ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE of TECHNOLOGY (AAiT)
SCHOOL of ELECTRICAL and COMPUTER ENGINEERING

Water Leakage Detection and Localization using Hydraulic Modeling and Classification Approach

By

Eliyas Girma

March 10, 2020

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE of TECHNOLOGY (AAiT)
SCHOOL of ELECTRICAL and COMPUTER ENGINEERING

Water Leakage Detection and Localization using Hydraulic Modeling and Classification Approach

By

Eliyas Girma

Advisor

Dr. Surafel Lemma

A thesis submitted to the School of Electrical and Computer Engineering in partial fulfillment for the Degree of Master of Science in Computer Engineering

March 10, 2020

Addis Ababa, Ethiopia

Water Leakage Detection and Localization using Hydraulic Modeling and Classification Approach

By:Eliyas Girma

Thesis Advisor	Examiner
Dr. Surafel Lemma	
Chairman of Department	Examiner
Dr. Yalemzewud Negash	

Submitted in Partial Fulfillment of the Requirements for Masters of Science in Computer Engineering

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

March 10, 2020

Declaration

I ,the undersigned, declared this thesis work is my original work. The work has not been presented elsewhere for assessment and all sources of materials and literary works referred for the thesis have been fully acknowledged.

Eliyas Girma

Date of Submission: March 10, 2020

Place: Addis Ababa,Ethiopia

Acknowledgment

I find myself in need of thanking my wonderful agent, the Almighty God, Allah for the strength and his blessing in completing this thesis.

I would also like to thank and appreciate my tireless advisor Dr. Surafel Lemma, who teaches me how to do research and helped me shape it into what you see now. I am so grateful for his consistent supervision and support. His invaluable help of constructive comments and suggestions throughout the problem formulation and thesis works have contributed to the success of this research.

In addition, I would like to acknowledge my seminar supervisors and fellow colleagues. I especially want to thank Mr. Menor Tekeba for his comments. I want to thank Mrs. Ethiopia Bisrat from the Department of Civil and Environmental Engineering, A.A.i.T, for answering my questions on Hydraulics related concepts. I also want to thank Addis Ababa Water and Sewerage Authority(AAWSA)crews for providing me what I inquired and arrange a visit on Legedadi reservoir.

Of course, this acknowledgment would not be complete without thanking My fiance Asiya Abdullahi, who had excellent advice to help me with the mysterious inner workings of young womans mind, and my priceless parents for their endless love, prayer, kindness, moral, and financial support for the exceptional accomplishment of this thesis.

Abstract

The distribution of treated water over a water distribution system faces different losses. In these distribution systems, a significant percentage of water is lost due to leakages. There are plenty of researches that were proposed to tackle this water resource wastage. Most of them, however, focus on large bursts and single leakage detection and localization.

This thesis proposed a leak detection and localization approach that uses hydraulic modeling and classification approaches for detection, and a statistical approach for localization. The general term used in identifying and locating leaks is Leakage Detection and Localization. The detection system aims at detecting small leakages (i.e., below 5% of the total supply). The localization phase of this thesis also aims at localizing multiple leakages.

The proposed approach is applied on Hanoi water network benchmark by doing different experiments. A realistic dataset is produced based on a benchmark dataset procedure (i.e., LeakDB). The detection and localization results under well-known calibrated hydraulic conditions are almost perfect (i.e., 100% for detection and $\geq$90 for localization). The introduction of demand and noise uncertainty, however, affects the accuracy of the detection of small leakages. The localization methodology has shown effectiveness in locating two and three leaks that are presented simultaneously in 9 of 20 cases on average accurately, and 11 of 20 cases at least one of them.

Keywords: Leakage Detection,Localization, Hydraulic Modeling, Classification, Combined residuals

Contents

List of Figures

List of Tables

Acronyms

LDL Leak Detection and Localization

GPR Ground Penetrating Radars

WDS Water Distribution System

NRW Non-Revenue Water

DMA District Metered Area

SCADA Supervised Control And Data Acquisition

CUSUM Cumulative Sum

AI Artificial Intelligence

MNF Minimum Night Flow

ANN Artificial Neural Network

KNN K-Nearest Neighbor

SVM Support Vector Machine

WNTR Water Network Tool for Resilience

LeakDB Leak Diagnosing Benchmark

IWA International Water Association

Chapter 1

Introduction

1.1 Motivation

Water is one of the natural resources that a human being cannot live without it. It has been common practice to distribute potable freshwater through long pipelines. In these distribution systems, a significant percentage of water is lost due to leakages from pipelines. Each year more than 32 billion m^3 of treated water is lost due to leakage from distribution networks [1] globally. The total cost was estimated at nearly 3 billion dollars per year due to WL. About 90 million people would be possible to be served by saving half of those losses [2]. The leakage is mainly due to the usage of old pipelines, inadequate corrosion protection, poorly maintained valves and sometimes due to mechanical damage. Mechanical damage is a damage caused during excavation for construction and building roads.Avoiding these causes to prevent pipeline damage is practically impossible. Hence, the focus of many researches and companies is on reducing the loss associated with water leakage by minimizing the detection of leaks and localization time.

Some researchers in [3] believe that it is not necessary for a detection system to entertain much smaller leakages in which the total cost for repairing such leaks exceeds the cost of leaving them. However such leaks will last longer until they become large leaks; and hence, could incur high cost. In addition to the silent water loss, different contaminants can intrude on the water distribution system. System

as some sewerage systems are also combined with the piping system. Therefore, we believe the issue of addressing small leakages is necessary.

The idea of localizing multiple leaks is an open challenge. Most of the research works focus on localizing a single leakage [3]–[6]. Therefore standing on the shoulder of these state-of-the-art works, it is tried to localize multiple leakage areas.

1.2 Background

The huge amount of water wasted due to leakages makes a significant economic loss for the water supply companies. The general term used in identifying and locating leaks is Leakage Detection and Localization (LDL). The early detection of these leakages is considered as conservation of water resources. However, it is not limited to conservation only but also environmental protection, health, and safety issue. Additionally, it is assumed from the perspective of water supply companies as an honor and reputation on asset management.

From the perspective of asset monitoring and water resource conservation, different techniques are being used by the water companies to detect and locate leakages. The techniques used for the purpose of detection and localization of water leaks on water distribution systems are categorized broadly as hardware-based and software-based methods. This classification is based on the mechanism used for detection and localization [7], [8].

1. Hardware-Based Methods -These methods are based on hardware devices that detect bursts or leakages on pipelines. These devices can be listening rods, leak correlators, leak noise loggers, Ground Penetrating Radars (GPR), gas injection and thermal infrared imaging devices. The hardware-based methods can be acoustic and non-acoustic. Acoustic leak detecting devices use the sound wave which is generated from leaking pipes to specify the exact location. Non-acoustic devices locate the leak through visual means that is either a radar image or a thermal infrared image.

Hardware-based methods are accurate in finding and locating leaks, but they are very expensive devices working in a limited section of the water distribution system at a time [7]. They are also labor-intensive i.e., they need a lot of skilled manpower for finding leaks in a large area. The other limitation of hardware-based methods is related to the distance they cover. The methods usually cover short distances; and hence, can be used to only a small section of the water distribution system. Table 1.2 summarizes the advantages and disadvantages of hardware-based methods.

Advantage	Disadvantage
• very high accuracy • Low error in localization	• very expensive • Labor-intensive (needs huge manpower) • Time-consuming • Detection range is limited • Remote monitoring is impossible

Table 1.2: Advantage and Disadvantage of HW based methods

Software-Based Methods - are methods that relied on some algorithm or some data analyzing tool on finding anomalies i.e., leaks in the data pattern. These methods are developed on tackling the limitation of hardware based methods. They can be used for monitoring the whole Water Distribution System (WDS). They are relatively cheaper than the hardware-based methods and they are not labor intensive [7].

Due to the development of high-resolution hydraulic sensors (pressure and flow), we can monitor the system state of the water distribution system. These hydraulic sensors are utilized in supervised control and data acquisition (SCADA) systems for monitoring water quality and hydraulic conditions. Extracting knowledge from these collected data becomes common nowadays. Different hydraulic sensors measure time series data of vital parameters like pressure, flow rate, and temperature. The time series sensed data is further analyzed to show any leakages or contamination level of the water in pipelines. Therefore, software based methods get their data from these hydraulic sensors.

The main theme for software-based methods is having predicted system states and comparing them with the current observation of the system state for finding abnormal events like leakages and contamination level. The approaches used in software based methods could broadly be classified as hydraulic model and data driven approaches. The works under these two categories are discussed in Chapter 3.

1.3 Problem Statement

Many of the state-of-the-art leak detection and localization works focus on single leak localization [3]–[6]. In reality, however, a water distribution system usually has multiple leaks. In addition, most research works focus on detecting and finding large bursts on pipelines [8]–[11]. Model-based leak detection approach that is proposed in Mashford et.al [3] tried to simulate small leakages up to 3.4 lit/hr. Mashford et.al's result showed that there were no difference in pressure from the no leak scenario in any of the simulations. The main reason for this is that they followed demand driven simulation option. In demand driven option the demands are changed irrespective of the pressure variation. In this case it is not suitable for a detection system to depend on pressure measurement. While the simulation option pressure driven demand gives a variation in pressure values due to leak demand variation. The lowest leakage rate with a significant pressure difference was recorded above 90 lit/hr(1.5 lit/min). Therefore, it is needed for the experiments on small leakage detection,(i.e., 100 lit/hr) to run on pressure driven demand simulations.

Therefore the purpose of this research is to address the two problems i.e., detection of small leakages and localization of multiple leaks. Particularly this research aims to address the two problems stated using a hybrid i.e., hydraulic modeling and data-driven approach. This study aims to answer the following research questions (RQ)s.

- **RQ1 [Small Leak Detection and Localization]**

Can we detect and localize small leakages($\leq$ 5% of total supply) using the proposed approach?

- **RQ2 [Robustness to Uncertainty]**
  What is the effect of adding demand and noise uncertainty on both detection and localization?

- **RQ3[Multiple Leak Localization]**
  Can we locate multiple leaks with the proposed approach?

1.4 Objective

1.4.1 General Objective

The main objective of this research is to come up with water pipeline leakage detection and localization mechanism that considers on detecting small leaks and localization of multiple leaks. The approach will try to combine hydraulic modeling with a data driven approach to detect water pipeline leakage and to localize the leak hot spot.

1.4.2 Specific Objective

- Use hydraulic modeling software to generate representative leak dataset.
- To develop leak detection approach that uses combined residual of pressure and flow data.
- To localize multiple leakages using statistical residual analysis.
- To evaluate the proposed leak detection and localization approach and compare it with previously proposed similar and benchmark approaches.

1.5 Research Methodology

In this research, the first task was doing literature survey on the state-of-the-art detection and localizing of leakages in water pipeline distribution networks. After analyzing the positive contributions from different methodologies, the proposed approach was designed.To evaluate the proposed approach and answer the research questions, different experiments were conducted. Finally, our experiment results are compared to previous related works and a conclusion is made based on the result.

1.6 Scope

This research aims is to develop leak detection and localization mechanism. The mechanism is centered on analyzing the leak pattern from a model-based hydraulic data. Due to unavailability of real data from current real water distribution systems, the methodology is tested according to a realistic Benchmark Dataset i.e., LeakDB [12]. Additionally the water network model that is used in this case study is Hanoi[13] benchmark water network.

1.7 Contributions

The main contributions of this thesis to the research field of water detection and localization are as follows:

1. A strong detection methodology that utilizes hydraulic modeling for extensive and representative leakage data generation and classification that separates observational data as leaky and non-leaky is proposed. Due to this capability, small leakages can be detected even in the introduction of slight demand uncertainty. However, a well-calibrated model is a precondition for this approach.

2. The previous hybrid approaches was only used for localization, while in this research it is used for detection of leakages. The proposed approach detects

small leakages that are less than 100 lit/hr (i.e.,less than 5% of the total supply).

3. For Training and classification purpose the residuals used in the detection methodology is the combination of available pressure and flow rate residuals. This makes it unique from other works and also improves the detection accuracy.

4. The localization approach is a statistical approach for searching extrema values in residual space for consecutive time steps after where the leak is detected. Using this approach the detection of multiple leaks has been solved.

1.8 Thesis Organization

The rest of this document is organized as follows. Theoretical backgrounds about leakages are explained in Chapter Two. Chapter Three presents summary of previous works proposed for leak detection and localization. Our proposed approach is explained in Chapter Four. Chapter Five elaborates the experimental scenarios, the results from the experiments and the comparison result with different related works. Finally Chapter Six presents conclusions from experimental observations and future works.

Chapter 2

Theoretical Background

This chapter discusses the theoretical concepts in water distribution systems (WDS), water leakage property, hydraulic modeling, and leakage detection schemes.

2.1 Water Distribution System

A Water Distribution System (WDS) is part of a water supply network with components that carry potable water from a treatment plant to consumers and different distribution pipelines. After water is collected from surface and ground sources, it is treated using chlorine and other chemicals to make it potable and is distributed using WDS.

A WDS contains a source (treatment plant and reservoir), pipelines, junctions (nodes in which different pipelines meet), pumps, and valves (flow control and pressure reducing valves). Pipelines are the links that transport water from any two junction nodes. There are three types of pipelines i.e., the transmission mains, distribution mains, and service lines. Large diameter water transmission mains called primary feeders are used to connect water treatment plants and service areas. Secondary feeders are connected between primary feeders and distribution mains. Distributors are water mains that are located near the water users, which also supply water to individual fire hydrants. A service line is a small-diameter pipe used to connect from a water main through a small tap to a water meter at the user's

location.

The other components are water storage tank facilities, or distribution reservoirs, that provide clean drinking water storage to ensure the system has enough water to service in fluctuating demands and to equalize the operating pressure [14]. Table 2.2 shows the main components of a Water Distribution System(WDS) with their function. Further description about a water distribution system could be found in [14].

Element	**Type**	**Primary Modeling Purpose**
Reservoir	Node	Provides water to the system
Tank	Node	Stores excess water within the system and releases that water at times of huge usage
Junction	Node	Removes (demand) or adds (inflows) water from/to the system
Pipe	Link	Conveys water from one node to another
Pump	Node/Link	Raises the hydraulic grade to overcome elevation differences and friction losses
Control Valve	Node/Link	Controls flow or pressure in the system based on specified criteria

Table 2.2: Common Water Distribution Elements. Taken from [14]

2.2 Leakages

The amount of water that is calculated from the difference between the supplied and consumed by the billed customers is termed as non-revenue water (NRW). In general according to International Water Association(IWA), water intended for consumption is segmented into 'Authorized Consumption', which corresponds to the billed or unbilled authorized consumption, and to the 'Water Losses', which corresponds to the 'Apparent Losses', due to unauthorized use, metering inaccuracies or calibration issues, and to the 'Real Losses', due to leakages, breaks, etc [15].

Leakages are real losses that are discharged from the water distribution system without affecting the total consumption or with slight interruption of water supply. While bursts are the unrecoverable water loss with high interruption of water supply to customers. Leakages can be classified in to active and passive. Passive leakages are leakages that are reported and visible on the ground while active (background) leakages are neither visible on the ground nor reported. According to leaking area coverage, leakages are also classified as abrupt and incipient [12]. Abrupt leakages are those leaks that have the same leaking area from the start to end. Incipient leakages are those that increase their flow gradually and cover more area over time.

Many water companies divide their WDS into zones by which they measure the inlet and outlet flow for leak assessment. We call this zonal section of the WDS a District Metered Areas(DMA). In a DMA, we have different pressure and flow sensors that are distributed. These DMA sensors combined with the SCADA systems are used with software based methods for leakage detection and localization. The software based methods for leakage detection and localization are classified in to hydraulic modeling based and data driven based approaches [7], [8].

The software-based methods aim at analyzing hydraulic data to find anomalies in the WDS. The occurrence of leak/burst increases the flow rate that enters the section of the WDS and decreases the pressure after the leak point. This event leaves a different signature from normal operational conditions. We call this change transient event due to leak. But transient events are not only caused by leakages/bursts. Boundary condition changes like valve/pump operational condition changes also create this effect. Therefore a typical LDL scheme must consider these effects.

2.3 Hydraulic Modeling

This section clarifys about hydraulic modeling. Hydraulic modeling is the process of using a mathematical representation of the real WDS based on energy and mass balance equations. Water network simulations, which replicate the dynamics of an existing or proposed system, are commonly performed when it is not practical for

the real system to be directly subjected to experimentation, or for the purpose of evaluating a system before it is actually built. Hydraulic simulations can be used to predict system responses to events under a wide range of conditions without disrupting the actual system. Using simulations, problems can be anticipated in proposed or existing systems, and solutions can be evaluated before time, money, and materials are invested in a real-world project [14].

EPANET[16] is a software application used to model water distribution systems. EPANET is developed by the United States Environmental Protection Agency (U.S. EPA) to solve nonlinear hydraulics of a WDS. EPANET uses the Global Gradient Algorithm (GGA), a variant of the Newton-Raphson method [17] to perform hydraulic analysis of any WDS section. With EPANET, users can perform extended-period simulation of the hydraulic and water quality behavior within pressurized pipe networks, which consist of pipes, nodes (junctions), pumps, valves, storage tanks, and reservoirs. It can be used to track the flow of water in each pipe, the pressure at each node, the height of the water in each tank, a chemical concentration, the age of the water, and source tracing throughout the network during a simulation period [18]. EPANET has a programmer's toolkit that allow developers to customize EPANET to their own needs. WNTR (Water Network Tool for Resilience) is one of the python packages that calls the libraries in EPANET toolkit.

The input for the hydraulic models are the pipe characteristic information including the pipe length, diameter, elevation, roughness coefficient of pipelines and demands on each node. In addition some boundary conditions like pump/valve closure information is given. The main WDS components and their properties of the WDS are listed in Table 2.3. By solving the mass balance and energy equations, the output of modeling gives the flow rate and pressure at a given pipeline or junction node. These simulation outputs can be taken as sensor inputs in SCADA systems in the absence of real time data for experimental purpose.

In WDS modeling the model used to predict WDS behavior under a number of operational conditions must be up-to-date [14]. The main goal of calibration (up-

dating WDS behaviors) is to change the WDS hydraulic parameters to reduce the error between the WDS hydraulic model state (outputs) and its corresponding system observations. The calibrated WDS hydraulic parameters (physical attributes) give information concerning the physical state of the WDS hydraulic model [8]. The common WDS hydraulic parameters that are calibrated are:

- Unmetered demand coefficients or nodal base demand
- Pipe roughness
- Valve setting or status
- Pump status or flow/head graph (pump curve)

The acceptable difference between the WDS hydarulic model states and observed values (flow and pressure) must be within 5% -10% in the water industry [19].

Componenets	**Static Properties**	**Dynamic Properties**
Storage-(Tanks/Reservoir)	Maximum capacity, Maximum height	Level (Pressure head)
Pipes(Links)	Length,diameter,roughness	Flow rate,Head loss
Nodes (junctions)		Pressure head,Nodal demand
Valves	Tau Value	Flow rate,Head loss, Tau Value
Pump		Pressure head,Pump Speed

Table 2.3: Water Distribution System Model Properties. Taken from [8]

Residual analysis

The detection of leakages with the hydraulic model is mostly based on either residual analysis between the WDS hydraulic model predictions and the WDS observations [4]–[6], [20]–[23] or evaluation of the pattern of WDS state estimates [24]–[26]. The focus of this paper is on residual analysis approach.

Residual analysis method is a kind of anomaly (novelty) detection. The residue is calculated as the difference between the observational (field measured) hydraulic data and a prior established baseline (i.e., either predicted data from historical data using data-driven approaches or using hydraulic modeling software). This residual vector, R, is determined by the difference between the measured pressure at inner nodes where sensors are installed and the estimated pressure at these nodes (see Equation 2.1).

$$R(t) = p(t) - p_o(t) \tag{2.1}$$

where p is measured pressure at inner nodes at time t; and, p_o is the estimated pressure at these nodes at the same time, t. p_o is obtained using the network model considering a leak-free scenario.

In some researches[23], [27] these residuals are evaluated against a threshold τ that is selected to take into account the measurement noise and model uncertainty. If the residuals are greater than this threshold, leak is detected and localization is instantiated. However, the choice of this threshold affects the quality of detection. Setting too high values makes the detection to miss leak events while setting the threshold value too low results in many false positives.

Hydraulic modeling highly depends on calibrated and accurate inputs for modeling purposes in order to represent reality. For this reason, some modeling techniques like adding demand uncertainty on the demand nodes and adding measurement noise on the pressure and flow data are considered [28]. Nowadays online hydraulic modeling is also being proposed[8] by integrating the modeling software with the SCADA output for getting perfect modeling prediction.

For the purpose of studying leakages different works used residual analysis method[4]–[6], [8], [20]–[23] . For the purpose of simulating leakages some of the researches [9], [29], [30]are used engineered events like opening fire hydrant. However, for the purpose of studying leaks in the absence of real field data, the practice of generating leak events using a modeling software is customary [4]–[6], [8].

One of the localization methods that uses residual analysis is the sensitivity method. The sensitivity method is a leak localization approach based on comparing the residual vector (obtained from the difference between measured and expected pressures of each sensor) against the leak sensitivity matrix that contains the effect of each possible leak in each residual. Equation 2.2 shows the sensitivity matrix S.

$$S = \begin{bmatrix} \frac{\partial P1}{\partial f1} & \cdots & \frac{\partial P1}{\partial fn} \\ . & \ddots & . \\ \frac{\partial Pn}{\partial f1} & \cdots & \frac{\partial Pn}{\partial fn} \end{bmatrix} \tag{2.2}$$

where each element s_{ij} of the sensitivity matrix S measures the effect of a leak f_j in the pressure of sensor p_i (i.e., the difference of pressure between the expected pressure and the one measured when a leak of magnitude f occurs in the node j). Each element is normalized according to the leak magnitude. The sensitivity matrix S has as many rows as sensors and as many columns as considered leaks. It is extremely complex to calculate S analytically in a real network since the model is based on a huge set of implicit non-linear equations. Instead, the previous works generate the sensitivity matrix by simulation software (as EPANET). First, the computation of the sensitivity matrix needs the construction of the non-leaky operation scenario of the network in a 24-hours time horizon, which allows to obtain the vector p(k) for the non-faulty pressure of each node of the network. Then, leak scenarios are considered in simulation by introducing a leak at a time in each node of the network. The pressures of the sensors in case of each considered leak scenario are stored in the matrix. Finally, using the non-leaky vector and leaky matrix , the sensitivity matrix S for each time instant of the horizon considered is computed. Thus, the sensitivity matrix is composed of (n x m) elements where each element is determined by computing the difference between the non-leaky and the leaky pressure obtained by simulation normalized with respect to the magnitude of the leak used to obtain the sensitivity matrix.

The model-based leak localization approaches that rely on the sensitivity-to-leak analysis [8], [22], [23], [27] matches the residual vector in Equation 2.1 against the columns of sensitivity matrix (Equation 2.2) using some metrics. More details about

the sensitivity-to-leak analysis can be found in [23]

The other commonly used data-driven localization approach is a statistical classifier. The statistical classifier takes in the residual vectors as an input and gives the location of the leak. As it is known machine learning algorithms depend on huge data for learning and pattern recognition. For this reason, hydraulic modeling is used to generate extensive leak residual data. Further researches that used this methodology will be explored in the literature review section.

Chapter 3

Literature Review

In this Chapter,we will discuss literatures that use software-based (SW) methods to detect and localize water leakage. The software-based methods are further categorized based on the approach they use to identify and localize leaks: hydraulic modeling and data-driven approaches. This Chapter is organized into three parts. The first part focuses on hydraulic modeling approaches. The second part is on the data-driven approach and the last part discusses the related works with the proposed approach (i.e., hybrid approach).

3.1 Hydraulic Modeling Based Approaches

Hydraulic modeling-based approaches use a hydraulic modeling software system. EPANET [16] and WATERGEMS [31] are the most common hydraulic modeling software systems. The detection and localization scheme relies on statistical comparison between the model prediction and the data (i.e., flows and pressures) from multiple hydraulic meters via SCADA systems. Here we review different papers that are based on hydraulic modeling and residual analysis.

Model-based leak detection and isolation techniques have also been studied starting with the seminal paper of Pudar and Liggett [20] which formulates the leak detection and location problem as a least-squares estimation problem. They used a non-linear derivative-based optimization method to detect burst/leak with a hy-

draulic model and artificial pressure and flow measurements. The sum of squared differences between predicted and artificial pressure measurements are used as the objective function. They concluded that an accurate hydraulic model parameter and a high number of measured data are required to improve the detection rate. They also used a sensitivity matrix for optimal sensor placement.

Pérez et al,(2009)[21] developed a methodology for pressure sensor placement for leak detection. The work utilized the sensitivity matrix approach that is proposed in [20]. The proposed approach was applied in three real case studies using Barcelona water network. A genetic algorithm was used to optimize the number of sensors to be used.

Pérez et al,(2010)[27] proposed a leakage localization method based on the pressure measurements and pressure sensitivity analysis of nodes in a network when a leak is present in a node. It used a model-based method that relies on pressure measurements and leak-sensitivity analysis. This methodology consists of computing online residuals, i.e., differences between the measurements and their estimations obtained using the hydraulic network model, and checking them against thresholds that take into account the modeling uncertainty and the noise. When the residuals violate their threshold they are matched against the leak sensitivity matrix in order to discover the leaking node. Although this approach has good efficiency under ideal conditions, its performance decreases due to the nodal demand uncertainty and noise in the measurements. The performance is also affected by the boundary condition changes. Under well calibrated model and absence of noise the result is perfect(100%) but when the level of uncertainty (addition of noise and demand fluctuation) increases the leaks were localized in a neighbor zone.

Ponce et.al[23] proposed a leak detection mechanism that is based on computing the difference (residual) between the pressure measurements against their estimation by means of hydraulic models. Five different ways of using the leak sensitivity matrix to isolate the leaks are described and compared. The five methods are binarized sensitivity method [27], angle between vectors method, correlation method, euclidean

distance method, and least square optimization method. The performance of these methods has been compared when applied to two academic water distribution networks i.e., Hanoi and Quebra water network. Finally, the three methods with better performance are applied to a district metering area (DMA) of the Barcelona WDN. Results have shown that the angle method increases the capability of isolating leaks in a great number of cases.

Okeya[8] proposed an online hydraulic modeling based leakage detection and localization methodology. The online hydraulic modeling is based on data assimilation coupled with short term demand forecasting methodology in a sampling time of 15 minutes. The online hydraulic modeling approach uses realistic data as input for modeling the next time step data. The detection system developed is a burst detection metric based on the moving average residuals between the predicted and observed hydraulic states (flows/pressures). For localization it is used sensitivity matrix approach. The methodologies for burst detection and localization were applied to two real District Metred Areas in the United Kingdom (UK) with artificially generated flow and pressure observations and assumed bursts. The results obtained this way show that the developed methodology detects pipe bursts in a reliable and timely fashion providing a good estimate of a burst flow and approximately locates the burst within a DMA. Their proposed approach, however, has difficulty detecting low burst magnitudes.

Summary

In the aforementioned approaches the sensitivity matrix method was used for localization. The sensitivity approach has some limitations. When the sensitivity matrix is generated, it considers some nominal leakage threshold. If the real leak size is different from the calculated one it would give wrong localization result. Additionally, it is affected by nodal demand uncertainties. The other point to notice is that most of the works depend on pressure measurements only.

Modeling approaches need high calibration of parameters i.e., accurate inputs are the preconditions that guarantee the modeling quality. The main advantage of using a modeling approach is that it models dynamic system conditions given accurate (calibrated) parameters.

3.2 Data-Driven Approach

Data-driven approach provides a mapping between the inputs and outputs of a given system, with the advantage of not requiring a detailed understanding of the processes that affect a system. It is emerging as an attractive option for the prediction of parameters and classification in water systems. The data-driven techniques analyze the field values from a SCADA system in near real-time to detect abnormal flows or pressures. These techniques do not make use of a hydraulic model, i.e., physically-based models to estimate hydraulic states (flow or pressure) of a WDS. This approach focuses on extracting meaningful information in time series data using different statistical [24], [32], [33] and artificial intelligence[10], [34], [35] approaches. According to Wu and Liu [36] data-driven approaches are classified as a classification method, prediction-classification method, and statistical methods.

The Minimum Night Flow (MNF) analysis is the simplest method employed by different water companies to detect pipe bursts and leakages. The authorized water consumption at the night time, from midnight up to morning 5 am, is usually less and the flow is recorded as a minimum on normal days. When leaks are present in WDS it is impossible to detect them in day time due to huge nodal demands. However, at night time the effect of leaks can be seen. The MNF monitoring technique involves comparing the current MNF to the acceptable MNF of the DMA on leak-free days. We cannot take the MNF analysis method as an accurate detection technique because the previous MNF data cannot guarantee if there is no leak.

Eliades and Polycarpou [37] used adaptive inflow approximation and leakage detection methodologies. They used CUSUM (cumulative sum) statistical method to detect changes in the residual space. They compared their work with a modi-

fied version of the MNF analysis method. However, the detection is dependent on threshold selection. The MNF analysis they used is described in the results section for comparison with the proposed approach.

Mounce et.al,(2003)[38] proposed the use of a Mixture Density Artificial Neural Network (MDANN) to predict flow value for the next 24 hours ahead based on historical flow data. The prediction is compared with the observed flow data. A classification module was used for analyzing the observed and predicted data. It gives the level of abnormality in the observed data. The problem of using these historical data to predict future values is dependent on the confidence of the historical data free from any abnormality (i.e., leakage or boundary change). The other problem of using conventional ANN for time series prediction is that they are only good predictors for problems that have static patterns but not for the dynamic non-linear systems. However, the use of temporal pattern recognition algorithms may reduce the problem of finding leakages in such dynamic systems.

Considering the dynamic pattern of WDS Mounce and Machell,(2006) [10] characterized the flow of water in a pipeline by a time series. A neural network was learned to recognize the signature of the flow under normal conditions and when leakage occurs. Two prototypes, static and time delay neural networks were developed and tested on data recorded in part from a real water distribution network. In terms of burst detection, the time delay network identified 75% of the test (unseen) burst compared to 4% for the static. This is because static neural networks doesn't have memory component. Static neural networks turn a temporal sequence into a spatial pattern encoded on the input layer.

Mounce and Boxall,(2011)[34] applied a Support Vector Regression (SVR) technique to detect pipe bursts using the unusual differences between WDS flow and pressure observations and predictions. Prior to applying the SVR to output flow and pressure predictions, historical flow and pressure observations collected over six months are used to train the SVR method. The technique makes use of both flow and pressure observations from real WDS without using a hydraulic model to detect

pipe bursts ranging in size from approximately 10% to 50% of the average daily flow. The technique is applied in real WDS data to detect burst offline. It is found that SVR methodology raises alarm faster than their previous method[39] when there is a burst of 10% to 50% . The system also detected engineered of known size (2 l/s) representing 6% of the maximum flow entering one DMA.

Carreno-Alvarado et.al[40] compares three machine learning algorithms i.e., principal component analysis (PCA), support vector machine (SVM), and relevance vector machine (RVM) for leak detection and isolation. They train the classifiers with different leak and non-leak data. After validating it, the classifier is be fed from SCADA data directly to perform binary classification. They used the Hanoi benchmark network to compare the two classifiers by taking two node's data for training and testing the classifiers. The authors did not clearly state their experimental setups rather they emphasized on selecting between SVM and RVM. The performance result for SVM and RVM was 12/17 and 11/17 respectively on leakage detection.

Summary

Data-driven approaches are good for parameter prediction [34] and classification [3]. In addition, we do not need to care about the system working principle. However, in applying to dynamic WDSs care must be given. Some data-driven approaches like static ANNs are not robust to boundary changes (pump/valve operational changes). Therefore to take advantage of these techniques, it is suitable to use a mixed/hybrid of model-based and data-driven approaches for leak detection and localization.

3.3 Related Work

Hybrid Approach

The problem of different dynamic hydraulic phenomena that could not be solved by some data-driven approaches has been solved with a calibrated hydraulic mod-

eling [8]. On the other side, data-driven approaches like classification have shown promising results in pattern recognition and anomaly detection [29]. Therefore by the combination of the two approaches, it is believed to fill the gap for localization of leakages [3]–[6].

The development of hydraulic modeling software systems on simulating leakages solve the problem of data shortage in training of the data-driven algorithms for prediction and classification.

Mashford et.al[3] used Support Vector Machines (SVMs) to analyze a collection of pressure data to obtain information concerning the location and size of leaks in the WDS. The SVM is trained on perfect pressure data which are generated from the EPANET model representing leaks by varying emitter coefficients. The leaking pressure values are given to the classifier to localize the nodes that are leaking. The results show that the location and size of leaks are predicted with a good degree of accuracy. However, the EPANET model and observation uncertainties are not included in the SVM hence the algorithm is unlikely to be implemented for practical purposes. The other limitation mentioned was detecting small leaks less than 100 l/hr or 2.5 l/s in d. Their result showed that there were no difference in pressure from the no leak scenario in any of the simulations. The main reason for this is that they followed demand driven simulation option. In demand driven option the demands are changed irrespective of the pressure variation. In this case it is not suitable for a detection system to depend on pressure measurement. While the simulation option pressure driven demand gives a variation in pressure values due to leak demand variation. The lowest leakage rate with a significant pressure difference was recorded above 90 lit/hr(1.5 lit/min). Therefore, it is needed for the experiments on small leakage detection,(i.e., 100 lit/hr) to run on pressure driven demand simulations.

Use of Classifier on Residuals

The limitation of the sensitivity matrix approach for localization of leakages that were mentioned earlier brought the idea of making the residual pressure data to be trained by a classifier for localization. Since in a geometric framework, the raw residuals do not provide fault confinement data and all the residuals are affected by the leaks. There is a high level of ambiguity that the nodal demands and leak size are unknown because of the noise affected the measurement taken [41]. Thus, the raw residuals are planned to be fed to the classifiers. In this situation, the use of the classifier for residual data does not encounter problems associated with purely data-driven methods . The use of residuals made the data dimensionality to be reduced.

In Ferrandez-Gamote et.al[4] a new leakage localization approach based on the combined use of models, pressure sensors and classifiers has been proposed. The hydraulic model of the network is used to generate residuals by comparing model predictions against the available measurements provided by sensors. Once residuals have been generated, they are analyzed using supervised classifiers in order to allow the leak localization. The classifiers are trained using data in leak scenarios in all the nodes of the network considering uncertainty in demand distribution, additive noise in sensor outputs and uncertainty in leak magnitude. Finally, the proposed approach has been tested using the well-known Hanoi [13] water network benchmark. For training, the pressure residuals are given as feature and the leak locations are given as a target for the classification. Multi-class classification is used by which the outputs were leaking nodes. For detection, the authors used MNF analysis. Results with training values have been computed with a leak magnitude value 10 l/s higher and lower than 60 l/s. They compared their work with Ponce's[23] angle method for localization using a sensitivity-matrix approach. The average efficiencies found are around 80% using the classifier method and around 50% using the angle method.

Soldevila et.al,(2016a)[5] continued on the above approach. In the first stage, residuals are obtained by comparing pressure measurements with the estimations

provided by a WDN model. In a second stage, a classifier is applied to the residuals with the aim of determining the leak location. The localization was proposed to be used after leakage detection using MNF analysis. A pressure model of the considered WDN is used in the first stage to compute residuals, i.e., differences between the measured (sensors) and estimated (model) values of the water pressure in nodes of the network, that are indicative of leaks. In a second stage, a classifier is applied to the obtained residuals with the aim of determining the leak location. The residual calculating and training the classifier is considered as off-line phase and the application of new calculated residuals for localization is considered as on-line phase.

This on-line phase relies on a previous off-line phase in which the network model is obtained and the classifier is trained with data generated in extensive simulations of the network. These simulations consider leaks with different magnitudes in all the nodes of the network, differences between the estimated and real consumer water demands and noise in pressure sensors. The underlying idea is to obtain a classifier that is able to distinguish the leak location independently of the unknown real leak magnitude and the presence of uncertainties associated with the water demands and the pressure measurements. KNN(k-Nearest Neighbor) algorithm is used for the classification task. For the localization to be enhanced they added temporal reasoning by taking an extended period of data samples. The proposed approach has been developed assuming a single leak only.

Soldevila et.al,(2016b)[42] used the same procedure to generate the residuals as their previous work [5] and then applied bayesian reasoning online to the available residuals to determine the location of leaks. The aim of the Bayesian reasoning is to determine the most likely leak that explains the mismatches between the measurements and the estimations. It is assumed that just one leak can be affecting the network in a given time instant (single leak assumption).

Soldevila et.al,(2017)[6] in a first off-line stage, the network model is simulated under different uncertainty conditions to obtain residual data for each potential leak

location (each network node). Then this data is used to calibrate probability density functions. Additionally, a study of the possible overlapped probability density functions based on the minimal statistic energy test is also proposed. In the online stage, a Bayesian classifier provides the time-dependent posterior probability of every possible leak. The effect of the time horizon in the decision has also been studied.

Summary

Modeling gives the detection and localization methodologies a reference to compare with the observed hydraulic data. On the contrary, classification approaches can guarantee on separating leaky and non-leaky scenarios. Hence, the hybrid approach helps to exploit advantage of each approaches. In line with the above works, we aim to use a hybrid approach to detect and localize water leakage.

However, the aforementioned state-of-the-art works focus on localization of single leakage. In addition, they all used the hybrid approach for localization only not for detection. The main reason for localizing single leak is that, the multi-label classification they used is resulting output one class at a time. Therefore it becomes a challenge localizing multiple leaks using a mixed classification approach in prior works. In this research, we aim to detect leakages that are small (i.e., $\leq$ 100 lit/hr) and localize multiple leaks.

Chapter 4

Proposed Methodology

In order to solve the gaps stated in the literature review summary a hybrid water pipeline leakage detection approach has been proposed. The approach followed for the detection and location of leaks uses residual analysis. Unlike prior works, the detection and localization phases are separated. After detecting a leak in the detection phase, the localization phase starts to localize the possible suspected leaking nodes.Below we present the details of the two phases.

4.1 Detection Phase

The first phase of the proposed approach is the detection phase. This phase utilizes both hydraulic modeling and data-driven approach. The hydraulic modeling is used to generate residuals that represent the system state. The generated residuals are fed to the classifier with leaky and non-leaky data for training. The major difference from previous works is that the proposed methodology uses a combination of pressure and flow residuals for training, validating and testing. We conjuncture that the combined residual enhances the detection rate. Generally, the detection phase consists of the following steps:

- Data generation
- Training and validating the binary classifier
- Detection

Data Generation

The chance of getting extensive real leak data is impossible. There is not enough leak data that could be used for training the classification model. Therefore, it has become common to generate leak residuals using modeling software [3]–[5] . However,using such data, care must be taken to capture the reality. The following concepts should be considered in the data generation step.

- Inputs for the model should be close to the reality; i.e., the modeling uncertainty, demand uncertainty, measurement uncertainty should be considered.
  - Modeling uncertainty means when designing a WDS using modeling software some asset properties can be changed through time. For example, when a broken pipe is replaced with the new pipe, the properties (internal diameter, roughness coefficient) also change. Such properties might not be updated on the model settings. Therefore when simulating using modeling approach it is useful to consider the modeling uncertainty by adding uncertainty on the prior modeling parameters.
  - Demand uncertainty describes the variations in demands (actual consumptions). In order for the model to be perfect, the demands given for the modeling software should be realistic enough. These demands can be regressed from historical real data using different prediction approaches. Currently, the use of deep learning for time series prediction becomes unbeatable [43], [44]. Either the use of on-line modeling [8] solves this problem by which real-time demand data from SCADA system can be used for modeling purpose.
  - Considering measurement noise from pressure and flow sensors is the other practical reason that makes the data realistic for representing the actual WDS. Some Gaussian noise is added to the modeling output before residual is generated.
  - Boundary conditions like valve/ pump closure or opening affect the system dynamics. Therefore, exact knowledge about the rules in which these

components work is very crucial. The rules that are given for the modeling software must match the reality of changing the operating conditions of these components.

- Considering the above conditions, two kinds of datasets are generated. The first is modeling the system in the absence of leak and the second is modeling the system in the presence of leakages. For leak modeling purpose we use the EPANET leak modeling equation that is proposed by Crowl and Louvar[45].

$$d_{leak} = C_d A p^{\alpha} \sqrt{\frac{2}{\rho}} \tag{4.1}$$

where d_{leak} is the leak demand m^3/s, C_d is the discharge coefficient (unitless), A is the area of the hole (m), p is the gauge pressure inside the pipe (Pa), α is the pressure exponent, and ρ is the density of the fluid. The default discharge coefficient is 0.75 (assuming turbulent flow), but the user can specify other values if needed. The value of α is usually set to 0.5 [46].

- In order for the data to be representative different leak magnitudes ,i.e., very small leakages(less than 10 lit/s) up to large leakages, using WNTR(Water Network Tool for Resilience)[46] tool is generated.

- Unlike all other previous works this thesis not only use pressure residuals but also flow rate residuals combined with the pressure residuals. Figure 4.1 illustrates the data generation phase diagrammatically.

Training and Validating the Classifier

The second step after generating representative leak residuals is training the classifier. The classifier used for detection is a binary classifier. The training residuals are labeled with two classes i.e., (0-no leak and 1-leak). The residual data in the leakage duration is labeled with 1 and labeled 0 where there is no leak in the data.

The pressure and flow residuals for different magnitudes of leak amount and leak location are combined as a single dataset. Then the combined dataset will be split

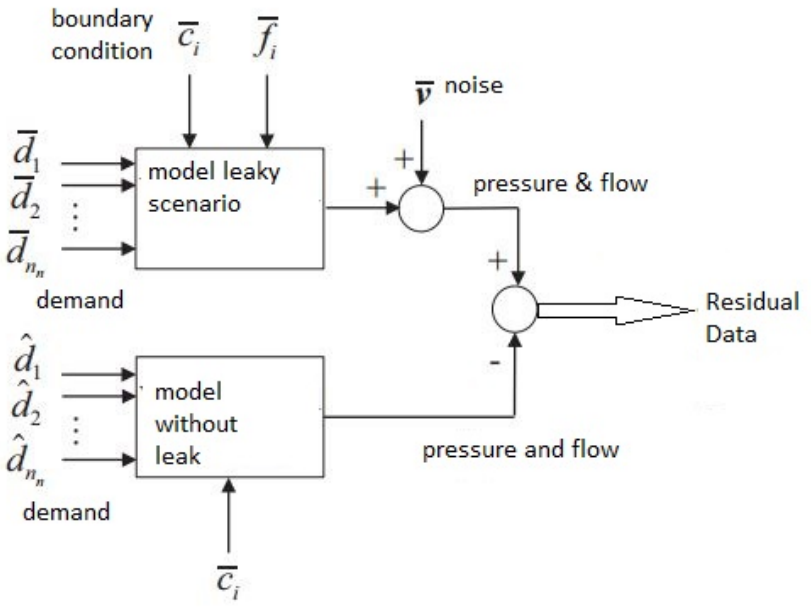

Figure 4.1: Residual Generation

for training and testing set. In this research, Support Vector Machine (SVM) is used for classification[47]. Given a training data, the support vector machine constructs a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples are maximized. The reason for choosing SVM is, it has an optimal hyperplane on separating different classes in minimizing the euclidean distance of the weight vector [48]. In addition the number of inputs are the combination of flow and pressure residuals so that SVM has the ability to treat high dimensional inputs. After the residuals of flow and pressure data are amalgamated they are normalized before training the classifier. For normalizing , a standard scaler from sklearn package is used.

We call this part of the detection phase as off-line part since residual generation, classifier training and validation are held before applying to real data. The detection part which is the on-line part is discussed next.

Detection Part

After training and validating the classifier it is ready to be applied on new observational residual data. This new observational residue data will be a difference between field measured pressure and flow data with the base , non-leaky data. After the residual calculated the detection module decides if the given residual is classified as leaky or non-leaky.

It is preferred to add another decision enhancing module in addition to the classifier. This new module is the anomaly detection module on the residual space. For this purpose kmeans clustering algorithm [47] which is a semi supervised anomaly detection algorithm is used. In the ideal condition, we assume the residual on the non-leaky point to be zero and non-zero if there is a leak. Based on this assumption, considering the demand uncertainty on different nodes and measurement noise we apply anomaly detection tests on the residual data.

From the two decision modules if the results show the possibility of leak on the residual data, we can conclude that there is a leak on the given data. The reason for enhancing the detection phase is that, building confidence in presence of leak before starting localization. Figure 4.2 shows the flow diagram of the detection module.

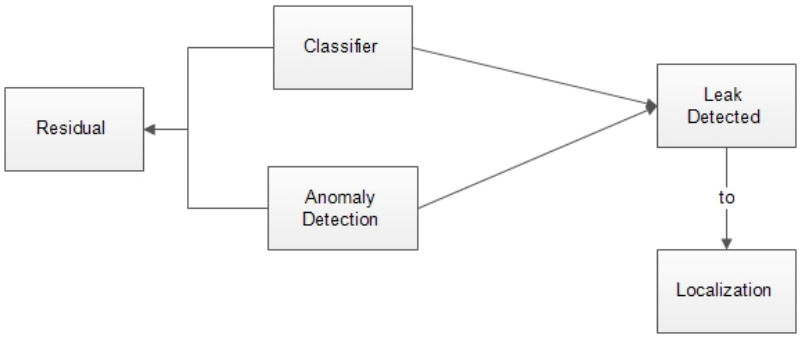

Figure 4.2: Detection Module

As an output, the detection module passes the time-stamp where the leak is first detected using the classifier module. This time value is where the residual pattern is turned from non-leaky to leaky. The localization phase starts to examine the

residuals after that time for identifying the leaking suspected nodes.

4.2 Localization Phase

The localization phase proposed here is a statistical method that takes the residual data and searches and outputs candidate leaking nodes.

The localization phase uses a statistical method to find the leak node after a leak is detected in the system. As it is known the occurrence of a leak increases the incoming flow and decreases the pressure of the downstream(i.e., nodes after the leaking node). The effect of one leak in any node can be seen on the other nodes, especially on nodes that are looped like the Hanoi network. However, the effect is not the same in every node.

The input for the localization phase is the time-stamp where the leak is first detected. The localization step follows the following step.

- First, we take the absolute value of each residual after the suspected time stamp. Taking the absolute value will make the deviation that is caused by pressure and flow rate to come in one dimension.
- In this step all the residuals show positive (non-zero value). However, their values are not the same. Neighborhood nodes to the leaking node show large residual values than the other nodes . However, the leaking node especially shows the highest peak value of all neighboring nodes. Therefore, finding a leaking node is turned in this step to finding a local maximum for multiple leak scenarios and finding a global maximum for a single leak scenario. In this thesis, it is chosen to put a threshold for selecting the candidate nodes. The threshold is not a discrete value but flexible with the data.
- A node is considered as candidate node when its residual value is greater than the threshold value τ. The threshold, τ, is the upper quartile (75% quartile) of the residual data at a given time-step. This process proceeds for all the

time-steps and the final candidate nodes will be the nodes that are selected at every time-step as a candidate. Algorithm 1 illustrates the pseudo-code for localization.

Algorithm 1 Localization Pseudo-code

```
1: t_0=starting timestamp
2: c_1, c_2, ... c_n//candidate for a time step
3: for t_0,...,TotalTimesteps do
4:     τ= 4^th quartile
5:     for all nodes do
6:         if node's residue ≥τ then
7:             C_i[x]=node
8:         end if
9:     end for
10: end for
11: Final Candidate =c_1 ∩ c_2 ∩ ... ∩ c_n
```

Chapter 5

Experiment

5.1 Dataset Description

Since the proposed approach follows a hydraulic modeling, we generate both the leak and the non-leak dataset using a hydraulic modeling software EPANET[16]. Due to the unavailability of real observational data, the on-line data for the detection and localization is also generated using EPANET. The leaks are simulated using a python package WNTR (water network tool for resilience) [46]. WNTR is a python library that uses the EPANET solver using a programming toolkit.

Vrachimis and Kyriakou [12] proposed a realistic leakage dataset, the Leakage Diagnosis Benchmark(LeakDB). Due to the limited access to real dataset, LeakDB is constructed based on the requirements of benchmark construction approaches. Hence, the generated data could be used as a common benchmark as well as for research reproducibility. LeakDB is open-source. The authors have also published the source code, that is created using the toolkit WNTR. The main motivation for using this benchmark dataset generation procedure is that the demands are artificially created based on historical real-data from water utilities in Cyprus where the authors had accessed. In this research, we manipulate the steps they followed to generate our experimental datasets.

The water network used in this case study is the well known Hanoi benchmark water network that is first used in Fujiwara and Khang,(1990)[13]. The network has 34 pipes, 31 nodes, and 1 reservoir. Figure 5.1 shows the Hanoi benchmark water network. The circles represent nodes by which water demand and also leaks are discharged. The directed arrows are pipelines. Node 1 is the reservoir while the rest are demand nodes. In EPANET, the simulation output is the pressure values in each node at a given time instant and the flow rate values in each pipeline. These output values are taken as sensor outputs in the simulation environment. In Hanoi water network the number of nodes , by which pressure data are taken, are 31 and the number of pipelines ,by which flow rate values are taken, are 34 links. The topology map of Hanoi network is depiected in Figure 5.1 The following data sets are generated according to different considerations.

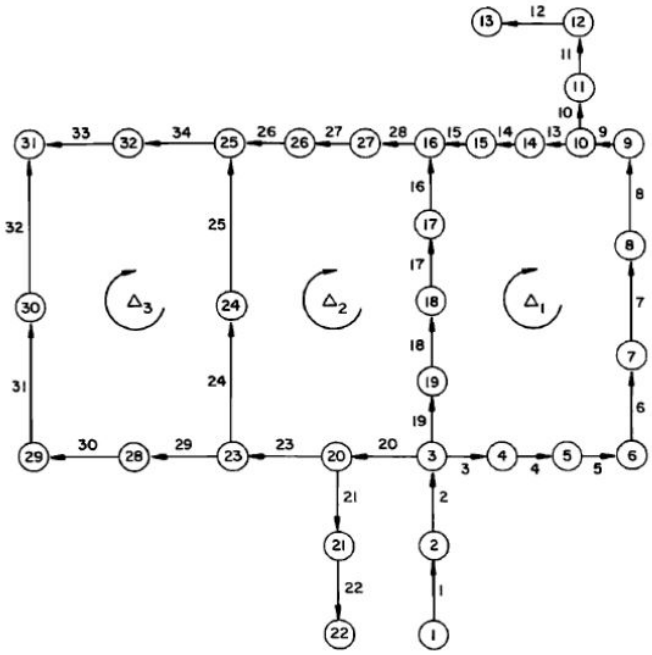

Figure 5.1: Hanoi Water Network. Taken from [40]

– **Simulating leaks on different sizes**- The leaks are created by considering a circular leaking area that has different diameters. The diameters are varied to show the intensity of leaks from small to large, ranging from

[0.1-1.5] mm with a difference of 0.1 mm. Similar to previous researches [4], [20], [21], [23], leaks in this research are assumed to occur in nodes. The amount of water leaking is also dependent on the pressure in the pipe. Considering the same network setting(i.e., pressure) by varying the leak area we tried to generate different leak hydraulic data. For each leak area, we have 2 similar scenarios that simulate leak but on a random node from the available 31 nodes. Totally we have 20 scenarios representing small to large leak sizes. For the purpose of training 50% of the scenarios (i.e., 10 scenarios for possible leak sizes) are used to train the classifier on the off-line process. The rest 50% is used to validate the classifier.

- **Demand uncertainty** - In order to represent the realistic case, it is created additional dataset that contains demand uncertainty. Here we vary the average nodal demand by 2% and 4% respectively . This demand variation is added as Gaussian noise, on the demands that we get from the LeakDB benchmark demand profile. The amount of noise uncertainty is chosen based on previous work that is found in [23] .
- **Noise uncertainty**- This category is similar as the previous dataset but it considers measurement noise on the pressure and flow measurements. The measurement noise is also added as Gaussian noise by varying 2% and 4% of the measured hydraulic data (flow and pressure) respectively.
- **Both Demand and Noise variation**- This category is generated by considering both demand variation and measurement noise on the first dataset setup.
- For the purpose of localizing multiple leaks, we prepared ten kind of datasets.In each case the number of leaking nodes is increasing by one.
- Finally for the purpose of comparing the proposed localization approach with other similar works a 250 scenario dataset is generated. These 250 data sets are prepared based on Soldevila et.al's work[5] . Considering 10 different leak sizes for 25 selected nodes on Hanoi water network a total of 250 datasets each for 1 month is generated.

The pipe properties for Hanoi benchmark used in this study is depicted in Table 5.1 . In this work, we assume the model uncertainty is zero i.e., we have full modeling information over the physical WDS. Other settings used while simulating the Hanoi network are presented below.

- Duration: 31536000 sec (1 month)
- Hydraulic timestep (sampling time): 1800sec (30 min)
- Headloss equation : Hazzen williams
- Specific gravity : 1.0
- Emitter exponent : 0.5
- Junction elevation :all of the nodes are 30m
- Demand data - taken from LeakDB

The data sets are created considering a one month duration in which the middle 10 days have leaks. For each leak area having a constant pressure the leak demand (leak flow rate) shows variation. The intuition for saying a leakage is small or not, is dependent on pressure not only on the leak size. For Hanoi network the leak flow rates are calculated by changing the leak size. Table 5.2 shows the recorded leak demands (flow rate) with their respective leak sizes. The leakages are in the range 1-13% from the total water supplied. In these thesis the leak sizes are denoted by their diameter considering circular leak area.

5.2 Experiment Tools

The simulation scripts and tools used in the experiment are implemented in Python using WNTR 0.2.1 [46]. WNTR 0.2.1 is a Python package for simulating hydraulic modeling. In addition, the classification and anomaly detection is carried out using a machine learning Python package, sciKit learn [47]. For manipulation on residuals a **pandas** [49] package is used. The hardware and software specification of the machine used for all experiments conducted in this thesis is given in Table 5.3.

Link	Length(m)	Diamete(m)	Link	Length(m)	Diamete(m)
1	100	1.016	18	800	0.6096
2	1350	1.016	19	400	0.6096
3	900	1.016	20	2200	1016
4	1150	1.016	21	1500	0.508
5	1450	1.016	22	500	0.3048
6	450	1.016	23	2650	1.016
7	850	1.016	24	1230	0.762
8	850	1.016	25	1300	0.762
9	800	1.016	26	850	0.508
10	950	0.762	27	300	0.3048
11	1200	0.762	28	750	0.3048
12	3500	0.6096	29	1500	0.4064
13	800	0.4064	30	2000	0.4064
14	500	0.4064	31	1600	0.3048
15	550	0.3048	32	150	0.3048
16	2730	0.4064	33	860	0.4064
17	1750	0.508	34	950	0.508

Table 5.1: Properties of Hanoi Water Network

5.3 Experimental Scenarios

In order to test the leak detection and localization approach, three groups of experiments are conducted using the generated datasets. The experiments are described as follows:

- Experiment 1: experiments on detecting and localizing small leakages.
- Experiment 2: experiments conducted to test the robustness of the proposed approach due to demand and noise uncertainty on detection and localization.
- Experiment 3: experiments conducted to test localization of multiple leakages.

Leak area	Leak flow rate(l/hr)
0.1 mm	0.76
0.2 mm	3
0.3 mm	6.8
0.4 mm	12
0.5 mm	18.8
0.6 mm	27
0.7 mm	36.7
0.8 mm	48
0.9 mm	61.7
1 mm	75
1.1 mm	90.9
1.2 mm	108
1.3 mm	126
1.4 mm	147
1.5 mm	170

Table 5.2: Leak Area vs Leak Flow Rate(lit/s)

Specification of Machine Used for Experiment	
Manufacturer	Acer
Model	NitroAN515-53
Processor	Intel core i5-8300H CPU @2.30GHZ
Memory	8GB DDR4
Operating System	64 bit Windows 10

Table 5.3: Machine specification

The proposed approach is compared with a state-of-the-art detection and localization methodologies. For comparison different works [5], [37], [40] are chosen. In state-of-the-art, the authors used MNF analysis for the detection of leaks. As mentioned in the literature review, for localization the authors used residuals and classifiers. Therefore, a comparison experiment is done both for detection and lo-

calization.

5.3.1 Experiment 1 [Small Leak Detection and Localization]

Experiments under this category aimed at evaluating the proposed approach on leak detection and localization. In this category we have the following set of sub Experiments:

- **Detection of Small Leakages**-In this scenario, 15 different leakage areas are created ranging from very small leak, 0.1 mm to 1.5mm. The leak size is increased by 0.1mm. These 15 datasets with a duration of 1 month are labeled as leaky with the middle 10 leak days. The 31 pressure node data and the 34 pipeline flow rate data are combined together to give 65 features and a target of binary value (i.e., 0 for leak-free, 1 for leaky). Half of the dataset is used to train the classifier and the rest is used for evaluation.

- **Detection Comparison**- Minimum Night Flow(MNF) analysis is the oldest and customary way for detecting leaks in a district metered areas. In this work a modified version of MNF which is proposed by Eliades and Polycarpou [37] is used. LeakDB [12] benchmark also uses this detection approach for comparison of newly proposed detection algorithms. The other experiment is the comparison of proposed combined residual approach with pressure residuals and raw pressure and flow rate sensed data. In this scenario the same dataset setup as other experiments is used.

- **Localization Comparison**-This experiment compares the proposed approach with similar hybrid approaches that are described in Carreno-Alvarado et.al[40] and Soldevila et.al[5]. First, it is tried to compare with Soldevila et.al[5] work that uses localization using pressure residuals and Multi-label classifier. They used the Hanoi network, and as an input, they only used two nodes as sensor data. The output of the classifier is the leaking node. 26 out of 31 nodes are the output classes. The details of their implementation are not mentioned clearly but it is tried to make a closer attempt with their work. Unfortunately, the attempt to re-do their approach did not give the results they reported. The

use of only two sensors from the available 31 nodes and a total class of 26 could not give the result they got.

5.3.2 Experiment 2 [Robustness to Uncertainty]

This experimental scenario was done to detect the robustness of the proposed approach to demand uncertainty, noise addition, and application of both uncertainties. In this scenario, 10 different leakage areas are created in the range between 1cm to 10cm each increasing by 1cm. During data generation phase an uncertainty ± 2% and ± 4% of the average daily base demand is added before simulation. The same amount of Gaussian noise is also added in the pressure and flow measurement to check robustness in the addition of noise. The same training and testing procedure is followed as Experiment 1.

5.3.3 Experiment 3 [Multiple Leak Localization]

To test the proposed localization approach, we conducted five experimental scenarios. In this experiment the number of leaking nodes is varied from single to five leakages at a time. This experiment is intended to show the ability of the proposed approach in localizing multiple leakages.

5.4 Evaluation Metrics

In terms of classification accuracy, the scoring algorithm uses the data labels provided with the datasets to calculate the standard classification metrics of the confusion matrix: the true positive (TP), false positive (FP), true negative (TN), and false-negative (FN). True Positive (TP) denotes when a correct leak is detected using both the classification and the anomaly detection modules. True Negative (TN) denotes when non-leaky residual is identified as non-leaky. False Positive (FP) denotes when leak is alarmed while there is no leak. False Negative (FN) denotes when it mistakenly identified leaky residual as non-leaky. Commonly used metrics such as precision, recall, and accuracy, can be computed based on the confusion

matrix. These metrics are described below.

Precision is a ratio of the number of correctly identified leaks relative to the total number of leaks identified by the detection module. Precision is computed using Equation 5.1.

$$prescision = \frac{TP}{TP + FP} \tag{5.1}$$

Recall also known as True Positive Rate, measures the proportion of True Positives i.e., identified leaks) to the actual total number of leaks. Recall is computed as shown in Equation 5.2.

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

Accuracy is a rate of positive classifications over the total classification samples and it is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.3}$$

F-measure is the harmonic mean of precision and recall. F-measure is computed as:

$$Fscore = \frac{2 * Recall * Precision}{Recall + Precision} \tag{5.4}$$

5.5 Results and Discussion

The previously stated experiments are organized here in three parts. The detection, the localization, and the comparison parts.

5.5.1 Experiment 1 [Small Leak Detection and Localization]

Results collected from conducted experiments to test small leakage detection and localization are presented as follow.

Small Leak Detection

Table 5.4 shows the results of this experiment scenario. The rows indicate the different leak sizes from very small to medium and large. The columns indicate the metrics that are calculated from the confusion matrix. The results inside the Table show two classifiers and one anomaly detection module in the order: KNN, SVM, Kmeans respectively. KNN is used to be compared with SVM (the proposed approach used SVM) . In this experimental scenario, only leak size is varied by making demand, noise, and model variations constant.

	Recall			Precision			F-measure			Accuracy		
	KNN	SVM	KMeans	KNN	SVM	KMeans	KNN	SVM	KMeans	KNN	SVM	KMeans
0.1 mm	100	100	98.9	95.8	100	100	97.8	100	99.4	98.5	100	99.6
0.2 mm	100	100	100	96	100	100	97.9	100	100	98.6	100	100
0.3 mm	100	100	100	96	100	100	97.9	100	100	98.6	100	100
0.8 mm	100	100	100	100	100	100	100	100	100	100	100	100
0.9 mm	100	100	100	100	100	100	100	100	100	100	100	100

Table 5.4: Detection Result of Small Leakage

As we can see from the results on Table 5.4, the classification result is presented. In terms of accuracy, SVM outperforms the other and Kmeans follows it with good accuracy. Kmeans clustering anomaly detection module also performs reasonably good relatively with the previous ones. Regarding to F-measure, for reasonably small leak size (50 lit/hr i.e.,less than 0.8 mm) SVM scores 100%. SVM result is affected by normalization of the training dataset. The normalization gives SVM high degree of generalization of the small leak datasets. However, in this experiment there is no demand and measurement fluctuation. It is by considering perfectly calibrated model as a baseline.

Detection Comparison

Comparison 1

The basic MNF analysis approach works by comparing the current nightly minimum flow with the previous nights for detecting leakages. This is by taking the average flow rates between low water demand hours usually between 1 am and 5 am. The intuition behind using the nightly flows is that, the flows during the night have smaller variations than during the day. Since the leakage losses will be larger because of the higher pressures in the system, it will be easier to detect leakages.Eliades and Polycarpou[37] modified the approach MNF detector that analyzes the minimum night flows during the night, to detect anomalies on the input flow of the network. A moving window w is defined in order to calculate the minimum MNF during those days; this is defined as $w(l)$ be the average night-flow measured for the l-th period and a threshold $\delta(l)$ is selected for a time window of M days.

$$\delta(l) = w(l) - min\{w(l-M), ..., w(l-1)\} \tag{5.5}$$

Let l be the day a leakage is detected, such that $l = argmin\ \delta(l) > h_w$, where h_w is a detection threshold which is selected off-line by using historical measurements, such that to minimize false positives and maximize true positives. This approach gives rise to certain trade-offs: setting h_w too low may cause a large number of false positive leakage fault alarms, while setting it too high may cause the detection algorithm to miss some leakages. In addition, due to the large uncertainties in the flow measurements, e.g. due to festivities or other events, this threshold could be exceeded even when no leakage has occurred in the system.

In this work, m= 3 days is chosen and a thresholds of h_w = 60,70,85, and 100 l/s are chosen. Table 5.5.1 shows the precision result of different scenarios for using MNF analysis in comparison with the proposed approach.

As shown in Table 5.5.1 the proposed approach detects all kind of leakages precisely. In MNF analysis, by taking threshold τ= 60 l/s it can detect small leakages

		Leak diameter (m)									
		0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Proposed˜Approach		100	100	100	100	100	100	100	100	100	100
MNF	$\tau = 60$	25	40	40	50	50	50	50	50	50	50
	$\tau = 70$	-	33.3	66.6	66.6	100	100	100	100	100	100
	$\tau = 85$	-	-	-	66.6	100	100	100	100	100	100
	$\tau = 100$	-	-	-	-	66.7	100	100	100	100	100

Table 5.5: Detection of Comparison Result[Precision]

with low score between leak diameter 0.01-0.03 m but it gives many false positives for the other large leak sizes. When we take τ= 70 l/s it can detect large leakages with leak diameter 0.05-0.1 precisely, but it has imprecision on detecting small leakages. A threshold of $\tau \geq 85$ l/s can detect large leakages, but small leakages are undetected with this approach. From this experiment we can take τ=70 l/s as balanced threshold.

Therefore from this experiment, it is preferable to use the classifier approach than the MNF analysis approach. From this, we can conclude that detection of small leakages, even very small leakages is possible with a well designed and calibrated model-based approach with a classifier approach. The proposed approach is free from thresholds that every leak sizes can be detected with the classifier automatically.

Comparison 2

The following result is the comparison of proposed combined (pressure and flow) residuals approach with separated residuals, and the use of raw (non-residual) flow and pressure data. Table 5.6 shows the result for leak area 1cm and Table 5.7 shows the result for leak size of 10cm.

As we can see the results from Table 5.6 and 5.7, the use of combined residuals enhance the detection accuracy. When the leak size increases the effect of using combined residuals clearly make a difference with the raw and separated features.

	RawData			*Residual*		
	Flow	*Pressure*	*combined*	*Flow*	*Pressure*	*combined*
Leak variation	72.3	73.6	70.5	77.2	99.9	99.9
Noise uncertainity	58.6	59.4	59.4	57.6	58.7	60
Demand uncertainity	61	57.5	60.2	58.5	61.3	61.3
All	57.4	55.3	56.3	56.5	58.6	57.3

Table 5.6: Summary Result for Leak diameter 1 cm [Accuracy]

	RawData			*Residual*		
	Flow	*Pressure*	*combined*	*Flow*	*Pressure*	*combined*
Leak variation	73.2	70.8	73.1	100	100	100
Noise uncertainty	55.6	79	74.4	56.7	94.6	94
Demand uncertainty	59.3	58.6	59.6	77.7	77.8	77.7
All	56.1	61.4	61.2	58.3	91.2	91.3

Table 5.7: Summary Result for Leak diameter 10 cm [Accuracy]

Additionally, where flow measurements are scarce we can use only pressure residuals in place of the combined residual approach because of their closeness with the combined approach.

Localization Comparison

In Carreno et.al [40], they used raw pressure residuals as input and the output classes are categorized into three as shown in Figure 5.1. The authors do similar work as in[4], [5] but they used raw pressure data instead of pressure residuals. In addition, they reduce the classes for classification to three. Therefore a comparison experiment has been made with them following the same procedure and using our combined residuals approach. For the purpose of comparison, the experiment is done by taking 24 hr time step data with a sampling time of 1 hr. The result in Table 5.8 shows considering three cases. The first case is the use of combined residuals (flow and pressure residuals)[proposed approach], the second case taking raw pressure values as in [40], and the last taking pressure residuals as in [5].

	Proposed Approach	Carreno-et.al, 2017 [40]	Soldevila et al., 2016a [5]
Taking 26 nodes	96.1	94.4	95
Taking two nodes	57.3	52.3	53.5

Table 5.8: Comparison Result for Single Leak Localization [Accuracy]

In Table 5.8, the two rows are representing number of sensors used for the classifier input. The results clearly depicted that taking all node data increases the localization accuracy. When the number of features decreases i.e., taking two nodes(nodes 16 and 23) the accuracy also declines. When we come to the result of the combined residuals, the proposed approach, it shows better accuracy with all the available nodes and also with the two nodes as a feature. In Carreno's approach, the use of raw pressure data gives the least accuracy than the others. When we come to the use of pressure residuals (Soldevila's approach) it gives a better result than using raw pressure values.

Summary

The results obtained from experiment 1 to answer RQ1[Small Leak Detection and Localization], demonstrated that in a well calibrated model small leaks can be detected and localized effectively. The proposed approach shows better detection accuracy by using combined residuals (flow and pressure residuals) than using raw residuals and pressure residuals approach. Therefore in the presence of plenty flow data, it is preferable to use the combined approach.

5.5.2 Experiment 2 [Robustness to Uncertainty]

This experiment is conducted to answer (in italics RQ2. What is the effect of adding demand and noise uncertainty on both detection and localization?). The results show robustness of the propopsed approach to different uncertainities.

Detection

As described in Section 5.3.2 detection of leakages is conducted in the presence of uncertainty. Tables 5.9 and 5.10 show the F-measure for 2% and 4% introduction of uncertainty for selected leak sizes respectively. As it can be clearly depicted, the proposed approach shows good robustness to demand variation for large leakages. The result for small leakages, however, are not robust to uncertainties.

	Leak diameter(m)								
	0.01			0.05			0.09		
Approaches	Noise	Demand	Both	Noise	Demand	Both	Noise	Demand	Both
Proposed Approach	3.1	43.5	41.5	72.5	97.2	65.3	95.6	100	95.9
MNF at $\tau = 70$	0	0	-	20	75	-	20	75	-

Table 5.9: Robustness to 2% Uncertainity [F-measure]

	Leak diameter(m)								
	0.01			0.05			0.09		
Approaches	Noise	Demand	Both	Noise	Demand˜	Both	Noise	Demand	Both
Proposed Approach	2	1.2	-	39.2	100	3.6	97.2	100	23.8
MNF at $\tau = 70$	0	0	-	66.6	75	-	66.6	75	-

Table 5.10: Robustness to 4% Uncertainty[F-measure]

For MNF analysis, addition of uncertainty gives poor result except for demand variation. This blanks in the tables above are Nan(undefined) values due to calculation of the F-measure. The other noticeable fact is that when the degree of uncertainty increases (i.e., form 2% to 4%) the detection score decreases. According to leak size increment the detection score increases i.e., large bursts are not much affected by the noise.

The results for the previous experiments show that under the perfectly modeled condition the detection of very small leakages is possible. However, when the degree of uncertainty increases the detection accuracy gets lower. But still, large leak sizes show resistance for noise and demand uncertainty. In a total stochastic condition, the result gets worse as we can see in the last experiment.

Localization

The localization results to test robustness of proposed approach are presented in this sub-section. Here it is considered localization of single leak along with the uncertainty additions.

We have four studies that are grouped as Experiments on considering only leak, Experiments on Demand variation, Experiments on Noise addition and Experiments on considering both demand and noise uncertainty. These experiments try to answer what the effect of adding demand and noise uncertainty is on localization.

	Located(20)	Mis-located(20)	Precision(100)
Leak variation	18	2	90
Noise uncertainty	14	6	70
Demand uncertainty	15	5	75
Both	13	7	65
Total	60	20	75

Table 5.11: Localization result of single leak under different scenario

As we can see from Table 5.11 , the result of localization under the first experiment (leak size variation), from a total of 20 scenarios 18 were found correctly. As for the two uncertainties, the number of correctly identified leaking nodes have reduced. One thing we observed in the experiment is that, when the leaking node is near to the reservoir (i.e., node 2 and 3) the localization results gives wrong localization. Since in the data generation phase the leaking nodes were selected randomly some leaking nodes were nodes 2 and 3. Therefore, the localization always misses to locate them.

When we consider single leak , the solution is the global maximum according to the localization proposed approach. However it is preferred to consider neighboring nodes that show higher residual values like the exact leaking node.

Summary

The results obtained from experiment 2 to answer RQ2[Robustness to Uncertainty], demonstrated that the proposed approach is robust for demand uncertainty. In reality, the most frequently faced uncertainty is demand fluctuation. Small leakages, however, are not robust to both noise and demand uncertainty. Along with increasing the uncertainty, the proposed approach shows better robustness than the MNF analysis in leak detection.

5.5.3 Experiment 3 [Multiple Leak Localization]

The greatest challenge of the previous works is to localize multiple leakage areas simultaneously. In this Experiment leaks are added on different nodes at the same time. We test 4 cases each having 20 scenarios.

	Localized leaks					
Leaks presented	None	One	Two	Three	Four	Five
Two	-	8	12			
Three	-	5	11	4		
Four	-	2	4	12	2	
Five	-	-	8	9	3	-

Table 5.12: Localization Result of Multiple Leaks

As we can see the result from Table 5.12, in the two leak scenario 12 of them are localized perfectly and in the remaining scenarios at least one of the leaks is identified. The reason for this is that other leaking nodes were node 2 and 3. As mentioned earlier the localization did not localize perfectly when leak occurred in node 2 and 3. For the three leak scenario it localizes 4 of them perfectly. In 11 of 20 scenarios, the proposed approach locates two leaks out of three and 5 of them only one out of three leaks. In the presence of four leaks, in two of the scenarios one leak and four leaks are localized respectively. In four scenarios out of 20 two leaks out of 4 leaks are identified and in 12 scenarios 3 leaks out of 4 leaks are identified. In

the presence of 5 leaks, in 8 scenarios two out of five, in 9 scenarios three out of five and in 3 scenarios four out of five leaks are localized.

Summary

The results obtained from experiment 3 to answer RQ3[Multiple leak Localization], demonstrated that at least two of the three leaks happening in the water network can be located confidently. However, when the number of leaks increased, the localization of all the leaks decreases. Therefore, from the experiment using statistical approach multiple leaks can be located using the proposed approach.

5.6 Threats to Validity

Threats to construct validity due to choice of parameters (classifier, sampling time, number of sensors): might affect experiments conducted with proposed approach. However, it is tried to choose the linear kernel for SVM classifier by trial and error from the available kernels. Regarding to sampling time it is considered to take LeakDB's way of sampling time. Therefore, for practical reason different realistic measurement sampling times should be considered.

In this thesis, the proposed approach is tested in Hanoi benchmark water network. The reason for choosing Hanoi is the data required for modeling is available and it is a looped network type. Therefore , in order for the approach to be generalized for other kind of water supply networks it needs a test on other networks.

Chapter 6

Conclusion and Recommendation

This Chapter provides the conclusion and recommendations for future works.

6.1 Conclusion

The distribution of treated water over a water distribution system faces different losses. These losses due to leakages make the water companies to loose huge revenue. We can not prevent leakages, but before the damage persist and flourished it can be detected and maintained. Early detection of leakage saves the wastage of water and also shows the capability of the water company on proper management.

There are a lot of leakage detection and localization mechanisms starting from hardware devices to software-based leakage detection and localization mechanisms using hydraulic analysis. This research proposed a hybrid(mixed) leak detection and localization approach that uses hydraulic modeling and classification approaches for detection and a statistical approach for localization. Previous hybrid approaches lack the ability to localize multiple leakages and also the ability to detect small leakages that are less than 10 lit/s. These two problems are tried to be addressed in this paper.

With the ability to simulate the system dynamics and generate extensive leak events, hydraulic modeling plays a great role in calculating residuals. Residual

analysis is a systematical way of reducing the observed data in the simple geometric way for finding leakages. The leaky and non-leaky residuals are fed to a binary statistical classifier for the detection phase. The localization phase starts its search for the leaking candidate nodes in the residual space. Nodes that show the same extrema effect in all the time-step across the residual are chosen as candidate nodes.

In order to test the proposed approach different experiments are conducted considering different scenarios including demand and noise uncertainty. From the detection and localization results, we can conclude the following points.

1. The detection and localization results under well known calibrated hydraulic conditions are almost perfect. Even very small leak amounts are able to be detected with the proposed hybrid approach.

2. The detection and localization accuracy decreases with increasing model uncertainties like noise addition and demand variation.

3. The detection of small leakages with the presence of uncertainty becomes unfruitful but the detection of medium and large leaks in the presence of uncertainty shows good accuracy. The reason is that all the analysis is done in the residual space so that uncertainties can be captured by the classifier.

4. The use of SVM classification algorithm benefits for capturing the residuals in the presence of demand uncertainty. The major uncertainty that we face on reality is demand stochasticity. The classifier is robust to demand variation.

5. The use of raw pressure and flow values, pressure and flow residuals and the use of combined residuals are illustrated in the experiment. The use of a combined residual approach which is unique from other works shows better results in detection.

6. The localization result is also perfect in the absence of uncertainty. However, the addition of demand and noise uncertainty affects the localization accuracy.

7. In the localization result where leaking nodes are near the reservoir i.e., in the case of Hanoi water network nodes 2 and 3, the result could not be located

correctly. The reason is that, the two nodes are located on the pipes from the reservoir to other distribution mains. Therefore, the reduction of pressure and water flow is considered as the reduction from the source(reservoir i.e., node1). As a conclusion, another approach should be considered to address the leaks in these nodes.

8. For the case of multiple leaks the proposed localization methodology has shown effectiveness on locating two, three and four leaks satisfactorily. However , when the leaks increase the localization accuracy decreases but still majority of the leaks could be localized. This achievement is further accompanied by using hardware-based methods for perfect localization. The localization gives a search space i.e., suspected leaking nodes and their neighbors. Further confidence increasing approaches like optimization or even hardware methods can be coupled to locate the target leaking spot.

9. For comparison of detection methodology, a famous MNF analysis is compared with the proposed approach. MNF analysis approach uses manual thresholds form practical observation, however, the threshold selection affects the detection of small leakages.

10. We also found that the proposed combined residuals approach performs better than other similar hybrid works with pressure residuals and raw residuals.

    .

6.2 Recommendation and Future Work

This thesis proposed a leakage detection and localization approach. Since all the modeling and testing phase carried on a synthetic dataset, water companies can taste it with a real dataset for external validity. In addition, as a future work we intend to test the proposed approach other academic water networks. In order to generalize the approach, it is recommended to test it other large networks.

The proposed approach works better with large transmission water networks. As a recommendation leakage detection and localization on service distribution water networks can be extended on the proposed approach. Sensor placement strategy(optimization) that is suitable with the proposed approach can be addressed.

The final result of the localization methodology is a candidate leaking nodes (i.e.,leaking nodes and its neighbors). Therefore, another search space reduction methodology can be extended on this work either using optimization or statistical process control method (use of control charts). As a recommendation, the storage requirement of the dataset generation in the LeakDB step is large. For the case of experimenting and using the dataset, storage space requirements also should be considered.

References

[1] USAID, *THE MANAGERS NON-REVENUE WATER HANDBOOK FOR AFRICA*. 2010.

[2] S. K. Bhagat, W. Welde, O. Tesfaye, T. M. Tung, N. Al-Ansari, S. Q. Salih, Z. M. Yaseen, *et al.*, "Evaluating physical and fiscal water leakage in water distribution system," *Water*, vol. 11, no. 10, p. 2091, 2019.

[3] J. Mashford, D. De Silva, S. Burn, and D. Marney, "Leak detection in simulated water pipe networks using svm," *Applied Artificial Intelligence*, vol. 26, no. 5, pp. 429–444, 2012.

[4] L. Ferrandez-Gamot, P. Busson, J. Blesa, S. Tornil-Sin, V. Puig, E. Duviella, and A. Soldevila, "Leak localization in water distribution networks using pressure residuals and classifiers," *IFAC-PapersOnLine*, vol. 48, no. 21, pp. 220–225, 2015.

[5] A. Soldevila, J. Blesa, S. Tornil-Sin, E. Duviella, R. M. Fernandez-Canti, and V. Puig, "Leak localization in water distribution networks using a mixed model-based/data-driven approach," *Control Engineering Practice*, vol. 55, pp. 162–173, 2016.

[6] A. Soldevila, R. M. Fernandez-Canti, J. Blesa, S. Tornil-Sin, and V. Puig, "Leak localization in water distribution networks using bayesian classifiers," *Journal of Process Control*, vol. 55, pp. 1–9, 2017.

[7] R. Li, H. Huang, K. Xin, and T. Tao, "A review of methods for burst/leakage detection and location in water distribution systems," *Water Science and Technology: Water Supply*, vol. 15, no. 3, pp. 429–441, 2014.

[8] O. I. Okeya, "Detection and localisation of pipe bursts in a district metered area using an online hydraulic model.," Thesis or dissertation, University of Exeter, 2018.

[9] S. R. Mounce, A. J. Day, A. S. Wood, A. Khan, P. D. Widdop, and J. Machell, "A neural network approach to burst detection.," *Water science and technology, 45(4-5), pp.237-246.*, 2002.

[10] S. Mounce and J. Machell, "Burst detection using hydraulic data from water distribution systems with artificial neural networks.," *Urban Water Journal, 3(1), pp.21-31.*, 2006.

[11] S. Mounce, J. Boxall, and J Machell, "Development and verification of an online artificial intelligence system for burst detection in water distribution systems," *J. Water Resour. Plann. Manage*, vol. 10, no. 1061, pp. 309–318, 2010.

[12] S. Vrachimis and M. Kyriakou, "Leakdb: A benchmark dataset for leakage diagnosis in water distribution networks.," *WDSA/CCWI Joint Conference Proceedings*, 2018.

[13] O. Fujiwara and D. B. Khang, "A twophase decomposition method for optimal design of looped water distribution networks.," *Water resources research, 26(4), pp.539-549.*, 1990.

[14] T. M. Walski, D. V. Chase, D. A. Savic, W. Grayman, S. Beckwith, and E. Koelle, *Advanced water distribution modeling and management.* HAESTAD PRESS, 2003.

[15] H. Alegre, J. M. Baptista, E. Cabrera Jr, F. Cubillo, P. Duarte, W. Hirner, W. Merkel, and R. Parena, *Performance indicators for water supply services.* IWA publishing, 2016.

[16] L. A. Rossman, *Epanet, users manual*, 1993.

[17] E. Todini and S Pilati, "A gradient algorithm for the analysis of pipe networks," in *Computer applications in water supply: vol. 1—systems analysis and simulation*, Research Studies Press Ltd., 1988, pp. 1–20.

[18] EPA, "Epanet," *Available at:https://www.epa.gov/water-research/epanet (Last Accessed Decmeber,23,2019)*, 2000.

[19] L. Ormsbee and S. Lingireddy, "Calibrating hydraulic network models.," *JournalAmerican Water Works Association, 89(2), pp.42-50.*, 1997.

[20] R. S. Pudar and J. A. Liggett, "Leaks in pipe networks," *Journal of Hydraulic Engineering*, vol. 118, no. 7, pp. 1031–1046, 1992.

[21] R Pérez, V Puig, J Pascual, A Peralta, E Landeros, and L. Jordanas, "Pressure sensor distribution for leak detection in barcelona water distribution network.," *Water science and technology: water supply, 9(6), pp.715-721*, 2009.

[22] R. Perez, G. Sanz, V. Puig, J. Quevedo, M. A. C. Escofet, F. Nejjari, J. Meseguer, G. Cembrano, J. M. M. Tur, and R. Sarrate, "Leak localization in water networks: A model-based methodology using pressure sensors applied to a real network in barcelona," *IEEE control systems magazine, 34(4), pp.24-36.*, 2014.

[23] M. V. Casillas Ponce, L. E. Garza Castañón, and V. P. Cayuela, "Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities," *Journal of Hydroinformatics, 16(3), pp.649-670.*, 2014.

[24] D. Jung and K. Lansey, "Burst detection in water distribution system using the extended kalman filter.," *Procedia Engineering, 70, pp.902-906.*, 2014.

[25] G. Anjana, K. S. Kumar, M. M. Kumar, and B. Amrutur, "A particle filter based leak detection technique for water distribution systems," *Procedia Engineering, 119, pp.28-34.*, 2015.

[26] N Majidi Khalilabad, M Mollazadeh, A Akbarpour, and S Khorashadizadeh, "Leak detection in water distribution system using non-linear kalman filter," *Int. J. Optim. Civil Eng, 8(2), pp.169-180.*, 2018.

[27] R. Pérez, V. Puig, J. Pascual, J. Quevedo, E. Landeros, and A. Peralta, "Leakage isolation using pressure sensitivity analysis in water distribution networks: Application to the barcelona case study," *IFAC Proceedings Volumes*, vol. 43, no. 8, pp. 578–584, 2010.

[28] J. Blesa and R. Prez, "Modelling uncertainty for leak localization in water networks," *IFAC-PapersOnLine, 51(24), pp.730-735.*, 2018.

[29] S. Mounce, "A comparative study of artificial neural network architectures for time series prediction of water distribution system flow data.," *Machine Learning in Water Systems - AISB Convention .*, 2013.

[30] G. Ye and R. A. Fenner, "Kalman filtering of hydraulic measurements for burst detection in water distribution systems," *Journal of Pipeline Systems Engineering and Practice (ASCE), , pp. 14-22.*, 2011.

[31] B. Jiang, F. Zhang, J. Gao, and H. Zhao, "Building a water distribution network hydraulic model by using watergems.," *In ICPTT 2012: Better Pipeline Infrastructure for a Better Life (pp. 453-461).*, 2013.

[32] M Romano, Z Kapelan, and D. Savic, "Bayesian-based online burst detection in water distribution systems.," *Integrating water systems, pp.331-337*, 2009.

[33] M. Romano, K. Woodward, and Z. Kapelan, "Statistical process control based system for approximate location of pipe bursts and leaks in water distribution systems.," *Procedia Engineering, 186, pp.236-243.*, 2017.

[34] M. R. Mounce S.R. and J. Boxall, "Novelty detection for time series data analysis in water distribution systems using support vector machines.," *Journal of hydroinformatics, 13(4), pp.672-686.*, 2011.

[35] K Aksela, M Aksela, and R Vahala, "Leakage detection in a real distribution network using a som.," *Urban Water Journal, 6(4), p. 279289.*, 2009.

[36] Y. Wu and S. Liu, "A review of data-driven approaches for burst detection in water distribution systems.," *Urban Water Journal, 14(9), pp.972-983.*, 2017.

[37] D. Eliades and M. M. Polycarpou, "Leakage fault detection in district metered areas of water distribution systems," *Journal of Hydroinformatics*, vol. 14, no. 4, pp. 992–1005, 2012.

[38] S. R. Mounce, A. Khan, A. S. Wood, A. J. Day, P. D. Widdop, and J. Machell, "Sensor-fusion of hydraulic data for burst detection and location in a treated water distribution system," *Information Fusion*, vol. 4, no. 3, pp. 217–229, 2003.

[39] S. Mounce, J. Boxall, and J Machell, "An artificial neural network/fuzzy logic system for dma flow meter data analysis providing burst identification and size estimation.," *Water management challenges in global change, pp.313-320.*, 2007.

[40] E. P. Carreno-Alvarado, G. Reynoso-Meza, I. Montalvo, and J. Izquierdo, "A comparison of machine learning classifiers for leak detection and isolation in urban networks," in *Congress on Numerical Methods in Engineering CMN*, 2017.

[41] A. A. A. Lah, R. A. Dziyauddin, and N. M. Yusoff, "Localization techniques for water pipeline leakages: A review," in *2018 2nd International Conference on Telematics and Future Generation Networks (TAFGEN)*, IEEE, 2018, pp. 49–54.

[42] A. Soldevila, R. M. Fernandez-Canti, J. Blesa, S. Tornil-Sin, and V. Puig, "Leak localization in water distribution networks using model-based bayesian reasoning," in *2016 European Control Conference (ECC)*, IEEE, 2016, pp. 1758–1763.

[43] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, 2019.

[44] A. Alexandrov, K. Benidis, M. Bohlke-Schneider, V. Flunkert, J. Gasthaus, T. Januschowski, D. C. Maddix, S. Rangapuram, D. Salinas, J. Schulz, *et al.*, "Gluonts: Probabilistic time series models in python," *arXiv preprint arXiv:1906.05264*, 2019.

[45] D. Crowl and J. Louvar, *Chemical Process Safety: Fundamentals with Applications*. Upper Saddle River, NJ: Prentice Hall, 720p., 2002.

[46] K. A. Klise, D. Hart, D. Moriarty, M. L. Bynum, R. Murray, J. Burkhardt, and T. Haxton, "Water network tool for resilience (wntr) user manual," *Washington, DC, USA*, 2017.

[47] scikit-learn developers, *scikit-learn user guide*. scikit-learn, 2019.

[48] S. Haykin, *Neural Networks and Learning Machines, 3/E*. Pearson Education India, 2010.

[49] *Pandas 0.25 documentation date: Jul 04, 2019 version: 0.25.0rc0.*