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, as seen in Chapter 2 with $\mathcal{V}$ as the set of vertices that represents the n connections between the components of the network (junctions, reservoirs, and tanks), and $\mathcal{E}$ are the edges, which represent the m links (pipes, valves, and pumps) in the network. The edges are associated with

a cost value based on the friction loss in pipes of the Hazen-Williams formula, that is, the pipe length divided by the pipe diameter, to guarantee a model closer to the real behavior of the water system. It will be consider that $\mathcal{S} \subset \mathcal{V}$ and therefore the number of elements ℓ of $\mathcal{S}$ fulfils $\ell \leq n$. The cluster problem can be defined as a set of observations nodes that aims to partition the ℓ observations into $n_s(\leq \ell)$ sets $\mathcal{C} = \{\mathcal{C}_1, ..., \mathcal{C}_{n_s}\}$ considering $\mathcal{S} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_{n_s}$ and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ if $i \neq j$. The Girvan-Newman (GN) [55] clustering method is proposed for STEP 1, GN clustering is an algorithm that focuses on edges mostly between communities, so clusters are defined by progressively removing edges from the original graph according to edge betweenness, which measures the importance of an edge in a network by counting the number of shortest paths that run through it.

For STEP 2, two methods were developed: the first being the model-based approach which uses hydraulic models formulated as an integer optimization problem. This approach can only be applied if the hydraulic model has high credibility. In addition, the second method is a data-driven approach to locating sensors at the topologically most essential nodes of each cluster, ensuring a spatially uniform distribution of sensors.

Model-based approach

To perform STEP 2 in the model-based is necessary to associate the nodes in the cluster $\mathcal{C}_i$ with an index value organized the components in ascending order. To generate this arrangement a new index variables $\hat{x}_i \in \{1, 2, ..., \ell\}$ and $i = 1, 2, ..., \ell$ are defined in such a way $\hat{x}_i$ contain the index to the physical node enumeration that has been rearranged as the i^{th} node. In this way, nodes are arranged with the new index in the cluster as follows

$$\begin{aligned} \hat{\mathcal{C}}_1 &= \{\hat{x}_1, \ldots, \hat{x}_{N_1}\} \\ \hat{\mathcal{C}}_2 &= \{\hat{x}_{N_1} + 1, \ldots, \hat{x}_{N_1} + \hat{x}_{N_2}\} \\ &\vdots \\ \hat{\mathcal{C}}_{n_s} &= \{\textstyle\sum_{i=1}^{n_s-1} \hat{x}_{N_i} + 1, \ldots, \ell\} \end{aligned} \tag{3.1}$$

where $\hat{x}_{N_j}$ represent the physical node that has been arranged as the last element of cluster $\hat{\mathcal{C}}_j$

for $j = 1, ..., n_s$, with $\hat{x}_1$ represents the physical node set to be the index number one of cluster $\hat{\mathcal{C}}_1$.

The goal in STEP 2 is to choose one node in each cluster $\mathcal{C}$ to be the place to install a sensor. The ATD performance index seen in Chapter 2 can be minimized to perform the optimal sensor placement by employing the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & x_i^l \leq x_i \leq x_i^u, \quad i = 1, ..., n_s \\ & x \in \mathbb{Z}^{n_s} \end{aligned} \tag{3.2}$$

where $\boldsymbol{x} = \{x_1, ..., x_{n_s}\}$ is the set of n_s sensors to be installed, with the constraint of $x_1 \in \hat{\mathcal{C}}_1, ..., x_{n_s} \in \hat{\mathcal{C}}_{n_s}$ and x_i^l, x_i^u is the lower and upper bounds based on the equation (3.1), i.e., $x_1^l = 1, x_2^l = \hat{x}_{N_1} + 1, \ldots, x_{n_s}^l = \sum_{i=1}^{n_s-1} \hat{x}_{N_i} + 1$ and $x_1^u = \hat{x}_{N_1}, x_2^u = \hat{x}_{N_1} + \hat{x}_{N_2}, \ldots, x_{n_s}^u = \ell$. And optimization function $f(\boldsymbol{x})$ the performance of the leak localization method considering sensor configuration $\boldsymbol{x}$ computed as the ATD performance index

$$f(\boldsymbol{x}) = \frac{\sum_{i=1}^{n_o} \sum_{j=1}^{n_o} \Gamma_{i,j} A_{i,j}}{\sum_{i=1}^{n_o} \sum_{j=1}^{n_o} \Gamma_{i,j}} \tag{3.3}$$

Once the optimization is complete, it is necessary the conversion of the indexes from clusters $\hat{\mathcal{C}}_i$ to clusters $\mathcal{C}_i$ to obtain the set of physical nodes where pressure sensors should be installed $\tilde{\boldsymbol{x}}$ for the optimal leak localization performance.

$$\tilde{\boldsymbol{x}} = \{\hat{x}_{x_1}, ..., \hat{x}_{x_{n_s}}\} \tag{3.4}$$

It should be noticed that simulations of leak scenarios in the inner nodes of the WDN, $n_o = n - n_I$, are necessary to compute the confusion Γ matrix defined in (2.1) and used in the optimization function. The burst analysis is recommended for this study since the objective is to have the lowest possible FPR value. The burst magnitude can be of the order of 5% to 10% of the average consumption of the WDN, and the time analysis will depend on the analyzed

WDN. Therefore, the performance of the leak localization will directly depend on the accuracy of the hydraulic model.

Data-driven approach

As explained in the previous section, STEP 2 aims to select nodes in each cluster of $\mathcal{C}$. The core idea of the present section is to locate sensors without using any hydraulic simulations since data availability is often limited or not suitable. In addition, it reduces the computational burden.

Thus, to define a criterion to approach the sensor placement problem, in the case of unavailable or incomplete hydraulic information on the network, the topology most central nodes of each cluster are considered suitable sensors locations. For this purpose, three indicators of the importance of the nodes were used to select the positioning of the sensors:

- Closeness centrality uses the inverse sum of the distance from a node to all other nodes in the graph, the more central a node is, the closer it is to all other nodes.

$$c_c(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)} \tag{3.5}$$

  where $d(i,j)$ is the distance between vertices i and j . And n is the number of nodes in the graph/clustering. Therefore for every cluster $\mathcal{C}_i$, Closeness centrality (3.5) would be computed for the N_i elements of the cluster and the sensor i would be the one that provides a minimum value.

- Betweenness centrality: measures how often each graph node appears on the shortest path between two nodes in the graph; for every cluster $\mathcal{C}_i$, closeness centrality is computed for the N_i elements of the cluster, and the sensor i would be the one that provides a maximum value of equation:

$$c_{st}(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{3.6}$$

  where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(i)$ is the

number of those paths that pass through i ;

- Eigenvector centrality uses the eigenvector corresponding to the largest eigenvalue of the graph adjacency matrix. A high eigenvector score means that a node is connected to many nodes who themselves have high scores, for every cluster $\mathcal{C}_i$, the eigenvector centrality is computed for the N_i elements of the cluster, and the sensor i would be the one that provides a maximum eigenvalue .

$$A_s x = \lambda_{max} x \tag{3.7}$$

where A_s is the adjacency matrix of the subgraph $\mathcal{G}$ of the cluster $\mathcal{C}_i$,and λ_{max} is the largest eigenvalue. It computes the centrality for a node based on the centrality of its neighbors, according to the coordinates $e_c(v)$ of the eigenvector $x = e_c$, associated with the largest eigenvalue of λ_{max} matrix.

3.2.1 Case study

The case study selected to test the performance is the reduced model of the real WDN of the Italian city Modena presented in Figure 2.4. EPANET hydraulic simulator was used to generate the data required to apply the sensor placement method with the following simulation condition:

- Each node has a leak scenario with a constant leak magnitude, randomly selected, with 12.3 to $24l/s$ representing 5% to 10% of the average consumption of the WDN.
- Each scenario has $72h$, samples were collected every 10 minutes and filtered to 24 hours values;

The minimization of the optimization defined in Equation (3.2) is carried out using Genetic algorithms (GA) based on principles of natural genetics and natural selection. The GA can be used in the context of sensor placement in WDN to find the near-optimal placement of these sensors for leak localization. In that case, a chromosome corresponds to the possible presence or absence of a sensor at a given node.

Table 3.1 shows the results obtained in three different scenarios: with $3, 5$, and 10 possibilities of sensor placement. The number of nodes chosen to have a sensor is exhibited in all scenarios for the GA solution and each node importance method.

To analyze the performance of the proposed approach, data with different conditions have been generated artificially. The following simulation conditions were used:

- A 72h with regular operation data with an uncertainty of 10% of the nominal demand value was considered;
- Each node has a leak scenario with a constant leak magnitude, randomly selected, with 3 to $6l/s$ representing 1 to 2.5% of the average consumption of the WDN. An uncertainty of 10% of the nominal demand value was considered in all scenarios;
- To reduce the uncertainty in the data, samples were collected every 10 minutes and filtered daily;

Table 3.1 show the ATD, F1 score and Kappa index performance to all scenarios. As the optimization of Equation (3.2) is based on the optimization of the ATD index, the results presented in Table 3.1 show that the solution obtained by the GA will be the best in any case, even if the proposed data-driven methodology is not the optimal solution of the sensor network, the obtained values are not far from those of GA. Note that even getting the best optimal value of the ATD, the solution obtained by the GA does not guarantee the best result for the F1 score and the kappa. The Eigenvector centrality is the results that present the worst results, principally in the scenario with three sensors, having the Cohen's Kappa with the worst value, inferior a 0.6.

On the other hand, the Betweenness and the Closeness centrality had a similar result. However, analyzing the ATD index, the Betweenness metric improves the scenario with 3 and 5 sensors. In the scenario with ten sensors, the Closeness centrality has the better performance. Whereas in a general case, the Betweenness centrality is the best choice in the data-driven methodologies.

Figure 3.1 shows the result of the second scenario with 5 sensors. Each cluster is highlighted

Table 3.1: Average evaluation metrics.

Criterion	Nodes with sensors	ATD	F1 score (%)	Kappa
3 sensors scenario				
GA	**88 109 207**	**6.82**	**82.43**	**0.74**
Closeness	91 147 207	7.50	82.01	0.71
Betweenness	7 63 109	7.37	79.63	0.70
Eigenvector	4 83 119	8.05	67.48	0.48
5 sensors scenario				
GA	**5 41 80 110 164**	**5.80**	**47.01**	**0.71**
Closeness	49 91 135 147 207	6.50	82.01	0.75
Betweenness	7 63 49 91 135	6.54	81.91	0.75
Eigenvector	4 83 92 119 134	6.70	75.64	0.65
10 sensors scenario				
GA	**10 31 47 78 129 171 129 183 225 258**	**4.45**	**28.97**	**0.62**
Closeness	11 31 49 83 91 105 137 129 180 258	4.70	73.36	0.68
Betweenness	1 11 31 35 49 83 91 129 135 180	4.68	69.60	0.64
Eigenvector	1 4 31 83 92 102 119 129 134 180	4.71	83.36	0.79

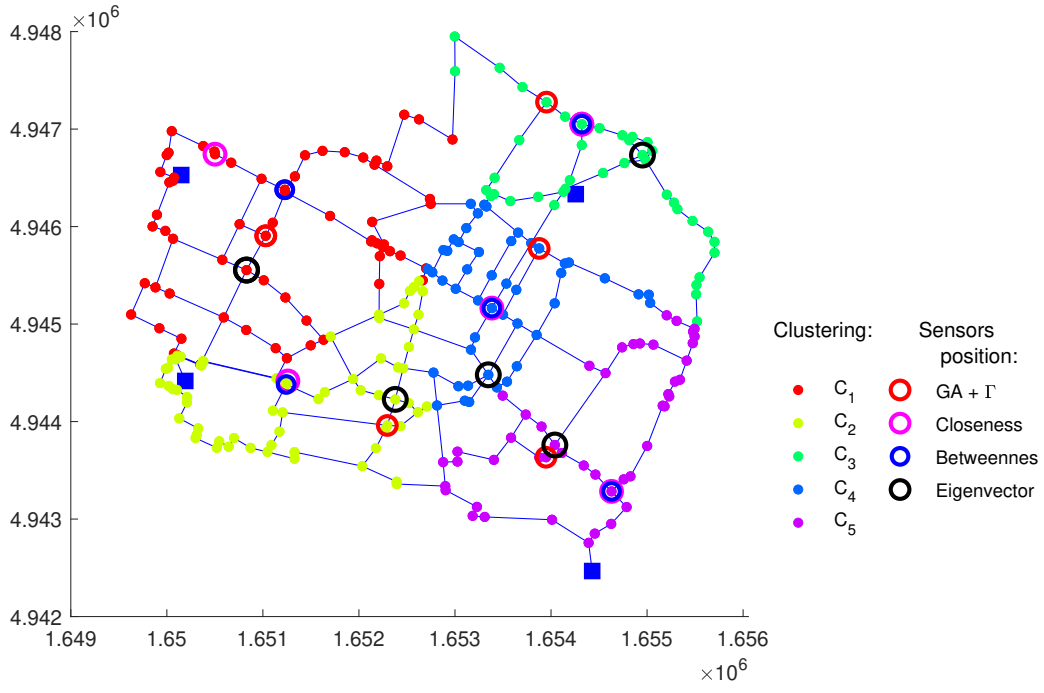


Figure 3.1: WDN of Modena and the four sensor layouts according to the three topological centrality metrics and GA solution.

with a different shade. The position of the sensors obtained with the proposed methods is shown with a circle with different colors. Clustering generated with GN provides a division focused on edges between communities that do not guarantee a homogeneous distribution of nodes. As the nodes are more communicated, the effect on the residue in a leak between these nodes will affect the sensor installed in that region more.

3.3 Sensor Validation

As the sensor positioning study is critical to guarantee the optimal result for developing the leak localization method in the WDN, a validation of the sensors is necessary during the leak localization process. Sensor validation is essential because if one or more sensors are not working correctly, the leak localization method will be compromised. This session will present the pressure sensor validation method in the WDN that must be executed when a leak is not detected by the leak detection method, as seen in the flowchart of Figure 1.2.

Anomalous values of pressure residuals r_i, $i = 1, ..., n_s$ defined in (2.15) can be used to detect sensor faults. In the same operating conditions, historical data of inner pressure sensors (leak-free data) can be used first to calibrate a pressure estimation model. Moreover, residual bounds $\underline{\varpi}$ and $\bar{\varpi}$ can be computed as maximum (positive and negative) residuals obtained with leak-free data. These bounds, allow the implementation

$$\begin{cases} r_i(k) \in \left[\underline{\varpi_i}, \bar{\varpi_i}\right] \Rightarrow \text{ No Fault in sensor } i(\phi_i(k) = 0) \\ r_i(k) \notin \left[\underline{\varpi_i}, \bar{\varpi_i}\right] \Rightarrow \text{Fault in sensor } i \ (\phi_i(k) = 1) \end{cases} \tag{3.8}$$

The accuracy of this fault detection method depends on the length of residual bounds $\underline{\varpi_i}$ and $\bar{\varpi_i}$ and, therefore, on the accuracy of pressure estimation. In order to increase the accuracy of the fault detection method, spatial residuals [11] between pressure residuals (2.15) can be computed

$$Sr_{i,j}(k) = r_i(k) - r_j(k) \quad \forall i = 1, ..., n_s - 1 \quad \text{and} \quad j = i + 1, ..., n_s \tag{3.9}$$

In the same way as pressure residuals, spatial residual bounds $\underline{\varepsilon}_{i,j}$ and $\bar{\varepsilon}_{i,j}$ can be computed using leak-free data, and the fault detection can be implemented as follows

$$\begin{cases} Sr_{s_i,s_j}(k) \in \left[\underline{\varepsilon}_{i,j}, \bar{\varepsilon}_{i,j}\right] \Rightarrow \text{No Fault}(\Phi_{i,j}(k) = 0) \\ Sr_{s_i,s_j}(k) \notin \left[\underline{\varepsilon}_{i,j}, \bar{\varepsilon}_{i,j}\right] \Rightarrow \text{Fault}(\Phi_{i,j}(k) = 1) \end{cases} \tag{3.10}$$

As model errors will affect in a similar way to close pressure sensors, it is expected that some spatial residual bounds will be smaller than pressure residual bounds. Therefore fault detection defined by (3.10) will be more sensitive to pressure sensor faults than the one defined by (3.8). The accuracy of the sensor fault detection can be increased by means of average computing residuals in a time window leading to smaller residual bounds.

Once a residual has been violated, i.e., at least one of the sensor faulty signals $\phi_i(k)\ i = 1, ..., n_s$ or spatial faulty signals $\Phi_{i,j}(k) \quad i = 1, ..., n_s - 1 \quad$ and $\quad j = i + 1, ..., n_s \quad$ is equal to one, the sensor fault isolation can be implemented in two stages as described in Algorithm 1.

Algorithm 1 Sensor validation search for sensor fault.

Stage 1: In case of the activation of one or more sensor faulty signals $\phi_i(k)\ i = 1, ..., n_s$, as these signals are uniquely related to sensors $s_i\ i = 1, ..., n_s$, the isolation is trivial and faulty sensors must be discarded for future leak localization, and the number of available healthy sensors n_s should be updated.

Stage 2: Only considers Spatial faulty signals $\Phi_{i,j}(k)$ of the n_s non-faulty sensors from $Stage1$. As these fault signals are potentially affected by two possible sensor faults s_i and s_j, the fault isolation can be implemented iteratively by the following steps:

1: **for** $i \leftarrow 1, n_s - 1$ **do**
2: **for** $j \leftarrow i + 1, n_s$ **do**
3: **if** $\Phi_{i,j}(k) == 1$ **then**
4:

$$\hat{\imath} = \underset{i \in \{1,...,n_s\}}{\arg\max} \{ \sum_{j=i+1}^{n_s-1} \Phi_{i,j}(k) + \sum_{j=1}^{i-1} \Phi_{j,i}(k) \} \tag{3.11}$$

5: Discard sensor $s_{\hat{\imath}}$, eliminate faulty signals related to this sensors, update n_s
6: **end if**
7: **end for**
8: **end for**

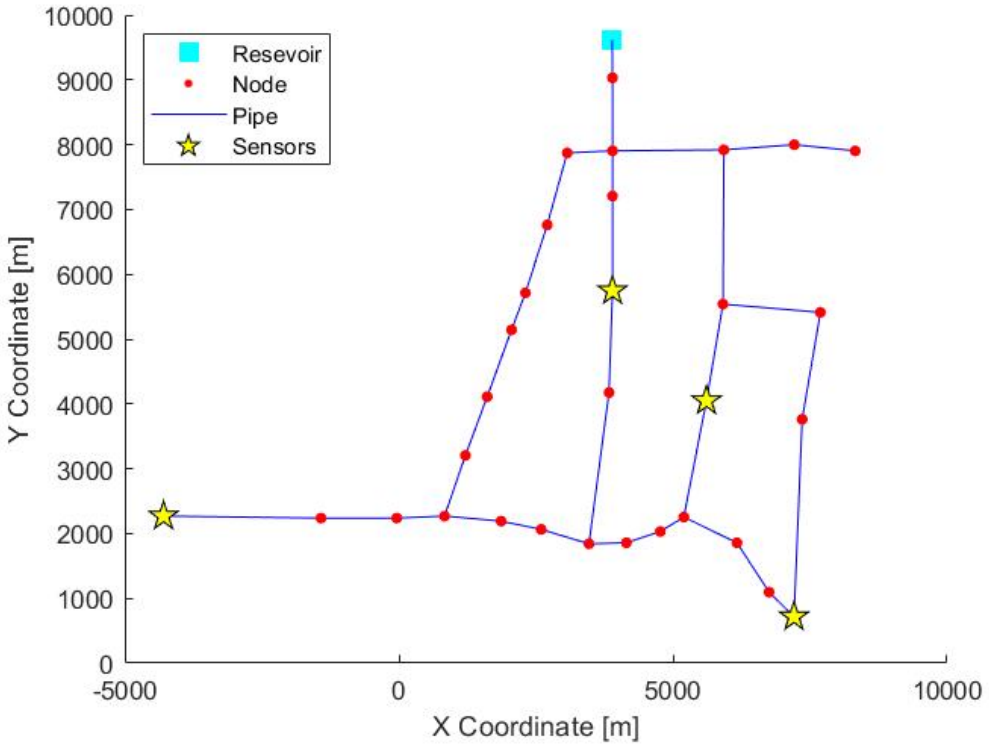


Figure 3.2: Configuration of four pressure sensors in Hanoi WDN

In case that two or more sensors obtain the same cost function in (3.11) and less than the maximum possible value $n_s - 1$, the computation of (3.11) should be done in a time window until new Spatial faulty signals are activated.

3.3.1 Case study

Two network are used for the case study to test the performance of the sensor validation approach. The first is the Hanoi network presented in Figure 2.3, and the second is the reduced model of the real water distribution network of the Italian city Modena displayed in Figure 2.4

Hanoi

To analyze the performance of the proposed approach, data with different conditions have been generated artificially using the EPANET hydraulic simulator. In order to consider realistic scenarios, some uncertainty has been added to the data [10]:

- A white noise has been combined to emulate the noise present in real measurements, and uncertainty of 10% (uniform distribution) was added in the nominal demand value.
- The sample rate is 10 min, but average hourly measurements are calculated to reduce uncertainties on the diagnostic stage.

A case study with four sensors in the nodes numbers $12, 17, 23, 29$ depicted in Figure 3.2 will be considered to illustrate the performance of the proposed sensor validation method. The four-sensor residuals computed by equation (2.15) have been considered in a time window of 24 hours using leak-free data leading to upper residual bounds equal to

$$[\bar{\varpi}_1, \bar{\varpi}_2, \bar{\varpi}_3, \bar{\varpi}_4] = [0.11, 0.06, 0.09, 0.11]$$

and the lower residual bounds equal to

$$[\underline{\varpi}_1, \underline{\varpi}_2, \underline{\varpi}_3, \underline{\varpi}_4] = [-0.14, -0.10, -0.10, -0.08]$$

In the same way, the six spatial residuals defined by (3.9) have been computed in the same conditions as sensor residuals leading to spatial residual bounds

$$[\bar{\varepsilon}_{1,2}, \bar{\varepsilon}_{1,3}, \bar{\varepsilon}_{1,4}, \bar{\varepsilon}_{2,3}, \bar{\varepsilon}_{2,4}, \bar{\varepsilon}_{3,4}] = [0.06, 0.07, 0.08, 0.04, 0.05, 0.03]$$

and

$$\left[\underline{\varepsilon}_{1,2}, \underline{\varepsilon}_{1,3}, \underline{\varepsilon}_{1,4}, \underline{\varepsilon}_{2,3}, \underline{\varepsilon}_{2,4}, \underline{\varepsilon}_{3,4}\right] = [-0.04, -0.06, -0.08, -0.06, -0.09, -0.03]$$

Figures 3.3 and 3.4 depict the evolution of sensor and spatial residuals with their respective residual bounds in a fault scenario of sensor 1 that corresponds to pressure sensor in node 12. The fault is a drift of $0.1[mcw]$ that starts on the 5th day. As shown in Figure 3.3, applying (3.8)

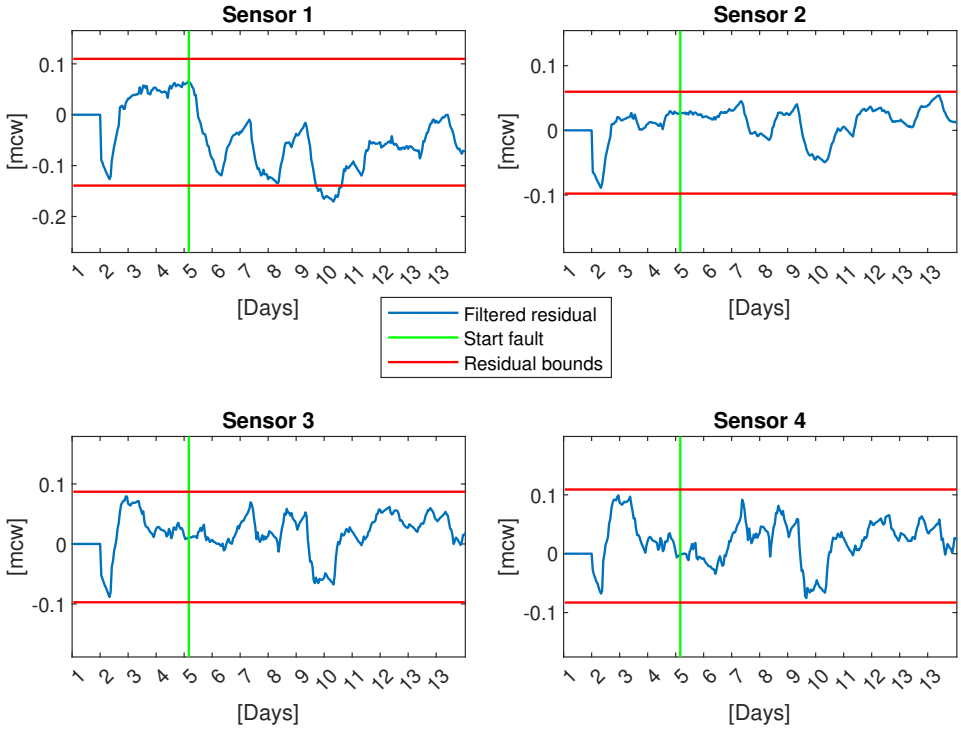


Figure 3.3: Graph of the filtered residual with a fault in sensor number 1 (a) 1st sensor (r_1), (b) 2nd sensor (r_4), (c) 3rd sensor (r_3), and (d) 4th sensor (r_2).

to sensor residuals it is impossible to detect the fault until the end of day 9 (i.e. 4 days later) when residual sensor 1 violates the bounds. However, applying (3.10) to spatial residuals it is possible to detect the fault in 10 hours: Sr_{s_1,s_2} violates its bounds in 10 hours, and Sr_{s_1,s_3}, Sr_{s_1,s_4} violate their bounds in 16 and 22 hours, respectively.

Modena Network

The Modena network was selected to illustrate the sensor validation in a large-scale network. EPANET hydraulic simulator was used to generate artificial data to analyze the performance of the proposed method. The following simulation conditions were considered:

- The leak scenario consists of data samples collected every 10 minutes and filtered to hourly values to reduce the uncertainty in the data.
- The uncertainty of demand is considered by introducing the uncertainty of $10(\%)$ of the nominal demand value. In addition, white noise is deemed to emulate the noise in the

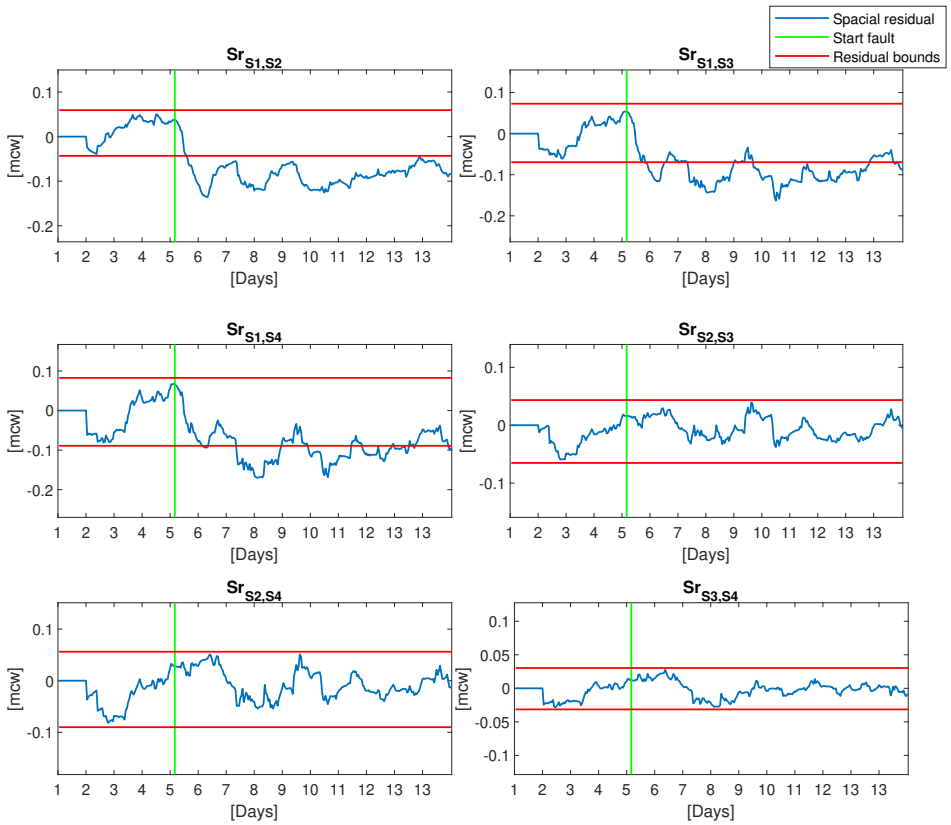


Figure 3.4: Graph of the spatial residual with a fault in sensor number 1

measurements.

The sensor bias, sensor drift, and abrupt sensor failure of sensor faults were proposed to analyze the sensor validation method. The sensor bias fault was simulated as a step change, and the drift fault was given as a time-varying ramp signal. In both cases, the fault magnitude was randomly chosen with a range of 0.1 to 0.2 [mwc]. The last fault was simulated by turning the sensor output to zero.

A total of 6000 scenarios were simulated with ten days each to evaluate the sensor validation method for the five sensor configurations depicted in Figure 5.14. Thus, 1000 scenarios were generated for each sensor with sensor bias, sensor drift, and abrupt sensor failure applied randomly, and the remainder 1000 without faults.

To calculate the residual and spatial residuals bounds, a 6-month leak-free scenario was

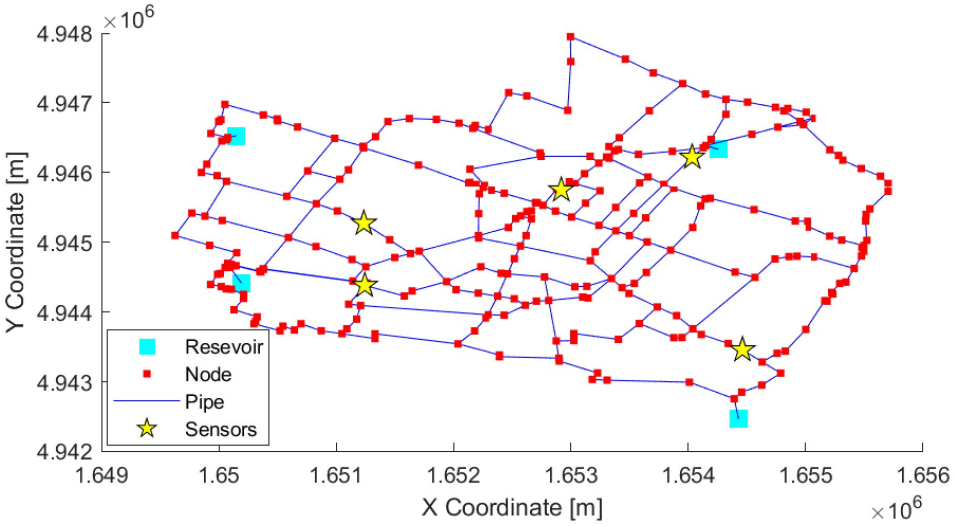


Figure 3.5: Configuration of pressure sensors in Modena WDN with 5 sensors.

generated. The five sensor residuals computed by equation (2.15) considering the time window of 24 hours and increasing 24% observed bounds. Leading to upper residual bounds equal to

$$[\bar{\varpi}_1, \bar{\varpi}_2, \bar{\varpi}_3, \bar{\varpi}_4, \bar{\varpi}_5] = [0.10, 0.06, 0.04, 0.01, 0.04]$$

and to lower residual bounds equal to

$$[\underline{\varpi}_1, \underline{\varpi}_2, \underline{\varpi}_3, \underline{\varpi}_4, \underline{\varpi}_5] = [-0.08, -0.05, -0.03, -0.01, -0.06]$$

Following, the ten spatial residuals defined by (3.9) were computed in the same conditions as sensor residuals leading to spatial residual bounds

$$[\bar{\varepsilon}_{1,2}, \bar{\varepsilon}_{1,3}, \bar{\varepsilon}_{1,4}, \bar{\varepsilon}_{1,5}, \bar{\varepsilon}_{2,3}, \bar{\varepsilon}_{2,4}, \bar{\varepsilon}_{2,5}, \bar{\varepsilon}_{3,4}, \bar{\varepsilon}_{3,5}, \bar{\varepsilon}_{4,5}] = [0.07, 0.08, 0.08, 0.10, 0.06, 0.05, 0.06, 0.03, 0.06, 0.05].$$

and

$$[\underline{\varepsilon}_{1,2}, \underline{\varepsilon}_{1,3}, \underline{\varepsilon}_{1,4}, \underline{\varepsilon}_{1,5}, \underline{\varepsilon}_{2,3}, \underline{\varepsilon}_{2,4}, \underline{\varepsilon}_{2,5}, \underline{\varepsilon}_{3,4}, \underline{\varepsilon}_{3,5}, \underline{\varepsilon}_{4,5}] =$$
$$[-0.07, -0.09, -0.08, -0.09, -0.05, -0.04, -0.06, -0.03, -0.04, -0.04]$$

For this study, the evaluation metric applied was classification accuracy. To this propose, the confusion matrix has been used, which presents the classification accuracy, and misclassification error, the horizontal axis of the confusion matrix describes the predicted labels of samples, while the vertical axis depicts the true labels of samples. The right side shows the percentages of correctly and incorrectly classified observations for each true class.

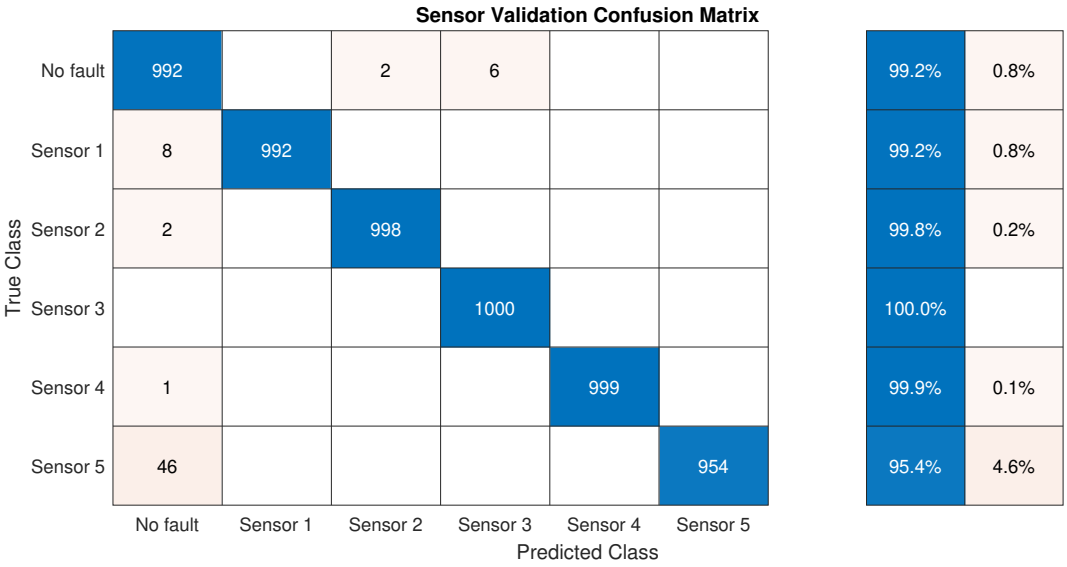


Figure 3.6: Confusion matrix for sensor validation method.

Figure 3.6 illustrates the result for the confusion matrix to all scenarios generated, as depicts the accuracy to detection fault in the sensor is very high, where the lowest accuracy is presented in fault sensor number five with an accuracy of $95,4\%$ and the highest in fault sensor number three with 100% of accuracy. Regarding the accuracy of the scenario with no-fault, 8 of the 1000 fault free scenarios presented one false alarm among the 240 samples of the scenario. Therefore, providing an average interval between false detections of $240000/8 = 30000$ hours.

3.4 Summary

This chapter has tackled two problems related to the pressure sensors installed in the inner nodes of the WDN. On the one hand, the optimal placement is to install inner pressure sensors in order to optimize the leak localization performance. On the other hand, the sensor validation allows the detection of possible pressure sensor faults that would affect the accuracy of leak localization. As a result of the chapter, a sensor placement method and a sensor validation strategy have been proposed.

The sensor positioning uses only the topological information from the WDN based on the high connection density of the graph to find the most important nodes of the network clusters has been presented. It employs only the topological characteristics of a WDN, combining the identification of clusters of nodes and topological centrality metrics for the design without carrying out any hydraulic simulation. It was proposed to provide a tool specially adapted for the frequent case where only partial information about the system is available. Following, a new approach to sensor placement that minimizes the average topology distance of leak isolability has been proposed, formulated as an integer optimization problem. However, the method uses a hydraulic model to simulate all node leaks based on the system's demand pattern and uncertainties.

In the following subsection, a sensor validation method based on the sensor pressure residuals is proposed. Historical data is used to calculate non-leak pressure estimations at sensed inner nodes. Residuals are generated using the comparison between these estimations and leak pressure measurements. The method is able to detect and isolate pressure sensor faults.

The proposed approaches have been explained, and an example is presented using the Hanoi and Modena Network as a case study. They demonstrate that the methodology for sensor placement using only the system topology information obtained a good result, ideal for cases where partial details on the system are available. In the analysis of the sensor validation, the results are satisfactory results.

In future research, a study of how different objective functions can improve the selection and the effect of other stressing conditions (i.e., sensor failures) in the network can change the

result of sensor placement.

CHAPTER 4

LEAK DETECTION

4.1 Introduction

This chapter will present the methodologies for leak detection developed during the thesis. Two methods will be demonstrated, both being data-driven methods that only depend on the information from the sensor's data.

The first method only uses information from the flow sensors installed in the inlets of the WDN, requiring free historical data and the current measurements of the sensors. The process uses the data fusion base seen in Chapter 2.4. The differential of this method is that the leak's location can occur during the entire duration of the day, as the method can combine the sensor measurements, reducing the measurement variance. This means that even if the leak happens during the period of greatest water consumption when the variance is more significant, the method can detect it within a few hours. The technique presented was applied to a DMA of the Barcelona WDN.

The second method is improving the development of leak detection, tackling the system's multi-leaks problem. The method uses data from the flow and pressure sensors, using the extra information from the sensor pressure to detect multi-leak. The fusion data of the flow and

pressure measurements are studied, thus obtaining the instant where the leak starts and if there is more than one simultaneous leak (multi-leaks) occurring in the network. As a result, the method was applied to the L-Town network.

The contributions of this chapter have been submitted in:

- [3] Alves, D., Blesa, J., and Duviella, E(2020) “Detecçao de vazamento de distribuiçao de água”. In: Conferência de Estudos em Engenharia Elétrica (CEEL) . DOI: 10.5281/2596-2221.xviiiceel.2020.548.
- Alves, D., Blesa, J., Duviella, E.,and Rajaoarisoa, L. (2022) Leak Detection in Water Distribution Networks Based on Water Demand Analysis . 11th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes - SAFEPROCESS. (to appear)
- Alves, D., Blesa, J., Duviella, E.,and Rajaoarisoa, L. (2022) Multi-leak detection and isolation in water distribution networks 2nd WDSA/CCWI Joint Conference.(to appear)

4.2 Methodology focus on single leaks

The proposed method to leak detection descends from the base of sensor fusion theory using the inlet flow measurement of the WDN to generate a virtual measurement. The technique can analyze if there is a leak in the system during all-day hours. Because of that, the leak detection will be faster than a method based on Minimum Night Flow (MNF) analysis that uses flow during night hours. Figure 4.1 shows the schema of the proposed method, which can be divided into two phases. First, the offline phase is the calibration of the parameters. Furthermore, the second online phase is where the evaluation of leak presence in the network is analyzed. We will explain the methodology in detail in the following.

The fundamental aspect of offline and online phases represents the WDN inlet flow, approximating the current and historical input flow. Therefore, the demand forecast in WDN is out of the scope of this work. However, it can be assumed that a demand forecast method calibrated

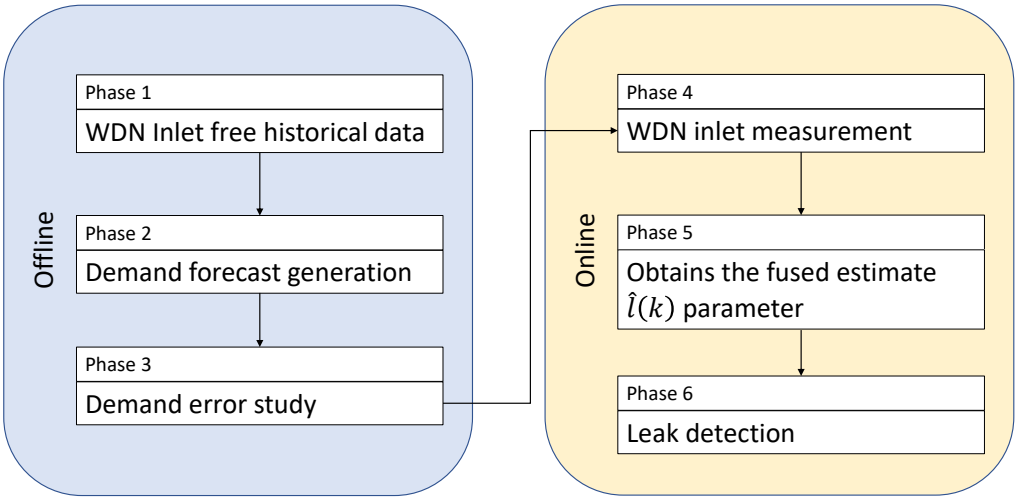


Figure 4.1: Overview of the proposed method

being historical data of the WDN [24] is available with the WDN input flow $y(k)$ at the instant k:

$$y(k) = \hat{y}(k) + e(k) \tag{4.1}$$

where $k = 0, 1, 2, 3, ...$ denotes the discrete time corresponding to time $0, T_s, 2T_s, 3T_s, ...$, being T_s the sample time of demand forecasting model, $\hat{y}(k)$ is the demand forecast and $e(k)$ is the error that for this study is considered adjusted by a normal distribution (Gaussian) [60] represented by the notation $\mathcal{N}(\mu, \sigma^2)$ with mean μ and standard deviation σ. The incoming demand estimation is more accurate in some periods of the days, thus a periodic variation in time T will be considered:

$$e(k) \sim \mathcal{N}(0, \sigma^2) \quad \text{with} \quad \sigma^2 = \sigma^2(k+T) = \sigma^2(k) \tag{4.2}$$

Let us consider $l(k)$ as the leak indicator, in the presence of a leak, $l(k) > 0$. Thus, Equation (4.1) can be rewritten as

$$y(k) = \hat{y}(k) + e(k) + l(k) \tag{4.3}$$

An approximation of the leak size $\hat{l}(k)$ can be given by the difference between the actual and the estimated inlet flow, with a leak estimation error equal to the demand forecasting error.

$$\hat{l}(k) = y(k) - \hat{y}(k) = l(k) + e(k) \tag{4.4}$$

It is possible to generate different leak estimations using a time window, W, taking into account the current inlet flow value and the previous values using the following equations:

$$\begin{aligned} &\hat{l}(k) = y(k) - \hat{y}(k) \\ &\hat{l}(k+1) = y(k+1) - \hat{y}(k+1) \\ &... \\ &\hat{l}(k-W+1) = y(k-W+1) - \hat{y}(k-W+1) \end{aligned} \tag{4.5}$$

Notice that leak estimations $\hat{l}(k-i)$ with $i = 0, ..., W-1$ have zero mean Gaussian errors with variance $\sigma^2(k-i)$. Considering slow leak variation in time window W, we get:

$$l(k) \approx \bar{l}(k) = \sum_{i=0}^{W-1} \frac{l(k-i)}{W} \tag{4.6}$$

an average leak estimation $\hat{\bar{l}}(k)$ can be computed at instant k applying the maximum Likelihood estimation method to the joint probability distribution of the W estimations fused in $\bar{l}(k)$. This joint probability distribution function will be denoted as $p(\hat{l}(k), ..., \hat{l}(k-W+1)|\bar{l}(k), \sigma_W^2)$ and can be expressed as

$$p(\hat{l}(k),...,\hat{l}(k-W+1)|\bar{l}(k),\sigma_W^2) = \prod_{i=0}^{W-1} \frac{1}{\sigma(k-i)\sqrt{2\pi}} e^{\frac{-(\hat{l}(k-i)-\bar{l}(k))^2}{2\sigma^2(k-i)}} \tag{4.7}$$

where σ_W^2 is the variance of the fused value $\bar{l}(k)$). The likelihood function L is defined as the logarithm of $p(\hat{l}(k),...,\hat{l}(k-W+1)|\bar{l}(k),\sigma_W^2)$, given by:

$$L(\hat{l}(k),...,\hat{l}(k-W+1)|\bar{l}(k),\sigma_W^2) = -\frac{W}{2}\log(2\pi) - W\sum_{i=0}^{W-1}\log\sigma(k-i) - \sum_{i=0}^{W-1}\frac{(\hat{l}(k-i)-\bar{l}(k))^2}{2\sigma^2(k-i)} \tag{4.8}$$

Maximizing the value of $L(\hat{l}_1,\hat{l}_2,...,\hat{l}_W|\bar{l}(k),\sigma_W^2)$, equaling to zero the derivative of $p(\hat{l}_1, \hat{l}_2,...,\hat{l}_W|\bar{l}(k),\sigma_W^2)$ with respect to $\bar{l}(k)$ [99], obtains the new virtual fused estimate measurement $\hat{\bar{l}}(k)$:

$$\hat{\bar{l}}(k) = \frac{\sum_{i=0}^{W-1}\frac{\hat{l}(k-i)}{\sigma^2(k-i)}}{\sum_{i=0}^{W-1}\frac{1}{\sigma^2(k-i)}} \tag{4.9}$$

that presents a zero mean estimation error

$$e_W(k) = \bar{l}(k) - \hat{\bar{l}}(k) \tag{4.10}$$

with variance

$$\sigma_W^2 = \frac{1}{\sum_{i=0}^{W-1}\frac{1}{\sigma^2(k-i)}} \tag{4.11}$$

The leak detection problem can be formulated as a change detection problem because, in a non-leak scenario, $\hat{\bar{l}}(k)$ will lead to small values but different from zero due to demand estimation errors, and in a leak scenario, its value will increase.

In the offline phase, the computation of the threshold Δ_W using historical free leak data will determine a value of $\hat{\bar{l}}(k)$ above which we can assume that a leak is present in the system. This threshold can be computed applying Equation (4.9) to historical free leak data, considering the worst-case scenario Δ_W will be equal to the maximum value of $\hat{\bar{l}}(k)$ computed for the whole historical non-leak data.

Given Δ_W, to reduce the number of false alarms, a n_d value can be stipulated, being the number of several following leak estimations bigger than the threshold that is necessary to trigger the leak detection.

In the online phase, the process of fused estimate leak size can be done, and the leak detection method can be computed by:

$$\begin{cases} \hat{\bar{l}}(k-i) > \Delta_W \Rightarrow \text{Leak} \quad , \quad \forall \text{ i=1,..,} n_d \\ \text{Otherwise} \Rightarrow \text{No Leak} \end{cases} \tag{4.12}$$

4.2.1 Case Study

The method presented was applied to a DMA of the Barcelona WDN. In particular, the set of historical inlet flows free-leak data depicted in Figure 2.2 presented in the Chapter 2.7 was available. For the leak detection analysis, different leak scenarios have been created considering the constant size of the leaks.

The scenarios have been tuned in the following way:

- The sample time, T_s, is one hour, and the period T equals 24 (1 day).
- The sample has 28 days. During the 14th first, the system is operating in regular operation. The leak started on day 14 at a random time that can start at 1 AM until 00 AM.

- The leak magnitude is constant for a given leak scenario but with different leak magnitudes of $0.5, 1.0, 1.5$ and $2.0[l/s]$ for different leak scenarios. Adding the uncertainty of Equation (2.8) with l^u of 1%.

Exist different ways to represent the demand forecast in the literature, as [34, 43]. However, in this case study, the demand forecast was considered only the periodicity of the demand extracted from the historical data to construct an estimate of the current water demand, as proposed in [24]. So, given a set of historical inlet flow free-leak data of N_d days sampled at T_s =1 hour, the demand forecast model will consist of 24 values (features) $\hat{y}_h$ with $h = 1, .., 24$ organized by the 1st feature equal to demand forecast at 1 AM, and the 24th is equal to 00 AM. These values are computed from historical data as follows.

$$\hat{y}_h = \frac{1}{N_d} \sum_{d=0}^{N_d-1} y(h + 24d) \qquad h = 1, ..., 24 \tag{4.13}$$

The first analysis made was concerning the election of the best amount of features, the leak detection proposed in the previous section considers inlet flow values from all the hours of the day (i.e., all the features), while most leak detection methods are based on Minimum Flow Analysis that only consider the flow at some hours during the night. So a general analysis was made considering time window $W = T = 24$ (i.e., one day) and maximum error threshold for fault detection (4.12) using a different number of features. In addition, simple thresholds $\delta_h \quad h = 1, ..., 24$ are computed as the maximum error of hourly demand estimations $\hat{y}_h$ computed by (4.13) considering the historical leak-free data. These thresholds show the lowest leak value detectable only considering hourly measurements. The following equation was used to generate δ_h:

$$\delta_h = \max_{d=0,...,N_d-1} |y_h(h + 24d) - \hat{y}_h| \tag{4.14}$$

This analysis is represented in upper Figure 4.2, which represents the hourly demand estimation (4.13) in blue and the upper red line the prediction $\pm\delta_h$ threshold. On the other hand, in the

lower Figure 4.2 represents hourly error variance σ_h^2. Note that the biggest variance σ_h^2 happens between 8 AM to 3 PM; this is already expected because they are the times of the biggest water consumption in cities. Consequently, we have a larger δ_h in those hours. With the same analysis, the smallest variance occurs during the night between 2 AM and 5 AM, as it is the last water consumption period, having the smaller δ_h. Remark that in the worst case of leak detection, a leak is produced in the hours with high σ_h^2.

Continuing the analysis to know how the data is affected by the number of features selected, they were stocked by the best variations in ascending order, i.e., the feature with the smallest variance is now the 1st feature, and the biggest is the 24th feature, showing in upper Figure 4.3 . In order to know how it would affect the threshold Δ_W, lower Figure 4.3 depicts the twenty-four errors that can be computed $e_{f|24}(k)$ with $f = 1, ..., 24$ using error Equation (4.10) but considering leak estimation in Equation (4.9) only with f features, being $e_{24|24}(k) = e_w(k)$. With every error obtained applying to leak-free data it can be computed a maximum error $\Delta_{f|24}$, being the $\Delta_{24|24} = \Delta_W$. In the analysis of the $\Delta_{f|24}$, it is noted that this value decreases smoothly when the number of features is bigger than five.

The second analysis was regarding the performance of the leak detection method. Four different leak magnitudes, 0.5l/s to 1.5l/s, representing 1.5% to 5.5% of the average system inflow, were generated with 10000 scenarios each, and they were applied to obtain the following parameters:

- True Positive Rate (TPR) is the percentage of leaks that are correctly identified as such.
- False Positive Rate (FPR) is the percentage of leak-free data that triggered the leak detection method.
- Difference Time Detection (DTD) is the time (in hours) from the leak appearance to the leak detection.

Tables 4.1, 4.2, 4.3 and 4.4 were created for the analysis of these parameters. The first choice of the features was 4, which are the first four features with the lower variance that included in

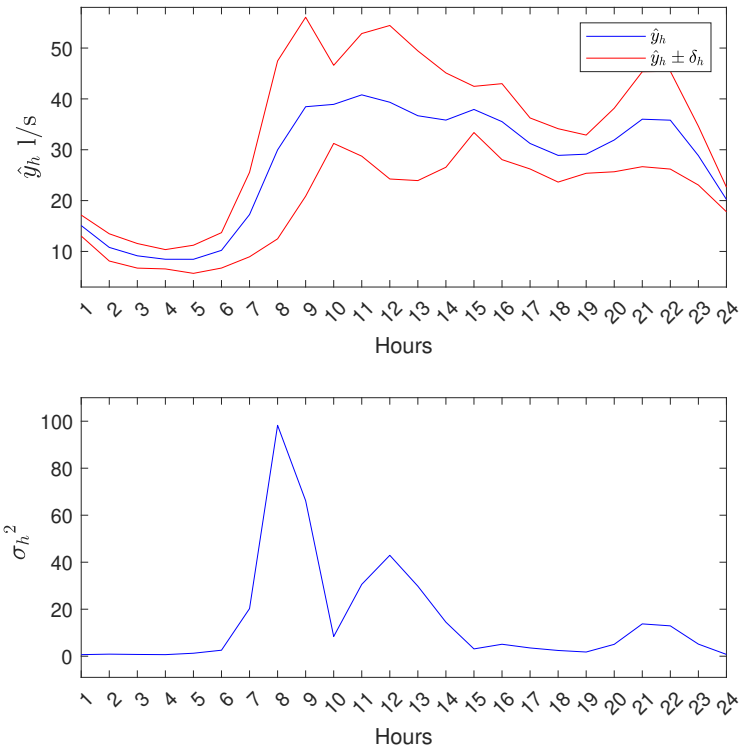


Figure 4.2: Hourly demand estimations $\hat{y}_h$ with the respect δ_h and variance values σ_h^2

the hours of the Minimum Night Flow Analysis. The second resource option was to use all available resources, which means the number of Features f equals 24. The importance of n_d was also examined, with Table 4.1 and Table 4.2 considering n_d equal to 1 and Table 4.3 and Table 4.4 considering n_d equal to 3.

Regarding n_d, a small increase in the detection time is noted when using $n_d = 3$ because more measures are needed to activate the detection. On the other hand, the number of FPR is

Table 4.1: Leak detection performance considering $n_d = 1$. Considering $f = 4$.

Leak magnitude (l/s)	FPR= 0.009	
	TPR (%)	DTD(hour)
0.5	17.7	162.872
1	90.3	111.374
1.5	100	36.888
2	100	18.998

Table 4.2: Leak detection performance considering $n_d = 1$. Considering $f = 24$.

Leak magnitude (l/s)	FPR= 0.004	
	TPR (%)	DTD(hour)
0.5	36.4	154.212
1	99.5	52.608
1.5	100	18.997
2	100	14.599

Table 4.3: Leak detection performance considering $n_d = 3$. Considering $f = 4$.

Leak magnitude (l/s)	FPR= 0.005	
	TPR (%)	DTD(hour)
0.5	8.7	160.332
1	78.3	125.795
1.5	100	44.439
2	100	21.243

Table 4.4: Leak detection performance considering $n_d = 3$. Considering $f = 24$.

Leak magnitude (l/s)	FPR= 0.001	
	TPR (%)	DTD(hour)
0.5	21.6	162.781
1	99.5	69.084
1.5	100	21.904
2	100	16.640

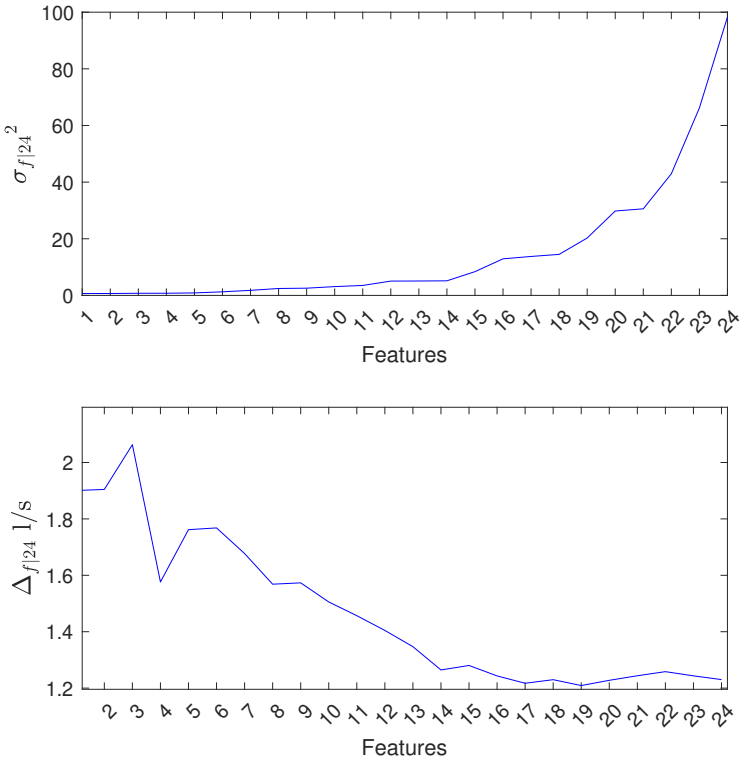


Figure 4.3: Sorted σ_h^2 by feature and $\Delta_{f|24}$ with respect to the number of Features f

significantly lower: more than 50% reduction when $f = 4$. In addition, it can be noted that it is already possible to detect leaks with a size of $0.5l/s$ but with a small TPR. Besides, it has 100% detection when the leak size is greater than or equal to $1.5l/s$.

For each network, it is necessary to do this type of analysis to know the value of the parameters to choose because each one has different behavior. The parameters must be selected according to each water distribution company's priorities. For example, it is also possible to manipulate data by dividing it into working and unworked days. However, it is necessary to have a wide range of data without leakage for this type of manipulation. For this data presented in Figure 2.2, it is not worth separating them, as the amount of information is not sufficient.

Table 4.5 was developed to compare the proposed method's efficiency with a method using only MNF. As already mentioned, this study is used more frequently for leak detection because, at this time of the day has the lowest water consumption and, consequently, the smallest variance. The same case study of the previous analysis was applied, only using the measurements from 2

Table 4.5: Leak detection performance using MNF measurement

Leak magnitude (l/s)	FPR= 0.010	
	TPR (%)	DTD(hour)
0.5	12.6	153.465
1	41.1	145.955
1.5	82.5	115.881
2	99.3	64.333

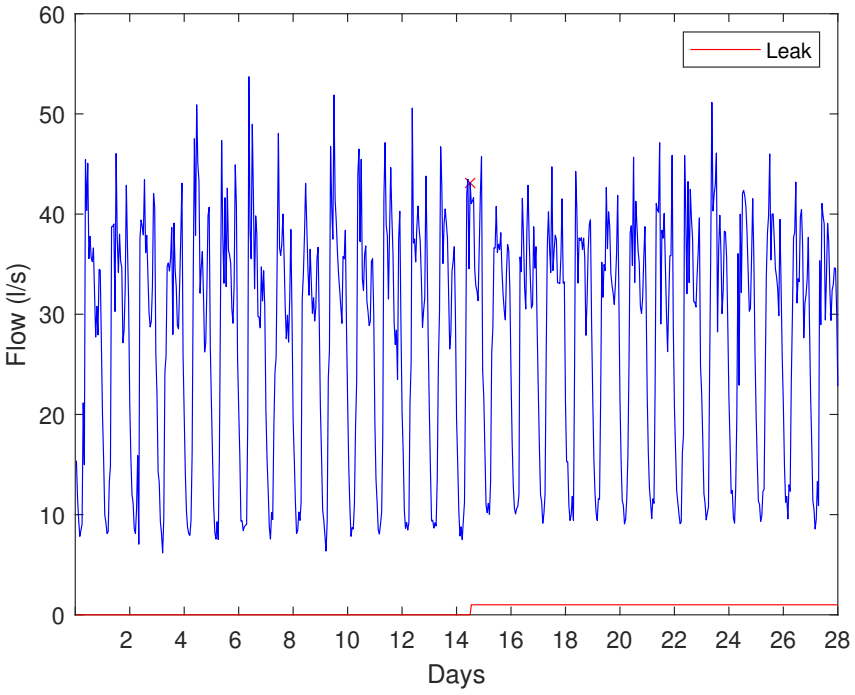


Figure 4.4: Inlet flow with a leak,start on the 14th day at 12PM

am to 6 am. The leak estimation error (4.4) was applied in the leak detection method (4.12) with $n_d = 1$. The FPR index was a reference to the threshold selection to improve the comparison. In this case, the FPR index of Table 4.1 was picked because it is an equivalent analysis that uses the four best features during the day that occurs during the night.

The study using only the MNF measurements shows that for an FPR result equal to 0.010%, the TPR is inferior to the result in Table 4.1 and the time detection is more significant because of the time delay of monitoring, that if a leak is not detected during the night or the leak started during the day, it is necessary to wait 24 hours to obtain new measurement information.

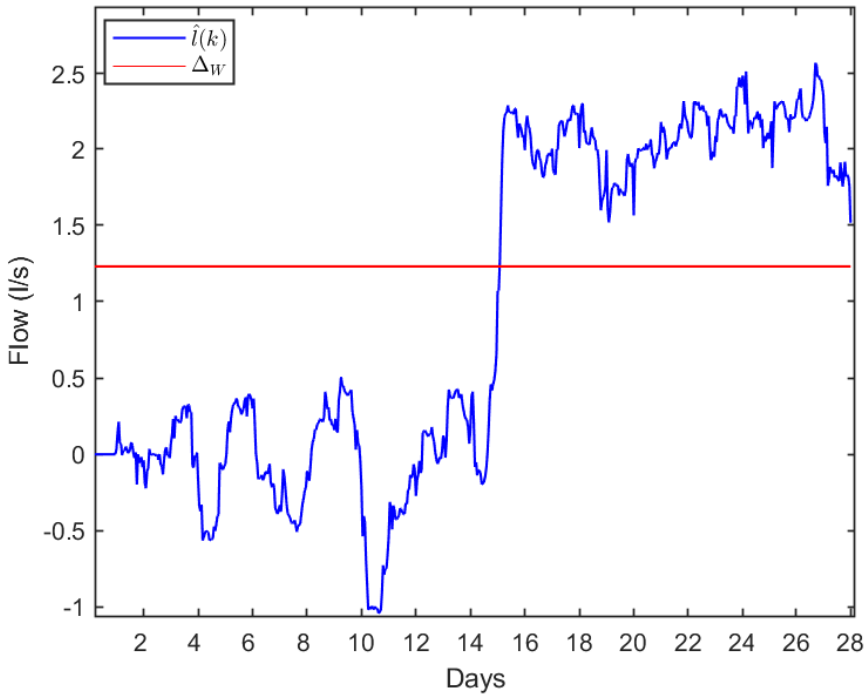


Figure 4.5: Error analysis

Figure 4.4 shows the model of artificial data created with 14 days without a leak, and 14 days with a leak of $2l/s$, the marker "x" and the red line show the exact instant that the leak was produced, in this case at 12 PM, remark that visually it is difficult to identify the leak. Figure 4.5 shows the error calculated with Equation (4.9) to the case simulation in Figure 4.4 . The analysis shows the leak detection when the error crosses the threshold. A second study can be done regarding the average error after detection, which is around $2l/s$ that is the leak magnitude.

4.3 Methodology focus on multi-leaks

The limitation of the previous leak detection method and leak magnitude estimation is the presence of a multi-leak scenario in the WDN. To detect multi-leaks using the information only of the inlet flow depends on the time distance between the leaks and if the leaks are abrupt bursts because a peak can be seen in the error analysis. Nevertheless, an external human review must analyze the error study. Moreover, in some situations, it is hard to differentiate when the network has a multi-leak scenario or if it is an incipient leak. To solve this type of problem, a complementary technique was developed with two steps: leak detection and Localization; this

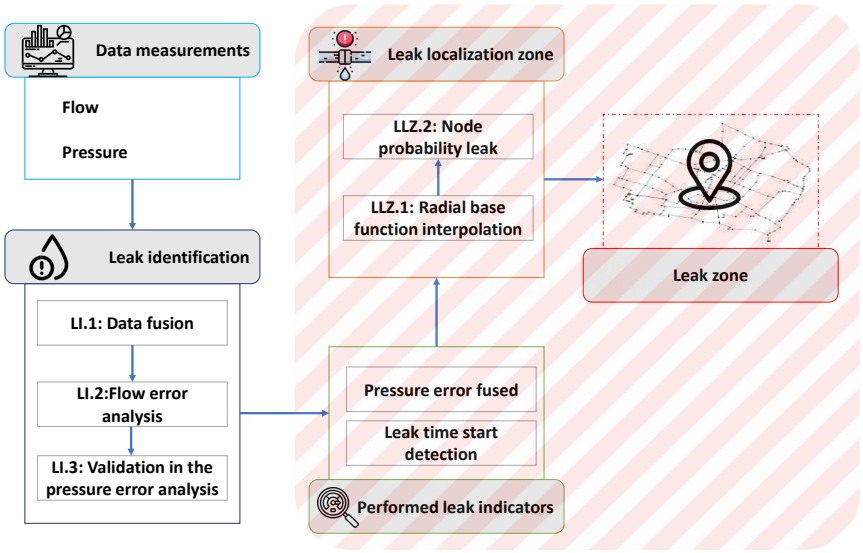


Figure 4.6: Flowchart of the leak detection and localization proposed method, highlighted the detection method

section is focused on multi-leaks detection.

Figure 4.6 describes the steps for obtaining the leak initiation time information and calculating the most likely zone to contain a leak. This chapter will explain the first leak detection phase. It descends from the base of sensor fusion theory using the inlet flow and the pressure measurements of the WDN to generate virtual measurements, able to detect the start time of the leaks in a multi-leak scenario. In the second phase, the fused pressure residual of all sensors and the longitude, latitude, and elevation of each node is applied in the radial base function (RBF) interpolation method to determine a network zone that has the fault that is explained in Chapter 5.

The fundamental aspect of the leak detection phase represents the WDN inlet flow and pressure, approximating the current and historical data. Therefore, the demand forecast and pressure forecast in WDN are out of the scope of this work. However, it can be assumed that a demand forecast method is calibrated using historical data of the WDN is available as in the previous section and that leak-free pressure estimations can be computed through historical leak-free available data.

The first step of leak identification, LI-1, is the development of the fusion of flow and pres-

sure data. This step transforms each hour of the day into different features, having 24 features, and their fusion improves leak detection thanks to reducing the uncertainties and noise in the measurement. The first fusion data addressed will be the flow measurement seen in the previous section, using Equation (4.9) to obtain the virtual fused estimate measurement and apply the leak detection method defined in Equation (4.12).

Furthermore, once $\hat{\bar{l}}(k)$ is above Δ_W is considered a disturbance in the system alarming to a probable presence of a leak that needs to be validated with the study of data fusion of pressure measurements, which will be explained in the next topic.

Data fusion of pressure measurements is performed by analyzing pressure residues generated by comparing internal pressure measurements and leak-free pressure estimations for each sensor, installed in the WDN, computed such as:

$$r_i(k) = \hat{p}_i(c(k)) - p_i(c(k)), \quad \forall i = 1, \ldots, n_s, \tag{4.15}$$

where $r_i(k)$, $\hat{p}_i(c(k))$ and $p_i(c(k))$ are the residual, leak-free pressure estimation, and pressure measurement at inner node i, $c(k)$ is the operating condition at given instant k defined by inlet measurements and n_s is the number of inner sensors installed in the WDN. In the same way as Equation (4.6), it is possible to generate n_s residuals analyses using a time window,W, (the same value of the leak estimations) considering the current residual pressure value and the previous values. The average of pressure residuals $r_i(k)$ can be computed at instant k applying the maximum Likelihood estimation method to the joint probability distribution of the W residuals analyses fused in:

$$\overline{r_i}(k) = \frac{\sum_{j=0}^{W-1} \frac{r_j(k-j)}{\sigma_j^2(k-j)}}{\sum_{j=0}^{W-1} \frac{1}{\sigma_j^2(k-j)}}, \quad \forall i = 1, \ldots, n_s \tag{4.16}$$

The finite difference will be applied to restrict the beginning and end of a leak in the system to

the daily data of residuals fused $\overline{r_i}(k)$. The finite difference corresponds to differential operation, an important concept in calculus commonly used to smooth non-stationary time series [28]. The backward difference has the expression of the form $\nabla_h[f](k) = f(k) - f(k-h)$. In this study, the finite difference is applied to Equation (4.16) calculated as follows:

$$\nabla_1 \overline{r_i}(k_{day}) = \overline{r_i}^{max}(k_{day}) - \overline{r_i}^{max}(k_{day} - 1), \quad \forall i = 1, \ldots, n_s \tag{4.17}$$

where $\overline{r_i}^{max}(k_{day})$ is the maximum value of fused residuals provided by Equation (4.16) in the day k_{day}: i.e. maximum of the 24 hourly values computed by means (4.16) in the day k_{day}.

When a leak occurs in the WDN, all the measurements of the pressure sensors will be affected, nevertheless, the sensors closer to the leak will show more disturbance; taking into account the network can be divided into α groups $\mathcal{C}$= $\{\mathcal{C}_1, \ldots, \mathcal{C}_\alpha\}$, with the region and the neighboring sensors as a parameter. To analyze the $\nabla_1 \overline{r_i}(k_{day})$ by group the index $\Delta\overline{r_{gi}}(k_{day})$ $i = 1, ..., \alpha$ can be calculated as the sum of the values provided by Equation (4.17) for the sensors that belong to the group $\mathcal{C}_i$:

$$\Delta\overline{r_{Ci}}(k_{day}) = \sum_{\forall j \in \mathcal{C}_i}^{\alpha} \frac{\nabla_1 \overline{r_j}(k_{day})}{\nabla_1 \overline{r_j}^{max}} \tag{4.18}$$

where $\overline{r_j}^{max}$ is the maximum value provided by Equation (4.17) for sensor j in leak-free historical data. In these analyses, a peak is produced in the signal when has a disturbance in the sensors, for example, when a leak starts or when it is fixed. To proceed with the leak detection method, a threshold Δ_{C_i} is computed for every group i in order to distinguish residual errors from leaks. Then, for each group $\mathcal{C}_i$ the leak detection method can be computed by:

$$\Delta\overline{r_{Ci}}(k_{day}) = \begin{cases} \Delta\overline{r_{Ci}}(k_{day}) & \Delta\overline{r_{Ci}}(k_{day}) > \Delta_{C_i} \\ 0 & \text{Otherwise} \end{cases}, \quad \forall i = 1, \ldots, \alpha \tag{4.19}$$

The $\Delta\overline{r_{Ci}}(k_{day})$ in Equation (4.19) is set to only present disturbances when a failure is similar to a leak in the system. To set the analysis for disturbances like a leak repair signature, the threshold of the first line must be set to $\Delta\overline{r_{Ci}}(k_{day}) < -\Delta_{C_i}$.

With the study of Equation (4.9), it is possible to analyse whether there is a leak, but it is limited to when there is only one leak in the system or when there are more leaks with time spaces of more than time window W. In other words, if multiple leaks co-occur or with a period smaller than W, the information from Equation (4.9) will only show the sum of the magnitude of all leaks. However, with the validation of the information with the Equation (4.19), it is possible to know when multiple leaks happen because it will present a peak in the analysis data, having a better result if the locations of the leaks are in different groups.

4.3.1 Case study

The Battle of the Leakage Detection and Isolation Methods is a challenge was provided by the organizers of the BattLeDIM [101], seen Chapter 2.7. The aim is to detect and locate several leaks in a hypothetical city created to this intent, as depicted in Figure 2.4.

In this challenge, the network can be divided into two distinct parts with different challenges: the first, Area A and Area B containing simultaneous leakage, and the second, Area C containing the AMR devices.

The leaks in Area A and B of the 2018 year will be addressed in this work. The data set of the BattLeDIM for this year contains the time and repair location of 9 pipe bursts that were fixed. Three types of leaks exist:

- Small background leaks with 1%–5% of the average inflow
- Medium pipe breaks with 5%–10%
- Large pipe bursts with leakage flow of more than 10% of the average system inflow ($\approx 50 l/s$)

The water utility corrects significant leaks with a flow rate above $4.5 l/s$ after a reasonable

time within two months. The leakages have two different time profiles: either abrupt bursts with constant leak flow rates or incipient leaks that evolve until significant outflow rates at which they remain constant. In [100] was presented the result of the leakage evolution historical Data-set in 2018, Figure 4.7 shows the 12 leaks in 2018, with outflow rates between 1.4 and $9.7l/s$ (5 and $35m^3/h$). Three leaks are not fixed, and nine leaks are repaired throughout the year that will be analyzed in this section in the highlighted order of n.1 to n.9.

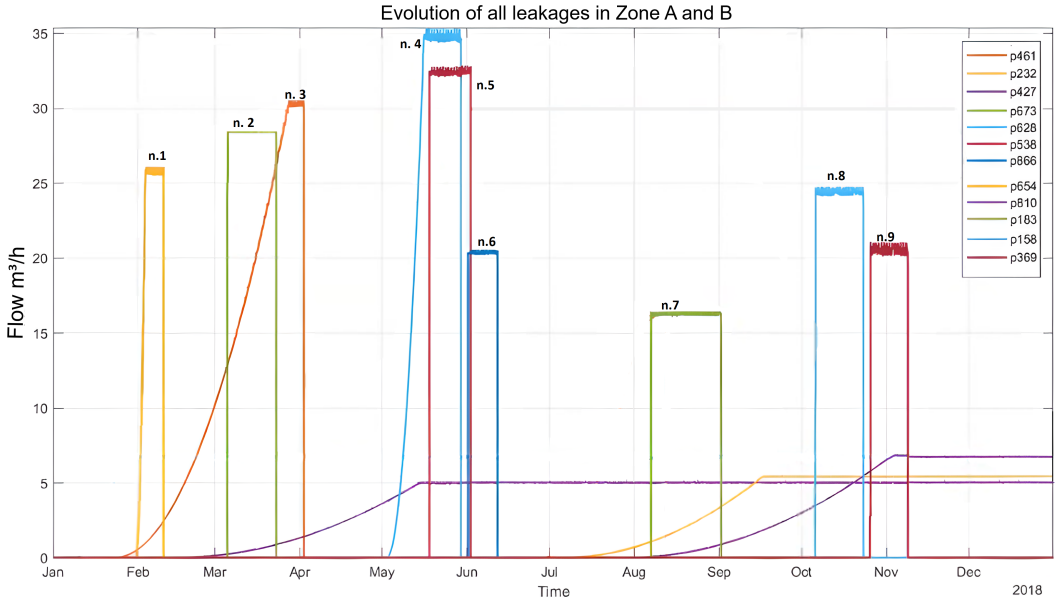
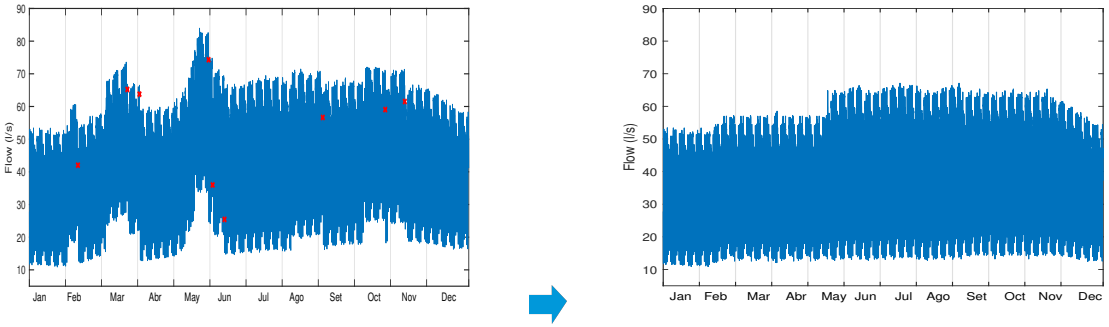


Figure 4.7: Evolution of leaks in $[m^3/h]$ in Area A and B during 2018, Source: [100]

To analyse the first step of the leak detection proposed in the Section 4.2 it necessary to do data manipulation to create a leak-free history data to calculate Equation (4.4). To perform this, it has been necessary to visually identify the beginning of the leak and remove the days that potentially contained the leak, then replace those days with days without leakage. The time and days of the week were taken into account to be modified the data, trying to make them as close as possible to a real free historical leak.

Figure 4.8 shows the manipulation done on the data. In Figure 4.8.(a), the red "x" indicators are the time when the leak was fixed, regarding that in March, May, and June there are simultaneous leaks. Figure 4.8.(b) is the manipulation done to generate a history of free-leak.

Measured data is available every 10 minutes, but it has been filtered every hour in order to obtain an hourly demand prediction (T_s=1 hour). In addition, daily periodicity has been consider



(a) Historical inlet flow given by the BattLeDIM (b) Adaptation for a leak free scenario

Figure 4.8: Flow inlet in Area A and Area B

i.e. $T = 24$. A polynomial has calibrated every hourly flow prediction (feature) variation throughout the year, to obtain the demand forecast more accurately, taking into account the variation in the behavior during the year, which can reach more than 5 l / s on the time of the day analyzed. In this step, it is possible to analyze the manipulation of data created, if any data replacement was done incorrectly, it could be seen. The following Equation represents the flow estimation:

$$\hat{y}_{h,d_{ay}} = \sum_{n=0}^{n_p} a_n^h {d_{ay}}^n \tag{4.20}$$

where $\hat{y}_{h,d_{ay}}$ is the demand estimation at hour $h = 1, ..., 24$ (first 1 AM and the last 00 AM) of day d_{ay}, a_n^h are the coefficients of polynomial at hour h and n_p is the order of the polynomial. An example of the demand forecast at 1 AM is presented in the Figure 4.9, the polynomial is not precisely in the middle because the water consumption is more constant during the week, having only the periodic peaks that are the weekends with the highest water consumption.

The leak detection has been done according to the presented method, considering $W = 24$, using the history of free leak data to calculate the maximum possible error and therefore create a threshold $\Delta_W = 2.5$. However, it is necessary to adapt the detection method to be able to identify several leaks, being essential information about the days of leak fixing; therefore, whenever a leak is corrected, the leak detection function can be restarted. In addition, it could create a delay in each leak fix information according to the window W chosen to restart the

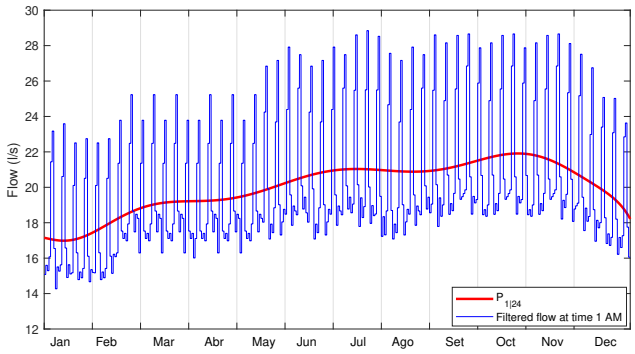


Figure 4.9: Filtered data by 1 AM and the new estimation $\hat{P}_{1|24}$

leak detection, consequently creating the prevention of false alarms. Figure 4.11. (a) shows the result obtained, from the leak estimation computation provided by Equation (4.9). As it can be see all leaks have been correctly detected, with the red markers "o" being the time of detection and the black markers "x" being the leak fix.

In Figure 4.11.(a) it is only possible to accomplish a multi-leak detection if the first leak is fixed; to proceed with the multi-leak detection, the approach's second step was proposed. It is necessary to develop a group of the n_s sensors in α groups. In this work, five groups were obtained by the heuristic approach considering the neighboring sensors and the distance between them. Figure 4.10 show the result of the grouping process. As it can be seen, the groups do not have the same number of sensors since group C_1 has more sensors concentrated in the same area. Another factor is using the pressure sensor data in more than one group because if a leak happens in the border zone between groups, the fault will be identified in more than one group analysis.

The five-group signal $\Delta\overline{r_{Ci}}$ is calculated with the pressure data using Equation (4.19) with the threshold value Δ_{C_i} equal of the number of sensor average of each group α . Figures 4.11.(b)-(f) presents the results of these five signals. A red circle is highlight for every time k that the leak detection of Equation (4.12) detect a leak in the system with a red line limited in the pressure analysis $\Delta\overline{r_{Ci}}$, Figures 4.11(b-f).

When these flow detections happen, it is necessary to validate with the $\Delta\overline{r_{Ci}}$ study of Equation (4.19). When abrupt burst faults begin in the WDN, it is possible to remark a peak in

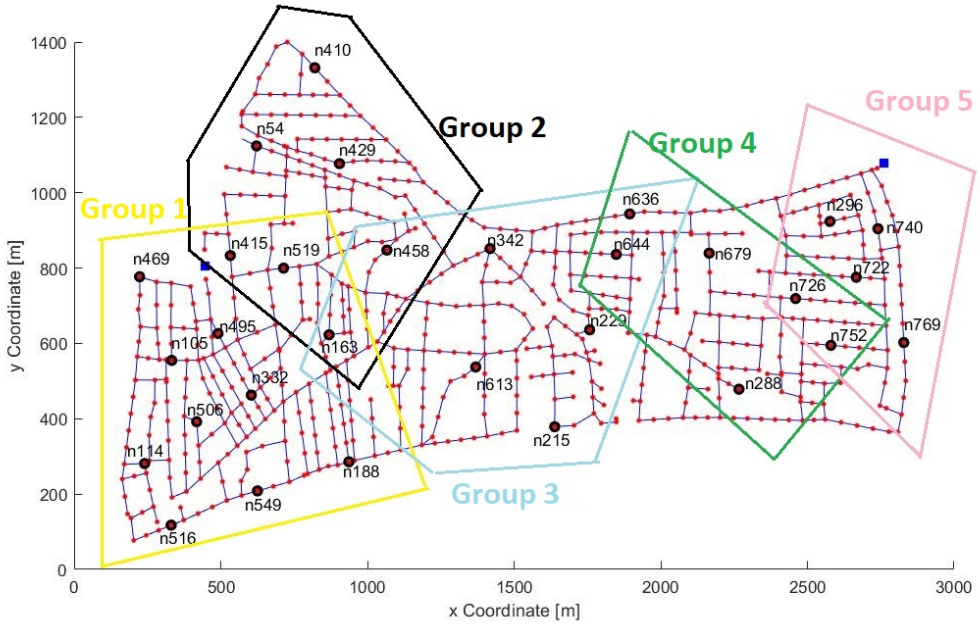


Figure 4.10: Division of sensors into α groups

the$\Delta\overline{r_{Ci}}$ analysis in the group more affected by the leak. This is the case of leaks n.1 and n.7, and it is possible to point out that leak number 7 started hours before the analysis of the $\hat{\bar{l}}$ alarm the fault. A careful analysis needs to be made in cases where a multi-leak exists, that is, leaks number 2-3, leaks number 4-6, e leaks number 8-9.

In leaks number 2-3, and an incipient leak begins in the pipe p427, which was not repaired, but the size magnitude is smaller than the other two, and it is impossible to detect it. The other two are the types of incipient and bursts. The bursts occur in Area B of the network and affect all $\Delta\overline{r_{Ci}}$ signal groups. However, this zone is an isolated area with just one sensor, and a study of it can be done, see [80]. A second peak can be detected that happens only in Group 1, indicating a probable second fault in this area: the incipient leak.

Leaks number 4-6 have an extra incipient leak in the pipe p427 that saturates in the meanwhile. The study of $\Delta\overline{r_{Ci}}$ of these times instant needs to have more attention because the leaks 4 and 5 are situated near each other in groups 1 and 3, and the leak in the pipe p427 is in group 2. Group 1 has five peaks at this period, with the two most prominent peaks identifying leaks 4 and 5. The other peaks are due to saturation in the pipe p427 and the proximity in the time when leaks start. Leak number 6 is in group 5, and it is easy to identify the start time because it only

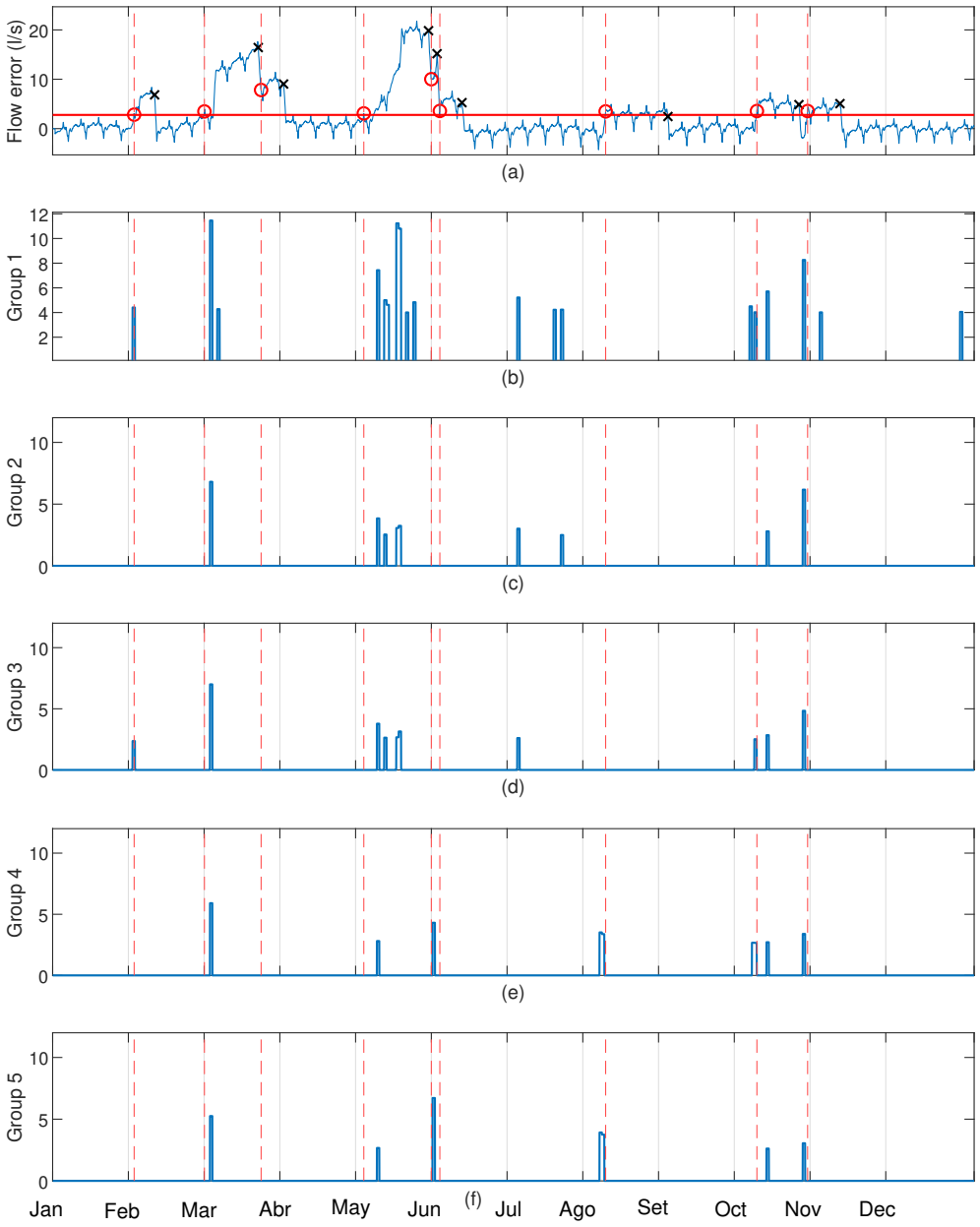


Figure 4.11: Result of leak detection, with a red line delimiting when $\hat{\bar{l}}$ exceeds the defined threshold W (a) leak estimation $\hat{\bar{l}}$ computation provided by Equation (4.9) (b-f) five-group signal $\Delta\overline{r_{Ci}}$ calculated with the pressure data using Equation (4.19)

affects groups 5 and 4.

The leaks 8-9 are not occurring together. However, the system has three saturated leaks in pipes p427, p654, and p610 that achieve the saturation moment during the leak 9. In the analysis of $\hat{\bar{l}}$ in this instant is possible only to identify the leaks 8 and 9. In the $\Delta\overline{r_{Ci}}$ examination, group 1 is the more affected, having five peaks, not making it clear at which time leaks 8 and 9 started but indicating a fault in the WDN.

The same analysis can be done when a leak is fixed. Figure 4.12 shows these results. Figure 4.12(a) is the same study as $\hat{\bar{l}}$ of Figure 4.11(a) but the black line that propagation to the other $\Delta\overline{r_{Ci}}$ signal is when a leak is fixed in the zone. In all analyses, a peak negative occurs due to a leak repair; the signal has more than 4 negative peaks caused by some uncertainties of measurements and their estimations.

4.4 Summary

This chapter has presented two complementary methods to leak detection and leak size estimation. The first is only using the information of inlet flow measurement, which is obtained with the flow sensors usually installed in the system, making possible the method implementation in most WDNs. And the second uses information on flow and pressure measurements and it is able to tackle the multi-leak problem.

The first method uses historical leak free data of the network to calculate a demand forecast combined with the sensor fusion theory to develop a leak detection that can analyze all measurements collected during the day, being an improvement due to most leak detection methods only using the analysis of the night flow. The leak detection method can be divided into two-phase, online and offline. The offline phase uses the historical data to generate a threshold that defines the regular operation of the system. In the online phase, the information of inlet flow measurement is processed and classified into normal or fault operations.

The case study presented was a real DMA of the Barcelona WDN with three months of free historical data. First, two analyses were carried out: a study of the demand forecast and

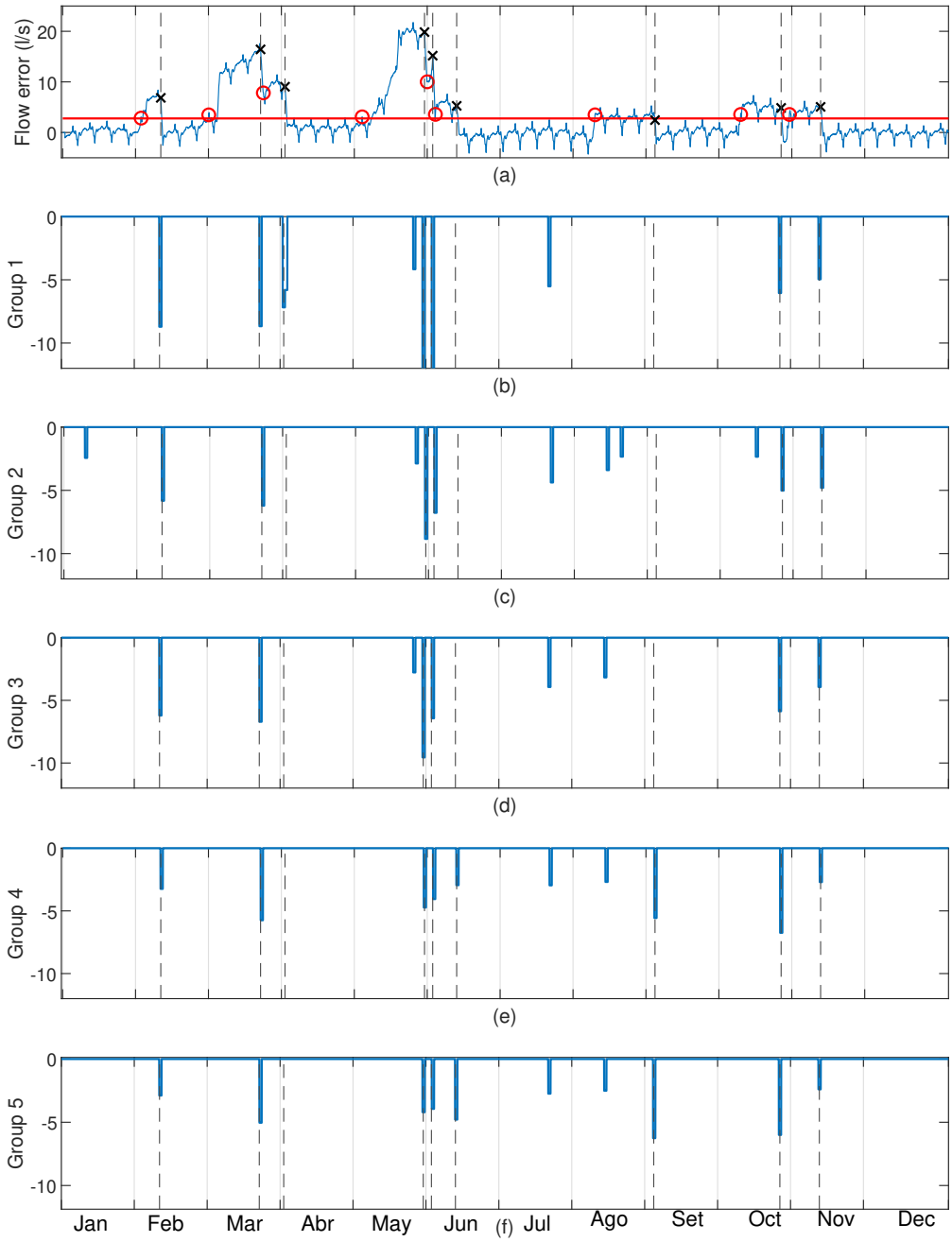


Figure 4.12: Result of leak detection, with a black line delimiting a leak fix report (a) leak estimation $\hat{\bar{l}}$ computation provided by Equation (4.9) (b-f) five-group signal $\Delta\overline{r_{Ci}}$ calculated with the pressure data using Equation (4.19) and the threshold equal a $\Delta\overline{r_{Ci}}(k_{day}) < -\Delta_{C_i}$

its variance in the respective hours of the day and the evolution of the threshold regarding the number of features. Then, different scenarios were generated with different magnitudes of leakage, ranging from 0.5 to 2 l/s, and the True Positive Rate, False Positive Rate, and Difference Time Detection were analyzed. In two scenarios, one using only nighttime measurements and the other using all available features, it is noted that when the leak is small, with the leak size value being similar to the threshold, it is not possible to detect 100%, still having a better result when it is used all available measurements.

The leak detection method capable of locating a single leak in the system is recommended for WDN with a small leak rate throughout the year, as the technique will help maintain a good network over time. The study can consider each leak unique if a network has leaked at long intervals. However, the method presented will not be effective for a WDN that has many leaks over a short period. This weakness encouraged us to develop the second leak detection method.

The second leak detection method is a multi-validate problem that starts with a study of the fused average flow and validates with the analysis of the fused average residual pressure divided by groups made by neighbors' sensors and the area of the WDN. The L-town network utilized in the Battle of the Leakage Detection and Isolation Methods has been used as a case study. The data studied were from the year 2018 with 12 leaks and only 9 repaired, having two different temporal profiles: burst pipe and incipient leaks that saturate in some instant. The data was open data published to the first phase of the challenge BatlleDIM [100]. The result of the leak detection demonstrates a good result when the leak is of the bursts type leak. On the other hand, detecting when the leak is incipient with a low growth rate is difficult because the method evolves with the data. Moreover, the method can detect simultaneous leaks.

A weak point of the proposed methods is using the threshold with a study of historical data to choose it. Although, for future work, a survey that replaces the threshold model would be ideal. It is possible the development of leak detection methods that provide a scale supporting the decision to classify the data in regular operation or fault. A study with the performance index ROC (Receiver Operating Characteristic curve), a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied, can be applied to analyze the performance of the chosen threshold.

CHAPTER 5

LEAK LOCALIZATION

5.1 Introduction

This chapter presents the approaches for leak localization in WDNs using pressure and flow measures. As seen in the flowchart Figure 4.1, this methodology is intended to be used after the leak has been detected, employing the method proposed in Chapter 4.

The chapter is divided into model-based and data-driven strategies. The first method is recommended to be implemented when the WDN has a reliable model, and in the developed approach, AMR is installed in the system, making it more accurate. The method also analyzes when the network has a multi-faults. Finally, the technique is validated using L-Town, presented in Chapter 2.7 in Areas A, B, and C, showing the importance of an accurate model.

Then the methods developed based on the data-driven will be presented having three techniques. The first method of leak localization is based on the unsupervised clustering techniques with the goal of leak localization in a group, and the method is validated using Modena WDN. The second approach studies the effect of the extra flow that the leak will generate in the WDN and how it affects the pressure sensors by analyzing the extra flow effect at a node level; the technique is validated using the Hanoi and Modena networks. Finally, the last leak localiza-

tion approach presented aims to locate a leak in a system that has simultaneous leaks, and it is validated using the Zone A and B of the L-Town network.

The contributions of this chapter have been submitted in:

- [80] Romero-Ben, L., Alves, D., Blesa, J., Cembrano, G., Puig, V., and Duviella, E. (2022) "Leak Localization in Water Distribution Networks Using Data-Driven and Model-Based Approaches". In: Journal of Water Resources Planning and Management 148.5 (2022), p. 04022016.

- [6] Alves, D., Blesa, J., Duviella, E.,and Rajaoarisoa, L. (2021). Robust data-driven leak localization in water distribution networks using pressure measurements and topological information. Sensors, 21(22), 7551.

- Alves, D., Blesa, J., Duviella, E.,and Rajaoarisoa, L. (2022). Data-driven leak localization in WDN using pressure sensor and hydraulic information. Integrated Assessment Modelling for Environmental Systems - 2nd IAMES 2022.(to appear)

- Alves, D., Blesa, J., Duviella, E.,and Rajaoarisoa, L. (2022) Multi-leak detection and isolation in water distribution networks 2nd WDSA/CCWI Joint Conference.(to appear)

- Romero-Ben, L., Alves, D., Blesa, J., Cembrano, G., Puig, V., and Duviella, E. - Leak detection and localization in water distribution networks: review and perspective. (to be submitted)

5.2 Leak localization based on Model-based

The proposed model-based leak localization method uses a hydraulic simulator to simulate theoretical pressure values caused by all potential leaks once a leak has been detected and its magnitude has been estimated. Simulated pressure values at different leak locations (hypothesis) are compared with the DMA measured pressure values to determine the most probable leak location. After the leak localization procedure, the hydraulic simulator is updated with the new extra

demand at the leak location, whose magnitude is the leak estimation value at the leak localization. This section aims to demonstrate the proposed leak localization method using model-based approaches associated with AMR devices. The methodology is applied to locate leaks at inner nodes of the network, but it can be extended to find pipe leakages by assuming the leak to be located in the pipe that connects the best node candidates.

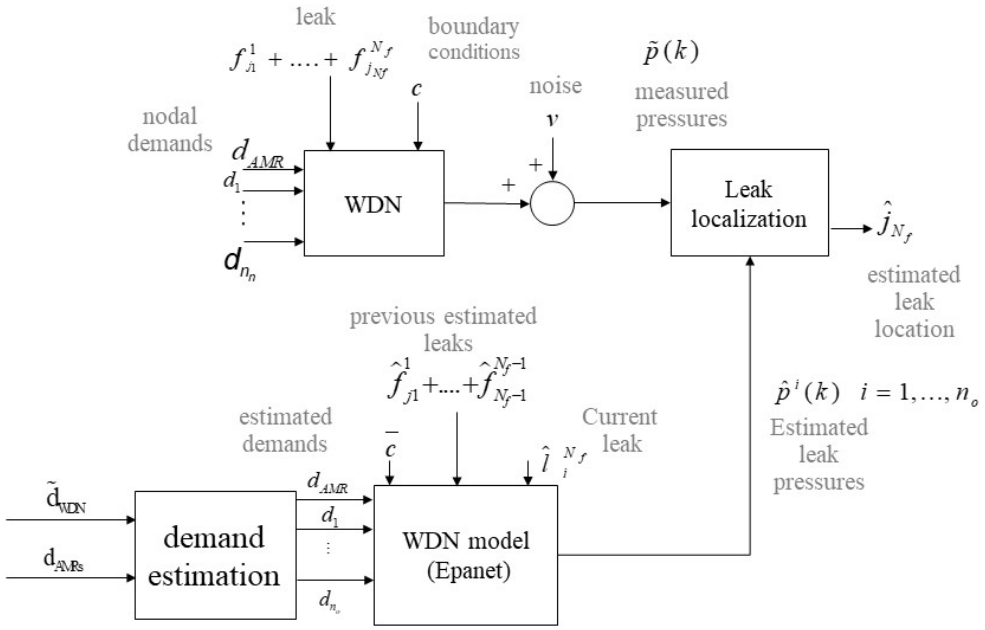


Figure 5.1: Leak localization scheme.

The model-based approaches use a simulating hydraulic network to simulate theoretical pressure values caused by all potential leaks and compare them with the DMA's measured pressure values.

The leak localization method relies on the scheme depicted in Figure 5.1. This scheme focuses on a solution with Automated Metered Readings (AMR) that provide the demand consumption of some users in the network, although nothing restricts using the method only with reliable forecast demands. The proposed method takes into account scenarios where the system has multiple leaks.

The model is fed with all available AMR measurement vector $\mathbf{q}_{AMR}(k)$ whose components are the measurements of user demands in some nodes at time instant k: $d_1(k), ..., d_{n_m}(k)$ with $0 \leq n_m \leq n_o$ being n_o the number of inner nodes in the network (i.e. the number of node minus the number of inlets, $n_o = n - n_I$). In addition it is necessary also an estimation of the demand in the rest of the nodes $(\hat{q}_{n_m+1}(k), ..., \hat{q}_{n_o}(k))$. This method is capable of analyzing data with multi-leaks $(f_{j_1}^1, f_{j_2}^2, \ldots, f_{j_{N_f}}^{N_f})$ that appear at time instants $k_1 < k_2 < .. < k_{N_f}$, i.e. sequentially, in nodes $j_1, j_2, \ldots, j_{N_f}$. At a time instant $k \geq k_{N_f}$ when leak detection algorithm detects the leak f^{N_f} and provides an estimation of its magnitude, it is necessary to have detected, estimated and localized the previous leaks $(\hat{f}_{j_1}^1, \ldots, \hat{f}_{j_{N_f-1}}^{N_f-1})$ with some degree of accuracy to proceed to the leak localization $\hat{j}_{N_f}$ as it is described in Figure 5.1. For each past leak the leak estimation magnitude is added to the corresponding node demand according to past leaks localization $(\hat{j}_1, \ldots, \hat{j}_{N_f-1})$. The WDN model is run, considering boundary conditions $\bar{c}(k)$ (for instance the position of internal valves and reservoir heads and flows), n_o times. At every simulation i with an additional demand equal to the current leak estimation ${f_i}^{N_f}$ in a different node $i = 1, \ldots, n_o$ providing a pressure vector $\hat{\mathbf{p}}^i(k) \in R^{n_s}$ with the theoretical pressure values in the n_s nodes where sensors are installed considering leak scenario i. The leak localization method consist in comparing the different theoretical pressures with the actual pressure measurements $\mathbf{p}(k) \in R^{n_s}$ that are also subject to the effect of sensor noise $\mathbf{v}(k)$. In order to increase the accuracy of the leak localization methods estimations and measurements are considered in a sliding window of length W.

For this purpose the theoretical pressure values in the time window W are stored in the pressure estimated matrices, $\hat{\boldsymbol{P}}^i(k) \in R^{n_s \times W}$ $i = 1, ..., n_o$ defined as

$$\hat{\boldsymbol{P}}^i(k) = \begin{pmatrix} \hat{\mathbf{p}}^i(k-W+1) & \ldots & \hat{\mathbf{p}}^i(k) \end{pmatrix}, \quad i = 1, ..., n_o \tag{5.1}$$

On the other hand, a matrix $\boldsymbol{P}(k) \in R^{n_s \times W}$ with the pressure measurement values in the time window W is built

$$\boldsymbol{P}(k) = \begin{pmatrix} \mathbf{p}(k-W+1) & \ldots & \mathbf{p}(k) \end{pmatrix} \tag{5.2}$$

Then, the leak location method is based on comparing the matrix $\boldsymbol{P}(k)$ with matrices $\hat{\boldsymbol{P}}^i(k)$ $i = 1, ..., n_o$ to find the most probable leak scenario. This comparison is made by applying:

$$\hat{j}_{N_f}(k) = \text{arg}_i \min \sum_{j=0}^{W-1} ||\boldsymbol{p}(k-j) - \hat{\boldsymbol{p}}^i(k-j)||_2 \qquad i = 1, ..., n_o \tag{5.3}$$

5.2.1 Case study

The proposed method was applied to the L-Town network, seen in Chapter 2.7. As explained, the network can be divided into three-zone: Zone C has AMRs installed in 89% of nodes being possible to generate an accurate hydraulic model. Zone B is connected to Area A by a pressure reduction valve (PRV), and there is only one pressure meter in the area, making it more challenging to locate the leak. And Zone A has a high density of pressure sensors (29 in total) distributed through the area, and the hydraulic model during the year is losing accuracy since the behavior of demand in the node change.

The proposed method was developed considering the extra information of the AMRs. Still, it is not limited by the same if the hydraulic model has high accuracy, the results of locating the leak will be satisfactory. The case study will analyze the three zones, starting with Zone C, which has the AMRs, and following with Zone A and B, which contains more uncertainties in the hydraulic model.

Area C

As seen in Chapter 2.7 the Zone C in the L-Town network has 92 nodes among these 82 with AMR devices, and three pressure sensor (see Figure 2.6). The information of the year 2018 it will be used in this case study. The first step that need to be done is the leak detection during the year.

The following information has been provided for 2018: the inlet water flow, the AMR measurements for 82 residences during the year, and the time for the leak fix on 12th August. The AMRs measurement corresponding to 89% of the residences, them a more accurate forecast de-

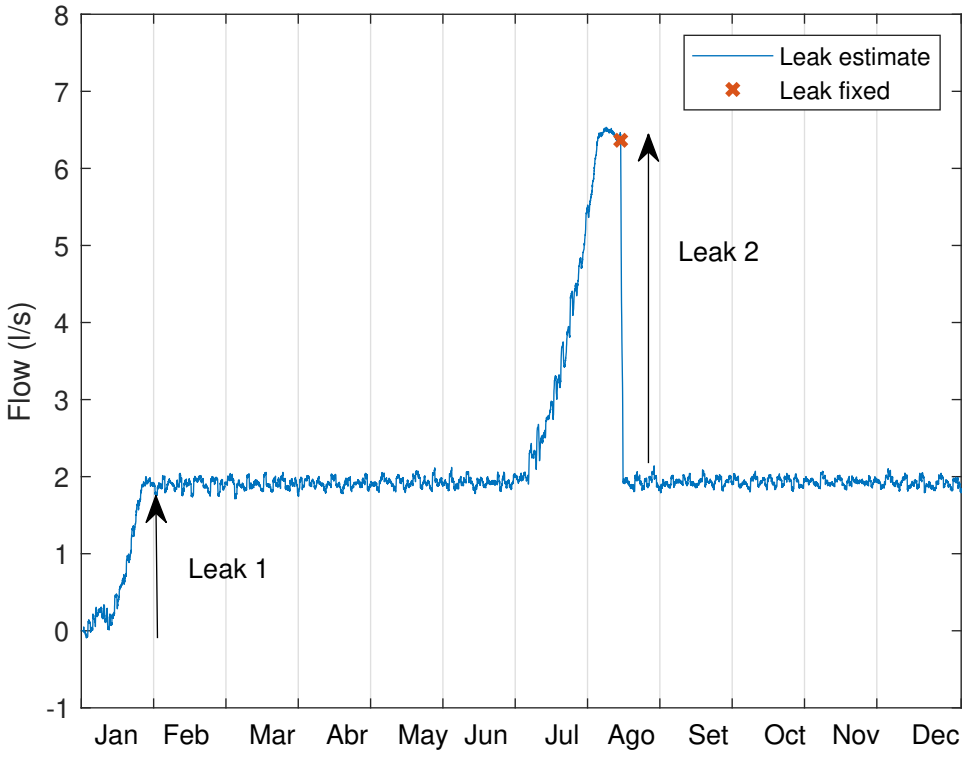
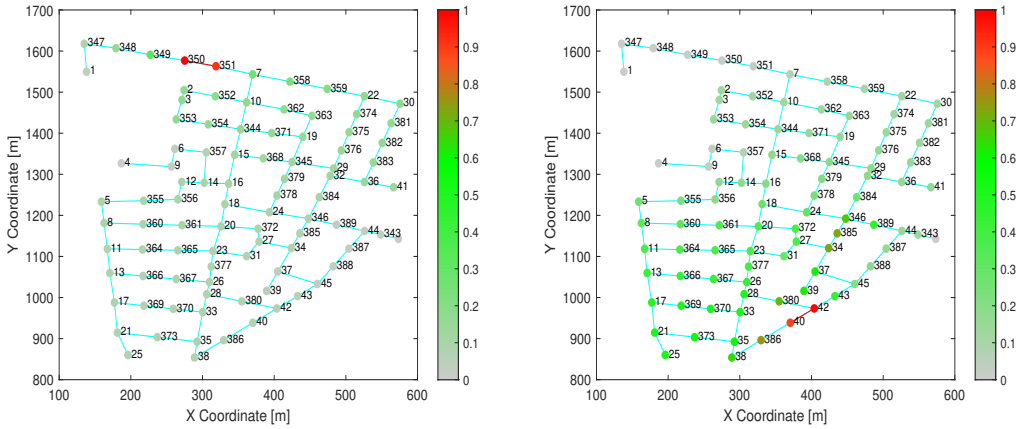


Figure 5.2: Analyze inlet flow Area C

mand, can be created. The leak detection seen in Chapter 4 is used to detect the moment of start leaking. Using the first week of the inlet flow and the measurements of the AMRs, a constant K has been created, being the percentage value between both flows. The following Equation shows the demand forecast for this area:

$$\hat{y}(k) = K \sum_{i=1}^{n_m} AMR_i(k) \tag{5.4}$$

where $AMR_n(k)\ n = 1, ..., n_m$ are the flow measurements at instant k of the n_m AMRs installed in Area C. The leak detection method has been applied considering $T_s = 1h$, $T = 24h$ and $W = 24h$, the estimated threshold Δ_W = 0.3 l / s that is much lower than the leaks present in Area C. Using the proposed method furnished the results shown in Figure 5.2: a leak of $2l/s$ at the beginning of the year, which was not fixed. And another leak at the beginning of July with a magnitude of $4.5l/s$ and fixed in August.



(a) First leak localization estimation (b) Second leak localization estimation

Figure 5.3: Leak localization candidates in Zone C

Individual analysis of each leak must always be performed, starting with the first leak detected. For the analysis made in the first leak, the interval between February 14th and March 13th was used, with the leak magnitude of $2l/s$. Figure 5.3.(a) shows the probability of leak location for each node by means of the index value $\gamma^i(k)$ defined as

$$\gamma^i(k) = \frac{||\boldsymbol{P}(k) - \hat{\boldsymbol{P}}^i(k)||_2 - \nabla P^{min}(k)}{\nabla P^{max}(k) - \nabla P^{min}(k)} \tag{5.5}$$

where the $\nabla P^{max}(k)$ is the maximum value of the comparing between matrix $\boldsymbol{P}(k)$ and matrices $\hat{\boldsymbol{P}}^i(k)$, $i = 1, ..., n_o$:

$$\nabla P^{max}(k) = \max ||\boldsymbol{P}(k) - \hat{\boldsymbol{P}}^i(k)||_2 \qquad i = 1, ..., n_o \tag{5.6}$$

and $\nabla P^{min}(k)$ is the minimum value, computed as:

$$\nabla P^{min}(k) = \min ||\boldsymbol{P}(k) - \hat{\boldsymbol{P}}^i(k)||_2 \qquad i = 1, ..., n_o \tag{5.7}$$

The greener is the lower $\gamma^i(k)$, and the redder is, the higher $\gamma^i(k)$ which is the increased chance of leaking. For the Figure 5.3.(a) leak, the most significant index of leak location is

between nodes 350 and 351; the two nodes connected with the pipe that has the leak.

Continuing to analyze the second leak detected, with an estimated leak magnitude of $4.5l/s$. Only a few days of analysis have been chosen between 14th to August 18th because the more constant the leak is more accurate the result will be; in this case, a visual selection has been used utilizing the Figure 5.2. As explained, it is necessary to add the previous non-fixed leaks in the estimated nodes; then, using the first analysis, the node 351 has been chosen to be added $2l/s$ to its demand. Figure 5.3.(b) shows the leakage index $\gamma^i(k)$ for the second leak. Pipe number 31 is indicated in red, which is the exact location of the leak; the most probable nodes are 40 and 42 with red color, which is the right position between pipe 31. In conclusion, the method had a satisfactory result for Zone C, which has an accurate hydraulic model.

Area A and B

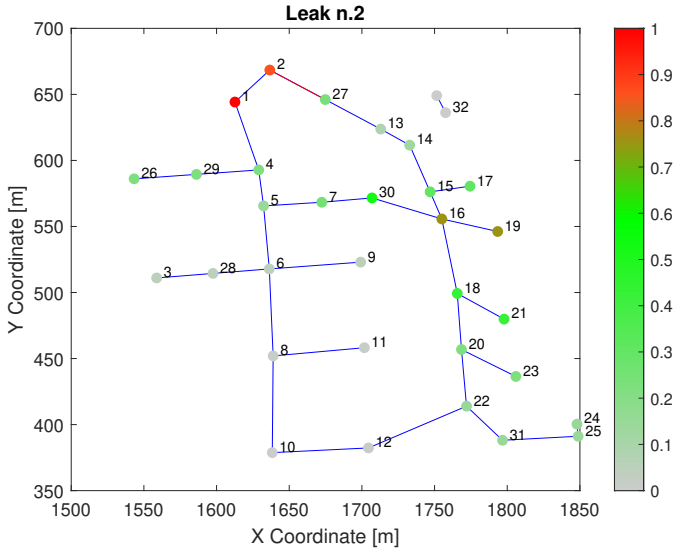
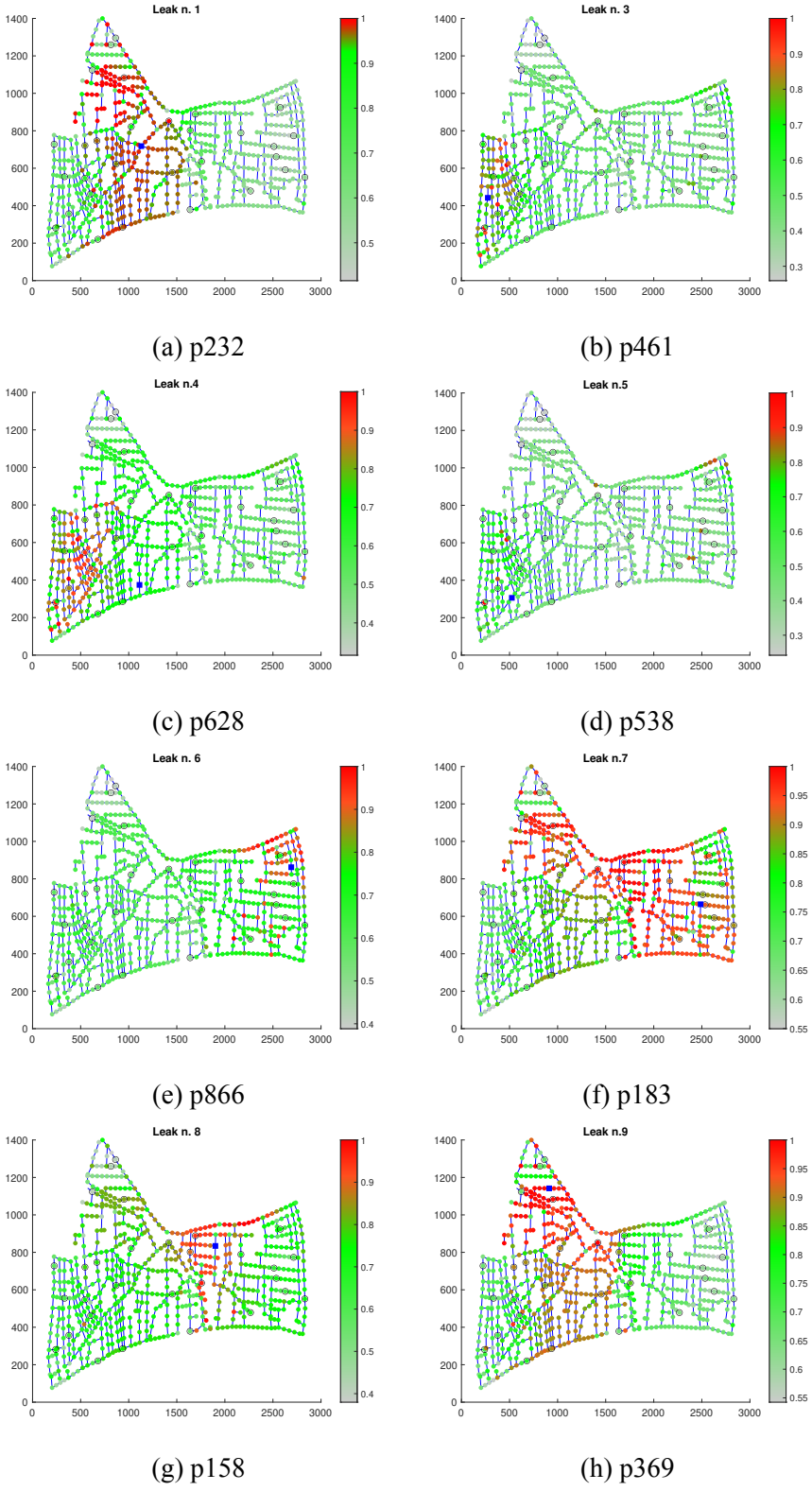


Figure 5.4: Leak localization in Zone B on pipe p673

Zone A and B are larger areas with 30 pressure sensors in total (one is located in zone B). The two water inlets to the network are located within this zone A, and a tank is filled from this area to provide water to Area C. The history of leak detection and estimation in these zones is shown in Figure 4.7, already studied in Chapter 4 on leak detection techniques. The proposed method will investigate the nine reported leaks to analyze the location. As there are no AMRs in these zones, the normalized proportional outflows in nodes throughout the year will not change;



(a) p232 (b) p461

(c) p628 (d) p538

(e) p866 (f) p183

(g) p158 (h) p369

Figure 5.5: The result of the leak localization for the six first leaks referring to the year 2018 in Zone A and B

using the information from the first week to estimate the these parameters for each node. The leak detection is studied in Chapter 4.3; the procedure of the leak localization is realized after the leak detection shown in Figure 4.11. Figure 5.5 shows the leak localization results using the proposed method. The blue squares indicate the location of the leak in Zone B (Figure 5.4) and Zone A (Figure 5.5), a highlighted red pipe represents the location of the leak.

Among all the leaks studied, the leak of Zone B has the best result represented in Figure 5.4, with the node most likely to leak the node connected to the fault pipe. Leaks occurring in Zone A have a larger node area with a probability of leaking because of the node demand uncertainty. Figure 5.5.(a) shows the leak n.1 with a fault in the pipe p232, the result presents a node area around the pipe that has the probability of leaking, but the most highlighted node is located one zone above the leak.

Figure 5.5.(b) refers to leak n.3 having few nodes with a high probability of being the leak's location, all in an area close to the failed pipe. Leaks n.4 and.5 are happening practically simultaneously; it is challenging to isolate the two. The result of Figure 5.5.(c)(d) shows that the leak in the p538 pipe affects the analysis more than the leak n.4, the worst outcome since the result was in a different area. The leak n.6 occurs in the p866 pipe display in Figure 5.5.(e), having a good result since the nodes with the highest chance of leaking are close to the defective pipe. The last six leaks are of burst type, depicted in Figure 5.5.(c-h), these leaks have small simultaneous leaks occurring in the same instant. Due to the difficulty of detecting and locating these small leaks, an additional error was added to the hydraulic model affecting the leak localization result, as seen in Figure 5.5.(f) that has the biggest red area.

The model-based leak location method has a weak point: if a leak is not detected and located correctly, the error will be propagated to the subsequent leaks, making the method less reliable.

5.3 Data-driven Leak Localization

The leak localization method based on a data-driven technique is formulated for the networks that do not have a hydraulic model, or despite having a hydraulic model, this is not accurate. As seen in the previous result, Zone A is a network with a large size and has uncertainty in demand

nodes. The accumulation of uncertainties in the hydraulic model causes the leak localization not to be satisfactory. For example, another type of uncertainty happens when the WDN has multi-leaks, and the analysis of leaks is not done in chronological order, or if previous leak locations and estimations are not provided with reasonable accuracy, the hydraulic model will accumulate these errors making the result of future leaks more doubtful.

For the reasons mentioned above, this section focuses on leak localization based on data-driven techniques using the information of the flow and pressure measurements and topological information of the WDN. The leak localization is done when a leak is detected, as seen in the flowchart in Figure 1.2. In addition, it is assumed that the leaks can only happen in the nodes of the network (as considered in [91], [72], or [16]), making the number of nodes equal to the number of a potential leak. The nodes correspond to water users, pipe junctions, and other structures like hydrants. However, if the number of nodes will not provide a representative discretization of the network, some artificial nodes could be considered. In this Section three data-driven method to leak localization are presented.

5.3.1 Leak localization based on cluster technique

The first data-driven leak localization method was developed to reduce the leak location area, which aims to select pre-defined zones with the highest chance of leaking. It requires hydraulic information such as pipe diameter and length, measurements from pressure sensors installed in the system, and some pressure sensors measurement obtained from leak events simulated in the field by the water utilities. The leak localization proposed in this section is divided into two parts: The first part focus on the generation of the clustering zone in the WDN. And the second part analyzes an index correlation calculated using sensor pressure measurements.

The WDN is represented by applying the Graph theory, as explained in Chapter 2.2.2. And the number of clustering will interfere with the result of the leak localization problem; if the WDN has more clustering, the best split will be the system nodes. Thus, the leak localization method presents two options for the cluster number. The first is the number of clusters equal to the number of pressure sensors n_s, $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{n_s}\}$. The second is a additional ρ clusters, in this way the number of cluster are $n_s + \rho$, $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_s, C_{(n_s+1)}, \ldots, \mathcal{C}_{(n_s+\rho)}\}$. The

contributions of this work are:

The proposed method has the following features:

- It is applicable to measurements temporal information using Bayesian time reasoning.
- Leak-free information can be obtained using historical data provided by the water company. In addition, nominal information can be used from the days prior to the appearance of the leak to deal with the uncertainty of water demand between different scenarios.
- It provides the cluster most likely where a leak is identified. It makes it easy to locate the leak in the real network.

Clustering

A water distribution network can be described by a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, as seen in Chapter 2.2.2. The edge e_{ij} is associated with a cost value related to the diameter and length of the pipe, as seen in work [88]:

$$w_{ij} = \frac{L_{ij}}{D_{ij}^5} \tag{5.8}$$

where w_{ij} is the cost associated to edge e_{ij}, L_{ij} and D_{ij} are the length and diameter of the pipe, connecting nodes i and j, both in meters [m].

The network clustering process is divided into two phases: The first cluster has the objective to divide the n (nodes) observations into $n_s (\leq n_o)$ sets $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{n_s}\}$ concerning the hydraulic distance of the sensors. For this, the set of data points $x = [x_1, ..., x_{n_s}]$ is generated for each node with the minimum distance from the node to the sensors:

$$x_j = d^W(i, j) \tag{5.9}$$

where $i = 1, ..., n_o$ and $j = 1, ..., n_s$ and the distance $d^W(i, j)$ is the minimum sum of weights

across all the paths connecting i and j.

The main objective of the clustering algorithm is to minimize the sum of distances between the points and their respective cluster centroid. The objective function is:

$$arg\min_{\mathcal{C}} \sum_{j}^{n_s} \sum_{x \in \mathcal{C}_j} ||x - \mu_j||^2 \tag{5.10}$$

where μ_j is the center of the clustering, which is the value of the distance from a sensor to the other sensors

$$\mu_j = [d^W(j,1), ..., d^W(j,n_s)], \qquad j = 1, ..., n_s \tag{5.11}$$

This first clustering makes it possible to analyze how many nodes there are in each zone. Of course, the ideal is to have clusters with homogeneous numbers of nodes. However, in real cases where pressure sensors are already installed in the network, similar numbers of nodes in the clustering may not be satisfied.

The second phase of clustering is designed to solve this problem. More ρ numbers of nodes are chosen to be a clustering center. So the goal is to divide n_o (inner nodes) observations into $n_s + \rho (\leq n_o)$ sets $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_s, C_{(n_s+1)}, \ldots, \mathcal{C}_{(n_s+\rho)}\}$. The extra ρ nodes are chosen on the frontier between clusters obtained in the first phase. The number of ρ nodes chosen has to be analyzed with the water company because the ρ selected in the most node containing sensors will need extra experiments information in the real network of events with leakage. Consequently, the exact number of ρ chosen will vary according to the water company's objective, as a previous cost-benefit study must be carried out, which will be explained better in the next section. The second clustering will be the same as Equation (5.10) with the new center:

$$\mu_j = [d^W(j,1), ..., d^W(j,n_s+\rho)], \qquad \forall j = 1, ..., n_s + \rho \tag{5.12}$$

Leak localization

The proposed leak location method aims to reduce the network area to the pre-defined regions with the highest chance of leaking. The method analyzes residuals from the pressure sensors already installed in the network. This process has an important role in maintaining easily the network and supporting the operator in its maintenance task.

The estimation pressure considering a leak-free scenario is done with the historical data analysis so that the network boundary conditions c are similar (e.g., reservoir pressures, flow, and consumer demands). A study of the pressure measurement of previous days or weeks where the system was considered without failures can be done. In this way, it can guarantee a better precision of the estimated measure of pressure. The residual pressure in internal nodes that contain a sensor, defined in (2.15), can be adapted as

$$r_i = \hat{p}_i(c) - p_i(c^{l_j}), \quad \forall i = 1, ..., n_s \tag{5.13}$$

where $\hat{p}_i(c)$ is the pressure estimation considering boundary conditions c in a leak-free scenario. On the other hand, $p_i(c^{l_j})$ is the pressure value measured by the inner pressure sensor i under boundary conditions c^{l_j} (the same heads and inflows in inlets as in c but with a leak in node j).

Following the ideas in [89] positive residuals can be obtained from the following transformation:

$$\bar{r}_i = r_i - \min(r_1, ..., r_{n_s}) \quad \forall i = 1, ..., n_s \tag{5.14}$$

Then, the likelihood index ∂_i is calculated as the normalization of the $\bar{r}_i$:

$$\partial_i = \frac{\bar{r}_i}{\sum_{j=1}^{n_s} \bar{r}_j} \quad \forall i = 1, ..., n_s \tag{5.15}$$

A simple leak localization method can be defined only with residual analysis of Equation

(5.14) (see [48, 79]). In this case, only the first phase of clustering is utilized, with the center of the clustering expressed in Equation (5.10). The selected area with the leak is the one that presents the component with the maximum size, i.e.,

$$\hat{C} = \underset{i \in 1,...,n_s}{\arg\max}\, \bar{r}_i \tag{5.16}$$

This simple method only needs the information from the pressure sensors being a good reference point to analyze the improvement of the method. The dependency on the sensor's availability and its positioning in the network is a limitation to this simple method. For example, if the sensors are not well distributed in the system, a clustering one area may contain many nodes and others few. With that in mind, extra data information can balance the number of nodes in each cluster.

The effect of a leak on a node causes correlation factor between sensors (see. [96]), making it possible for objects within the same cluster to be as similar as possible (i.e., high intra-class similarity), while objects from different clusters are as different as possible (i.e., low inter-class similarity). Knowing this, it is possible to select ρ strategic points in the network (nodes) to be a new clustering center and thus balance the number of nodes in the cluster. Therefore, it is necessary to have leak-data scenarios in each node with sensors and selected centroids.

The experiment of leak-data scenarios produced in each node with sensors and selected centroids is called a leak event, and a signature vector of events $\boldsymbol{\nu^e}$ can be generated for each event $e \in 1, ..., n_s + \rho$. Figure 5.6 shows the sensor correlation scheme for each leak event generated with the signature vector of events $\boldsymbol{\nu^e} \in \mathbb{R}^{n_s}$. Being $\boldsymbol{\nu^e} = [\nu_1^e, ..., \nu_{n_s}^e]$ with ν_i^e normalized likelihood index ∂_i of Equation (5.15).

For a given measured residual $r_1, \ldots, r_{n_s}$ computed by Equation (5.13), the vector $\partial = [\partial_1, ..., \partial_{n_s}]$ is computed and the Euclidean distance is used to analyze the distance between the measurement and the centroids of each clustering:

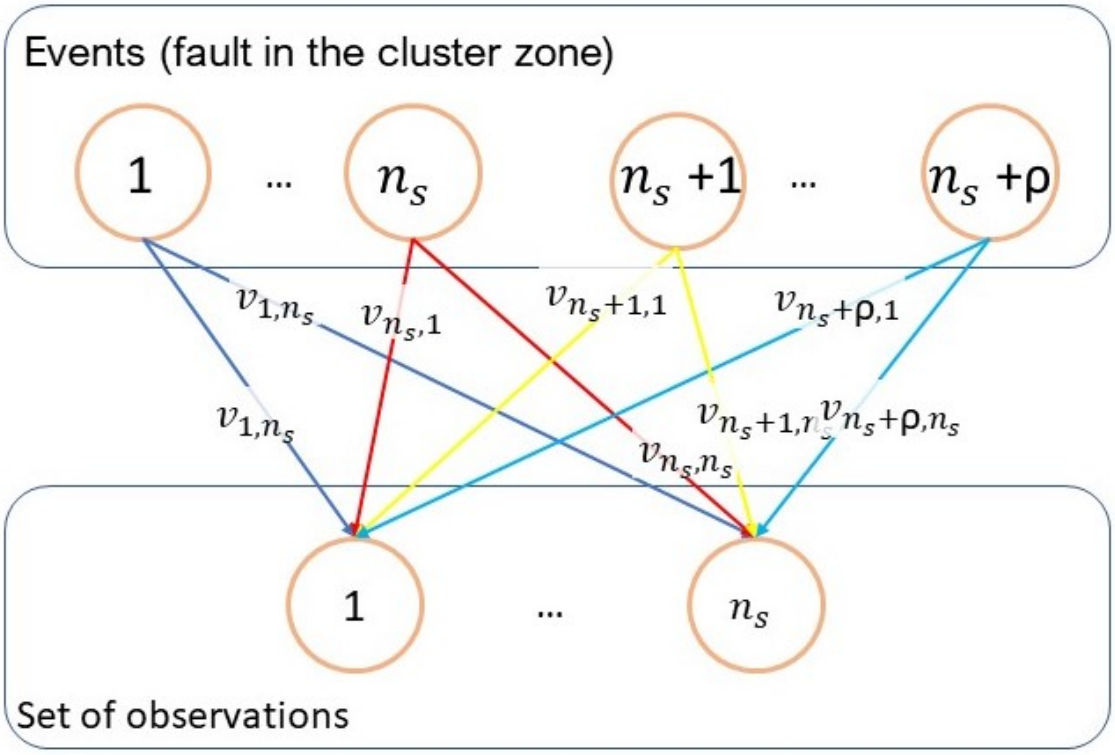


Figure 5.6: Schema of the correlation between the sensor and the leak events

$$\theta^e = \sqrt{(\partial_1 - \nu_1^e)^2 + ... + (\partial_{n_s} - \nu_{n_s}^e)^2} \tag{5.17}$$

The most probable cluster is determined as the one that provided the minimum distance of Equation (5.17)

$$\hat{C} = \underset{e \in 1,...,n_s+\rho}{\arg\min}\ \theta^e \tag{5.18}$$

Until now, only the single time instant analysis was studied, to improve the performance of the method and make it possible to analyze in time series the normalized θ^e information at different time instants k can be considered by applying Bayes' rule as:

$$P^e(k) = \frac{P^e(k-1)\theta^e(k)}{\sum_{l=1}^{n_s+\rho} P^l(k-1)\theta^l(k)} \tag{5.19}$$

where $P^e(k-1)$ is the prior probability whose initial value has to be determined (for example $P^e(0) = 1/(\rho + n_s)$). Then, the leak node localization can be estimated by using posterior leak correlation by:

$$\hat{C}(k) = \underset{e \in \{1,\ldots,s+\rho\}}{\arg\max} \{P^e(k)\} \tag{5.20}$$

5.3.1.1 Case Study

The Modena Network simplified version of the real WDN from the Italian city Modena is represented in Figure 2.4 in Chapter 2.7. This large-scale network comprises 268 junctions (nodes) connected through 317 pipes and served by four reservoirs.

To test the evaluation of the proposed leak location method, artificial data were generated under different conditions using a hydraulic simulator Epanet 2.

Using the hydraulic simulator, data was generated with uncertainties related to the consumer's nodal demand created by applying a random value uniformly of $10[\%]$ of the nominal demand value in the default value of the Modena benchmark in the Epanet repository and considering that the exact size of the leak is unknown, but is contained in the range of 5 and $50[l/s]$, representing 2% to 20% of the network consumption.

For each leak scenario, of the 268 nodes, a leak is simulated lasting 72h. The sampling rate is 10 min, but measurements are filtered hourly to reduce the impact of uncertainties in the diagnostic phase.

Six scenarios were generated with different sensors number to analyze the previously explained method. The first column of Table 5.1 displays the scenarios only with sensor information installed on the network. The second column is the information of the corresponding number of nodes that comprise the cluster clustering. Equation (5.10) and (5.11) were used to generate the clustering. The number of sensors installed in the network is 3, 5, and 8, with two different cases demonstrating the effect on the result concerning the positioning of the pressure sensors.

Table 5.1: Scenarios using only pressure sensors

Case	Nodes with sensors	Number of nodes in the clustering
1	**11 50 80**	64 109 95
2	**10 44 93**	83 74 111
3	**9 65 94 109 247**	63 103 45 23 34
4	**10 63 113 247 250**	55 45 90 45 33
5	**10 23 45 62 64 94 119 259**	19 24 48 28 25 53 47 24
6	**18 35 63 100 153 158 248 236**	45 32 31 27 30 30 34 39

Table 5.2 shows the scenarios demonstrated in Table 5.1 plus the extra nodes, ρ, which will be the additional center of clustering. The centroids are the n_s values of sensors installed in the n_s network, highlighted with the text in bold, and the ρ selected nodes are listed after the sensors, with $\rho + n_s$ being the centroids referred to in Equation (5.12). The value of ρ varies from one scenario to another to analyze how ρ affects the leak localization result. In addition, the second column represents the number of nodes of each cluster after applying Equation 5.10 and (5.11). It is not always possible to balance the number of nodes in the areas, but the number of nodes in the areas is smaller.

Table 5.2: Scenarios using $n_s + \rho$ centroids

Case	Node set to be the cluster center	Number of nodes in the clustering
1	**11 50 80** 120 160	69 44 88 34 33
2	**10 44 93** 37 111 151 224	63 41 48 12 14 56 34
3	**9 65 94 109 247** 196 246	55 49 41 43 25 22 33
4	**10 63 113 247 250** 130 189 231 267	38 34 41 29 21 28 30 17 30
5	**10 23 45 62 64 94 119 259** 113 139 245 222	19 20 22 22 26 17 18 26 33 28 20 17
6	**18 35 63 100 153 158 248 236** 85 113 178 166 214 232	20 25 25 22 16 14 21 17 17 21 11 20 25 14

As noted before, the leak location method using the maximum residual variance seen in

Equation (5.13) is a good point of comparison of method improvement. Therefore, the ATD was calculated applying the maximum residual method with the scenarios of Table 5.1. Then, the proposed leak localization method was applied to the scenarios in Table 5.2.

In the simplified WDN of Modena, in the example of Case 2 of Table 5.1 where three pressure sensors are considered, the computed clusters are depicted in Figure 5.7.(a). Following the clustering of the network in Case 2 of Table 5.2 contains the same pressure sensor position with four additional centroids. It is depicted in 5.7.(b). The nodes containing pressure sensors are highlighted with a red circle, and the ρ extra centroid is highlighted with a black circle. The clustering area with sensors is reduced by up to 56%, allowing a more homogeneous division of areas in the WDN.

Figure 5.8 shows the result of the ATD (node) of the two analyses, using the Bayes temporal reasoning in both cases with evolution of 48h. Figure 5.8.(a) is the result of the scenarios of Table 5.1 using the maximum residual approach, and Figure 5.8.(b) is the result of Table 5.2 scenarios applying the proposed method.

The analysis of the cases pairs that contain the same number of sensors in Table 5.2, Cases $\{1, 2\}, \{3, 4\}$, and $\{5, 6\}$, (results in Figure 5.8.(a)) demonstrates the importance of a previous study for sensor placement. The ATD result improves with the same number of sensors in the network, which can reach up to one node difference. Moreover, cases $\{2, 3\}$ have a similar result even though the two have a difference of 2 sensors installed in the network.

The results in Figure 5.8.(b) display the evolution of the ATD of the presented method. Comparing the correspondent Case in Figure 5.8. (a) an improvement in the ATD index is noticed in almost all cases. Only Case 1 presents better results using the maximum residual method due to the sensor placements on the network. Cases $\{1, 2\}, \{3, 4\}, \{5, 6\}$ have the same number of sensors with only varying values of ρ. It is shown it is possible to obtain better results by increasing the value of ρ.

The same analysis was made analyzing the ATD with the distance in kilometers. This study is critical because as the distances between the pipes are not uniform, the variation of the ATD in the node may not indicate the actual improvement. Figure 5.9.(a) shows the result of the

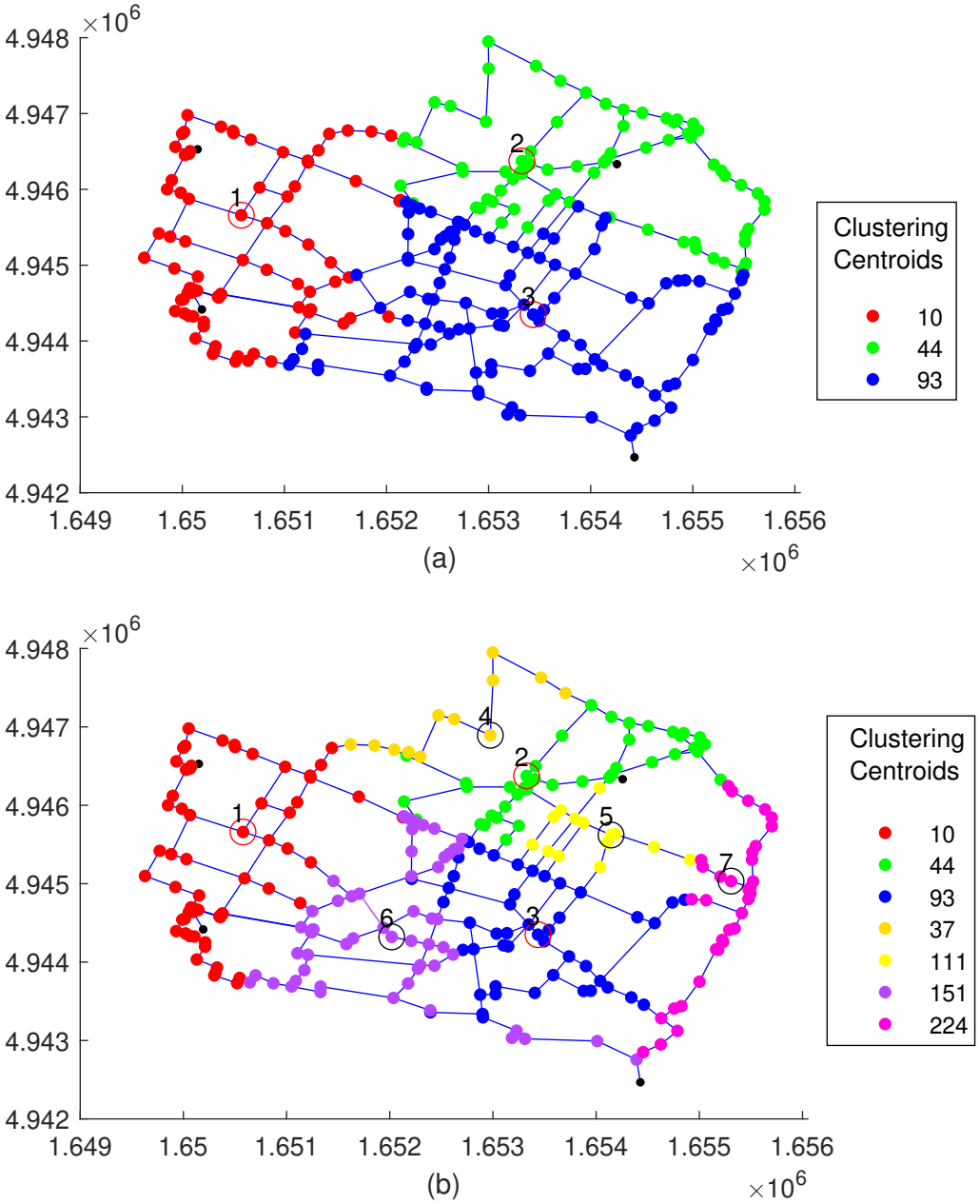


Figure 5.7: Clustering of Case 2: (a) Table 2, with only pressure sensors highlighted with a red circle (b) Table 3, with sensors and 4 extra centroids, highlighted with a red and black circle

scenarios of Table 5.1 using the maximum residual approach. It demonstrates the importance of comparing the ATD in nodes and kilometers. The cases $5, 6$ have a similar value when they are compared in kilometers and in nodes, the case 6 has a slight improvement. Figure 5.9.(b) is the result of Table 5.2 scenarios applying the proposed method. This result shows that in Case

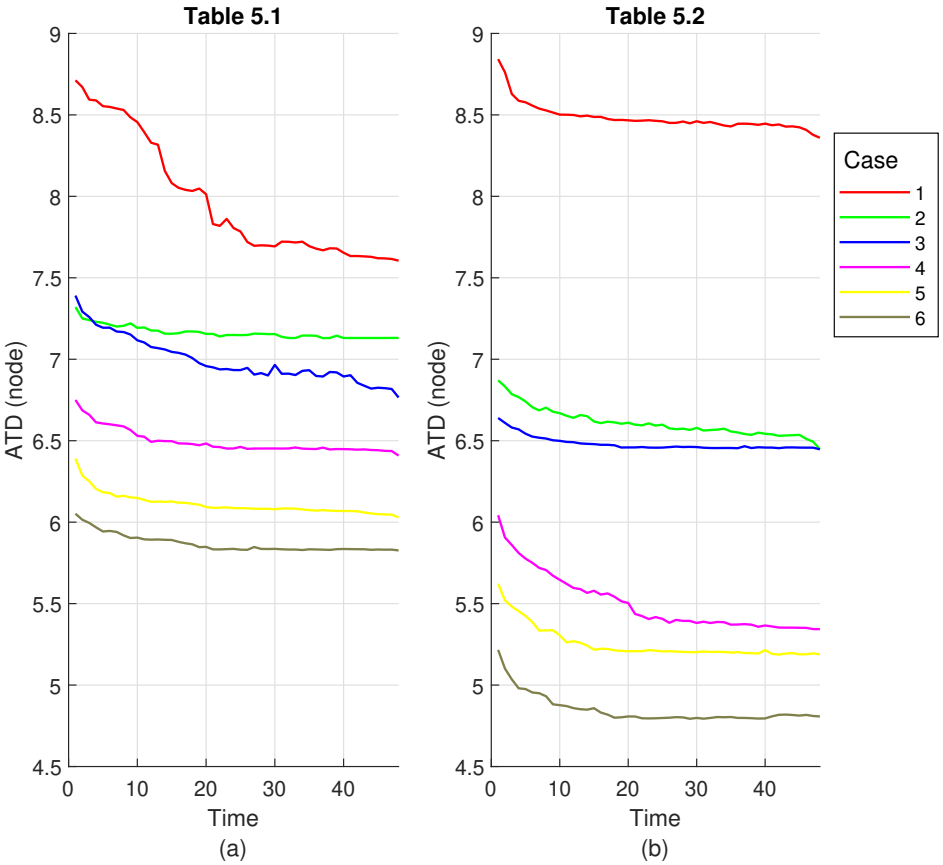


Figure 5.8: Evolution of ATD (node) when using the Bayes temporal reasoning (a) scenarios of Table 2 using the maximum residual approach (b) scenarios of Table 3 using the proposed method

1, even with similar results compared to the maximum residual, it remains more constant with a better result in the time series. In the cases $\{2, 3\}$ that has an improvement in kilometers distance present a similar value in the comparison of the ATD in node.

As remarked, the positioning of the pressure sensors in the network affects the results dramatically. Therefore, a pre-analysis of the sensor placement, setting the best nodes to contain sensors and the best node to be the clustering centroids, can improve the results illustrated.

5.3.2 Leak Localization the common path study

The second leak location methodology aims to move from the cluster study to a node selection with more probability of leaking. This section presents a data-driven leak location using

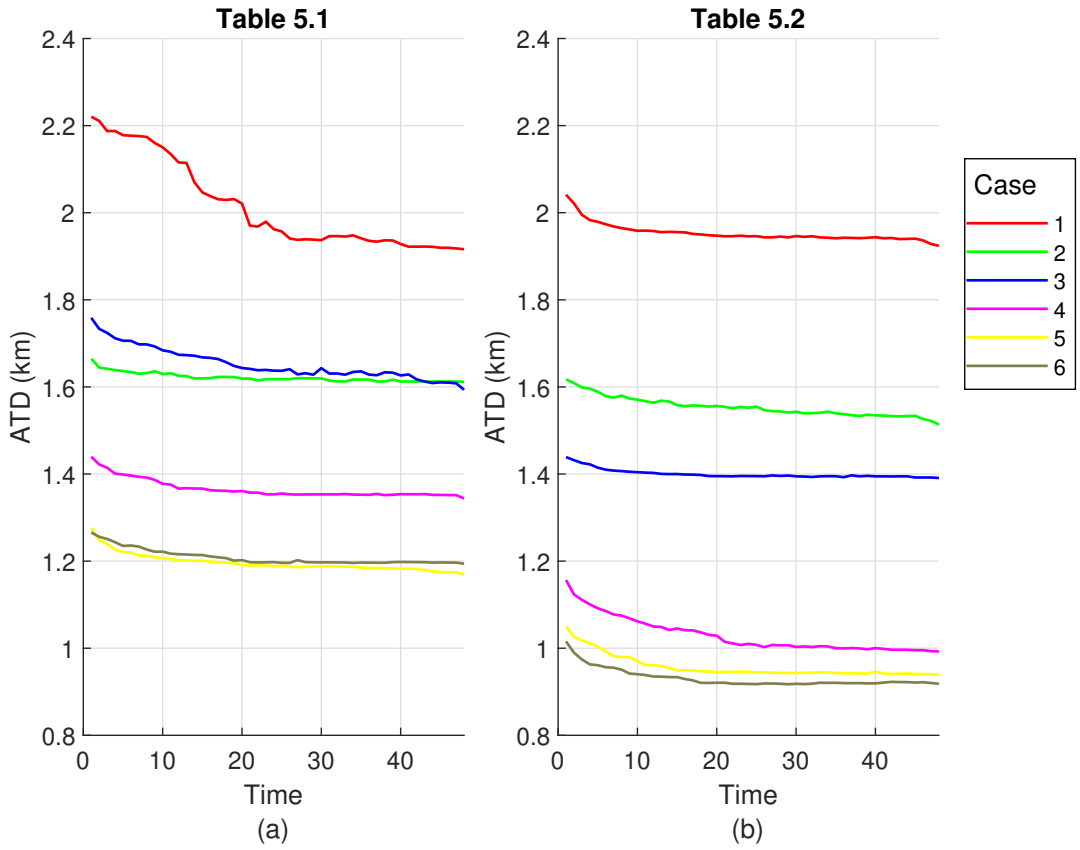


Figure 5.9: Evolution of ATD (km) when using the Bayes temporal reasoning (a) scenarios of Table 5.1 using the maximum residual approach (b) scenarios of Table 5.2 using the proposed method

topological information that provides the most likely paths for extra flows produced by leaks. A new incidence factor from every combination of nodes and sensors is computed with this information. Every incidence factor determines how a leak in a particular node affects a specific pressure sensor. On the other hand, historical data is used to calculate non-leak pressure estimations at sensed inner nodes. Residuals are generated using the comparison between these estimations and leak pressure measurements. Incidence factors are integrated with residuals in likelihood indexes to give the most probable leak node in a leak scenario.

The network can be represented as a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ explained in Chapter 2.2.2 with $\mathcal{V} = \{v_1, \ldots, v_n\}$ is the set of vertices that represents connections between the components of the network, additionally the last $\{v_{n_o+1}, ..., v_n\}$, represent the vertices of the system's input, being n_I the number of the inlets, with $n_I \geq 1$. The elements of the set $\mathcal{E} = \{e_1, \ldots, e_m\}$ are

the edges, which represent the m pipes in the network. Two leak localization methods will be proposed. The first one will only use available measurements, and its diagnosis will point out to one of the inner pressure sensors installed in the WDN. Therefore, the detected leak should be in an area around this sensor (cluster). The second method will combine the information of the first method with the topological information: characteristics of the pipes and connections between the nodes of the WDN, in a likelihood index that will allow the leak localization at the node level.

Leak localization at cluster level

As stated before, the proposed leak localization is applied after the leak detection. In addition to inlet pressure and flow sensors, it is assumed that n_s pressure sensors are installed at different inner nodes. Consider a leak l_j acting on the node j of the network, and the used measurements are assumed to be captured under a leaky situation. Also, admitting leak-free historical data of all the sensors is available. The residual pressure in internal nodes that contain a sensor, can be defined as Equation (5.13) and the positive residuals can be obtained applying Equation (5.14), as seen in the previous section.

Then, as the leak localization can be achieved by determining the residual pressure component with maximum size (see [48, 79]), leak localization can be formulated, similar to Equation (5.3.1) defined in previous section, as:

$$\hat{j} = \underset{i \in \{1,\dots,s\}}{\arg\max} \{\bar{r}_i\} \tag{5.21}$$

Notice that the result of leak localization method (5.21) is one of the n_s pressure sensor locations.

Then, the leak localization results in $\hat{j}$ point out not only to sensor location j but also to the nodes that produce a higher incidence to this sensor than the other sensors (cluster j).

Leak localization at node level

Considering Hazen-Willians Equation (2.11) for every pipe (edge e_z) it can be defined a resistance R_z:

$$R_z = \frac{10.7 \cdot L_z}{\rho_z^{1.852} \cdot D_z^{4.87}} \tag{5.22}$$

Among the multiple pipe path that can connect every pair of nodes ij, it can be computed a path $\mathcal{P}_{ij}^{min}$ with a minimum total resistance R_{ij} by means of :

$$\mathcal{P}_{ij}^{min} = \arg\min_{\mathcal{P}_{ij}^{(k)} \in \mathcal{P}_{ij}} \sum_{e_z \in \mathcal{P}_{ij}^{(k)}} R_z \tag{5.23}$$

where $\mathcal{P}_{ij} = \{\mathcal{P}_{ij}^{(r)}, \ldots, \mathcal{P}_{ij}^{(e)}\}$ denotes the set of paths connecting nodes i and j.

On the other hand, the minimum path from the n_I inlets to a node j , $\mathcal{I}_j^{min}$, can be obtained by applying the computation of the minimum paths from the n_I inlets to node j by means of (5.23) and determine what is the one with the minimum resistance among the n_I paths.

When a leak is produced in node j, $\mathcal{I}_j^{min}$ is the most probable path for the extra flow produced by the leak. So the effect of a leak in node j to sensor n_{si} depends on the intersection of the paths from inlets to node j and the node where the sensor is located n_{si}: $\mathcal{I}_j^{min}$ and $\mathcal{I}_{n_{s_i}}^{min}$. To quantify the degree of incidence of the leak to the sensor, an incidence factor $g_{j,n_{si}}$ is defined as:

$$g_{j,n_{si}} = R_{j,n_{si}}^{c} \bar{g}_{j,n_{si}} \tag{5.24}$$

where $R_{j,n_{si}}^{c}$ is the resistance of the path defined by $\mathcal{I}_j^{min} \cap \mathcal{I}_{n_{si}}^{min}$, the superscript c refers to the common path between node and sensors, and $\bar{g}_{j,n_{si}}$ is a normalization factor that takes into account the inverse of the resistance from the node j to the different sensors:

$$\bar{g}_{j,n_{si}} = \begin{cases} \frac{\frac{1}{R_{j,n_{si}}}}{\sum_{l=1}^{s} \frac{1}{R_{j,n_{si}}}} & \text{if } j \neq n_{si} \\ 1 & \text{if } j = n_{si} \end{cases}$$

The n_s incidence factors associated to a leak in node j, $g_{j,n_{s_i}}, i = 1, ..., n_s$ can be normalized:

$$\lambda_{j,n_{si}} = \frac{g_{j,n_{si}}}{\sum_{l=1}^{n_s} g_{j,n_{si}}} \tag{5.25}$$

where coefficient $\lambda_{j,n_{s_i}}$ determines the relative incidence of a leak in node j to sensor n_{s_i} regarding all the n_s sensors and need to fulfill:

$$\sum_{i=1}^{s} \lambda_{j,n_{si}} = 1 \quad \forall j = 1, \ldots, n_o \tag{5.26}$$

For every node $j = 1, ..., n_o$ the most sensitive sensor to a leak in this node can be computed as:

$$\hat{j} = \underset{i \in \{1,\ldots,n_s\}}{\arg\max} \{\lambda_{j,n_{si}}\} \tag{5.27}$$

The n_s clusters used in leak localization defined in (5.21) can be computed using the set of nodes that provide the same value of $\hat{j}$. The following equation is the definition of the cluster associated with the sensed node l:

$$C_l = \{v_j \in \mathcal{V} | \underset{i \in 1,\ldots,n_s}{\arg\max} \{\lambda_{j,n_{si}}\} = l\} \tag{5.28}$$

where $l = 1, ..., n_s$. Topological information of $\lambda_{j,n_{s_i}}$ and the measurement information of residuals $\bar{r}_{n_{s_i}}$ can be integrated in a parameter θ_j defined as:

$$\theta_j = \frac{1}{\bar{\theta}} \sum_{i=1}^{n_s} \lambda_{j,n_{s_i}} \bar{r}_{n_{si}} \tag{5.29}$$

where $\bar{\theta}$ is a normalization factor. Then, θ_j can be interpreted as a likelihood index, and the leak localization at cluster level defined in (5.21) can be formulated at node level as:

$$\hat{j} = \underset{j \in \{1,\ldots,n_o\}}{\arg\max} \{\theta_j\} \tag{5.30}$$

In order to improve the performance of the leak localization method, the information of the residuals at different time instants k can be taken into account applying the Bayes rule as:

$$P_j(k) = \frac{P_j(k-1)\theta_j(k)}{\sum_{l=1}^{n_o} P_l(k-1)\theta_l(k)} \tag{5.31}$$

where $P_j(k-1)$ is the prior probability whose initial value $P_j(k-1)$ has to be determined (for example $P_j(0) = 1/(n_o)$). Then, the leak node localization can be estimated by using posterior leak probabilities by:

$$\hat{j}(k) = \underset{j \in \{1,\ldots,n_o\}}{\arg\max} \{P_j(k)\} \tag{5.32}$$

5.3.2.1 Case Study

Hanoi WDN

The Network used for the case study is a reduced city's network model from Hanoi's WDN. It is composed of one inlet (reservoir), 34 pipes, and 31 nodes, represented by Figure 2.3.

To analyze the performance of the proposed approach, data with different conditions have been generated artificially using the EPANET hydraulic simulator. In order to consider realistic scenarios, some uncertainty has been added to the data [10]: the magnitude of the leak is random with a range of 25 to $75[l/s]$, i.e., between 1% to 2.5% of the average inlet flow of the WDN.

Furthermore, white noise has been combined to emulate the noise present in real measurements, and uncertainty of 10% (uniform distribution) was added in the nominal demand value.

The daily water consumption pattern used for the calibration of Equation (2.14) is shown in Figure 5.10, having four days of operation.

The sample rate is 10 min, but average hourly measurements are calculated to reduce uncertainties on the diagnostic stage.

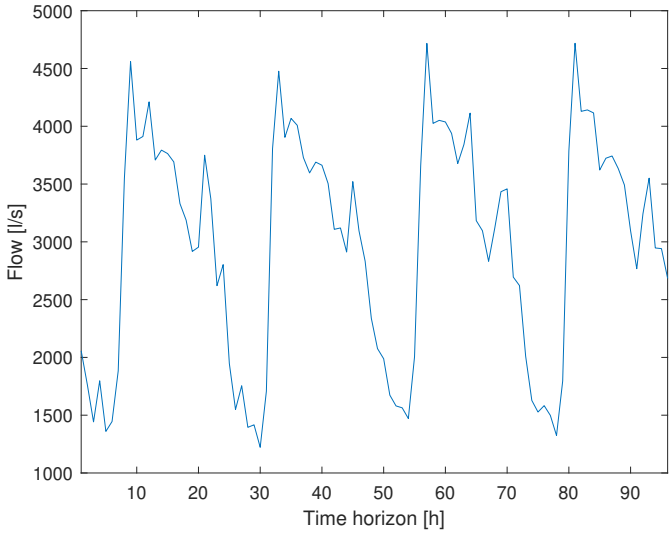


Figure 5.10: Flow consumption.

The evaluation of the performance of the proposed leak localization method at node level defined in Equation (5.32) will be analyzed utilizing Average Topological Distance (ATD), presented in Chapter 2.6.

Four cases have been considered with different quantities of sensors in the network to analyze how this affects the final result. Table 5.3 presents the distribution of the selected nodes to contain a sensor. As seen in [92], the positioning of the sensors produces different results. In this section we will not discuss the adequate sensors arrangement, they were chosen to consider an improvement in the results regarding the number of sensors.

Using the inlet flow data and non-leak historical pressure measurements of the selected sensors, the β_i and the α_i with $i = 1, ..., 31$ in (2.14) have been identified (notice that $n_I = 1$).

Table 5.3: Nodes with Sensors

Case	Nodes with sensors
1	12, 17, 23, 29
2	6, 12, 17, 23, 29, 21
3	6, 12, 15, 17, 23, 21, 27, 30
4	6, 9, 12, 15, 17, 24, 21, 22, 28, 29, 31

With these parameters, the pressure estimations under a non-leak condition in the network can be calculated considering inlet measurements using Equation (2.14) and posteriorly applied to calculate the residuals (4.15) with measured pressures in leak scenarios. In addition, non-leak pressure measurements and estimations are used to generate fault-free pressure residuals $r_{s_i}(k)$ and bounds $\underline{\sigma_i}$, $\bar{\sigma}_i$ $i = 1, ..., n_s$ as well as spatial residuals $Sr_{s_i,s_j}(k)$ and bounds $\underline{\varepsilon}_{i,j}$, $\bar{\varepsilon}_{i,j}$ $\forall i = 1, ..., n_s - 1 \quad \text{and} \quad j = i + 1, ..., n_s$.

For every sensor configuration, normalized incidence factors (5.25) have been computed with topological information: node connections and pipe characteristics (length, diameter, and roughness). Figure 5.11 compares the information on the incidence of single leaks to pressure sensors obtained by a hydraulic model with the one obtained by topological information. The nodes selected to have sensors are defined in the first case in Table 5.3. The colors symbolize the clustering referring to the sensor: the yellow color represents the sensor n. 12, the violet color the sensor n. 17, the red sensor n. 23, the green sensor n. 29, and the black is the undefined nodes. Figure 5.11. (c) shows the clustering that groups the nodes that produce the produces a maximum pressure deviation in a specific pressure sensor. The black nodes produce a similar variation of pressure (difference of variation less than 0.1 [mwc]) in at least two different pressure sensors. A hydraulic model to compute the difference between non-leak and leak pressures in all the nodes for the different leaks is required to obtain this information. On the other hand, Figure 5.11. (a) shows the clustering that takes into account the shortest weighted pipe length (hydraulic distance), that is, the sum of $(L_z/D_z^{4.87})$ for all edge e_z in the path to the sensors, being the smallest one used to define the most resemblance to the sensor, used in work [89]. Finally, clustering depicted in Figure 5.11.(b) is the one defined by Equation (5.28) that is based on the common resistance path. These two last clusters that only require topological information could be used in the leak localization method at the cluster level defined in Equation (5.21). It is

important to emphasize that the clustering is based on the common resistance path proposed in this paper and depicted in Figure 5.11. (b), resembles much more than the clustering based on the actual leak effect in the network (given by the model) depicted in Figure 5.11. (c) then the clustering is based on the hydraulic distance depicted in Figure 5.11. (a). Therefore, the clustering proposed in this paper provides more accurate information for leak localization purposes than the one based on the hydraulic distance. For example, as shown in Figure 5.11. (c), when a leak is present in nodes 3,4,5,6,7,8, or 9, the most affected sensor by the leak is the sensor in node 12. This information is the same as the one provided by the clustering depicted in Figure 5.11. (b) and computed only with topological information. However, using the clustering of Figure 5.11. (a) based on the hydraulic distance between nodes and sensors, the closest sensor to these nodes is the sensor in node 17.

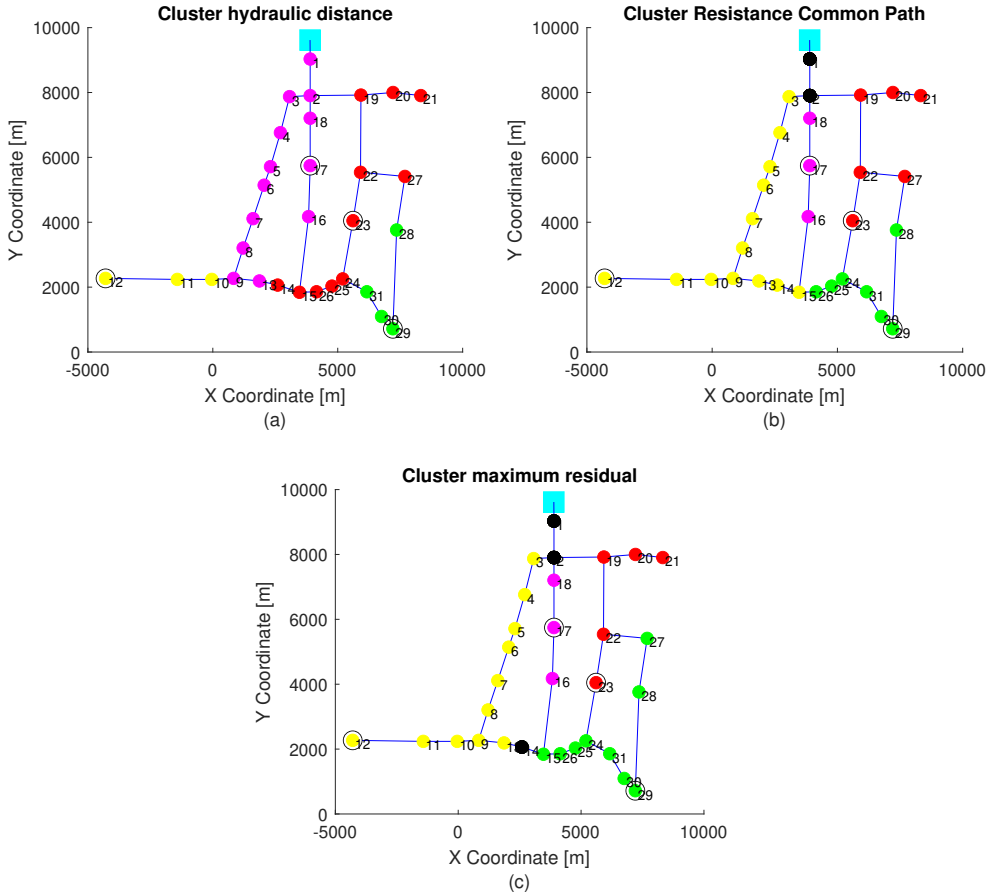


Figure 5.11: The s clustering generated with the aspects : (a) shortest weighted pipe length, (b) The resistance take into account the common path R^c_{j,s_i}, (c) the maximum residual .

Figure 5.12 shows the correlation analyses of the relative incidence index λ_{j,s_i} defined in Equation (5.25) for all the nodes $j = 1, .., 31$ depicted in every subplot for every sensor s_i $i = 1, ..., 4$. As this index is normalized, its values are in the range [0,1). The nodes with the higher index (more brown color) are the ones that produce a higher effect in the pressure sensor s_i.

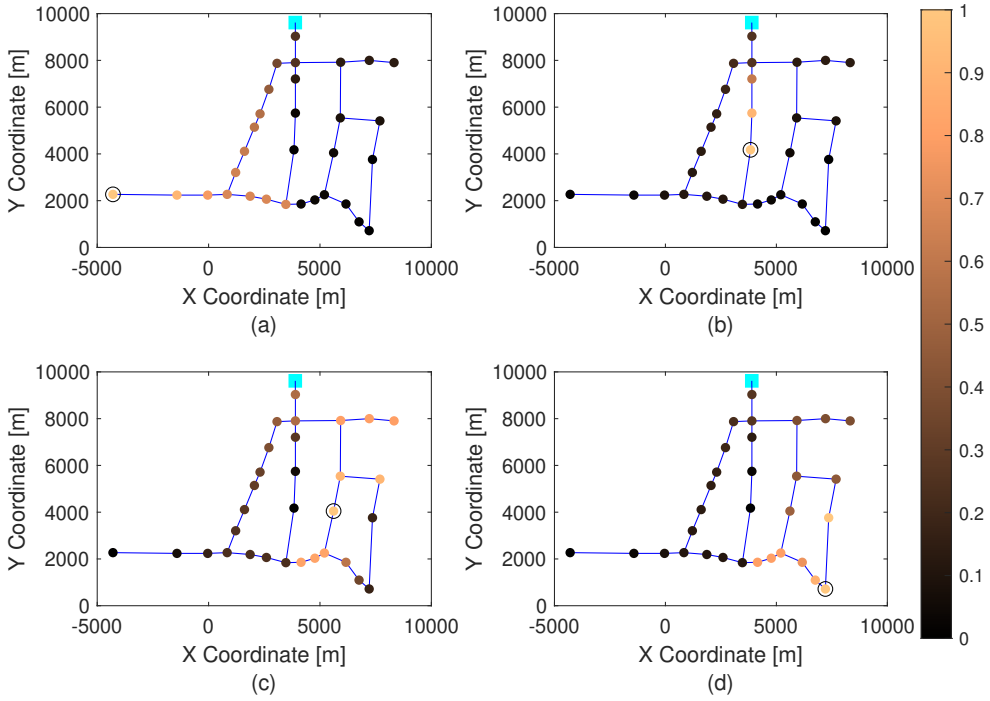


Figure 5.12: Relative incidence index λ_{j,s_i} for all the nodes ($j = 1, ..., 31$), corresponding to: (a) 1st sensor ($i = 1$), (b) 2nd sensor ($i = 2$), (c) 3rd sensor ($i = 3$), and (d) 4th sensor ($i = 4$).

Figure 5.13.(a) displays the evolution of the ATD (in nodes) obtained by the leak localization method based in Kriging spatial interpolation methodology presented in [89] with the time horizon (in hours) used recursively by the Bayes rule in (5.31). Four different sensor configurations are considered with 4, 6, 8, and 10 sensors placed in an optimal way in order to maximize the performance of the leak localization proposed [89]. The performance can be compared with the one obtained by the new leak localization method proposed in this paper at node level defined in Equation (5.30) with the same data set and the same sensor configurations as in [89], depicted in 5.13.(b) and with the sensor configurations shown in Table 5.3, depicted in 5.13.(c).

Figure 5.13.(a) shows that the leak detection performance of the Kriging method improves

significantly from four to eight sensors and more moderately compared to ten sensors, still having a good result since even with noise data managing to reach an ATD equal to 2.5 node. When compared to the new proposed leak localization method as it can be seen in Figures 5.13.(b) and 5.13.(c), the new leak localization method always over performs the Krigring method. Even in the case of using the sensor configurations proposed in [89] that were computed to optimize the performance of the Kriging method. Figure 5.13.(c) shows that the sensor configurations proposed in [89] are not optimal for the proposed method but the performance can be improved changing the sensor configurations, in this case manually.

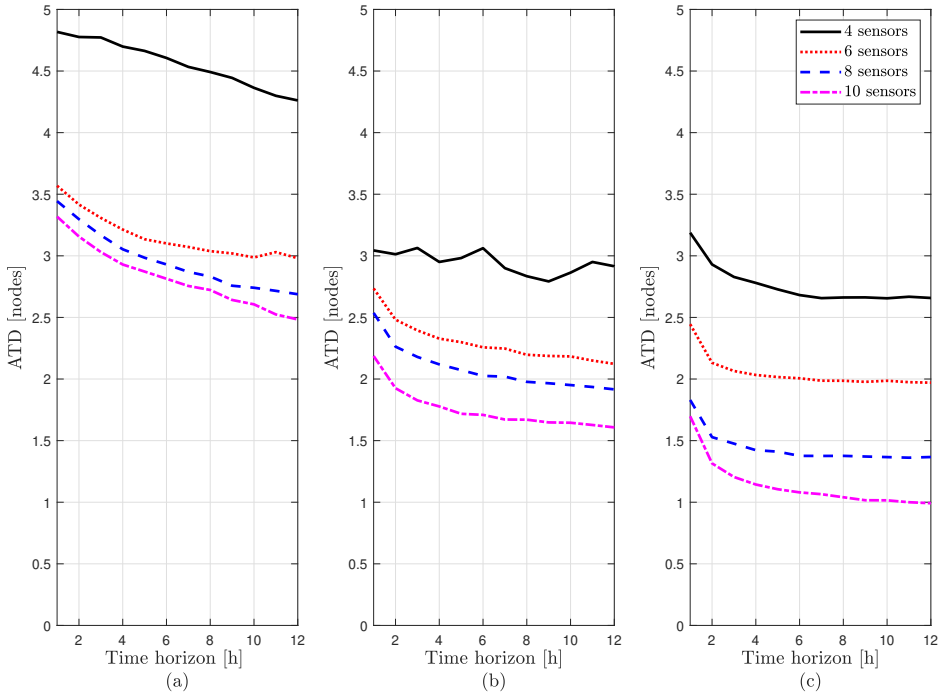


Figure 5.13: Evolution of the ATD between the methods:(a) using the Kriging interpolation method presented in [89], (b) using the new leak localization method with the same sensor configurations as in [89] and (c) using the new localization method with sensor configurations of Table 5.3.

Modena WDN

The second case study selected to test the performance is the reduced model of the real water distribution network of the Italian city Modena. This large-scale network comprises 268 junctions (nodes) connected through 317 pipes and served by four reservoirs. Displayed in Chapter

2.7.

EPANET hydraulic simulator was used to generate artificial data to analyze the performance of the proposed method. The following simulation conditions were considered:

- The leak scenario consists of data samples collected every 10 minutes and filtered to hourly values to reduce the uncertainty in the data.
- The uncertainty of demand is considered by introducing the uncertainty of $10[\%]$ (normal distribution) of the nominal demand value. In addition, white noise is deemed to emulate the noise in the measurements.
- The leak size is randomly selected with a range of 3 to $6[l/s]$, representing 1 to 2.5% of the network consumption.

As applied in the previous case study, the Average Topological Distance (ATD) was used to assess the performance of the proposed leak localization method at node level defined in (5.30). Two scenarios have been considered with five and ten pressure sensors that are presented in Figure 5.14.(a) and 5.14.(b) respectively. As emphasized in the last section, performance in the leak localization task is highly dependent on the number of sensors installed in the network.

Figure 5.15 shows the result of ATD evolution as defined in (2.22) applied with Bayes posterior time reasoning (5.31) to represent the leak location performance of the proposed method. This figure shows that leak localization performance reach an ATD of 8 and 5.5 nodes with 5 and 10 inner pressure sensors installed in the network respectively. Considering that the proposed leak localization method only requires topological information and non-leak historical data in available measurements, the obtained performance is reasonably good.

5.3.3 Leak localization based on interpolation method

The last data-driven leak localization aims to study the multi-leak problem. This section complements the technique seen in Chapter 4.3 of leak detection that will focus on the leak detection part, seen in Figure 5.16. The leak location method is based on creating leak probability zones in the WDN capable of performing the multi-leak study in the WDN.

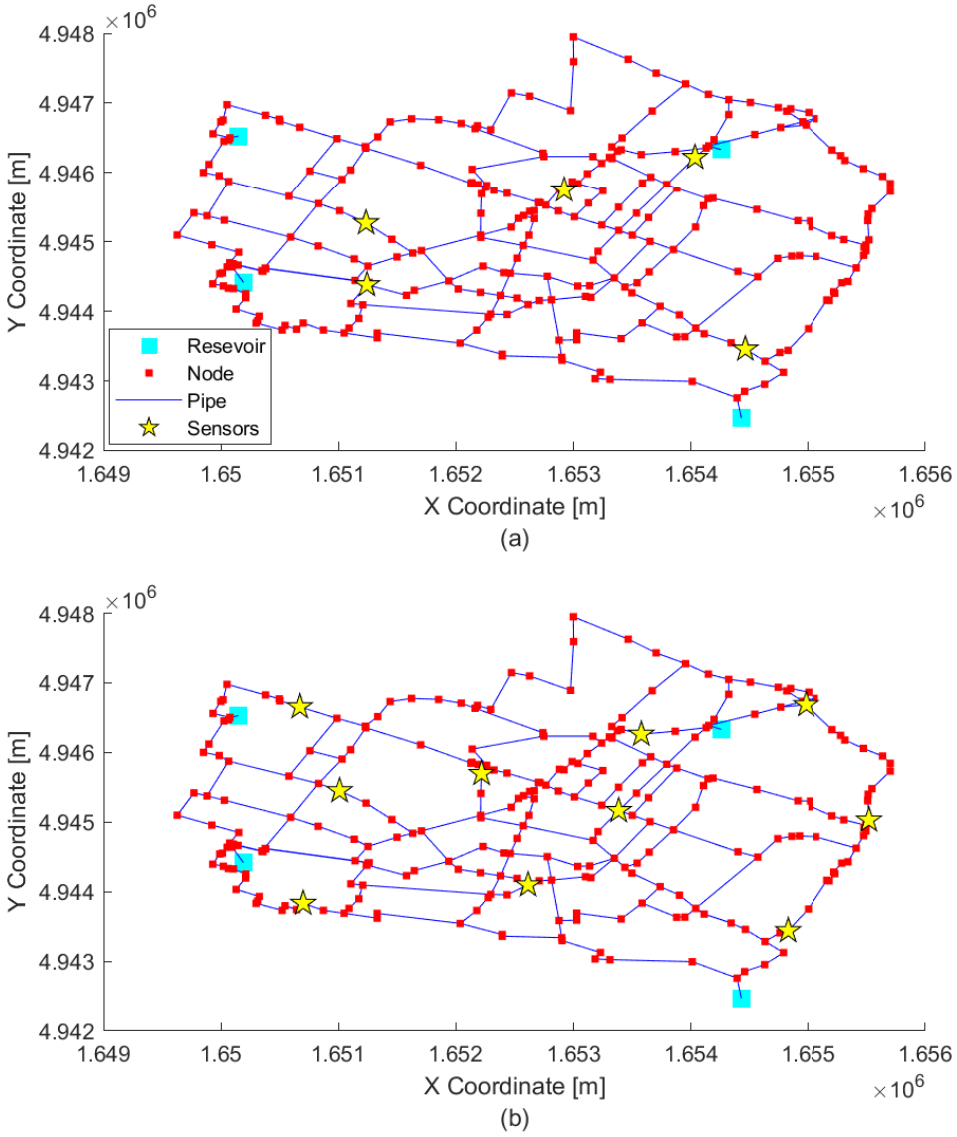


Figure 5.14: Configuration of pressure sensors in Modena WDN: (a) 5 sensors, (b) 10 sensors.

The method is based on the interpolation of data for the WDN. It has already been studied in other works [45, 96]. Still, as questioned in work [96], the interpolation of measured pressure to the nodes that do not have sensors trying to identify the fault at a node-level still has a long way to develop. However, the interpolation of leak indicators to determine the zone close to the sensors that have a leak is of great help for water companies as it will reduce the system zone for the leak's location.

To predict zones with unmeasured nodes the method will use the following information:

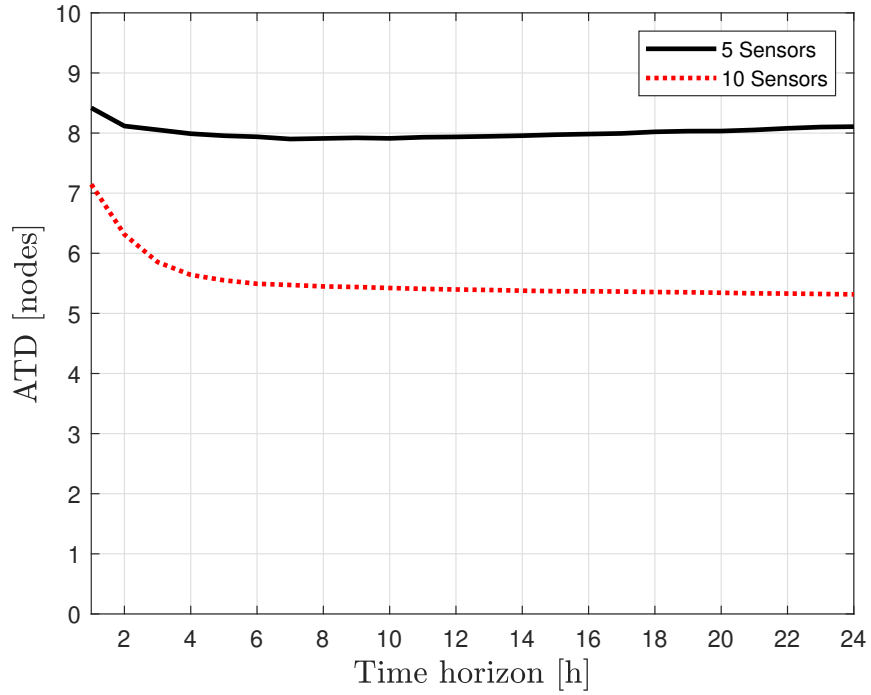


Figure 5.15: Evolution of the ATD

- (i) the average pressure residuals $\overline{r_i}(k)$ of Equation (4.16) available from the installed sensors $i = 1, ..., n_s$, which is developed in more detail in the Chapter 4.3:

- (ii) the topological information of the nodes in the network, and

- (iii) the Radial basis function (RBF) interpolation technique, see in Chapter 2. The RBF interpolation can be used in any dimension; in this work, the dimensions used are the latitude, longitude, and elevation of each node in the WDN, and the average pressure residuals of Equation (4.16) are the values to be interpolated.

In particular, interpolated residuals $\mho_i(k)$ at every inner node $i = 1, ..., n_0$ are computed from average pressure residuals $\overline{r_i}(k)$ of sensors $i = 1, ..., n_s$ as:

$$\mho_i(k) = \sum_{j=1}^{n_s} \lambda_j(k)\phi(||x_i - \underline{x_j}||) \tag{5.33}$$

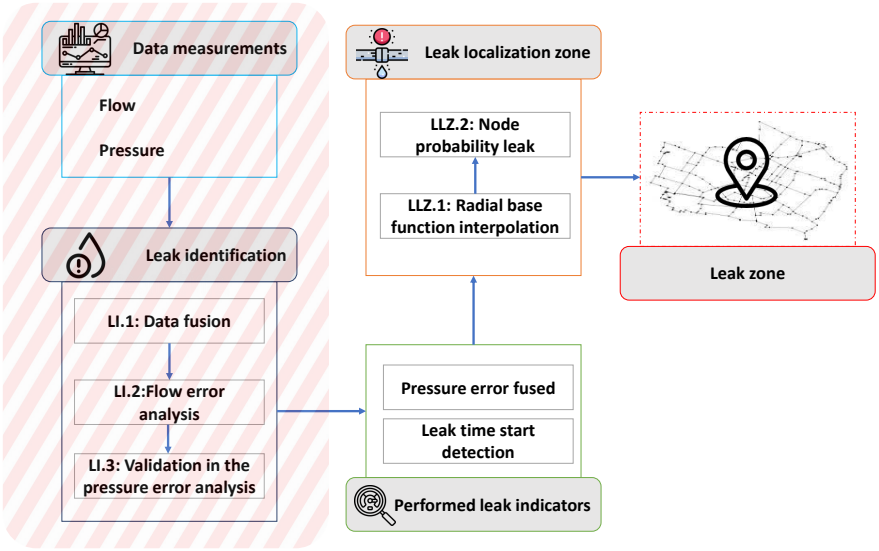


Figure 5.16: Flowchart of the leak detection and localization proposed method, highlighted the localization method

where $x_i \in \mathbb{R}^3$ is the coordinate (latitude, longitude and elevation) of inner node i ($i = 1, ..., n_0$) and $\underline{x_j} \in \mathbb{R}^3$ is the coordinate of sensor j ($j = 1, ..., n_s$) (center points for the RBFs), $\lambda_j(k)$ are coefficients determined by the n_s average pressure residuals at instant k and sensor coordinates and $\phi(r)$ is a radial basis function, set as :

$$\phi(r) = \sqrt{1 + 4r^2} \tag{5.34}$$

Once interpolated residuals $\mho_i$, $i = 1, ..., n_0$ have been compute, a probability index $\gamma^i(k)$ similar to the one defined in (5.35) can be defined as

$$\gamma^i(k) = \frac{\mho_i(k) - \mho^{min}(k)}{\mho^{max}(k) - \mho^{min}(k)} \tag{5.35}$$

where $\mho^{max}(k)$ and $\mho^{min}(k)$ are the maximum and minimum values of (5.33) computed for all the nodes $i = 1, ..., n_0$.

5.3.3.1 Case study

The case study will be the continuation of the case study in Chapter 2.7, the L-Town network. As explained before, the leaks in Area A and B of the 2018 year will be addressed in this work. The data set of the BattLeDIM for this year contains the time and repair location of 9 pipe bursts. Three types of leaks exist:

- Small background leaks with 1%–5% of the average inflow
- Medium pipe breaks with 5%–10%
- Large pipe bursts with leakage flow of more than 10% of the average system inflow ($\approx 50 l/s$)

The water utility corrects significant leaks with a flow rate above $4.5 l/s$ after a reasonable time within two months. The leakages have two different time profiles: either abrupt bursts with constant leak flow rates or incipient leaks that evolve until significant outflow rates at which they remain constant. Figure 4.7 shows the 12 leaks in 2018, with outflow rates between 1.4 and $9.7 l/s$ (5 and $35 m^3/h$). Thee leaks are not fixed, and nine are repaired throughout the year, which will be analyzed in this paper in the highlighted order of n.1 to n.9.

To perform the step of the proposed leak localization approach, the time instant of each leak begins in addition to the time they are repaired was used to calculate an average of the residues in Equation (4.15) to apply the RBF interpolation method. Figure 5.17 shows the results of the nine fixed leaks. The zones quoted to leak vary according to the location of the fault and how it affects the surrounding sensors, but for all leaks retaining the apex in red in the leak region, calculated with Equation 5.35. Leaks reference has a large leak probability zone since the WDN occurs in multiple leaks in different positions. These leaks will affect the pressure sensors, thus increasing the red area at risk of leaking.

The result of Figure 5.17 can be compared with Figures 5.4 and 5.5 that, are the result of the model based approach on the same case study presented in this section.

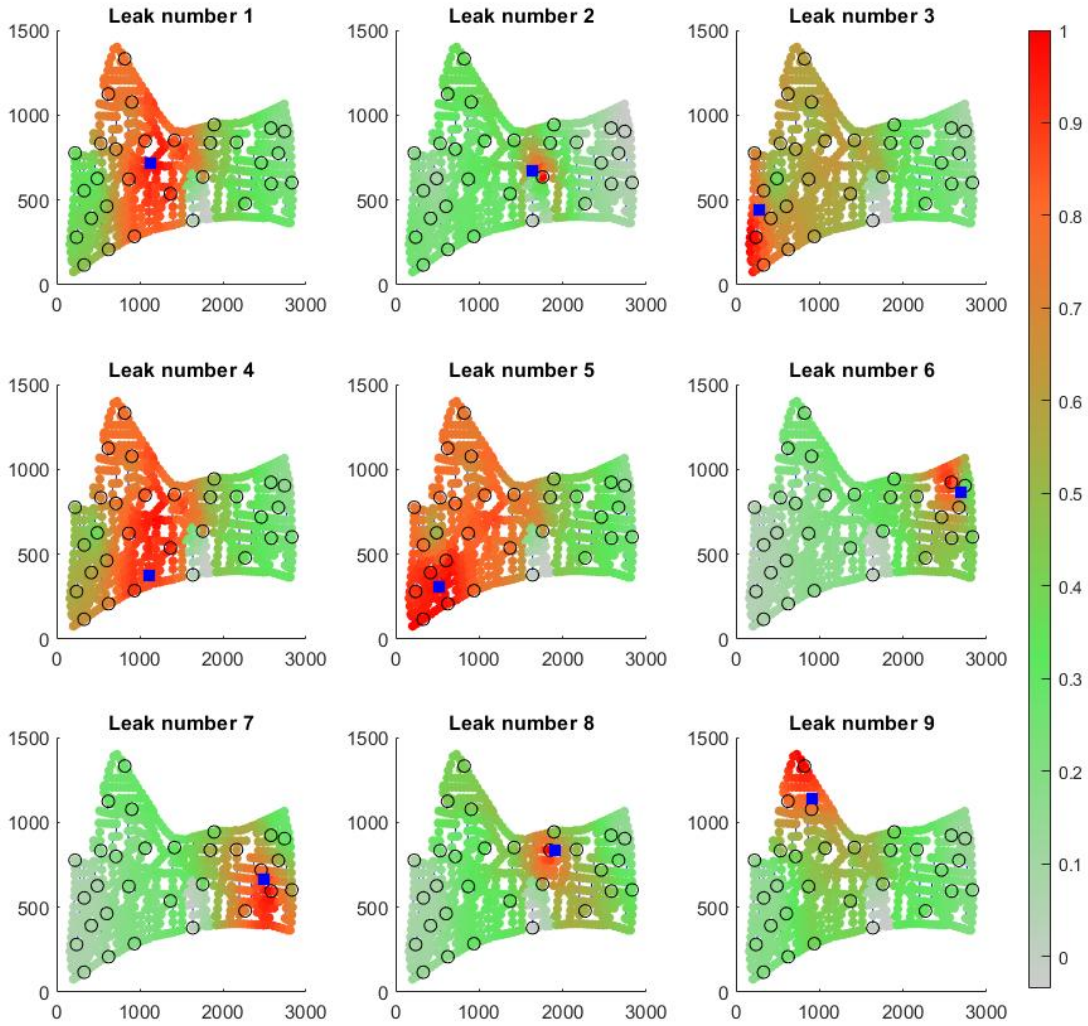
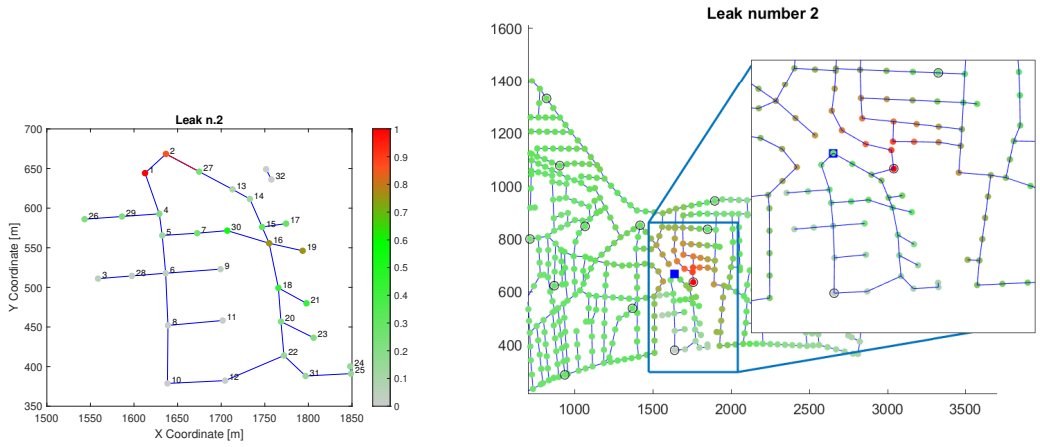


Figure 5.17: Graphical comparison of the interpolated states for the nine leaks in the WDN

The first possible comparison is in Zone B with leak number 2 (n.2), where it is delimited with a PRV and has only one pressure sensor installed in the area. The limited number of sensors complicates the result of leak location methods. Figure 5.18 shows the comparison of results using the based on model-based approach (Figure 5.18.(a)) and the result applying the method presented in this section, with a zoom in Zone B (Figure 5.18.(b)). As seen in Figure 5.18.(b), the result using the hydraulic model has a better effect; even though the nodes most affected by the leak in Zone B are the nodes present in Zone A in the border of the two zones.

When applying the data-driven method, leaks that occur in Zone A are better located. For example, leaks number 4 and 5 (n.4 and n.5) that appear at the same time has a more significant difference comparing the model-based results (Figure 5.5) and data-driven results (Figure 5.17). Since the results of the model-based approach of locating leaks point to the nodes that do not have leakage, as seen in Figure 5.5.(c-d) and the result using the data-driven method in Figure 5.17 shows the highest leakage rate γ^i at the location where the leakage is, thus having the best result.



(a) Model-based approach (b) Data-driven approach

Figure 5.18: Leak localization in Zone B on pipe p673

5.4 Summary

This Chapter has presented different methodologies to leak localization; one focuses on a model-based approach, and three are based on a data-driven approach. The leak localization method based on the model-based approach uses the hydraulic model to compare the different leak scenarios to select the scenario with more similar pressure results to the actual pressure measurements. The method covers the multi-leak problem by studying each leak chronologically and updating the model with the leak location estimation. The case study presented was the L-Town network with the 2018 data; the analysis has been divided into two areas: Zone C has the AMRs devices providing an accurate hydraulic model with good results. And Zone A and B have uncertainties in the node demand forecast and the leak position during the year. The development in Zone B has a better result because the area is small and has only a single leak; however, in

Zone A, the result worsens as the year goes on because the uncertainty keeps increasing. If one has to point out the weakness of this method, one can say that if the leaks are not correctly introduced in the hydraulic model, errors will accumulate in the model, making the results less reliable.

In the leak localization using a data-driven technique, the goal is to locate the leak using the topological information, flow, and pressure measurements. This Chapter has presented three different approaches full data-driven methods to leak localization problems and the method is triggered when a leak is detected.

The first technique in WDN based on distance clustering in association with the residual analysis distance of the pressure sensor. The proposed approach has been explained, and an example is shown using the Modena Network simplified version of the real WDN as a case study. The distance clustering approach has been introduced using the Euclidean distance with the centroids set to be the pressure sensor's location and additional ρ strategical nodes in the network. An extra data collection needs to be done in the real WDN, simulating a leak event in the network in each location to be a cluster center.

The second technique aims to move from a cluster study to an node analysis. The leak localization in WDN is based on historical non-leak data, and the topological information of the network is proposed. The method is based on the evaluation of residuals generated by leak pressure measurements in some inner nodes and the estimation of leak-free pressures in these nodes utilizing a reduced-order model and historical data. Topological information is used to compute a new incidence factor that considers the most probable path of water from reservoirs to pressure sensors and potential leak nodes. The proposed incidence factor combined with residual information generates a likelihood index that allows the leak localization at the node level. The proposed method's general performance for leak location is evaluated in reduced models of the Hanoi and Modena water distribution networks.

The last approach to leak localization is a complementary study of leak detection introduced in the Chapter with a two-phase methodology: the first is leak detection, and the second is leak localization presented in the Chapter 4.3. The second phase is the leak localization zone that

applies the Radial basis function to interpolate the average residual pressure for each sensor to all the nodes in the network, resulting in the zone most likely to have the fault. The L-Town network utilized in the Battle of the Leakage Detection and Isolation Methods has been used as a case study. The data studied were from the year 2018 with 12 leaks and only 9 repaired, having two different temporal profiles: burst pipe and incipient leaks that stature in some instant. The result is good because for all leaks studied, the leak probability area is always around the defective tube, and the size of the area varies from leak to leak and depends on whether there are simultaneous leaks in the WDN.

CHAPTER 6

CONCLUDING REMARKS

6.1 Conclusions

In this thesis, several theoretical contributions and application results on leak detection, leak localization, sensor placement and sensor validation have been presented. Specifically, the conclusions are summarized concerning the envisaged thesis objectives as follows:

- ***Objective 1*** *Propose leak detection and estimation methods able to detect incipient leaks and distinguish multi-leak scenarios in WDNs*

  It has been shown in Chapter 4 that two studies of leak detection have been proposed. The first study uses information from flow sensors installed in the WDN inlets. The method uses the fusion sensor base to create a virtual measurement that will be used to detect and estimate the leaks magnitude in the system. The second study is a complementary analysis to the first method to analyze multi-leak problems using information from inlet flow and pressure sensors. The first method can obtain the data of the sum of the leaks happening simultaneously in the system. The second method uses the information of the pressure sensors to define the moment of the beginning of the leak.

- ***Objective 2*** *Develop a leak localization scheme that combines hydraulic models and leak*

detection analysis to tackle the multi-leak problem.

The method presented in Chapter 5.2 shows the study of the multi-leak localization in the WDN using a hydraulic model. The hydraulic model has to be well-calibrated with a minimum of uncertainty. The method introduces the use of the automatic meter reading devices installed on some nodes to know the exact demand consumed in these nodes, thus making the model more accurate. The method proposes to study the leaks in chronological order, and for each leak, an update is necessary with the information on the estimated magnitude and location of the leak in the hydraulic model. The study of multi-leaks in the WDN is a challenge when applied to model-based because any uncertainty applied to the hydraulic model can affect the localization result, which was shown in the T-town case study.

- ***Objective 3*** *Propose new data-driven approaches that tackle the problem of leak localization at cluster and node level in WDN.*

  Two strategies have been shown in Chapter 5.3 to obtain the full data-driven leak location developed for the unique leak problem. The first method introduced aims to reduce the area of the WDN in clustering to facilitate the search for the leak in the field. The second leak localization technique seeks to move from the cluster study to the node study. Both methods use topological information to generate the WDN representation using the Graph theory.

- ***Objective 4*** *Develop a new data-driven approach that tackles the multi-leak problem.*

  A method has been presented in the Chapter 5.3 to tackle the problem of multi-leak using a data-driven technique. The method uses Radial basis function interpolation to interpolate a leak index for all WDN nodes resulting in an area with a leak probability. The technique had a better result than the model-based method because when the WDN has several leaks happening simultaneously, it is challenging to develop an accurate hydraulic model.

- ***Objective 5*** *Tackle practical problems of using pressure sensors: Sensor validation and optimal sensor placement.*

  Regarding leak supervision in WDN, it is necessary to carry out more studies than leak

detection and location, such as sensor placement and sensor validation. In the study of sensor placement, it is important to guarantee the optimal leak localization result since sensor placement affects the outcome of the leak location analysis. And the sensor validation is crucial to ensure the leak localization robustness. A new technique for sensor placement and sensor position has been presented in Chapter 3, demonstrating the importance of these two steps to improve the results of leak localization methods.

6.2 Findings & Contributions

As mentioned in the objectives, the main contribution of these theses was methods focused on model-based and data-driven leak detection and localization for better monitoring of leaks in the WDN, in contrast to using flow and pressure sensors. And the validated sensor and optimal sensor position methods help the leak localization method to obtain an optimal result. Following are the contributions and findings of this thesis:

- In chapter 3, new sensor validation methods and position methods were introduced. The contribution of the sensor validation is its easy application in WDN that have pressure sensors to guarantee the good functioning of the sensors, ensuring robustness for the leak supervision methods that use the pressure sensors. The method develops a correlation between pressure sensor measurements. The optimal sensor position method was developed to support the leak localization methods that use the pressure residue study; two methods were presented, the first one focused on model-based, being an excellent option when the hydraulic model is reliable but requires a long process time.

  The second method presents three centroid models contributed to the analysis when the WDN has only hydraulic information, such as node positioning and pipe information.

- Chapter 4 presented the leak detection methods; the two methods presented are based on data-driven, being easy to implement in the WDN as they use flow and pressure information obtained from sensors already installed in the network. The objective of the methods was to detect leaks in the order of 2% to 5% of the average network inflow. The first method contributes to the analysis of single leaks, having good results when the demand

forecast is well-calibrated. However, this method must be adapted for WDN that have multiple leaks in the system in a short period of time.

The second method uses flow and pressure information, contributes to the analysis of multi leaks problems, providing information on which pressure sensor is most affected by a given leak, and assisting the water company in starting the tracking for the leak in the surroundings of the most affected sensor.

- In Chapter 5, four methods of leak localization were presented. The main objective of these methods was to use only flow and pressure information and possibly use ATD equipment. The first method is model-based, contributing to the analysis of multiple leaks in the system and the option of using the ATD to complement the study. This method has the same weak point as the methods that use the model-based base, which is the high sensitivity to the hydraulic model; however, it is an excellent option for WDN that have ATD and a reliable hydraulic model, presenting excellent results.

  The three data-driven methods contribute to the study of leakage that does not have a hydraulic model. The first two methods contribute to the analysis of single leaks, or when leaks occur with a large gap. The Leak localization based on the cluster technique method uses extra pressure data from several nodes, contributing to the reduction of clusterings to pinpoint the leak's location. Leak Localization of the common path contributes to a new cluster analogy relating to the path that the water travels to reach the sensors; it is more indicated in the WDN that has only one inlet, requiring information from pressure sensors and hydraulic information. The Leak localization based on the interpolation method has as an innovation the study of multi-leaks with a new interpolation model of the beginning of pressure sensor data to start an area with more chances of leaking.

6.3 Future Work

There are still some open issues regarding the presented problems in Chapters 3-5. From the summary of each chapter, several improvements have been introduced. Generally speaking, some interesting ideas for future directions derived from this thesis are suggested as follows:

- A comparative study between the developed leak localization/ detection methods and literature methods. In addition, complementary research focuses on the BattleDIM challenge [100] as a basis for comparison of the multi-leakage methods.

- Applications of all the presented methods to real WDN are interesting. Among them, new challenges in the implementations of these approaches could be met and useful solutions could be demonstrated.

- The leak detection study could be extended to deal with the validation of the water consumption forecast. For example, the seasonal effect can cause drifts in the total water consumption and can lead to false positives if the drift has a positive slope.

- A study to estimate the node demand throughout the year could be developed to improve the techniques that use the hydraulic model. This study can reduce the uncertainties present in the hydraulic model and thus improve the results of techniques that use model-based

- A detailed study of the interpolation methods could be done with the main objective of knowing which one best adapts to the WDN, and can be implemented in the leak localization methodology focused on data-driven.

BIBLIOGRAPHY

[1] A Abdulshaheed, F Mustapha, and A Ghavamian. “A pressure-based method for monitoring leaks in a pipe distribution system: A Review”. In: *Renewable and Sustainable Energy Reviews* 69 (2017), pp. 902–911.

[2] Kazeem Adedeji et al. “Towards achieving a reliable leakage detection and localization algorithm for application in water piping networks: An overview”. In: *IEEE Access* 5 (2017), pp. 20272–20285.

[3] Débora Alves, Joaquim Blesa, and Eric Duviella. “Detecçao de vazamento de distribuiçao de água”. In: *Conferência de Estudos em Engenharia Elétrica (CEEL)* (2020).

[4] Débora Alves et al. “Data-driven leak localization in WDN using pressure sensor and hydraulic information”. In: *IFAC-PapersOnLine* 55.5 (2022), pp. 96–101.

[5] Débora Alves et al. “Leak Detection in Water Distribution Networks Based on Water Demand Analysis”. In: *IFAC-PapersOnLine* 55.6 (2022), pp. 679–684.

[6] Débora Alves et al. “Robust data-driven leak localization in water distribution networks using pressure measurements and topological information”. In: *Sensors* 21.22 (2021), p. 7551.

[7] Georg Arbesser-Rastburg and Daniela Fuchs-Hanusch. “Serious Sensor Placement—Optimal Sensor Placement as a Serious Game”. In: *Water* 12.1 (2019), p. 68.

[8] M Bakker et al. “Heuristic burst detection method using flow and pressure measurements”. In: *Journal of Hydroinformatics* 16.5 (2014), pp. 1194–1209.

[9] Joaquim Blesa, Fatiha Nejjari, and Ramon Sarrate. “Robust sensor placement for leak location: analysis and design”. In: *Journal of Hydroinformatics* 18.1 (2015), pp. 136–148.

[10] Joaquim Blesa and Ramon Pérez. “Modelling uncertainty for leak localization in Water Networks”. In: *IFAC-PapersOnLine* 51.24 (2018). 10th IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes SAFEPROCESS 2018, pp. 730–735. ISSN: 2405-8963.

[11] Joaquim Blesa et al. “An Interval NLPV Parity Equations Approach for Fault Detection and Isolation of a Wind Farm”. In: *IEEE Transactions on Industrial Electronics* 62.6 (2015), pp. 3794–3805.

[12] Ken Brothers. “Leak detection practices and techniques: a practical approach”. In: (2003).

[13] *Cadreanalytic cadreanalytic website*. `https://www.cadreanalytic.com/cadreflo.htm`. Accessed: 23-05-2022.

[14] Myrna Casillas, Luis Eduardo Garza Castañón, and Vicenç Puig Cayuela. “Extended-horizon analysis of pressure sensitivities for leak detection in water distribution networks”. In: *IFAC Proceedings Volumes* 45.20 (2012), pp. 570–575.

[15] Myrna Casillas, Luis Garza-Castañón, and Vicenç Puig. “Extended-horizon analysis of pressure sensitivities for leak detection in water distribution networks: Application to the Barcelona network”. In: *2013 European Control Conference (ECC)*. IEEE. 2013, pp. 401–409.

[16] Myrna Casillas et al. “Leak signature space: An original representation for robust leak location in water distribution networks”. In: *Water* 7.3 (2015), pp. 1129–1148.

[17] Teck Kai Chan, Cheng Siong Chin, and Xionghu Zhong. “Review of current technologies and proposed intelligent methodologies for water distributed network leakage detection”. In: *IEEE Access* 6 (2018), pp. 78846–78867.

[18] Roya A Cody, Pampa Dey, and Sriram Narasimhan. “Linear prediction for leak detection in water distribution networks”. In: *Journal of Pipeline Systems Engineering and Practice* 11.1 (2020), p. 04019043.

[19] Roya A Cody, Bryan A Tolson, and Jeff Orchard. “Detecting leaks in water distribution pipes using a deep autoencoder and hydroacoustic spectrograms”. In: *Journal of Computing in Civil Engineering* 34.2 (2020), p. 04020001.

[20] Andrew F Colombo, Pedro Lee, and Bryan W Karney. “A selective literature review of transient-based leak detection methods”. In: *Journal of hydro-environment research* 2.4 (2009), pp. 212–227.

[21] Pep Cugueró-Escofet et al. “Assessment of a leak localization algorithm in water networks under demand uncertainty”. In: *IFAC-PapersOnLine* 48.21 (2015), pp. 226–231.

[22] Shantanu Datta and Shibayan Sarkar. “A review on different pipeline fault detection methods”. In: *Journal of Loss Prevention in the Process Industries* 41 (2016), pp. 97–106.

[23] Narsingh Deo. *Graph theory with applications to engineering and computer science*. Courier Dover Publications, 2017.

[24] Emmanuel A Donkor et al. “Urban water demand forecasting: review of methods and models”. In: *Journal of Water Resources Planning and Management* 140.2 (2014), pp. 146–159.

[25] Mohamed Fahmy and Osama Moselhi. “Automated detection and location of leaks in water mains using infrared photography”. In: *Journal of Performance of Constructed Facilities* 24.3 (2010), pp. 242–248.

[26] Malcolm Farley and Stuart Hamilton. “Non-intrusive leak detection in large diameter, low-pressure non-metallic pipes: are we close to finding the perfect solution”. In: *Proc. IWA World Water Congr.* 2008, pp. 1–9.

[27] Zahra Fereidooni, Hooman Tahayori, and Ali Bahadori-Jahromi. “A hybrid model-based method for leak detection in large scale water distribution networks”. In: *Journal of Ambient Intelligence and Humanized Computing* 12.2 (2021), pp. 1613–1629.

[28] Philippe Flajolet and Robert Sedgewick. “Mellin transforms and asymptotics: Finite differences and Rice’s integrals”. In: *Theoretical Computer Science* 144.1-2 (1995), pp. 101–124.

[29] Okitsugu Fujiwara and Do Ba Khang. “A two-phase decomposition method for optimal design of looped water distribution networks”. In: *Water resources research* 26.4 (1990), pp. 539–549.

[30] Caspar VC Geelen et al. “Monitoring support for water distribution systems based on pressure sensor data”. In: *Water Resources Management* 33.10 (2019), pp. 3339–3353.

[31] Zhiqiang Geng et al. “A novel leakage-detection method based on sensitivity matrix of pipe flow: case study of water distribution systems”. In: *Journal of Water Resources Planning and Management* 145.2 (2019), p. 04018094.

[32] Simon N Gosling and Nigel W Arnell. “A global assessment of the impact of climate change on water scarcity”. In: *Climatic Change* 134.3 (2016), pp. 371–385.

[33] James-A Goulet, Sylvain Coutu, and Ian FC Smith. “Model falsification diagnosis and sensor placement for leak detection in pressurized pipe networks”. In: *Advanced Engineering Informatics* 27.2 (2013), pp. 261–269.

[34] Guancheng Guo et al. “Short-term water demand forecast based on deep learning method”. In: *Journal of Water Resources Planning and Management* 144.12 (2018), p. 04018076.

[35] Shikha Pranesh Gupta, Abha Mahalwar, and P Udaykumar. “Analysis of different techniques for locating leaks in pipes in water distribution system using WSN”. In: *2014 Innovative Applications of Computational Intelligence on Power, Energy and Controls with their impact on Humanity (CIPECH)*. IEEE. 2014, pp. 173–177.

[36] Erfan Hajibandeh and Sara Nazif. “Pressure zoning approach for leak detection in water distribution systems based on a multi objective ant colony optimization”. In: *Water Resources Management* 32.7 (2018), pp. 2287–2300.

[37] Stuart Hamilton and Bambos Charalambous. *Leak detection: technology and implementation*. IWA Publishing, 2020.

[38] Xuan Hu et al. “Novel leakage detection and water loss management of urban water supply network using multiscale neural networks”. In: *Journal of Cleaner Production* 278 (2021), p. 123611.

[39] Zukang Hu et al. “Review of model-based and data-driven approaches for leak detection and location in water distribution systems”. In: *Water Supply* 21.7 (2021), pp. 3282–3306.

[40] Pingjie Huang et al. “Real-time burst detection in district metering areas in water distribution system based on patterns of water demand with supervised learning”. In: *Water* 10.12 (2018), p. 1765.

[41] Osama Hunaidi and Peter Giamou. “Ground-penetrating radar for detection of leaks in buried plastic water distribution pipes”. In: *International Conference on Ground Penetrating Radar*. 1998.

[42] Osama Hunaidi et al. “Detecting leaks in plastic pipes”. In: *Journal-American Water Works Association* 92.2 (2000), pp. 82–94.

[43] Christopher Hutton and Zoran Kapelan. “Real-time burst detection in water distribution systems using a Bayesian demand forecasting methodology”. In: *Procedia Engineering* 119 (2015), pp. 13–18.

[44] *InfoWorks WS InfoWorks WS website*. `https://www.innovyze.com/en-us/products/infoworks-ws-pro`. Accessed: 23-05-2022.

[45] Mohammadreza Javadiha et al. “Leak localization in water distribution networks using deep learning”. In: *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE. 2019, pp. 1426–1431.

[46] Tom Nørgaard Jense et al. “Plug-and-play commissionable models for water networks with multiple inlets”. In: *2018 European Control Conference (ECC)*. IEEE. 2018, pp. 1–6.

[47] Hector A. Jensen and Danko J. Jerez. “A Bayesian model updating approach for detection-related problems in water distribution networks”. In: *Reliability Engineering & System Safety* 185 (2019), pp. 100–112.

[48] Tom Nørgaard Jensen and Carsten Skovmose Kallesøe. “Application of a novel leakage detection framework for municipal water supply on aau water supply lab”. In: *2016 3rd Conference on Control and Fault-Tolerant Systems (SysTol)*. IEEE. 2016, pp. 428–433.

[49] Javier Jiménez-Cabas et al. “Localization of leaks in water distribution networks using flow readings”. In: *IFAC-PapersOnLine* 51.24 (2018), pp. 922–928.

[50] Donghwi Jung and Kevin Lansey. “Water distribution system burst detection using a nonlinear Kalman filter”. In: *Journal of Water Resources Planning and Management* 141.5 (2015), p. 04014070.

[51] Donghwi Jung et al. “Improving the rapidity of responses to pipe burst in water distribution systems: a comparison of statistical process control methods”. In: *Journal of Hydroinformatics* 17.2 (2015), pp. 307–328.

[52] Jiheon Kang et al. “Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems”. In: *IEEE Transactions on Industrial Electronics* 65.5 (2017), pp. 4279–4289.

[53] Fatma Karray et al. “Earnpipe: A testbed for smart water pipeline monitoring using wireless sensor network”. In: *Procedia Computer Science* 96 (2016), pp. 285–294.

[54] Jack P.C. Kleijnen. “Regression and Kriging metamodels with their experimental designs in simulation: A review”. In: *European Journal of Operational Research* 256.1 (2017), pp. 1–16. ISSN: 0377-2217.

[55] J Richard Landis and Gary G Koch. “An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers”. In: *Biometrics* (1977), pp. 363–374.

[56] Daniele Laucelli et al. “Detecting anomalies in water distribution networks using EPR modelling paradigm”. In: *Journal of Hydroinformatics* 18.3 (2016), pp. 409–427.

[57] Rui Li et al. “A review of methods for burst/leakage detection and location in water distribution systems”. In: *Water Science and Technology: Water Supply* 15.3 (2015), pp. 429–441.

[58] Sren Lophaven, Hans Bruun Nielsen, and Jacob Sndergaard. “Automatic mapping of monitoring data”. In: (2005).

[59] Dália Loureiro et al. “Water distribution systems flow monitoring and anomalous event detection: A practical approach”. In: *Urban Water Journal* 13.3 (2016), pp. 242–252.

[60] Obaid Malik. “Probabilistic leak detection and quantification using multi-output Gaussian processes”. PhD thesis. University of Southampton, 2016.

[61] SR Mounce, JB Boxall, and J Machell. “Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows”. In: *Journal of Water Resources Planning and Management* 136.3 (2010), pp. 309–318.

[62] Stephen R Mounce, Richard B Mounce, and Joby B Boxall. “Novelty detection for time series data analysis in water distribution systems using support vector machines”. In: *Journal of hydroinformatics* 13.4 (2011), pp. 672–686.

[63] JM Muggleton et al. “A novel sensor for measuring the acoustic pressure in buried plastic water pipes”. In: *Journal of sound and vibration* 295.3-5 (2006), pp. 1085–1098.

[64] KiJeon Nam et al. “An efficient burst detection and isolation monitoring system for water distribution networks using multivariate statistical techniques”. In: *Sustainability* 11.10 (2019), p. 2970.

[65] Ali Nasirian, Mahmoud F Maghrebi, and Siavash Yazdani. “Leakage detection in water distribution network based on a new heuristic genetic algorithm model”. In: (2013).

[66] *NextGen Simulation Suite NextGen Simulation Suite website*. `https://www.greggeng.com/software-solutions/nextgen-simulation-suite/`. Accessed: 23-05-2022.

[67] Isaac Okeya, Christopher Hutton, and Zoran Kapelan. “Locating pipe bursts in a district metered area via online hydraulic modelling”. In: *Procedia engineering* 119 (2015), pp. 101–110.

[68] Isaac Okeya et al. “Online burst detection in a water distribution system using the Kalman filter and hydraulic modelling”. In: *Procedia Engineering* 89 (2014), pp. 418–427.

[69] CV Palau, FJ Arregui, and M Carlos. “Burst detection in water networks using principal component analysis”. In: *Journal of Water Resources Planning and Management* 138.1 (2012), pp. 47–54.

[70] Ramon Pérez et al. “Accuracy assessment of leak localisation method depending on available measurements”. In: *Procedia Engineering* 70 (2014), pp. 1304–1313.

[71] Ramon Pérez et al. “Leak localization in water networks: a model-based methodology using pressure sensors applied to a real network in Barcelona”. In: *IEEE control systems magazine* 34.4 (2014), pp. 24–36.

[72] Ramon Pérez et al. “Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks”. In: *Control Engineering Practice* 19.10 (2011), pp. 1157–1167.

[73] *Pipeflow Pipeflow website*. `https://www.pipeflow.co.uk`. Accessed: 23-05-2022.

[74] Raido Puust et al. “A review of methods for leakage management in pipe networks”. In: *Urban Water Journal* 7.1 (2010), pp. 25–45.

[75] Joseba Jokin Quevedo Casín et al. “Leakage location in water distribution networks based on correlation measurement of pressure sensors”. In: (2011).

[76] Marcos Quiñones-Grueiro et al. “An unsupervised approach to leak detection and location in water distribution networks”. In: *International Journal of Applied Mathematics and Computer Science* 28.2 (2018).

[77] Michele Romano, Zora Kapelan, and Dragan Savić. “Burst detection and location in water distribution systems”. In: *World Environmental and Water Resources Congress 2011: Bearing Knowledge for Sustainability*. 2011, pp. 1–10.

[78] Michele Romano, Zoran Kapelan, and Dragan Savić. “Automated detection of pipe bursts and other events in water distribution systems”. In: *Journal of Water Resources Planning and Management* 140.4 (2014), pp. 457–467.

[79] Michele Romano, Kevin Woodward, and Zoran Kapelan. “Statistical process control based system for approximate location of pipe bursts and leaks in water distribution systems”. In: *Procedia Engineering* 186 (2017), pp. 236–243.

[80] Luis Romero-Ben et al. “Leak Localization in Water Distribution Networks Using Data-Driven and Model-Based Approaches”. In: *Journal of Water Resources Planning and Management* 148.5 (2022), p. 04022016.

[81] Lewis A Rossman. “EPANET 2: Users Manual”. In: (2000).

[82] Francisco Javier Salguero, R Cobacho, and MA Pardo. “Unreported leaks location using pressure and flow sensitivity in water distribution networks”. In: *Water Supply* 19.1 (2019), pp. 11–18.

[83] Gerard Sanz et al. “Leak detection and localization through demand components calibration”. In: *Journal of Water Resources Planning and Management* 142.2 (2016), p. 04015057.

[84] Ramon Sarrate et al. “Sensor placement for leak detection and location in water distribution networks”. In: *Water Science and Technology: Water Supply* 14.5 (2014), pp. 795–803.

[85] Yu Shao et al. “Time-series-based leakage detection using multiple pressure sensors in water distribution systems”. In: *Sensors* 19.14 (2019), p. 3070.

[86] S Shimanskiy, T Iijima, and Y Naoi. “Development of microphone leak detection technology on Fugen NPP”. In: *Progress in nuclear energy* 43.1-4 (2003), pp. 357–364.

[87] Alex Simpkins. “System identification: Theory for the user, (ljung, l.; 1999)[on the shelf]”. In: *IEEE Robotics & Automation Magazine* 19.2 (2012), pp. 95–96.

[88] Adria Soldevila et al. “Leak localization method for water-distribution networks using a data-driven model and Dempster–Shafer reasoning”. In: *IEEE Transactions on Control Systems Technology* 29.3 (2020), pp. 937–948.

[89] Adrià Soldevila et al. “Data-driven approach for leak localization in water distribution networks using pressure sensors and spatial interpolation”. In: *Water* 11.7 (2019), p. 1500.

[90] Adrià Soldevila et al. “Leak localization in water distribution networks using a mixed model-based/data-driven approach”. In: *Control Engineering Practice* 55 (2016), pp. 162–173.

[91] Adrià Soldevila et al. “Leak localization in water distribution networks using Bayesian classifiers”. In: *Journal of Process Control* 55 (2017), pp. 1–9.

[92] Adrià Soldevila et al. “Sensor placement for classifier-based leak localization in water distribution networks using hybrid feature selection”. In: *Computers & Chemical Engineering* 108.Supplement C (2018), pp. 152–162. ISSN: 0098-1354.

[93] Sophocles Sophocleous, Dragan Savić, and Zoran Kapelan. “Leak localization in a real water distribution network based on search-space reduction”. In: *Journal of Water Resources Planning and Management* 145.7 (2019), p. 04019024.

[94] David Steffelbauer and Daniela Fuchs-Hanusch. “Efficient sensor placement for leak localization considering uncertainties”. In: *Water resources management* 30.14 (2016), pp. 5517–5533.

[95] David Steffelbauer et al. “Pressure-leak duality for leak detection and localization in water distribution systems”. In: *Journal of Water Resources Planning and Management* 148.3 (2022).

[96] Congcong Sun et al. “Leak localization in water distribution networks using pressure and data-driven classifier approach”. In: *Water* 12.1 (2019), p. 54.

[97] SWAN. *Stated NRW (Non-Revenue Water) Rates in Urban Networks. Technical report, Smart Water Networks Forum*. 2011.

[98] *Synergi Pipeline Simulator Synergi Pipeline Simulator website*. `https://www.dnv.com/software/services/pipeline/synergi-pipeline-simulator-index.html`. Accessed: 23-05-2022.

[99] Spyros G Tzafestas. *Introduction to mobile robot control*. Elsevier, 2013.

[100] Demetrios Vrachimis Stelios; Eliades. “The battle of the leakage detection and isolation methods 2020: Overview and results”. In: *Zenodo DOI* 10 (2020).

[101] Stelios G Vrachimis et al. “BattLeDIM: Battle of the Leakage Detection and Isolation Methods”. In: (2020).

[102] Qi Wang et al. “Two-objective design of benchmark problems of a water distribution system via MOEAs: Towards the best-known approximation of the true Pareto front”. In: *Journal of Water Resources Planning and Management* 141.3 (2015), p. 04014060.

[103] Xiaoting Wang et al. “Burst detection in district metering areas using deep learning method”. In: *Journal of Water Resources Planning and Management* 146.6 (2020), p. 04020031.

[104] *WaterGEMS WS WaterGEMS WS website*. `https://virtuosity.bentley.com/`. Accessed: 23-05-2022.

[105] Yipeng Wu and Shuming Liu. “A review of data-driven approaches for burst detection in water distribution systems”. In: *Urban Water Journal* 14.9 (2017), pp. 972–983.

[106] Yipeng Wu et al. “Burst detection in district metering areas using a data driven clustering algorithm”. In: *Water research* 100 (2016), pp. 28–37.

[107] Yipeng Wu et al. “Using correlation between data from multiple monitoring sensors to detect bursts in water distribution systems”. In: *Journal of Water Resources Planning and Management* 144.2 (2018), p. 04017084.

[108] Zheng Yi Wu, Paul Sage, and David Turtle. “Pressure-dependent leak detection model and its application to a district water system”. In: *Journal of Water Resources Planning and Management* 136.1 (2010), pp. 116–128.

[109] Xiang Xie et al. “Leakage identification in water distribution networks with error tolerance capability”. In: *Water Resources Management* 33.3 (2019), pp. 1233–1247.

[110] Weirong Xu et al. “Disturbance extraction for burst detection in water distribution networks using pressure measurements”. In: *Water Resources Research* 56.5 (2020).

[111] Guoliang Ye and Richard Andrew Fenner. “Kalman filtering of hydraulic measurements for burst detection in water distribution systems”. In: *Journal of pipeline systems engineering and practice* 2.1 (2011), pp. 14–22.

[112] Guoliang Ye and Richard Andrew Fenner. “Weighted least squares with expectation-maximization algorithm for burst detection in UK water distribution systems”. In: *Journal of Water Resources Planning and Management* 140.4 (2014), pp. 417–424.

[113] Dina Zaman et al. “A review of leakage detection strategies for pressurised pipeline in steady-state”. In: *Engineering Failure Analysis* 109 (2020), p. 104264.

[114] Qingzhou Zhang et al. “Leakage zone identification in large-scale water distribution systems using multiclass support vector machines”. In: *Journal of Water Resources Planning and Management* 142.11 (2016), p. 04016042.

[115] Mengfei Zhou et al. “An integration method using kernel principal component analysis and cascade support vector data description for pipeline leak detection with multiple operating modes”. In: *Processes* 7.10 (2019), p. 648.

[116] Xiao Zhou et al. “Deep learning identifies accurate burst locations in water distribution networks”. In: *Water research* 166 (2019), p. 115058.

[117] Bhagvat Zolapara and Maulik Joshi. “Designing Water Supply Distribution Network using Loop Software for Zone-I of Village Kherali”. In: (2015).