Acoustic Monitoring for Leaks in Water Distribution Networks

by

Roya Cody

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Dr. Michael Todd Professor, Dept. of Structural Engineering, University of California San Diego
Supervisor:	Dr. Sriram Narasimhan Professor, Dept. of Civil & Environmental Engineering, University of Waterloo
	Dr. Bryan Tolson Associate Professor, Dept. of Civil & Environmental Engineering, University of Waterloo
Internal Members:	Dr. Monica Emelko Professor, Dept. of Civil & Environmental Engineering, University of Waterloo
	Dr. Giovanni Cascante Professor, Dept. of Civil & Environmental Engineering, University of Waterloo
Internal-External Member:	Dr. Jean-Pierre Hickey Assistant Professor, Dept. of Mechanical & Mechatronics Engineering, University of Waterloo

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The contents in sections 3.2.7 and 5.3 of this thesis have been published in the following journal article co-authored by myself, Dr. Jinane Harmouche, and my supervisor.

Cody, R., Harmouche, J., and Narasimhan, S. (2018). Leak detection in water distribution pipes using singular spectrum analysis. *Urban Water Journal, 15(7)*, 636-644.

The methodology was jointly developed by myself and Dr. Harmouche. All experimental development and data collection was done by me. Dr. Harmouche provided technical guidance in singular spectrum analysis.

Chapter 4 and section 5.5 has been included in the following journal article co-authored by myself, Dr. Pampa Dey (Assistant Professor, Laval University), and my supervisor.

Cody, R. A., Dey, P., and Narasimhan, S. (2020). Linear prediction for leak detection in water distribution networks. *Journal of Pipeline Systems Engineering and Practice, 11(1)*, 04019043.

The development of the methodology and numerical and experimental studies were conducted by me. Dr. Dey provided technical guidance in verifying the processing and provided editorial assistance in writing the manuscript.

Abstract

Water distribution networks (WDNs) are complex systems that are subjected to stresses due to a number of hydraulic and environmental loads. Small leaks can run continuously for extended periods, sometimes indefinitely, undetected due to their minimal impact on the global system characteristics. As a result, system leaks remain an unavoidable reality and water loss estimates range from 10%-25% between treatment and delivery. This is a significant economic loss due to non-revenue water and a waste of valuable natural resource. Leaks produce perceptible changes in the sound and vibration fields in their vicinity and this aspect as been exploited in various techniques to detect leaks today. For example, the vibrations caused on the pipe wall in metal pipes, or acoustic energy in the vicinity of the leak, have all been exploited to develop inspection tools. However, most techniques in use today suffer from the following: (i) they are primarily inspection techniques (not monitoring) and often involve an expert user to interpret inspection data; (ii) they employ intrusive procedures to gain access into the WDN and, (iii) their algorithms remain closed and publicly available blind benchmark tests have shown that the detection rates are quite low.

The main objective of this thesis is to address each of the aforementioned three problems existing in current methods. First, a technology conducive to long-term monitoring will be developed, which can be deployed year-around in live WDN. Secondly, this technology will be developed around existing access locations in a WDN, specifically from fire hydrant locations. To make this technology conducive to operate in cold climates such as Canada, the technology will be deployed from dry-barrel hydrants. Finally, the technology will be tested with a range of powerful machine learning algorithms, some new and some well-proven, and results published in the open scientific literature.

In terms of the technology itself, unlike a majority of technologies that rely on accelerometer or pressure data, this technology relies on the measurement of the acoustic (sound) field within the water column. The problem of leak detection and localization is addressed through a technique called linear prediction (LP). Extensively used in speech processing, LP is shown in this work to be effective in capturing the composite spectrum effects of radiation, pipe system, and leak-induced excitation of the pipe system, with and without leaks, and thus has the potential to be an effective tool to detect leaks. The relatively simple mathematical formulation of LP lends itself well to online implementation in long-term monitoring applications and hence motivates an in-depth investigation. For comparison purposes, model-free methods including a powerful signal processing technique and a technique from machine learning are employed. In terms of leak detection, three data-driven anomaly detection approaches are employed and the LP method is explored

for leak localization as well. Tests were conducted on several laboratory test beds, with increasing levels of complexity and in a live WDN in the city of Guelph, Ontario, Canada.

Results form this study show that the LP method developed in this thesis provides a unified framework for both leak detection and localization when used in conjunction with semi-supervised anomaly detection algorithms. A novel two-part localization approach is developed which utilizes LP pre-processed data, in tandem with the traditional cross-correlation approach. Results of the field study show that the presented method is able to perform both leak-detection and localization using relatively short time signal lengths. This is advantageous in continuous monitoring situations as this minimizes the data transmission requirements, the latter being one of the main impediments to full-scale implementation and deployment of leak-detection technology.

Acknowledgements

There are many people who have made this thesis, not only possible, but a journey I will always remember fondly. Attempting to thanking all of them would be next to impossible. However, I would like to take this opportunity to thank a number of people.

To start I would like to express how deeply grateful I am to my advisor, Prof. Sriram Narasimhan, for all the guidance and support he's shown me throughout my Ph.D. His support has not only been academic but in all my pursuits. He has been a great mentor. His encouragement and integrity is something I will keep with me. I am also grateful to my co-supervisor, Dr. Bryan Tolson, for his insight and support throughout my Ph.D.

I am grateful to my PhD defense committee members, Prof. Monica Emelko, Prof. Giovanni Cascante, Prof. Jean-Pierre Hickey, and Prof. Michael Todd for reviewing my dissertation.

I would also like to take this opportunity to acknowledge the Natural Sciences Engineering Research Council of Canada, through their Strategic Project Grants program, for the funding support which has made this research possible. I also want to thank my industry partners, Dennis Mutti, Tim Sutherns and Don Plouffe, who not only helped make the field test possible, but made every meeting an enjoyable one.

I would like to give my biggest Thank You to Terry Ridgway, without whom the development of the laboratory test bed would not have been possible. Over the years Terry has been much more than just a big help, he has become a dear friend, who was always willing to help no matter the task, and always there to listen when needed.

My heartfelt thanks goes all the members of the Structural Dynamics, Identification and Control (SDIC) lab; to name a few: Dirk, Pampa, Jinane, Nina, Stan, Nick, Evan, Marshal, Piyush, Kevin, Dylan, and Guru. You have all contributed to how amazing this experience has been for me. I want to thank Dirk Friesen and Nina Feng in particular for all those late nights we stayed out collecting data. A special thanks are also due to my friends who have been the best support system I could have asked for: Thomas Czerniawski, Erfan Amiri, Billy Li, Suman Giri, Seun Aremu, Jon Francis and Gillian Finnie. Most of whom were unfortunate enough to have studied a sufficiently similar field to myself and were frequently probed with questions at all hours.

Finally, I would like to thank my family: my sister, Tina Cody, whose drive has always inspired me; my cousin, Anahita Bigtashi, who, in every way, is my best friend; my cousin, Pooya Baktash, who has always been like a brother to me, I am so proud of everything you've accomplished; and last but not least, my parents for their unconditional love. I only want to make you all proud.

Dedication

This is dedicated to my parents (Tom and Gina Cody) and my Daee (my late uncle Mahmoud Bigtashi).

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and motivation

The worldwide population is growing by roughly 80 *million* people each year [Worldometer, 2019]. The effects felt by increased population, and thereby water resources demand, are compounded by the effects of climate change which produce a significant decrease in the maximum annual spring river flows, as well as the frequency and extent of rainfall [Gupta, 2013]. The United Nations estimates that by 2025 30% of the world's population residing in 50 countries will face water shortages [The World Counts, 2020]. Similarly, by 2024, the EPA predicts 40 out of 50 states will face similar shortages under average conditions in some portion of their states [EPA, 2015]. In 2005, Canada withdrew approximately 42 km^3 of water for economic and household activities; 90% of this water went to support economic activity while 9%, 3.8 km^3, went to the residential sector [Canada, 2013]. An important factor adding to the significant amounts of water which is withdrawn, and in turn the predicted shortages, is the loss of water via undetected leakages.

A significant portion of water is lost between treatment and delivery. Canadian municipalities lose 13.3% of treated water [Canada, 2011], with some municipalities losing as much as 22%, prior to delivery to residents [Canada, 2011]. Such non-revenue water, due to leaks and bursts, is one of the main contributing factors to this increase [Canada, 2011]. Leaks also pose public health risks [Kirmeyer and Martel, 2001, Fox et al., 2015] due to intrusion of contaminants from the surrounding environment [Deng et al., 2011]. Early detection and remediation of leaks can prevent small leaks deteriorating into large bursts, thus mitigating significant water loss and the associated risks thereafter. Reliable and robust long-term continuous leak monitoring strategies which can detect and locate leaks

in the initial stages so that early interventions can be put into action, are hence urgently needed.

Detection of large bursts associated with pressure drops and visible consequences (e.g., surface flooding) is relatively straightforward. Such events generally produce large fluctuations in the global parameters such a system pressure, which can be detected over large distances with relatively sparsely located sensors. Sometimes, they can also be detected through water balance calculations or simply through citizen reporting. On the other hand, smaller burst events and leaks in water distribution systems are more difficult to detect. They can remain underground and unnoticed for long periods of time. The general approach to deal with such leaks today is to periodically inspect pipe networks for signs of leaks using what is commonly known as leak surveys. Such inspections, while effective, also tend to be labour intensive, slow, and mostly used to react to known leaks or bursts. To date, relatively little work exists in the literature in terms of being able to detect and locate leaks using automated long-term monitoring technology, where the onset of new leaks can be detected relatively quickly, as soon as they occur, with little to no user intervention.

The main aim of this dissertation is to develop and study the performance of a novel long-term monitoring system which can detect leaks in water mains, while being robust to changes or fluctuations due to other factors such as operational changes or seasonal variations. The main scientific contribution of this work is in the development, testing and evaluation of a new hydrant-mounted sensor system along with a suite of novel data-driven decision-support tools for the purposes of leak detection. The hydrant-mounted monitoring system was specifically selected for this application as this presents the most viable means to access the hydraulic conditions using existing access points in the network and can operate year-round in cold climates such as in Canada.

In terms of its impact, this work has strategic importance to Canada and beyond. Even though Canada holds nearly 20% of the world's fresh water supply, only 7% of this is renewable, and a vast majority of that is not easily accessible as it is retained in lakes, underground, or in glaciers. As a result, shortages still arise due to drought, infrastructure problems, and increased demand [NWRI and Meteorological Service of Canada, 2004]. Urban centers in Canada and worldwide, however, do not have such an abundance of available water. While the primary cause of leaks in water distribution pipelines are largely speculative, it is widely assumed that the main contributing factors include temperature (seasonal freeze-thaw), water demand stress, the occurrence of hydraulic transients, and pipeline deterioration, as well as corrosion. As many municipalities work with water systems which are aging and deteriorating rapidly, this work is both timely and highlights the strategic importance of minimizing water loss in water distribution networks (WDNs).

1.2 Working principle of WDN

In a WDN, water is collected from various sources, such as wells, underground aquifers and lakes and is subsequently transported to and held in outdoor water reservoirs. Upon treatment of this water (now potable), it is then transported to ground level storage chambers (primarily service reservoirs). This water is then distributed to various water towers (i.e.elevated storage tanks) via transmission mains and subsequently disseminated to subsections of the WDN using a combination of gravity and pumps through the water distribution mains. From the water distribution mains, services lines connect to individual homes and buildings to deliver the water. Figure 1.1 illustrates a typical WDN.

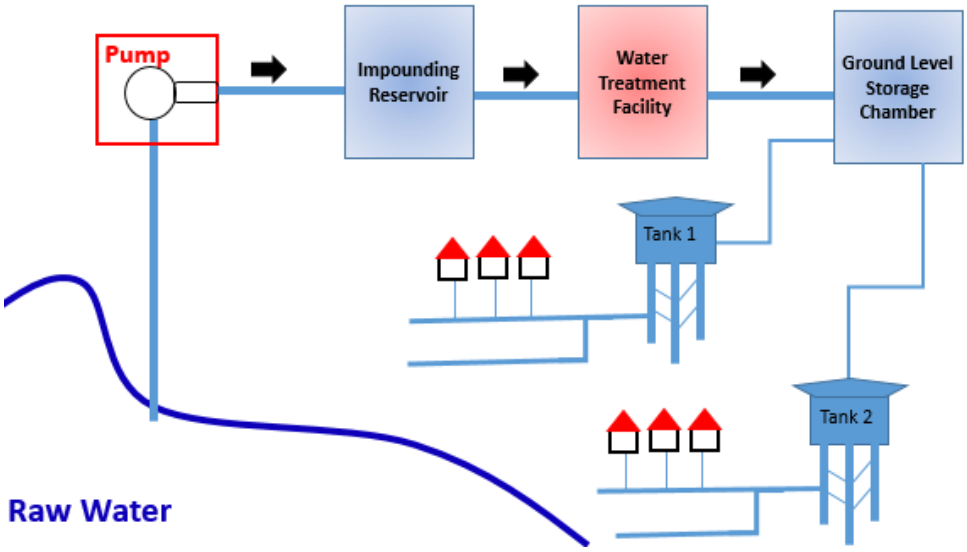


Figure 1.1: Sample layout of a city's water distribution network

Different cities have different pressure requirements for the distribution mains depending upon the gradient from the water towers and the number of pumps in the system. There are distinct minimum requirements for fire safety for different cities as well. The minimum requirement for fire safety is the amount of pressure required in the water network for the supply of adequate water pressure at each fire hydrant for fire fighting purposes. For example, the Fire Code in the province of Ontario specifies a minimum water flow of 140 kPa pressure [Ministry of the Solicitor General, 2016]. Municipal fire codes also require that fire hydrants be situated at minimum intervals: no more than $400 - 600$ $feet$ between them, depending on zoning, and within 90 $meters$ of the front of each house [National Fire

Protection Association, 2015]; as well, fire hydrants must be placed in an unobstructed position within 45 *meters* of where a fire truck would park [Building Code, 2018].

1.3 The need for monitoring WDNs

WDNs provide a convenience that most people in the developed world could not imagine living without. Humans have been successfully channeling the flow of water since the Neolithic age (c. 5700 - 2800 BCE), beginning with primitive agricultural irrigation systems [Mays, 2010]. However, water is extremely difficult to contain, and system leaks have been an ever-present complication. Frontinus (c. 40 - 103 CE), a Roman senator and engineer, left detailed measurements of water flow in the aqueduct systems of ancient Rome. His notes reveal a dramatic difference in intake and delivery, a phenomenon attributed to "evident leaks" in the system. Frontinus dictated that the leaks and infrastructure damage were caused by "the accumulation of deposit. . . the unlawful behavior of nearby owners, the force of the elements, or faultiness of construction" [Evans, 1997]. Essentially, the state of WDNs has not changed. The authors of a World Bank study estimated that in developing countries, roughly 45 million cubic meters of water are lost daily, with an annual economic value of over USD \$3 *billion* [The World Bank, 2016].

This loss of water, besides having an impact on economy, also leads to a variety of undesirable consequences. One of the main causes of water loss—deteriorating or damaged infrastructure—not only leads to water contamination and health concerns, but could also lead to a drop in pressure in that section of the network. The pressure drop causes significant inconvenience to the residences experiencing its effects and likely filing complaints, as well as, and much more importantly, causes a safety hazard as far as fire safety is concerned [Yves Filion et al., 2004]. Not only this, excessive leakage adds to the overall cost and challenges in sourcing, abstracting, treating, and distributing water. These four dimensions are further detailed in Table 1.1. These impacts include the waste of energy resources in treating water after initial abstraction and again, once the water leaks and infiltrates the storm water system, which unnecessarily strains the natural ecosystems and degrades relationships between consumers, water utility operators, directors, and shareholders, as well as government and regulatory bodies [European Commission, 2015]. Extreme cases of excess leakage may also result in intermittent supply and, ultimately, a complete failure by water utility to continue to provide the required service to the customers.

Table 1.1: Impacts of Water Distribution System Leakage [European Commission, 2015]

Environmental	Political and Societal	Economic	Legal and Regulatory
• Unnecessary removal of water from ecosystems impacts biodiversity and increases concentrations of water pollutants. • Wastes electricity and chemicals used for unnecessary water treatment, emitting ozone and greenhouse gases.	• The current levels of leakage are perceived by the regulators, the public, and the media as too high for most water utilities.	• Depletes water supplies, limiting infrastructure capacity development. • Higher cost of production and distribution, higher chemical usage, higher energy usage, higher costs of treatment for waste disposal. • Leak monitoring, localization, and repair are costly investments that must be balanced with the other impacts of leakage.	• Customers wish to see water utility operating efficiently so that it does not charge them excessively for leakage. • Economic regulators expect operating and investment costs to be justified. • Directors and shareholders expect water utility to run efficiently. • Environmental regulators seek to avoid undue abstractions of raw water. • National government departments aim to safeguard future water supplies.

1.4 Impact

While leakage in WDNs can be associated with severe consequences as discussed above, it is also important to understand what factors affect WDNs and the intensity of their impact. The environmental impact on water distribution systems is a widely assessed area, since the life cycle of the system is used to characterize its long-term sustainability. The operation phase typically has the highest potential for inflicting environmental damage on the system's life-cycle [Bonton et al., 2012]. The pumping of raw and treated water is directly correlated with water demand, and these processes are the most energy intensive and greenhouse gas (GHG) intensive parts of the life cycle process [Vince et al., 2008]. The processes of raw water treatment and pumping are the most directly correlated with water demand and are the most intensive, in terms of energy usage and greenhouse gas emission. [Vince et al., 2008]. The transportation and use of treatment chemicals in the system are also noteworthy. However, their impact is minuscule in comparison to the pumping of water. Consequently, the improvement of failing infrastructure can help reduce the energy burden of water treatment systems [Racoviceanu et al., 2007]. The impact of the factors listed can lead to the contamination of water, which might further lead to negative consequences, the most important being the impact on the health of the consumers.

Significant importance should be given to the potential contamination of potable water in the distribution system in the event of a leak. In cases where small leaks occur and are present for extended periods of time, the possibility of contaminants entering the potable water supply through leaks and being transported through the network to consumer taps is of great concern. It is especially alarming given that the materials surrounding water pipes can often contain harmful contaminants, including viruses and fecal bacteria [University of Sheffield, 2015]. The intrusion of the contamination in the area surrounding a leak is amplified with the occurrence of pressure transients, in which the negative pressure in the system then pulls contaminants in from the area surrounding the pipe fracture [LeChevallier et al., 2003]. This makes the timely identification and detection of a leak even more important.

1.5 Leak detection technology landscape

In the context of leaks, large or small, most methods in use today are inspection techniques, not methods intended for long-term monitoring. Leaks are often difficult to detect with traditional methods. Large lengths of pipe are routinely excavated in order to find and

repair small defective sections, often as a result of complaints, once a leak (or burst) surfaces and is evident visually.

While effective, most current inspection methods tend to be manual, depend on an expert to interpret data, and hence not conducive to long-term monitoring. The need to rely on expert judgement is one which incurs added cost and limits wide-scale autonomous deployment. Moreover, many of these techniques rely on the metal pipe surfaces to transmit acoustic energy, which suffers from detection issues in materials such as plastic. Most publicly available evaluation studies do not offer convincing results using commercially available technologies in blind tests and field conditions. Hughes and Venkatesh [2016] produced a study of Echologics water monitoring prototype, *EchoShore*, which is a system that most closely resembles a long-term continuous monitoring strategy. This report found issues related to battery-life and dependability, as replacements were needed for 15% of the systems that were installed. However, directly relevant to this dissertation, of the 13 leaks that occurred throughout the duration of their study, only two of the leaks were detected and reported by their system. The first was detected by inspectors who were on site that day, and also stated that the leak was audible, so there was uncertainty surrounding the systems ability to detect this leak without the presence of the inspectors. The second leak was detected using inspection-level detection methods, after the leak had already been reported and repaired. As stated in their executive summary, based on this 10 month long field deployment study, they were unable to conclude that there was significant opportunity of major savings based on the implementation of this system. A similar study was produced by Anguiano et al. [2016] using Echologics' hydrophone suited *LeakFinderRT* system which yielded a 54% weighted accuracy (with an 85% TPS and a 22% TNR) in a laboratory test bed setting. However when it was tested in the field, Anguiano et al. [2016] concluded there not only for *LeakFinderRT* but for *ZoneScan Alpha* as well, there was not sufficient information to evaluate the systems performance since the reported leaks went either unverified or the cause of observed water was never identified.

Aside from the two studies mentioned, to the author's knowledge, there exists very little in the public domain that documents and evaluates the performance and failure rates of inspection technologies in blind tests and in live WDN, and none exist for long-term monitoring technologies. In order for events—including leaks and bursts—to be recognized as soon as they occur without an expert in the decision-making loop, long term monitoring is the only feasible option. Furthermore, given the nature of the relatively closed product landscape with very limited third-part validation tests available in the open scientific literature, there is an urgent need to develop and evaluate new technology in a rigorous and objective way in the open scientific domain.

1.6 Long-term monitoring solution

Long-term continuous monitoring for events requires a fundamental re-thinking of both technology as well as the application procedure. Such factors as, weather, WDN system-integration, and connectivity and maintenance of the monitoring system must be taken into account. Any alterations to existing water distribution infrastructure must be done judiciously, so as not to disrupt the pipe system. The ongoing need for system modification and growth makes the use of model-based techniques inefficient, as the system models would need to be accurately be configured and tested for similarity. Though the effectiveness of model-based techniques has been proven in literature [Moser et al., 2015, Goulet et al., 2013], ease of wide system-integration has always been a limitation. There are many issues with purely data-driven techniques as well, the most important of which being that they do not naturally obey physical constraints (that we may know to be true, due to domain knowledge) due to their limited observation of the environmental dynamics; hybrid approaches (e.g., graphical models, knowledge-injection via constrained neural optimisation, etc.) have become popular for this reason. Most importantly, such a system should be able to operate year-round, especially in cold climates such as in Canada. In general, in cold-climate regions, even relatively large leaks that may be visible during the warm-temperature months, may not surface during colder months due to the frozen ground surface; preventing such leaks from being detected for extended periods of time. While the cost may be higher for the initial installation compared to one-time inspection methods, the long-term costs are significantly reduced when we consider the detrimental effects of the aforementioned events and the cost of spot-inspections throughout the system are taken into account.

In order to verify the effectiveness of different long-term monitoring methodologies, testing and validation is an important step. Many of the current WDN laboratory experimental setups are too simplistic and do not capture the complexity of actual WDNs. Many laboratory experimental test-beds lack fundamental field representations. While some of these systems are highly representative, such as by burying extended lengths of pipes [Covas et al., 2006], others find ways to be representative indoors. Some account for elbows and junctions, while others completely simplify to straight lengths. None of these systems take all factors into account, such as pressure, elbows and Ts, realistic materials and pipe diameters. While actual field conditions can be difficult to replicate in laboratory settings, an attempt to better represent them still needs to be made. It is not reasonable to consider simplified test setups to be representative of field conditions in order to validate proposed methodologies —field deployment is the ideal validation case study.

1.7 Research objectives and scope

The main objective of this dissertation is to develop a long-term, continuous, passive, WDN monitoring and decision support system to detect leaks in live WDN. In principle, the same technology can also aid in the detection of other events such as pressure transients and vandalism, however these are not pursued in this dissertation. The development will focus on a dry-hydrant based system which can operate year-round, which is crucial for operating in cold climates such as here in Canada. The overarching objectives of this dissertation are as follows:

- develop a long-term monitoring system requiring minimal modifications to and capable of continuously monitoring a WDN;
- develop the attendant decision support tools which can detect newly developed leaks and bursts events based on acoustic sensor data from the developed monitoring system, with minimal human expert intervention;
- test and validate the system using both laboratory experiments and field tests on a live WDN.

The scope of this dissertation is limited to the detection of newly developing leak and burst events assessing only changes in baseline conditions, and will not address the presence of pre-existing leaks. The important distinction between the system developed during this dissertation and existing inspection methods for leak detection is that the current system is envisioned to be deployed as a long-term monitoring tool and not a short-term supervised inspection system. As well, the work done in this dissertation is limited to those events which can be detected using acoustic data. Unlike many previous studies, the use of hydrophones in this dissertation to monitor acoustic data is motivated by the fact that sound signatures have been shown to travel further inside water mains relative to the pipe wall, even in traditionally challenging materials such as plastic. The main focus of this dissertation is to evaluate the system experimentally in a live WDN to study, quantify and validate the approach. To the knowledge of the author, this is the first effort where a long-term monitoring system capable of being mounted from a fully functional hydrant has been developed, tested, and validated in a live WDN. It is also important to acknowledge that fouling and bio-films are inevitable in WDNs [Batté et al., 2003], potentially resulting in affecting acoustic vibration characteristics within the system, e.g., due to a change in diameter [Vassiljev et al., 2005, Lansey et al., 2001] and roughness along the pipe wall [Shulemovich, 1986], or due to dislodging. However, such sources are not considered within the scope of this dissertation.

1.8 Organization of dissertation

The dissertation contains 7 chapters and is organized as follows:

- **Chapter 1** provides a brief introduction to the need for water distribution system event monitoring. It presents the overarching research goal and and specifies the significance of the study, including its limitations.
- **Chapter 2** presents a literature review on various traditional and current leak detection methodologies for water distribution networks and reviews the types of leak which can occur in the system. The research gap areas are identified and specific research objectives are outlined.
- **Chapter 3** provides background on relevant concepts which will be useful throughout the dissertation.
- **Chapter 4** presents the underlying principle of Linear Prediction and its analytical application to leak signals.
- **Chapter 5** presents results from the laboratory test bed for leak detection and localization methodologies.
- **Chapter 6** presents results for the field test bed for leak detection and enhanced localization methodologies.
- Finally, a number of conclusions resulting from the dissertation work are discussed in **Chapter 7**, followed by several recommendations for future study.

Chapter 2

Literature Review

The main objective of this chapter is to present an overview of the different types of detection and monitoring methods which exist today, from a technology standpoint, to detect water main leaks and bursts. An effort is made to classify the technologies based on their underlying working principle, with a relatively heavy focus on those methods employing acoustic signals. It is important to note here that acoustic techniques refer to methods that rely on accelerometers (vibrations) and sound propagation inside the fluid (hydrophones). This chapter starts with a review on the types and causes of leaks in WDN, followed by a discussion on the leak detection technology landscape today and concluding with a brief summary of the gaps in both the literature and the technology. For the purposes of this dissertation and the literature review, only a qualitative distinction is made between leaks and bursts; often controlled tests are called leaks, although in practice leaks are slow to develop while a burst is considered a sudden and dramatic loss of pipe integrity. Qi et al. [2018] defined bursts are creating conditions such that nodal demands can not be satisfied, defined by flows greater than 50 L/sec. Although pertinent to the problem being studied, a review of the literature related to the economic impact of damage incurred by various leaks, as well as other types of events such as contaminant intrusion, is relatively limited in this chapter.

2.1 Types and causes of leaks in WDNs

Water distribution systems are complex conveyance systems that undergo varying pressures, stresses, strains, and temperatures. As a result, they are a challenge to design,

build, and maintain. Despite the best efforts in system design, leaks remain an unavoidable reality. A common misconception about water distribution leaks is that the majority of water is lost as a result of large main breaks and catastrophic main failures. This is because their dramatic impacts on the network and high flow rates receive most of the attention from a public visibility standpoint, as such events cause the most disruption to our daily lives and require disproportionate resources in order to address them. However, well-run systems suffer the majority of their losses as a result of background leakage, long-running unreported leaks, and reported leaks where the repair is delayed [Thornton, 2002]. Therefore, finding the sections of the network that contain leaks and then pinpointing the exact locations of the leaks for repair, is central to leakage reduction as part of an active leakage control policy. Although, for the purposes of this dissertation, it is not very important to delineate exactly which type of a leak we are dealing with, it is important to put them into context when we discuss the relevant technology in this chapter.

Leaks are broadly classified into three categories [Thornton, 2002]:

- background (undetectable) leakage—low flow rate with perpetual duration; tends to increase with increasing age of the network;
- reported breaks—high flow rate with a short duration, typically brought to the attention of the water utility by the general public when they surface or cause a disruption in supply; and
- unreported breaks—moderate flow rate, the duration depends on the intervention policies applied by the utility; many go undetected without some form of active leakage control.

Different types of leaks result in different levels of impact on the system. This variation makes some leak types easier to detect than others. Bursts yielding immediate and significant impact, while leaks, allowed to persist, can create equally significant impact over longer periods of time. However with effective leak detection, leaks can be discovered more quickly, thus limiting their overall impact. The most severe leak type, immediately detected and causing the most obvious impact, being a main break caused by pipe fracture, is easily detectable, by utilizing system or line pressure. A pinhole leak, caused by corrosion or stress by stones after poor back-fill, can be harder to detect as they typically induce only local changes in the hydraulic conditions. Seepage is another common type of leakage caused by deteriorated asbestos cement pipes. Leakage can also be caused at system opening, joints and appurtenances. Adding to these, a detailed study of leakages and their causes can be found in Table 2.1.

Table 2.1: Types and Causes of Water Distribution System Leaks [Thornton, 2002]

Main break of pipe fracture	Used to describe a catastrophic pipe failure caused by pipe deterioration, fluctuating or excessive pressure, ground movement, or a combination of these factors. Main breaks are relatively easy to locate as these failures usually become quickly and visually apparent at ground surface level due to the massive volume of water released.
Crack	A pipe failure mechanism occurring as circumferential or longitudinal failure that usually results from pipe deterioration or ground movement.
Pinhole	Small circular failures in a pipeline usually caused by corrosion or stress by stones after poor backfill. Steel pipes installed in a corrosive environment without appropriate protection are particularly vulnerable. Pipelines should always have some protective layer, or at the very least a backfill layer of sand.
Seepage	Most commonly observed on deteriorated asbestos cement pipes where the pipe wall becomes semi-porous and water escapes slowly.
Leakage on packing glands of pumps and valves	Caused by deterioration as the system ages and usually occurs when a valve is used after a long period of inactivity. Easily detected visually at pumps or by valve chamber that is full of clear potable water.
Pipe joint leaks	Many older couplings and weld joints are not corrosion-protected and therefore deteriorate long before the pipe itself. When ground movement occurs, most of the strain is experienced at the pipe joints, often resulting in leakage and, eventually, a fracture.
Leaking service connection pipe	Service connection leaks are the most common type of leak. Between the water main and the customer water meter there are often more than one change in pipe size and material, which requires many joints that are especially vulnerable. Service connections are also often installed shallow in the ground in close proximity to disturbance from the traffic load above.
Leaking fire hydrants, air valves, and scour valves	System openings and appurtenances also occasionally leak water.

2.2 Leak Management

Broadly speaking, leak management consists of two main approaches: leak prevention and leak detection. Leak prevention is performed mainly through managing the overall pressures within the system through control devices and maintaining the overall physical health of the WDN through capital investments and maintenance. Clearly, the effectiveness of this approach depends on our ability to predict failure causes correctly (which is not usually possible for other than pressure related) and the availability of capital budgets and resources. Both these factors are not rooted in the technology aspect and hence not central to this dissertation. Of particular interest in this dissertation is the technology necessary to detect leaks that are otherwise not detectable within the system, which belongs to the second category of leak management.

The management of excess pressure including pressure transients [Williams and Kuczera, 2014, Wu et al., 2010, Silva et al., 1996] requires an accurate and extensive knowledge of system pipe layouts (this is explained in Section 2.3.4), as well as limiting the duration of all detected leaks. In a WDN the pressure is not constant, it can vary across a system. It varies throughout a given day, across a given week, and throughout the year. At times this variation can include sudden changes, caused by changes in demand, sudden opening or closing of valves, sudden shutdown of a pump in the system, among various other causes [Budris, 2014]. The change in the steady state of motion of water can cause pressure transient waves to arise and propagate through the system, also known as a "water hammers" [Thorley, 1968]. The kinetic energy of the liquid moving through the fluid manifested in the form of a pressure wave can cause significant damage to pipes and to pumps or fittings [Budris, 2014]. Generally, damage to a large system occurs over repeated cycles of such transients and hence if a system is adequately monitored, the repetition of these damaging events can be mitigated through appropriate control devices, thereby limiting the potential for system damage.

Pressure management can be achieved through techniques such as variable speed pump control, tank regulation, and the implementation of pressure-regulating valves. Higher system pressures are typically associated with higher levels of system leakage, and so the central objective of pressure management, besides providing adequate service to customers, is the reduction of background leakage by minimizing system pressures [Vicente et al., 2016]. In addition to minimizing background leaks, the reduction of excess pressure and pressure transients prevents avoidable bursts and mitigates the costly process of locating and repairing them. For this purpose, the fixed and variable area discharge (FAVAD) principles [May, 1994b], the burst and background estimates (BABE) concept [Lambert, 2004], as well as, the more recent, N1 exponent relationship [Thornton, 2002, Lambert,

2004] are all popular tools for determining optimal pressures based on related leak flow rates and burst probabilities. These are essentially management tools and can be used to effect improvements in leakage management plans. A brief description of these concepts is provided next for the sake of completeness.

The FAVAD principle, simply put, is an equation [May, 1994a] that describes the pressure-leak relationship,

$$Q = C_d\sqrt{2g}(A_0 h^{0.5} + mh^{1.5}), \tag{2.1}$$

in which Q is the flow rate through an orifice, C_d is the flow coefficient, g is acceleration due to gravity (m/S^2), A_0 is the initial leak opening without any pressure in the pipe, h is the pressure head (m), and m is the slope of the pressure head-area [Deyi et al., 2014]. This was the established 'best practice' form of the equation for pressure. It states that the leakage ratio will increase proportionally with the increase in pressure to the power of 0.5. This method was later expanded upon [van Zyl and Cassa, 2014], and a dimensionless leakage number (L_N) was introduced,

$$L_N = \frac{N1 - 0.5}{1.5 - N1}. \tag{2.2}$$

This created the N1 exponent relationship in which, instead of raising pressure to the power of 0.5, it is raised to the power of N1. An accurate assumption for N1 will thus influence the reliability of the predictions. Fanner et al. [2007] showed that N1 values close to 0.5 is representative of leaks with small round holes, while N1 values closer to 1.5 is representative of small leaks (typically undetectable) from corroded or misaligned joints and fittings, as they are more sensitive to pressure. This concept is used to model the sensitivity of different leaks to pressure in a system. It is effective for gaining insight into the behavior of a system, however is not a detection tool.

The BABE concept is essentially a standardization for categorizing leaks in a system. Broadly speaking, it considers water loss in three categories: background leakage (undetectable and therefore unreported), reported leaks and bursts (i.e., $> 0.5\ m^3/hr$ [Al-Washali et al., 2019]), and unreported leaks and bursts (but detected by a City's employed manpower) [Lambert et al., 1999]. The concept states that leaks consist of multiple leak events, each of which is a function of the average flow rate and run-times [Taha et al., 2016, Al-Washali et al., 2019]. From this, the annual volume of water lost can be determined.

2.3 Leak Detection

WDN leak-detection methods fall into five main categories: water balance [May, 1994b, Wallace, 1987], acoustic leak-detection, non-acoustic leak-detection, hydraulic, and data-drive, as depicted in Figure 2.1. This classification is arrived at solely for the purposes of this dissertation and to the knowledge of the author there is no accepted taxonomy for leak detection methods in the literature. The main principles underpinning each of these methods and the relevant literature are explained in detail next, while emphasizing data-driven techniques utilizing acoustic or other sensor means, which form the core of this dissertation.

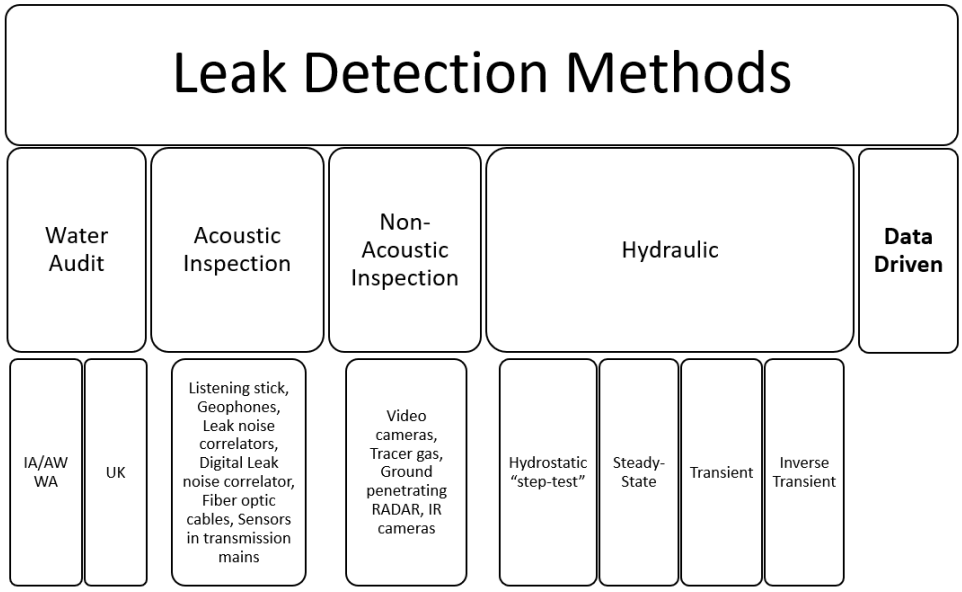


Figure 2.1: A classification of leak detection methods described in literature.

Within these five identified categories, there are general benefits and limitations associated with each. Water balance methods, while an intuitive approach, requires municipalities to introduce district metered areas (DMAs) and real-time acquisition of basic hydraulic parameters, in order to study the flow of water in and out of a system or sub-system, which is not always feasible for older systems. However, for newer municipalities it can be a simple and cost effective way to identify general regions of significant water loss. In order to achieve a more local identification of the source of water loss inspection methods can be deployed. While highly effective, they can be very costly and time consuming, and have

been proven to be most effective only if the presence of a leak is known and simply its location is of interest. Leak detection equipment covers a wide range of technologies and capabilities as summarized in Table 2.2. This table summarizes the effectiveness of finding leaks in different types of mains in the WDN, the trunk main representing the central line of the WDN, leading back to the city water supply, while the distribution mains are the smaller pipe networks which creates a web providing water to all areas, and finally the service pipe lines connections the distribution mains to the commercial and residential houses for water consumption. Many of these inspection equipment require experienced operating personnel to effectively execute and interpret the detection results. A good understanding of the nature and occurrence of real leakage losses is crucial in order to make effective choices about appropriate technology applications. Even highly sophisticated and expensive leak detection equipment cannot solve a utility's leakage problem if the utility does not understand the real extent and nature of leakage occurrences in its distribution system [Thornton, 2002]. Beyond the quality and sophistication of the equipment available, the most important factor for success in detecting leaks is the experience of the leak detection team in using the technology and interpreting the results received from the equipment [Thornton, 2002].

Table 2.2: Leak-Detection Methods

Leak-Detection Methods		Effective for finding leaks in		
		Trunk mains	distribution mains	service pipes
Acoustic Methods	Leak Noise Correlator	Yes	Yes	
	Noise loggers		Yes	
	Listening stick		Yes	Yes
	Multi acoustic sensor strip		Yes	Yes
	In-pipe sounding	Yes		
Non-Acoustic Methods	Gas Injection		Yes	Yes
	Ground penetrating radar	Yes	Yes	Yes
	Infrared photography	Yes		
	In-Pipe hydraulic plug			Yes

Hydraulic leak detection methods use hydraulic characteristics, such as flow and pressure, to locate leaks in pipeline. While they have proven effective in many laboratory and field tests, the typically require significant knowledge of the system layout and thus are can not be easily deployed on a large scale. On the other hand, data driven leak detection methods, given a general training period, can be deployed on a much larger scale without many of the previously listed limitations. This chapter goes into more details on each method.

2.3.1 Water Balance or Water Audit

The amount of water lost in a distribution system can be quantified using a water balance, which is a tool used to study the flow of water in and out of a system or sub-system. There are two dominant balance methodologies used for quantifying water losses: (1) the IWA/AWWA standardized water balance methodology [Valentine, 2009], explained next and (2) the UK water balance methodology [Lambert, 1994], as seen in Table 2.3. These water balance methodologies were based on work performed by May [1994b] and the Water Research Foundation [Wallace, 1987].

Table 2.3: The UK water balance methodology [Lambert, 1994]

<table>
<tr><th colspan="5">Distribution Input (DI)</th></tr>
<tr><th colspan="4">Water Taken (WT)</th><th>Distribution Losses (DL)</th></tr>
<tr><td colspan="3">Water Taken (WT)</td><td>Distribution Operational Use (DOU)</td><td>Distribution Losses (DL)</td></tr>
<tr><td colspan="2">Water Delivered Through Supply Pipes (WDS)</td><td>Miscellaneous Water Taken (WTM)</td><td>Distribution Losses (DL)</td></tr>
<tr><td>Measured (WDSM)</td><td>Unmeasured (WDSU)</td><td>Unmeasured Supply Pipe Losses (WDSL)</td><td>Miscellaneous Water Taken (WTM) [Legally and illegal, Meter under-registration]</td><td>Distribution Losses (DL)</td></tr>
</table>

The IA/AWWA water balance methodology takes the system input volume (corrected for known errors, such as possible inaccurate meter readings or inaccurate projected consumption patterns) and divides this into two components [Mutikanga et al., 2013]: (1) authorized consumption, and (2) water losses. The authorized consumption is then divided into: (1a) billed authorized consumption (this encompasses all revenue water), which is essentially made up of billed metered consumption, and billed unmetered consumption; and (1b) unbilled authorized consumption, which is made up of unbilled metered consumption and unbilled unmetered consumption. The water losses are divided into: (2a) apparent losses, which are made up of unauthorized consumption, and costumer metering inaccuracies and data handling errors; and (2b) real losses, which encompass leakage on transmission and distribution mains, losses at utility's storage tanks, and leakage on service connections up to customer metering point. The total non revenue water (NRW) is

all unbilled authorized consumption, apparent losses and real losses.

The challenges associated with measuring water flows in and out of the entire system as well as the development of the aforementioned water balance methodologies lead to the implementation of district meters areas (DMAs) within a municipalities water distribution network (WDN), in which utilities partition their distribution systems into smaller, more manageable sub systems. DMAs typically represent between 500 and 3000 properties [Morrison et al., 2007]. The flows into and out of DMAs are closely monitored to determine leakage as excess flow beyond legitimate customer usage. This flow monitoring usually occurs in the middle of the night when legitimate customer use is at a minimum, network pressures are high, and leakage is at its maximum percentage of total DMA inflow [Mutikanga et al., 2013]. If a water system has already been established, implementing a DMA system is not something that can easily be incorporated with existing infrastructure, it is a costly and difficult task; while for newer systems it is a simple change in the pipeline layout to implement.

2.3.2 Acoustic Leak Detection

Acoustic leak detection depends on the vibration (or, sound) generated by water leaking from an orifice in a pressurized pipe. The vibration is transmitted through both the pipe materials [Leslie-Milbourne et al., 2004] and the water within the pipe [Khalifa et al., 2010, Khulief et al., 2011]. The frequencies produced by leaks vary depending on the type of leak, pipe, and backfill. It is also important to note that not all leaks produce a detectable vibration.

Qualitatively speaking, three types of vibration are generated from leaks in buried pipes [Hennigar, 2013]: (1) friction vibration, (2) fountain vibration, and (3) impact vibration. Friction vibration is a result of water forcing its way through the pipe wall and emanating vibration through the pipe. It tends to be higher in frequency, ranging between 300 to 3000Hz. Fountain vibration is the lower frequency vibration (10 to 250Hz) generated by turbulent, circulating water around the leak site. Impact vibration (10 to 250Hz) is generated by the collision between the exiting jet of water and the backfill surrounding the pipe [Hennigar, 2013]. Most studies have shown that frequencies above 300 Hz are rarely recorded in acoustic measurements (passive) as the attenuation is generally high and especially so in non-metallic pipes. Hence, most of the acoustic energy is concentrated below 300 Hz and in most applications.

Factors Affecting Leak Vibration Quality

A great degree of variation is possible within the pipe and the surrounding environment when considering the details of an appropriate system to use. As a result, the acoustic energy, quality and propagation of the emitted leak vibration are different from case-to-case. A few of the factors that affect the quality of acoustic energy generated by leaks are listed below:

- **Water pressure within the pipe**

  Leak induced acoustic energy is proportional to pressure: the higher the pressure, the better the leak induced energy intensity [Thornton, 2002]. Therefore, there is an advantage to performing acoustic leak detection at night, when the water distribution system pressure is typically at its peak.

- **Type of pipe**

  Summarily, the harder the pipe material and the smaller the diameter, the better the quality of the leak induced energy [Thornton, 2002]. On average, the pipe materials and their associated sound qualities are as follows: *good for leak noise sounding*: cast iron, steel, copper; *average for leak noise sounding*: ductile iron, asbestos cement. *poor for leak noise sounding*: PVC, MDPE, HDPE, internally lined/externally wrapped.

- **Type of backfill covering the pipe**

  Cavities and moisture in backfilled soil diminish the transmission of leak vibration [Thornton, 2002]. Sandy soils and asphalt conduct vibration well whereas clay and concrete do not.

- **Sources of interference with leak noise**

  There are many sources of interfering noise in the process of acoustic leak detection that may obfuscate the signal. These include: automobiles, aircraft, trains, pressure-reducing valves, partially closed valves, and other vibrating mechanical equipment.

- **Type of leak**

  As covered in Section 2.3.2, there are many types of leaks that could occur in a WDN. Each type of leak typically produces a noise of a particular quality. Smaller leaks often have a high frequency "hissing" sound and larger leaks are characterized by a low frequency "rumble" [Thornton, 2002]. The details of the noise quality for the leak types are covered in Section 2.3.2 as described next.

Expected leak noise quality differs depending on the type of break. Main breaks caused by pipe fracture create a low frequency rumble caused primarily in two cases: (1) substantial leak rate reduces the pressure in the pipe, weakening the acoustic energy (e.g., vibration of the pipe shell or acoustic waves within the water column), (2) the stationary water pocket created at the location of the leak site could dampens the leak-induced acoustic energy. Cracks and pinhole leaks typically cause a large variety in noise quality, but usually have a high audible frequency. Seepage will typically cause very poor noise quality, therefore these leaks are usually categorized as undetectable background leakage. On the other hand, leakage on packing glands of pumps and valves cause highly audible frequency, detectable by direct sounding at the valve spindle. Pipe joint leaks cause a large variety in quality, mainly dependent on pipe material. Finally, leaking service connection pipes and leaking fire hydrants, air valves and scour valves are usually easy to detect and access since close-proximity direct sounding is possible [Thornton, 2002].

Hunaidi and Chu [1999] performed an experimental investigation into the acoustic signatures of different simulated leaks in plastic water distribution pipes and found that most leak noise is concentrated at low frequencies. In their study the authors concluded that the spectral region of interest for leaks in plastic pipes in water distributions systems is between 5 and 200 Hz Distinct peaks are found at the low end of the spectrum from 5-40Hz which Hunaidi and Chu [1999] attributed to the simulated leak. Hunaidi and Chu [1999] stated that typically little information is gained by including above 45Hz and thus recommends low pass filtering at this point. As a note, it is important to recognize that these recommendations may or may not applicable to general situations and should be followed with caution.

Acoustic Leak Inspection Equipment

In order to perform acoustic leak detection, there are several types of inspection equipment in use today. Each of these key approaches are discussed in the context of the literature landscape, next.

- **Mechanical or electrical listening stick**

  The listening stick is a traditional instrument used to systematically sound all mains and service connections. The most common type is a simple steel shaft connected to an ear piece with or without a sound amplifier. The listening stick is placed onto a fitting, whereby any leak noise is transferred from the pipe, to the steel shaft, and finally to the ear piece for interpretation by the technician. This poses a number

of problems, but the most limiting of these is that of human listening. While the human audible range is from 20-20,000Hz, humans have limitations in listening at the lower end of the frequency spectrum. The frequency created by most reasonably sized leaks range from approximately 20Hz to 250 Hz [Mutikanga, 2012]. While this falls within the range, human error is more likely to occur, with people of slightly older ages, within this range.The effect is amplified when plastic pipes are used, due to the stronger attenuation rate of vibration, and most leak frequency signals are below 50 Hz [Hunaidi et al., 2000]. The pipe burial depth also effects the accuracy of this method's detection capabilities, since pipes deeper beneath the ground generally lead to more vibration attenuation.

- **Ground microphone or geophones**

  Ground microphones are used to listen for leaks from the ground's surface where direct system contact points such as valves, hydrants, and service connection curb-stops are spaced far apart, making other investigative tools less reliable. Ground microphones can also be used to pinpoint the exact location of a leak using cross correlation, described below.

  This is a newer technology built off of the same principle of listening rod systems, incorporating amplifiers to help alleviate the limitation of the human acoustic listening range [Smith et al., 2000]. However, this still leaves the limitations of system effectiveness to the operator's experience and subjectivity.

- **Leak noise correlator**

  A leak noise correlator is comprised of a correlator unit made up of a receiver and a processor, and two sensors equipped with radio transmitters. The two sensors must be placed on exposed fittings, straddling a suspected leak. The leak noise radiates out through the pipe and eventually collected at each sensor straddling the leak. This process differs from the aforementioned methods because it uses the speed of sound as opposed to acoustic energy level created by the leak. The correlator uses the time difference between the two arrival times at the two sensor locations, information about the pipe material, size, and the distance between the two sensors to calculate the location of the leak [Gao et al., 2005]. The accuracy of the process is highly sensitive to the operator's inputs and the physical characterization of the piping system. Tee and branching lines can similarly cause problems [El-Abbasy et al., 2016, Bracken and Cain, 2012, Gao et al., 2004]. However, beyond a proper assessment of the deployment location, the operator's skill level required is relatively low [Li et al., 2015]. This is by far the commonly employed method to localize leaks within a pipe run where a leak is known to be present.

Hunaidi [2000] found that while commercially available leak noise correlators can find the location of leaks in controlled tests when the presence of a leak is known, the detailed knowledge of the parameters for the site specific investigation need to be measured at the time of the test to ensure accuracy. This type of system however has only been proven effective experimentally in cases where a leak is known to exist. Blind experimental test cases are limited, and with mixed results [Hughes and Venkatesh, 2016, Anguiano et al., 2016].

- **Digital correlating leak noise logger**

  A new distributed form of acoustic leak detection combines acoustic logging and leak noise correlation, such as Eureka Digital [ADS, 2009] or ZCorr [SubSurface Leak Detection, 2019]. It essentially consists of multiple acoustic loggers installed in the vicinity of the suspected leak, on flooded hydrants, valves or other exposed or accessible piping. A controller of these loggers is then used with a special software, proprietary to the company of choice, to determine the location of the leak. It can then create a multi-dimensional map to locate the leak more accurately than the two point correlator systems. This method is also automated, minimizing the need for experienced operators. These developed systems utilize proprietary algorithms for leak detection and localization. Based on background knowledge of these types of systems, it is likely they utilize cross correlation and automatic estimated wave velocity in the system.

  While these methods typically work effectively for larger metal pipes [Li et al., 2015], it has yet to be proven in literature that smaller leaks in larger diameter plastic pipes can effectively be detected. This is due, in part, to the low frequencies created by smaller leaks in medium to large sized pipes. The current state of this detection equipment's accuracy is dependent on the number and arrangement of loggers, with diminished accuracy in non metallic pipes.

- **Sensor inserted into the transmission main**

  The difficulty with executing the acoustic detection of leaks on water mains is that there is usually a long distance between fittings that can be used as sounding contact points. With these long distances and the pipes' large diameter, most of the sound (vibration) energy from a leak will dissipate before it can be detected. One type of leak detection equipment developed for transmission mains inserts a sensor into the transmission main that then travels along with the flow in the pipe, picking up any noise generated by a leak [Kurtz, 2007, 2006]. These free swimming acoustic leak detection devices can more effectively survey much longer lengths of pipe with a

simple deployment. Kurtz [2006] began review of this methodology with an isolated field experiment, Khulief et al. [2011] continued further validating the procedure with an simplified laboratory experiment, consisting of short lengths of pipes in sequence attached to a metal configuration meant to hold it in place.

While this method certainly has merits, results from independent academic studies are not available in literature to assess their performance in field studies. Furthermore, its effectiveness is directly tied to knowledge of its exact location at the exact point in time, also known as odometry. Furthermore, due to their size and the need to deploy roughly tennis-ball sized objects into pipes, their deployment is generally considered practical for large diameter mains.

- **Fiber optics**

  Another method for detecting leaks in water mains uses acoustic fiber optics [Jia et al., 2015]. Fiber optics for leak detection rely on different types of light backscattering in optical fibers in the presence of a leak within a system [Stajanca et al., 2018]. A continuous fiber optic cable is installed along the main and connected to a data acquisition system that allows permanent real-time acoustic monitoring. There are two types of fiber optic monitoring systems, temperature sensing and acoustic sensing. The changes in the optical characteristics due to the temperature changes caused due to coming in contact with water is utilized for leak detection, while the optical time delay is used for localization. Acoustic sensor cables (which rely on the Rayleigh-based waves [Stajanca et al., 2018]) work in a similar way except, instead of temperature differential, vibration induced by the occurrence of a leak is detected [Jia et al., 2015] using a time-domain reflectometry (TDR). TDR is used for detection by assessing the partially back-scattered laser pulses within a fiber-optic cables which occur in the presence of a leak.

  While these types of systems rely on fairly well established technologies, deployment at scale in geographically distributed systems is associated with prohibitively high installation and maintenance costs.

2.3.3 Non-Acoustic Leak Inspection

In addition to acoustic leak detection methods, there are also a number of non-acoustic methods that have been used by utilities and researchers. However, none of these methods are considered for wide scale use based on the current level of research and reported effectiveness. A number of non-acoustic inspection methods are described herein.

- **Video cameras**

  Tethered inline video inspection, while requiring a 2 *inch* or large tap, allows for real time CCTV inspection. A parachute is attached behind the camera which carries it down the pipeline. While results are reported mostly with regards to collecting useful information rather than leak detection [Laven et al., 2010, Kuntze and Haffner, 1998, Kirkham et al., 2000], it is an inspection method which can be utilized if a problem is known but the cause is unknown. A major limitation lies in the inspection length of the technology being limited by how far this flow can carry the hydrophone and cable through the pipe before friction stops it [Laven et al., 2010].

  Other researchers such as Nassiraei et al. [2006], Rome et al. [1999], use untethered robots for inspection, however, many still require system maps for navigation [Kirchner and Hertzberg, 1997]. They are typically used for sewer inspection [Nassiraei et al., 2006, Rome et al., 1999]

- **Tracer gas**

  A water-insoluble gas is injected into a pipe system and then detected at the surface using a sensitive gas detector. The gas used is typically hydrogen, due in part to it being lighter and easily detectable. This light-weight quality allows for small leaks to be more easily detected [Hunaidi et al., 2000]. The gas detector is sensitive to even small amounts of this gas. Due to the costly nature of this method, its application is typically limited to small leaks, since this is the area in which it has out-performed other more cost effective methods [Li et al., 2015].

- **Ground penetrating RADAR**

  This is a non-invasive, non-destructive testing method in which a continuous cross-section profile of the area is produced using high frequency electromagnetic waves, with instrumentation such as the pulseEKKO radar [sensoft, 2020] system. These radar waves are introduced into a system to map the location of the pipe. The mapping is based on the reflected signal. Leaks typically create voids, which lend themselves well to be detectable using radar. Alternatively, a reading of a change in pipe depth may be detected where there is none, which would also be indicative of a leak [Hunaidi and Giamou, 1998]. When the subsurface is frozen, as is typical seasonable in cold climate locations, this inspection method has not yet been proven effective [Eyuboglu et al., 2003].

- **Infrared technology**

  This method uses thermal infrared (IR) cameras to display emitted IR radiation from pipe systems. The IR camera measures the emitted IR radiation of an object. Thermal contrast exists and is detectable when a leak is present below the surface [Fahmy and Moselhi, 2010]. This inspection method is highly effective and requires minimal operational experience due to the rather simple visual nature of the captured images. However, in cold climates and surface conditions can greatly affect the quality of the image and thus the accuracy of this inspection method [Li et al., 2015].

The use of acoustic measurement sensors is more cost-effective and provide a more time effective response of the occurrence of system events.

2.3.4 Hydraulic Leak Detection

Hydraulic leak detection methods (i.e., non-acoustic, non-inspection methods) utilize hydraulic characteristics such as flow and pressure as a means to detect, locate and quantify leaks in pipelines. A brief description of such methods found in the literature is summarized next.

- **Hydrostatic**

  Pressure management techniques have been cited as being included amongst the primary factors required for leak management [Vicente et al., 2016]. This is common practice in DMAs for leak detection using a process called "step-test" [Mutikanga, 2012], which essentially involves the sequential closing of valves within a section of the pipe, and then reviewing the corresponding effect of flow on the meter. A large reduction would be indicative of that section of pipe harboring the leak. This method is limited by the need for DMAs to be deployed, and the great inconvenience it would cause to the locals in the surrounding area during periods of water shut off to their region. While this method can accurately detect if there is unaccounted for water being consumed within a DMA, identifying where and how it is being lost (theft, leaks, etc.) is beyond the scope of this methodology.

- **Steady State**

  Steady state hydraulic leak detection begins with the study of pressure across WDNs and how the presence of leak signatures manifest under steady state conditions [Pudar and Liggett, 1992]. Simply put, the expected pressure at a location is compared with

the measured pressure, and if the discrepancy between the two becomes too large a leak is identified as being in the vicinity. Pudar and Liggett [1992] addressed the problem of localization by analytically solving the inverse problem of equivalent orifice areas which could be creating the pressure signatures. This analytical method was later replaced by the use of EPANET2 which, with the aid of computer numerical derivatives, is able to obtain leak signatures [Perez et al., 2009].

Casillas Ponce et al. [2013] worked on an extended time horizon analysis of pressure sensitivity using this method. However he reported significant uncertainty can be present with the application of this method.

A major limitation with model based fault diagnosis methods largely centers around the lack of accuracy in leak localization. As well, the accuracy of detection relies heavily on sensor placement. Both of these limitations are heavily impacted by the variation in nodal demand values[Jahanpour, 2019]. Jahanpour [2019] addressed these issues however the methodology was never tested in a field setting.

- **Transient based detection**

  Transient based leak detection has received significant attention—both with and without attendant models—in both theoretical and experimental studies [Gong et al., 2016, Ferrante et al., 2013, Brunone, 1999], and field [Jackson et al., 1977, Papadopoulou et al., 2008, Nguyen et al., 2018] for over a decade.

  The underlying principle is that a pressure change caused by a burst or leak creates a transient (water hammer) to propagate from the leak location. The transient response of the system without a leak is used to model the system in its ideal state; transients are then induced to verify if a leak is present or the system is in the same state, since the presence of a leak would manifest as a change in its transient response. This methodology however requires accurate knowledge of the system pipe layout, since these waves are reflected at boundaries, as well as the wave speed within the system. Without accurate knowledge of all of these, parameter leaks can not be effectively located [Ferrante et al., 207]. The magnitude of the pressure wave is indicative of the magnitude of the leak, while arrival time of the reflected signal is related to its location [Brunone, 2001].

  One well studied method for leak localization is TDR, in which the time of arrival of partially reflected signals are used to determine leak location. When an induced transient signal meets a leak, and part of the energy is reflected back, the presence of a leak is exposed and can be used for localization [Brunone, 1999]. This method is rather simple conceptually, however detecting small changes in pressure caused by

this leak reflection can be difficult if the change is not significantly large enough. Furthermore, the presence of background transients and instrument noise can significantly disrupt the signal [Colombo et al., 2009]. Brunone [1999] determined the presence of these leaks by visual inspection of the change in transient response between leak and non leak cases, while other researchers continued reviewing this method in order to automate this analysis process. This was done to automatically detect the change in response without the need for visual inspection [Lee et al., 2007]. Misiunas et al. [2005a] demonstrated good accuracy with the application of the cumulative sum (CUSUM) algorithm [Basseville and Nikiforov, 1993] which detects changes in the assumed mean of the signal in the ideal system, and the slightly increased mean of the signal when a leak is present [Eliades and Polycarpou, 2012]. This procedure was tested with field experiments. However for the most part, this method has only been validated on very simple, single pipe set ups in laboratory settings.

These methods have been further expanded by Lee et al. [2005] with the development of frequency domain analysis methods, in which transient analysis was done in the frequency domain as opposed to previous works focused in the time domain.

- **Inverse transient**

  This method generally involves inducing an acceptable sized transient, then measuring the systems' response at a selected location in the network in order to compare the results with simulated pressure responses until a best fit is determined. The simulated case which most closely matches the actual data is then representative of the most likely scenario to be occurring in the network. Pudar and Liggett [1992] was among the first to advocate for transient based methods, and was the first to discuss the application of an inverse problem, in which simulated responses are used as comparisons to determine the actual state of the system. This method was considered an improvement on the transient method since a baseline was no longer required.

  While this method works well in theory, the generation of accurate models and knowledge of system details and representative leak simulations is a limiting factor. Researchers have experimentally validated inverse transient methods [Soares et al., 2011], however this is at a small scale and controlled laboratory conditions, in which all parameters are known and easily calculated.

2.3.5 Data Driven Leak Detection

The aforementioned inspection and monitoring methods all essentially rely on the data collected to infer leaks. Once pertinent data has been amassed, it must be analyzed, either

after the fact, or for immediate event detection. In general, the problem of data-driven leak detection can be addressed as supervised [Terao and Mita, 2008, Rashid et al., 2015] or semisupervised problems [Mounce et al., 2010]. In the supervised case, both leak data samples and non-leak samples are needed to build a classification model, which is often hard to obtain in full-scale field applications. Semi-supervised approach (also known as anomaly detection) only requires the normal state (or non-leak) data to train the detection model and hence is a more practical alternative in continuous leak detection monitoring scenarios.

Literature pertaining to data-driven learning methods with water system applications is limited. A number of machine learning methods can be applied to leak detection in pipes, the most common and promising of which includes Support Vector Machines(SVM) [Sato and Mitra, 2007, Mounce et al., 2011], artificial neural networks (ANN) [Mounce et al., 2010, Aksela et al., 2009, Romano et al., 2011, Jin et al., 2010, Mounce et al., 2006, Caputo and Pelagagge, 2003, Mounce and Machell, 2006], and Bayesian Learning[Poulakis et al., 2003]. These methods yield effective classification, however their application is typically limited to fully supervised training sets. These methods require specification of a finite number of system states (i.e., classes), limiting applicability to case studies where data sets during leak events are available. Pressure and water demand data are the predominant data types used for leak detection studies [Misiunas et al., 2005b]. While these sensor types are well suited for low resolution sampling, making them ideal for long term deployment, this limits their detection to larger leaks at a greater spatial scale. The application of a classification method involves knowing the number of classes prior to the simulation, which involves an extensive supervised training period in order to implement these methodologies in the field. However, to that extent these methods have been validated.

A number of studies, e.g., [Mounce et al., 2010], have applied artificial neural networks (ANNs) to detect pipe bursts. This process yields effective timely classification, however, significant historical data is generally needed for the training process. This is not always available, limiting its application, or requiring an extended training period. Aksela et al. [2009], Romano et al. [2011], Jin et al. [2010], Mounce et al. [2008], Caputo and Pelagagge [2003] also reviewed the application of ANN method and yielded promising results for pipe burst detection, however the same limitations were found. ANN is a good method for obtaining reasonable predictions, however when applied to water distribution networks, extensive data history and extended training periods, generally on the order of many months, is required, making this process very computationally expensive. Furthermore, ANNs need constant updating to maintain their accuracy. In the realm of model-based data-driven methods, Ye and Fenner [2011] found that the adaptive Kalman filtering improved the performance of ANNs while reducing the training period time.

SVMs as binary classification method has been reviewed by many researchers, including Salam et al. [2014], Mashford et al. [2012], Zhang et al. [2016], Aksela et al. [2009]. Others, such as Terao and Mita [2008], who reviewed traditional two-class SVM in the time domain, in which leaks and non leak cases are trained and tested, yielded high accuracy of 97%. This was done in a field case study situation and used an interesting PCA based feature selection method. Rashid et al. [2015] reviewed the data in both time and frequency domain, yielding similar results of 78%-94% accuracy when applying KNN, SVM and GMM models for binary classification of leak and non-leak cases. The features selected in the time domain were expected value, variance, gradient, and Kurtosis. The frequency domain features included the selected pseudo spectrum, entropy, power spectral density, and percentage of energy.

Another non-numerical modeling method, namely Bayesian inference, involves the probabilistic classification of a current state belonging to one of the previously known cases; the case in which the current state shares the highest probability. This has been reviewed extensively by Leu and Bui [2016], Poulakis et al. [2003], Puust et al. [2008].

Data-driven methods relying on machine learning tools for leak detection in WDNs have largely been applied to pressure or water demand measurements. For example, Wu and Liu [2017] summarize a table of 21 past data-driven studies in their review paper and all of these studies only utilize either demand or pressure data. Past studies using hydro-acoustic data (collected in the water column) to detect leaks in pipelines are less common and include Khulief et al. [2011], Almeida et al. [2014], Martini et al. [2017], Gao et al. [2018]. Almeida et al. [2014], Gao et al. [2018] both review different techniques using cross-correlation applied to leak detection in PVC pipe systems. Martini et al. [2017] uses a hydrophone and accelerometer to denote the difference between the *leak* and *leak-free* cases. While Khulief et al. [2011] applied a similar baseline deviation novelty detection assessment based on the Fourier domain approach, they replaced deep learning for novelty detection with simple statistical measures such as root mean-square (RMS). This reliance on simple statistical data summaries requires the system to have minimal baseline variability and noise. Similarly, the use of the cross-correlation approach in leak detection requires a fairly quiet system with a number of variables known a-priori. Martini et al. [2017] also based their novelty detection on basic statistical feature analysis.

A major challenge in all the aforementioned classification techniques is the necessity of known classes. While studies have effectively classified cases with high accuracy, they all require knowledge of all possible cases during the training period. This is not always readily available. Furthermore, it is not robust or easily adaptable.

Anomaly Detection

The application of anomaly detection algorithms involves modeling the normal state of the system, enabling deviations from the known norm to be detected, i.e. the detection of an event which strays from the system's normal state. This is a benefit from traditional classifications algorithms as it can detect previously unknown events. One drawback to this, and any classification methodology, is the need for a training period. Anomaly detection algorithms can be broadly classified as a semi-supervised learning algorithm.

There are a number of well established anomaly detection methods, including K nearest neighbor (KNN) [Zhang and Zhou, 2005], correlation based outlier detection [Koh et al., 2007], one class SVM [Cody et al., 2017, 2018], and many more. Mukkamala et al. [2002] showed that for intrusion detection, the OCSVM methodology developed by Scholkopf et al. [2001] generates models which not only process much shorter training periods but often outperform ANNs. While a number of methods could successfully work for anomaly detection, reviewing all of these is not feasible. Anomaly detection methods, however, have been minimally applied to water distribution networks and their application has thus far been primarily limited to water quality anomaly detection. Mounce et al. [2011] reviewed machine learning based techniques for anomaly detection in time domain. However, the study of instance-based, data driven methods for anomaly detection in water distribution systems is limited.

2.4 Testing and validation

In order to verify the effectiveness of different long-term monitoring methodologies, testing and validation is an important step. Many of the current WDN laboratory experimental setups are too simplistic and do not capture the complexity of actual WDNs. While actual field conditions can be difficult to replicate in laboratory settings, an attempt to better represent them still needs to be made. Many laboratory experimental test-beds lack fundamental field representations, which may include:

- typical field pipe diameters should be utilized, as opposed to much smaller pipes, such as those used by Jia et al. [2015]
- representative materials and pressures,
- representative means of pressurizing the system [Khalifa et al., 2010] —that is avoiding the use of pumps to pressurize the system as it saturates the low end of the

frequency spectrum (unlike Jia et al. [2015], Khulief et al. [2011]); the use of reservoirs [Soares et al., 2008, Lazhar et al., 2013, Mpesha et al., 2001, Lee et al., 2005] can also cause an unrealistically quiet system,

- realistic hydrophone mounting within the test bed as opposed to mounting the sensor in the middle of the pipe section along the main flow [Khulief et al., 2011, Ferrante et al., 2013],
- sufficient complexity with the inclusion of bends and tees [Jia et al., 2015],
- and longer distances [Ferrante et al., 2013, Soares et al., 2008, Covas et al., 2006] as opposed to very short pipe segments [Khalifa et al., 2010].

The use of laboratory test-beds as a means of methodological validation, however, often present many challenges including: pressurizing the system via representative means; a need for adequate space, resulting in over-simplified networks; and the required use o realistic materials, since different materials possess different physical properties which can greatly effect the outcome of a test. While laboratory test-beds are an adequate first step in methodological validation, field deployment is the ideal validation case study.

2.4.1 Field implementation

Relatively few studies exist in the open scientific domain dealing with field implementation. PIPENET [Stoianov et al., 2007] and WaterWise [Whittle et al., 2010, 2013] are notable examples of such field implementations, especially dealing with monitoring technologies in WDNs. These studies employed pressure and accelerometer measurements to detect large transient events along with water-quality monitoring. The effectiveness of pressure sensors are limited to the detection of large events associated with pressure transients; also, while accelerometers provide more accuracy when in close proximity, they are not effective on plastic pipes due to the attenuation associated with acoustic waves travelling on plastic pipe walls. Moreover, their systems are focused solely on low frequency detection. While this is consistent with the findings of Hunaidi and Chu [1999], due to the large amount of noise inherent to any large scale system, the review of a larger portion of the spectrum is imperative to ensure accurate results. In the event the noise in the system saturated a region of the frequency spectrum, harmonics of the leak signatures can still be reviewed. As well if a pump or electrical system is located in the vicinity of the sensors, the low region of the frequency spectrum will be saturated.

Sadeghioon et al. [2014] developed SmartPipes as a pressure monitoring system, dependant on force-sensitive resistors that are mounted on the surface of the pipes. This system measures the changing diameter of the pipe which results from internal pressure changes. Thus, it requires excavation to expose pipe regions for installation. Also, Sadeghioon et al. [2014] do not address the leak size that would cause a measurable pressure change that is noticeably different from the pressure change associated with simple consumption variability.

Current commercial acoustic technologies, such as Echologics, are generally limited to metal pipes [Bracken and Cain, 2012, Wang et al., 2010, Hughes and Venkatesh, 2016]. As WDNs are updated, cities are more frequently opting for PVC pipes as opposed to metal ones, limiting the effectiveness of technologies based on accelerometers and pressure sensors. As well the majority of these commercially available technologies, when evaluated in blind field tests were inconclusive or concluded low success rates [Hughes and Venkatesh, 2016, Anguiano et al., 2016].

2.5 Gaps in existing literature

The identified gap areas in the current state of event detection in water distribution systems and the expected contributions in terms of specific research objectives from this dissertation are summarized below.

1. Existing water distribution event detection systems such as those discussed previously in this chapter are deployed either when there is already knowledge of a problem, or more specifically the existence of a leak, or must be deployed in specific known regions since they are only accurate when traced over every portion of the region of interest. Long-term passive monitoring, in many cases, is either too expensive to install or have unproven benefits over previously implemented maintenance procedures.

2. Nearly all leak detection laboratory test systems are unrepresentative of field conditions, such that any results for systems described in the literature cannot reasonably be compared to field conditions. There is often the use of either unrepresentative materials, very small ratios between the leak size and pipe diameter, as well a generally very short pipe segments.

3. Fire hydrants remain the most common access points in a WDN; however, they have only been used under flooded conditions to detect leaks. Such a method does

not extend to long-term monitoring situations, especially in cold climates. To the author's knowledge, this is the first time where dry-barrel hydrants have been used for leak detection and localization.

4. While the detection of anomalies using machine learning is well-established, current passive monitoring systems utilize either feature and threshold methods for event detection, without a good justification for the use of either, or methods far too computationally intensive for realistic field deployment.

5. Many methods in the literature have employed acoustic methods for metal pipes and those results cannot simply be extended to field settings, especially for PVC pipes, which remain a commonly used material in WDNs today.

2.6 Specific Objectives

Based on the identified gap areas, the specific research objectives of this dissertation are as follows:

1. To develop the sensor hardware and software specifications for a hydrant mounted leak monitoring system. This entails:
   (a) developing the requisite hardware with adequate capabilities for data sampling and collection, and
   (b) developing the software in order to enable passive data collection during desired times of interest.

2. To develop an algorithmic framework to implement leak event detection based on the acquired sensor data from the developed hydrant mounted system. This entails:
   (a) reviewing the sensitivity of the prediction method to leak induced signals,
   (b) developing a detection and localization methodology which is scalable to long-term monitoring and to field conditions, and is robust to the natural variability existing in the environment.

3. To experimentally test and validate the system using both laboratory experiments and field tests on a live water distribution network. This entails:

(a) developing a laboratory test best for proof of concept validation of the proposed methodology which is reasonably representative of field conditions, and

(b) experimentally deploying the proposed system in a live water distribution network and examining its performance.

Chapter 3

Background

In line with the research objectives presented in the previous chapters, this chapter provides relevant background on concepts that will be useful throughout the dissertation. In addition to a brief theory (linear) of sound propagation in pipes, this chapter includes concepts within signal processing and dimensionality reduction method which are used in the data processing phases of the analysis later presented; the theory behind the classification algorithm used is described in detail; as well as a detailed description of the class separability metric and various classification performance measures.

3.1 Sound Propagation in Water Pipes

Linear theory governs much of our understanding today regarding the propagation of sound in water. The confinement provided by the pipe material has a significant effect on this propagation, both in terms of the modes and attenuation. This section will briefly review the basic equations of motion within a pipe along with the equation for a fluid borne wave, which is expanded to derive the equation for the speed of sound in water used within this dissertation.

Acoustic waves attenuate inside pipelines is largely due to the intrinsic absorption properties of the material the pipeline is constructed with. As such, larger diameters and more flexible pipes (e.g., plastic) tend to attenuate acoustic energy significantly more compared to their rigid counterparts such as cast iron. Previous studies have shown that signals in the low-frequency end of the spectrum are the most reliable for leak detection suggesting that these frequencies are both excited by the leak and propagate most effectively. Previous studies, e.g., by Hunaidi and Chu [1999], concluded that the region of interest for leaks

in plastic pipes in water distributions systems is between 5 and 200 Hz (up to 1 kHz for metal pipes [Ma et al., 2019]). Higher frequencies are attenuated through damping present in the pipe walls and connections; as well, due to the coupling between the fluid and the pipe wall in the radial direction is also responsible for the significant dampening which occurs in plastic pipes [Muggleton et al., 2004]. Hence, only low frequency waves remain and propagate in the pipes for long distances.

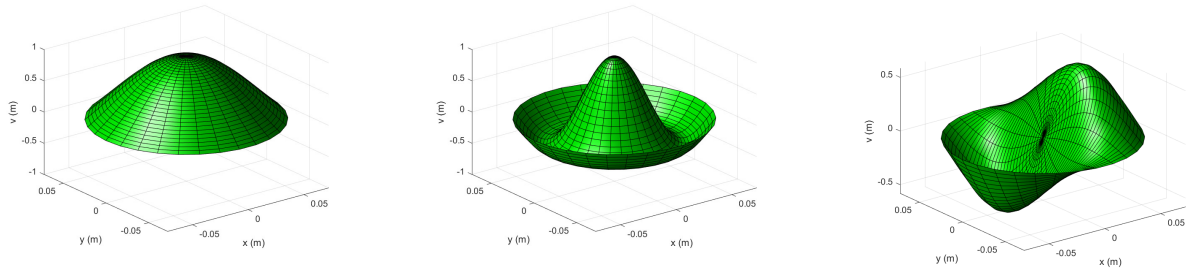
3.1.1 Wave Equation of Motion

The wave equation in one dimensional space can be derived using several physical analogs, such as a vibrating string. Similarly for two dimensional space, the motion of a thin membrane such as drum-head that is stretched uniformly in all directions can be used to motivate the mathematical formulation.

The physical motion of (m, n) modes is derived in Appendix D [Kinsler et al., 1999] as expressed in equation 3.1, for the fundamental mode shapes are shown in Figure 3.1.

$$u_{r,\theta,t} = A_{mn} J_m(k_{mn} r) cos(m\theta + \gamma_{mn}) cos(\omega_{mn} t + \phi_{mn}), \tag{3.1}$$

where A is the amplitude of the mn mode, $J_m(kr)$ is the *Bessel function* of order m of the first kind, k is the wave number, m is mass, θ an angel for the equation in cylindrical coordinates, ϕ is the phase angle, ω is the angular frequency, γ is the azimuthal phase angle and t is time.



(a) Mode(0,1) (b) Mode(0,2) (c) Mode(1,1)

Figure 3.1: Modes of wave propagation in a fluid inside a rigid pipe.

In Figure 3.1, the modes are denoted by the pair (m, n), where m represents the number of radial nodal lines whereas the second integer n controls the number of nodal circles. The

minimum value of n is 1 which corresponds to first mode with the nodal circle at fixed boundary.

The rigid boundary at $r = a$ also means that the normal component of the velocity vector is equal to zero,

$$J'_m(j'_{mn}) = \frac{\partial}{\partial r}[J_m(k_{mn}a)] = 0. \tag{3.2}$$

For a wave to be a plane wave in the case of free-field (open water), the acoustic variables have to be of constant amplitude and phase on any plane perpendicular to the direction of propagation. For the case of fluid-filled pipes, mode $(0, 1)$, which is the first mode, is considered as the plane wave mode because the particles involved in the fluid motion orthogonal to the direction of propagation are all in phase. The frequency below which this mode occurs can be obtained by using the extrema of the first kind of Bessel's function [Kinsler et al., 1999] and this frequency is termed as the cut-off frequency. The first few roots of Bessel function and its derivatives are listed in standard texts, including Kinsler et al. [1999]. Therefore, the cutoff frequency obtained using extrema of Bessel's function can be obtained using,

$$fc_{mn} = \frac{1}{2\pi}\frac{j'_{mn}c}{a}. \tag{3.3}$$

It is safe to assume that leaks induce energy in the lower spectral regions. While there is little in the published literature about what exactly the near field frequencies are, generally it is believed that leak induced energy is in the hundreds of hertz range and not in the kHz, especially a few centimeters away from the leak location. Hence, the modes associated with (m, n) of (1,1) or (0,2) cannot theoretically exist in this case. This is because the lowest cutoff frequency fc_{mn} corresponds to $(1, 1)$ and is equal to $1,863\ Hz$ for the $15\ cm$ pipes used in this study, and studies show that most leak energy is far lower than this value. Stated differently, the only mode that can exist in this system is the $(0, 1)$ mode and no others. This is also evident from the the axial wave number given by,

$$k_{amn} = k_{mn}\sqrt{1 - \left(\frac{fc_{mn}}{f}\right)^2}. \tag{3.4}$$

which becomes imaginary for frequencies greater than the frequency associated with the $(0, 1)$ mode.

3.1.2 Expression for a fluid borne wave

The pipe equation for $n = 0$ axisymmetric wave motion begins with the equilibrium of forces in the axial and radial directions. These are given, with reference to Figure 3.2, in equations 3.5 and 3.6 [Muggleton et al., 2002], respectively.

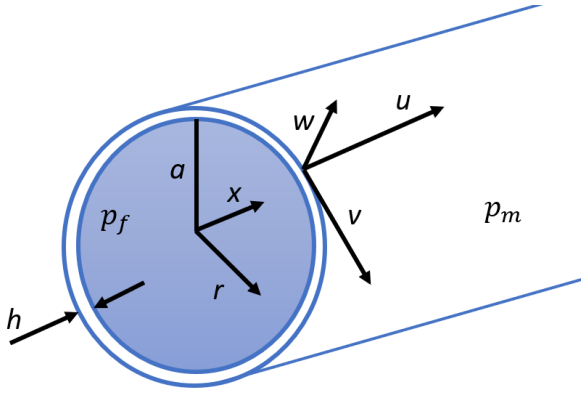


Figure 3.2: The co-ordinate system for a fluid filled pipe, surrounded by an infinite elastic medium. The shell displacements are u, v and w in the axial (x), circumferential (θ), and radial (r) directions respectively.

$$\rho\ddot{u} = \frac{\partial \sigma_x}{\partial X}, \tag{3.5}$$

$$(p_f(a) - p_m(a))(a/h) = \sigma_\theta + \rho a\ddot{w}, \tag{3.6}$$

in which it is assumed there is no circumferential variation, evaluated at $r = a$. Where σ is the stresses, ρ is the density of the shell material, and a and h are the radial and thickness of the shell wall, respectively $(h << a)$.

The travelling wave solutions in equation 3.7 [Kinsler et al., 1999] can be used to describe the displacements,

$$u = U_s e^{j(\omega t + k_s x)},\ w = W_s e^{j(\omega t + k_s x)}, \tag{3.7}$$

in which ω is the angular frequency, and k_s is the axial wavenumber for the s wave. This coupled with the shell equations can further be expanded to derive the expression for the s wavenumbers [Muggleton et al., 2002].

The expression for the wavenumber, k_1 (plane wave mode) of the fluid borne wave of motion for a fluid filled pipe surrounded by an infinite elastic medium given by Muggleton et al. [2002] and can be expressed as a ratio of the impedance of the fluid ($z_{fluid} = -2iB_f/(a\omega)$ to the pipe wall ($z_{pipe} = i(\rho h\omega - Eh/(a^2\omega))$ and surrounding medium (z_{rad}) [Muggleton et al., 2004], as shown in equation 3.8. Each impedance is the resistance defined as a function of both the wavenumber in the external medium (frequency) and the radial component of that wavenumber (wave angle).

$$k_1^2 = k_f^2(1 + \frac{z_{fluid}}{(z_{pipe} + z_{rad})}), \tag{3.8}$$

in which

$$z_{rad} = R_{rad} + i\omega M_{rad} = \sum_m \frac{-i\rho_m c_m k_m}{k_{m1}^r} \frac{H_0(k_{m1}^r a)}{H_0'(k_{m1}^r a)}, \tag{3.9}$$

where M_{rad} and R_{rad} are the mass and resistance components of the radiation impedance of the surrounding medium at the pipe wall; m is each wave type present in the surrounding medium, ρ_m, c_m and k_m are the density, wavespeed and wavenumber, respectively, for all wavetypes present. $(k_{m1}^r)^2 = k_m^2 - k_1^2$ is the radial component of the wavenumber in the surrounding medium. Based on the assumption that the surrounding medium is infinite (and thus no incoming waves are present), H_0 is a Hankel function of the second kind, representing outgoing waves (when the $e^{i\omega t}$ time dependence is adopted), the prime denotes the differentiation [Muggleton et al., 2004].

Equation 3.8 can be expanded to,

$$k_1^2 = k_f^2(1 + \frac{\frac{2B_f}{a}}{\frac{Eh}{a^2} - \omega^2(\rho h + M_{rad}) + i\rho R_{rad}}), \quad (h << a) \tag{3.10}$$

where k_f is the contained fluid wavenumber, B_f is the bulk modulus of the contained fluid, a and h are the radius and thickness of the shell wall, respectively, E is the shell material Young's modulus, ω is the angular frequency, ρ is the density of the shell material [Muggleton et al., 2004].

The Young's modulus may be complex if the material is lossy ($E \to E(1 + i\eta)$ where η) is the material loss factor), which is the case for PVC pipes [Knight, 2007].

For the in-vacuo case, z_{rad} goes to zero, and thus equation 3.8 simplifies to [Muggleton et al., 2004],

$$k_1^2 = k_f^2(1 + \frac{z_{fluid}}{z_{pipe}}) = k_f^2(1 + \frac{2B_f/a}{(Eh/a^2 - \omega\rho h)}, \tag{3.11}$$

which at low frequencies simplifies to,

$$k_1^2 = k_f^2(1 + \frac{2B_f/a}{Eh/a^2}). \tag{3.12}$$

3.1.3 Speed of sound in water

The sound propagation velocity (also known as phase speed) is the speed at which a pressure wave travels in a given medium, as a function of the fluid's density and bulk modulus. This parameter is useful when determining how long it will take a wave to propagate through a system.

In order to derive the equation for the speed of sound from equation 3.12, the wavenumber becomes a function of the angular frequency and the speed of sound, $k_1 = \omega/c$; similarly $k_f = \omega/c_f$ where c_f represents the free field speed of sound. Thus using the free field speed of sound as $c_f = \sqrt{K/\rho}$, and radius as $a = 2D$, equation 3.12 results in equation 3.13. However equation 3.12 is based on the assumption that the pipe is thinned walled and thus free to expand throughout, i.e. $D/e > 10$, therefore $\psi = 1$ and thus the term is omitted.

The speed of sound in circular elastic pipes is calculated using [Gao et al., 2004, Pinnington and Briscoe, 1994]:

$$c = \sqrt{\frac{1}{\rho(\frac{1}{K} + \frac{D\psi}{Ee})}}, \tag{3.13}$$

in which $\rho = 1000\ kg/m^3$ is the density of the fluid, $K = 2.18 * 10^9\ Pa$ is the isothermal bulk modulus (K_t) of fluid fresh water, and $E = 3.069\ GPA$ is the elastic tensile modulus of PVC pipes. D is the inner diameter of the pipe, while e is the pipe wall thickness, and ψ is the pipe support factor.

3.2 Signal Processing Concepts

Signal processing broadly refers to the analysis, modification and synthesis of signals such as vibration, sound, images, etc. Signal processing techniques can be used for many applications, such as to improve storage efficiency, improving some aspect of the quality of the signal such as de-noising, and isolating, emphasizing or detecting certain components within the measured signal. The field of signal processing is mature and is rich with tools

that are able to extract pertinent information from time series data such as periodicity, fundamental components and non-stationary elements to name a few.

Specific to the analysis covered within this dissertation is the use of discrete time signal processing. That is, all signals reviewed are discrete time series which are obtained by sampling a continuous acoustic signal at uniformly spaced time intervals, as depicted in Figure 3.3. This uniformly spaced time interval is referred to as a sampling rate, that is the number of samples taken from a continuous signal per one second interval. The ability to represent the underlying continuous, but un-observable, function using a series of discrete values opens up the potential to utilize extensive computing resources to interrogate them in ways previously considered impossible. However, the process of digitization in itself results in several numerical artifacts, which has to be understood and dealt with appropriately as explained throughout this background section.

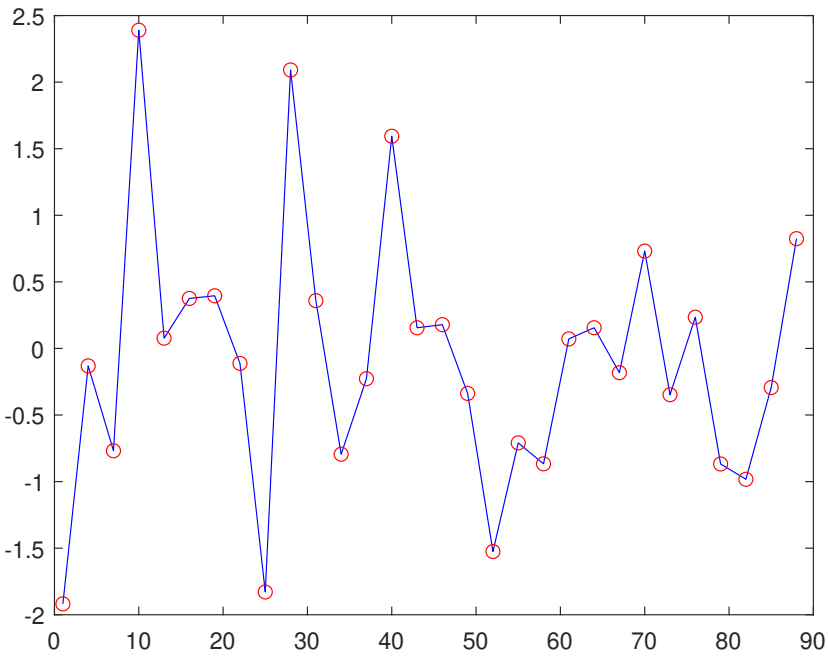


Figure 3.3: Continuous acoustic signal sampled at uniformly spaced time intervals.

3.2.1 Time Series Modelling

Although strictly not viewed only in the context of signal processing, fundamentally, this dissertation deals with time series data and hence modeling is described here. Time series data can be dealt with directly without an attendant model, or with an accompanying model. This section deals with the modeling approach, which will come in handy later on during the development of linear predictive models. The fundamental process in time series modeling to model a deterministic system is an Auto Regressive Moving Average model (ARMA model). This model is a combination of two simpler models, the Auto-Regressive model (AR model), and the Moving Average model (MA model). Therefore the ARMA model utilizes the flexibility of both these simpler models in order to model more complex systems; utilizing both auto-regressive and moving average inputs.

The AR model describes a system in which each data point is a weighted combination of its past values. It can be expressed as a time dependant linear function of a finite set of its weighted past values and a white noise term.

$$AR(p) : y(t) = \varphi_1 y(t-1) + \varphi_2 y(t-2) + \cdots + \varphi_p y(t-p) + a(t), \tag{3.14}$$

in which $y(t)$ is the data point being modeled and $a(t)$ is the white noise term [Hipel and McLeod, 1994].

The MA model describes a system which is represented by a series of identically distributed random variables. The MA model is expressed as a linear aggregation of previous white noise.

$$MA(q) : y(t) = a(t) - \theta_1 a(t-1) - \theta_2 a(t-2) - \cdots - \theta_p a(t-p). \tag{3.15}$$

White noise, or shock, is described as a normally distributed random variable with mean 0 and variance σ^2.

The ARMA model can therefore be expressed as a combination of these two models,

$$\begin{aligned} ARMA(p,q) : & y(t) - \varphi_1 y(t-1) - \varphi_2 y(t-2) - \cdots - \varphi_p y(t-p) \\ & = a(t) - \theta_1 a(t-1) - \theta_2 a(t-2) - \cdots - \theta_p a(t-p). \end{aligned} \tag{3.16}$$

Using the backshift operator (applied element wise to produce the previous element) the ARMA model can be simplified to, $\varphi(B)(z_t - \mu) = \theta(B)a(t)$ in which $y(t) = (z_t - \mu)$, where z_t is any given data point and μ is the mean of the series.

In order to identify the (p, q) values for the ARMA model, i.e. the model order, the given data must be generally explored, i.e. exploratory analysis must first be done. This involves:

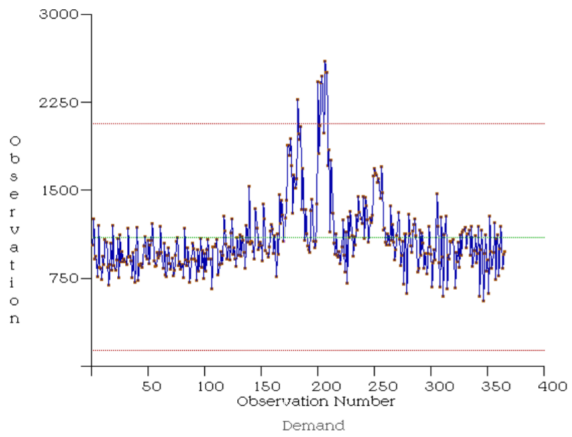
- checking for *normality* of the data set;
- ensuring *stationarity* in the data set, i.e. ensuring a mean of zero by differencing if needed;
- ensuring *homoscedasticity*, i.e. removing any change in variance by applying the Boxcox transformation [Box and Cox, 1964] to the data set if needed;
- and interpreting the autocorrelation plots. To do this the sample autocorrelation function (ACF), sample partial autocorrelation function (PACF), sample inverse autocorrelation function (IACF) and sample inverse partial autocorrelation function (IPACF) must be calculated and visually represented from the data set for interpretation.

The Boxcox transformation is a way of eliminating heteroscedasticity (creating homoscedasticity), that is non-constant variance. It transforms non-normal dependent variables into a normal shape [Hipel and McLeod, 1994]. The transformation has the form,

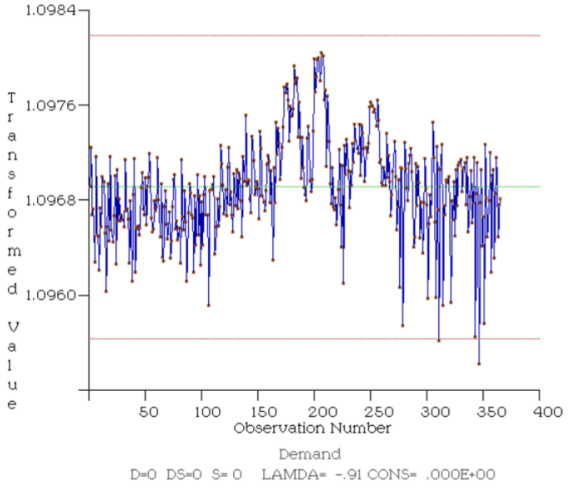
$$z_t^\lambda = \begin{cases} \frac{(z_t+C)^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 x \geq 1; \\ \ln(z_t + C), & \text{if } \lambda = 0, \end{cases} \tag{3.17}$$

where $z_t + C > 0$.

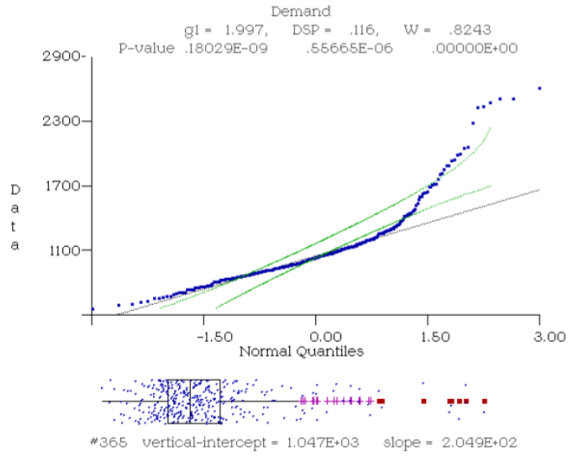
A sample data set and it's equivalent Boxcox transform can be seen depicted in Figure 3.4 (a) and (b), respectively. As well, their respective normality plots can be seen shown in Figure 3.4 (c) and (d), respectively. The latter show that the transform corrects for the non-constant variance of the data which more correctly aligns with the underlying model assumptions.The basic interpretations of the aforementioned graphs are outlined in Table 3.1.



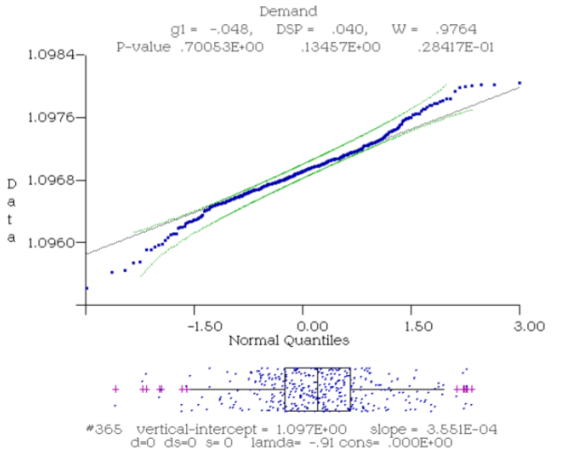
(a) Trace plot of the daily water demand data in the town of Ayr, Ontario for the year 2001.



(b) Trace plot of the daily water demand data in the town of Ayr, Ontario for the year 2001 with boxcox transform of $\lambda = -0.91$ applied.



(c) Normality distribution of data set, without any transform applied.



(d) Normality distribution of data set with $\lambda = -0.91$.

Figure 3.4: Normality distribution of data set.

Table 3.1: Interpreting correlation function plots.

Model	ACF	PACF	IACF	IPACF
AR	Attenuates	Truncates	Truncates	Attenuates
MA	Truncates	Attenuates	Attenuates	Truncates
ARMA	Attenuates	Attenuates	Attenuates	Attenuates

Sometimes, it is insightful to observe time series without inherent seasonalities, e.g., daily or seasonal. In order to remove such seasonal effects (i.e. temporal dependence) in the time series data additional techniques such as data differencing can be undertaken. For example, the previous observation is subtracted from the current observation, resulting in a series of differences. This can be seen demonstrated in Figure 3.5, where the distinct temporal dependence is removed in the differenced data.

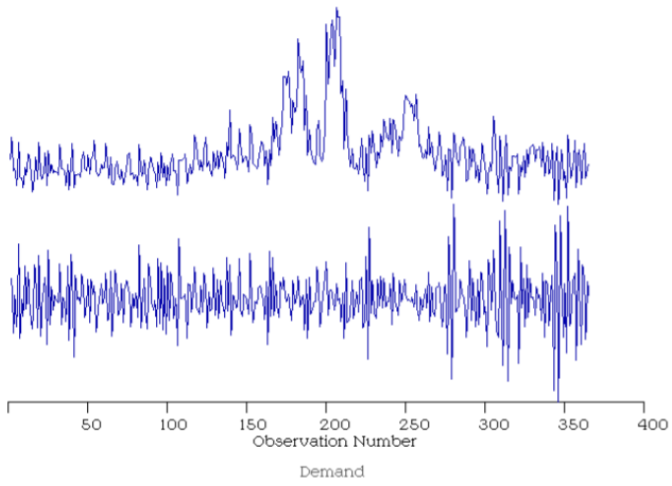
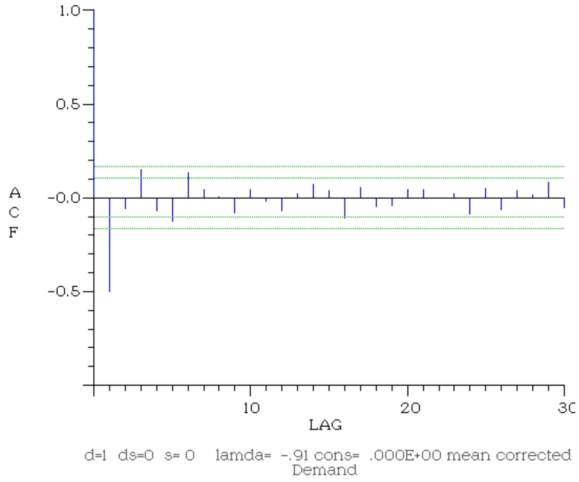
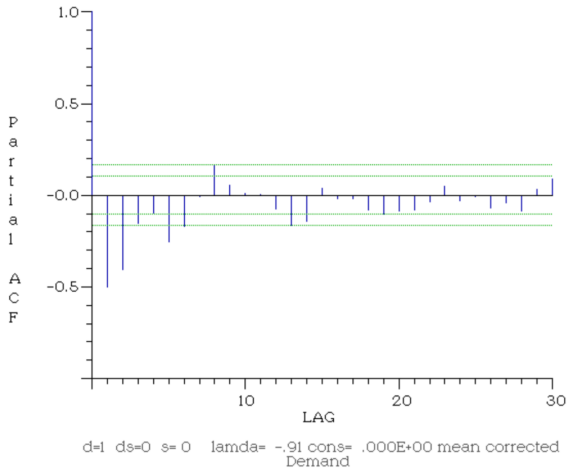


Figure 3.5: Bivariate trace plot of the boxcox transformed daily water demand data, i.e. $\lambda = -0.91, d = 0$; and the differenced water data, i.e. $\lambda = -0.91, d = 1$.

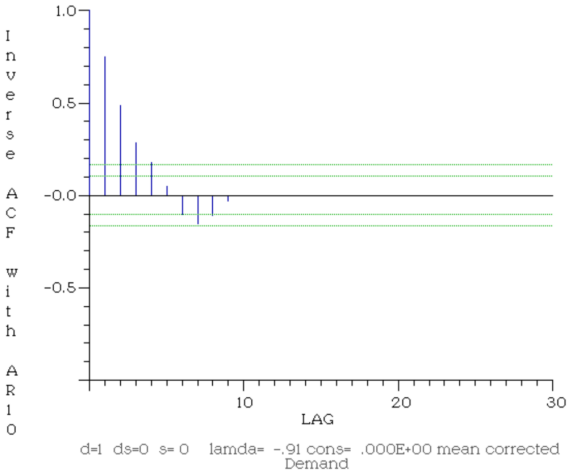
These plots can be seen depicted in Figure 3.6, for the sample demand data normalized depicted in Figure 3.5, with a $\lambda = -0.91, d = 1$. The IACF is simply the ACF with its p, q parameters switched and is used to verify the PACF; while the IPACF is the PACF with its p, q parameters switched and is used to verify the ACF. Table 3.2 summarizes the interpretations that can be made from these plots shown in Figure 3.6.



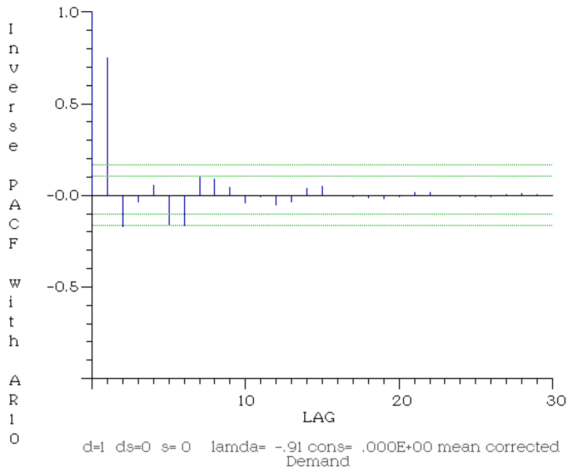
(a) Sample ACF.



(b) Sample PACF.



(c) Sample IACF.



(d) Sample IPACF.

Figure 3.6: Water demand, sample ACF, PACF, OACF, IPACF, $\lambda = -0.91, d = 1$.

Table 3.2: Exploratory data analysis - ACF, PACF, OACF, IPACF.

Plots	Observation	Interpretation
ACF	Truncates at lag-1(Attenuates steeply)	MA(1)(AR(1))
PACF	Attenuates until lag-5	MA(5)
IACF	Attenuates until lag-3	MA(3)
IPACF	Truncates at lag-1	MA(1)

Multiple interpretations regarding adequate models can be concluded from the plots in figure 3.6. While an initial interpretation can be seen in Table 3.2, the decision of which model is the optimal choice, taking into account number of parameters and well as accuracy of the model, is usually determined based on the minimum output Akaike Information Criterion (AIC), calculated as follows,

$$AIC = -2\ln(ML) + 2k, \tag{3.18}$$

in which $\ln(ML)$ is the maximum log likelihood function for the model fit to the given data set, and k is the number of model parameters. Since the optimal model will have the minimum AIC this decision criteria tends towards models with an increased fit, but penalizes the use of additional parameters, this is ideal in order to avoid simply over fitting the data with too many parameters in order to ensure better fit. Another option for decision criterion the number of parameters are selected based off of is the Bayes Information Criterion (BIC), explained later in Section 3.3.2.

After adequate exploratory data analysis has been completed, the general approach to time series modelling involves three primary steps: (1) model identification, (2) model parameter estimation, and (3) diagnostics check. Once the appropriate model is identified and the model parameters which best fit the data in question have been selected the adequacy of the selected model should be verified, typically by assessing the residual correlation function in order to ensure the residuals are white, and by reviewing statistical properties such as skewness and kurtosis to determine normality.

3.2.2 Fourier Treatment

Fourier analysis of signals is by far the most commonly used signal representation in signal analysis. The basic idea in this approach is to represent the signal in terms of periodic

bases, specifically trigonometric bases consisting of sines and cosines. One can view Fourier analyses as the outcome of correlating an underlying continuous function with a set of sines and cosines with varying frequencies. The results of this correlation are in terms of amplitudes, which can be seen as the degree to which individual sines and cosines are correlated to the given signal.

The Fourier series is a representation of a periodic function, $f(t)$, in terms of harmonically related sines and cosines, which when combined by a weighted summation will reproduce the data set. The signal is represented with respect to frequency rather than time. Fourier series make use of the orthogonality relationships of the sine and cosine functions, and can be presented within a periodic time interval T_o as follows,

$$f(t) = a_o + \sum_{n=0}^{\infty} \left(a_n \cos\left(2\pi t/T_o\right) + b_n \sin\left(2\pi t/T_o\right)\right), \quad t_1 \leq t \leq t_1 + T_o, \tag{3.19}$$

where T_o is the period of the signal $f(t)$,

$$\begin{aligned} a_o &= \frac{1}{T_o}\int_{t_1}^{t_1+T_o} f(t)dt \\ a_n &= \frac{2}{T_o}\int_{t_1}^{t_1+T_o} f(t)\cos 2\pi t/T_o dt, \quad n = 1, 2, \cdots \\ b_n &= \frac{2}{T_o}\int_{t_1}^{t_1+T_o} f(t)\sin 2\pi t/T_o dt, \quad n = 1, 2, \cdots \end{aligned}$$

in which a_0 represents the bias or the offset term.

Since sinusoids can also be represented by complex exponential functions, equation 3.19 can be described using complex exponential bases,

$$f(t) = \sum_{n=-\infty}^{\infty} D_n e^{j\omega_o n} \tag{3.20}$$

where $\omega_o = 2\pi/T_o$ and,

$$D_n = \frac{1}{T_o}\int_{-T_o/2}^{T_o/2} f(t)e^{-j\omega_o n}dt. \tag{3.21}$$

The Fourier transform is the extension of the Fourier series to a non-periodic function. In order to represent the non-periodic signal the limits of the integration in equation 3.21 become $\pm\infty$ as opposed to $(-T_o/2, T_o/2)$, and the Fourier transform in term becomes,

$$F(n\omega_o) = \int_{n=-\infty}^{\infty} f(t)e^{jn\omega_o t} \tag{3.22}$$

and,

$$D_n = \frac{1}{T_o}F(n\omega_o). \tag{3.23}$$

Discrete Fourier Transform

The discrete Fourier Transform (DFT) is a non-parametric frequency analysis technique (i.e. it does not require any *a priori* information about the signal), used to represent the original sequence in the frequency domain [Brandt, 2011b]. It is a method used to convert measured samples of a function into a same-length sequence of complex valued function of frequency.

Mathematically, it differs from a Fourier transform as it is computed from a finite number of samples [Brandt, 2011b]. The DFT treats the data as if it were periodic, i.e. $f(N)$ to $f(2N-1)$ is the same as $f(0)$ to $f(N-1)$. As such the DFT equation is evaluated for the fundamental frequency ($\frac{1}{NT}\ Hz$, $\frac{2\pi}{NT}\ rad/sec$) and its harmonics (not forgetting the D.C. offset of $\omega = 0$). The equation for a finite DFT is defined as [Brandt, 2011b],

$$F(n) = \sum_{k=0}^{N-1} f(k)e^{-i2\pi kn/N}, \tag{3.24}$$

where $f(k)$ is now a sampled version of the function consisting of a sequence of N complex numbers. This can be rewritten as the sum of the real and imaginary parts, respectively, as seen in equation (3.25), this ensures the result will only be nonzero if some frequency content exists in *x(n)* [Brandt, 2011b].

$$F(n) = \sum_{k=0}^{N-1} f(k)\cos\left(2\pi kn/N\right) - i\sum_{k=0}^{N-1} f(k)\sin\left(2\pi kn/N\right) \tag{3.25}$$

The sampling rate at which the function is sampled is $\delta f = \frac{1}{T_o}$, in which T_o is the sampled period in seconds. The sampling rate is chosen by first determining the highest

frequency of interest present within the signal, and sampling at at least twice that rate. This is referred to as the *Nyquist* frequency. The Nyquist theorem states that the signal needs to be sampled at twice the rate of the highest frequency of interest in order to accurately reproduce the signal that is being sampled. Typically oversampling a signal is suggested in order to improve resolution and signal-to-noise ratio, as well as assist in avoiding aliasing (and phase distortion when anti-aliasing filters are applied). A signal should be sufficiently over-sampled to guarantee an accurate measure of amplitude. A signal is said to be over-sampled by a factor of N if it is sampled at N times the Nyquist frequency. Undersampling leads to aliasing, which is the distortion that results when the signal reconstructed from samples is different from the original continuous signal. It is the effect that causes different signals to become indistinguishable, aliases of one another, when sampled. This can be seen depicted in Figure 3.7.

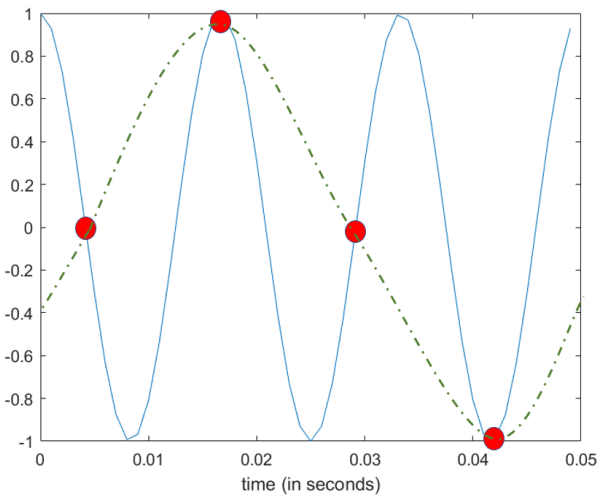


Figure 3.7: Aliasing- actual signal and aliased signal.

Several computationally efficient algorithms exist today to evaluate discrete Fourier transforms and readers are referred to standard texts [Oppenheim et al., 2008] for a complete treatment on the subject.

3.2.3 Filtering

In digital signal processing the process of filtering is used to either partially or completely suppress the presence of unwanted components in a signal. This most often means removing the relevant frequency content from the spectrum. It should be noted that while filters block

energy in the band of frequencies that they are designed for, they also attenuate signals in the pass band frequency range a swell. The most typical desired frequency response of the signal can be classified into four basic band forms describing which frequency bands the filter passes (the passband) and ones which it rejects (the stopband):

- **Low-pass filter**

  This filter passes signals with a frequency lower than a selected cutoff frequency, and attenuates signals with frequencies higher than the cutoff frequency. Digital low pass filters are often used to smooth data sets, removing the short-term fluctuations while maintaining the long-term trend. The ideal form of this filter completely eliminates all frequencies above the the cutoff frequency, while passing all those below unchanged, essentially lacking any transition region. However in practical implementation, because the sample time series is not infinite, this is not the case, and this ideal form is approximated and includes some roll-off. This filter is most commonly used to remove the effect of aliasing.

- **High-pass filter**

  The compliment of the low-pass filter, the high-pass filter passes signals with a frequency high than a selected cutoff frequency, and attenuates signals with frequencies lower than the cutoff frequency. This filter is most commonly used to remove DC noise.

- **Bandpass filter**

  When the high-pass and low-pass filters are used in conjuction they produce the bandpass filter. This filter passes frequencies within the two cutoff frequencies, and attenuated frequencies outside of this range. These filters are often used to isolate, and thus amplify, desirable frequency ranges.

- **Stopband filter**

  The compliment of the bandpass filter, the stopband filter passes most frequencies unaltered, but attenuates those within the two cutoff frequencies. A notch filter is the most extreme version of this, with a very narrow stopband, typically $1 - 2\ Hz$.

The design of these linear digital filters involves information on the desired filter response, as different filter families exist. Filter families can be selected based on maximal phase response, steepest cutoff, etc., including anti-aliasing filters which provide a tradeoff between bandwidth (freq range) and aliasing. Anti-aliasing filters are used before a signal

to restrict the bandwidth to satisfy the Nyquist theorem over the band of interest. The **Butterworth filter** is the most commonly used as it produces a maximal flat frequency response. More details about the filter and its design can be found in standard texts, e.g., [Porat, 1997, Tan and Jiang, 2018].

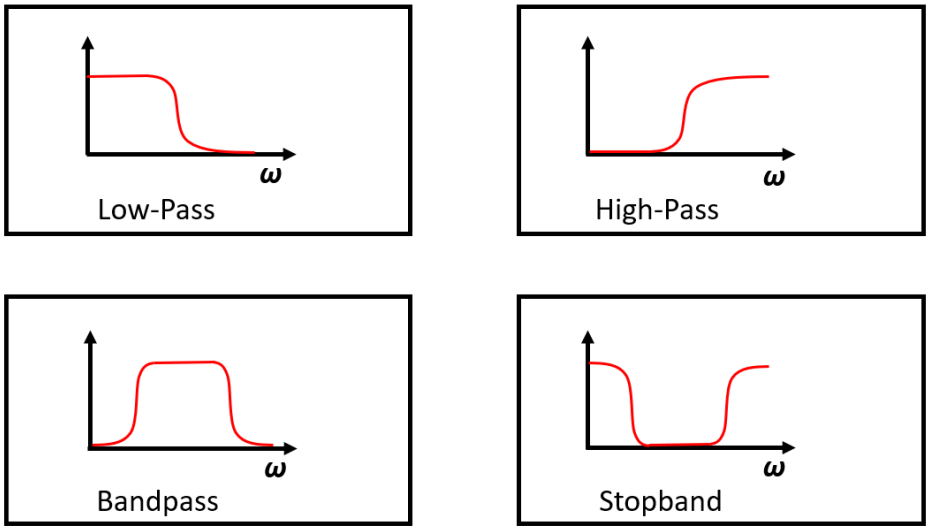
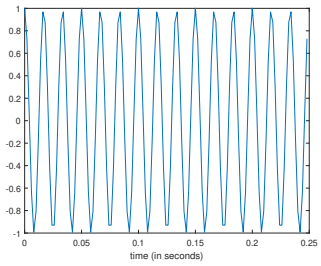


Figure 3.8: Schematic of DSP Butterworth filters.

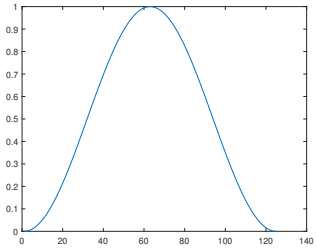
3.2.4 Windowing

Leakage occurs when the measured signal occurs at a frequency which does not lie on the frequency increment, i.e. the signal frequency falling between two spectral lines in the DFT, as well as any sinusoids with non-integer number of periods [Brandt, 2011b]. This causes the frequency content to be allocated to other spectral regions and such artificial discontinuities show up in the FFT as high-frequency components and can be much higher than the Nyquist frequency and are aliased between 0 and half the sampling rate. Hence, the resulting spectrum is not the actual spectrum of the original signal and appears as if energy at one frequency leaks into other frequencies. This phenomenon is known as spectral leakage. It is important to note that this leakage is an artifact introduced by the DFT operation, which computes the FT in discrete frequency bins.

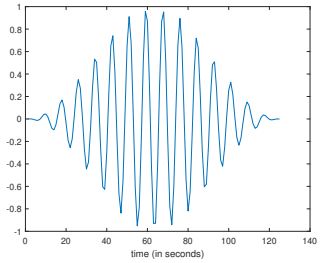
Windowing is a technique used to minimize spectral leakage. To understand how windowing works, it is easier to interpret its effect in the frequency domain. The multiplication



(a) Sample 60 Hz signal time series, f(x).



(b) Window function (W) —Hanning window.



(c) Windowed time series, W(f(x)).

Figure 3.9: Windowing applied to sample time series.

operation in the time domain is convolution in the frequency domain. The FT of a window resembles a sync function which, when convolved with a FT of a signal, accentuates the main frequency, while attenuating the leaked frequencies.

Windows could involve multiplying the time signal with rectangular or a non-rectangular weighting functions prior to digitally implementing the FT, which is called windowing. The effect of time-windowing is that it causes the start and the end of the finite time signal to fade towards zero [Brandt, 2011b]. The commonly used window for vibration signals is called a *Hanning window* [Brandt, 2011b], this is depicted in Figure 3.9. The Hanning window's Fourier transform has a main lobe that is wider than that of the rectangular window, which causes the energy in the leaked frequency regions to dissipate more [Wickramarachi, 2003].

For the Hanning window the data is weighted higher in the middle than at the ends following the function in equation (3.26):

$$w(n) = \frac{1}{2}(1 - \cos(2\pi n/N - 1)) \tag{3.26}$$

where, N is the number of samples in the window.

The Hamming window is often the preferred window function for speech processing (the traditional application of linear prediction which is used in this dissertation), as well as generally preferred in signal processing literature, due to its significant suppression of the first side lobe [Patel et al., 2013] and thus is less likely to cause individual peaks to be lost in the spectrum. The equation for the Hamming window is given by,

$$w(n) = 0.54 - 0.46\cos(2\pi n/N)) \tag{3.27}$$

in which the window length is $L = N + 1$. This can be seen visually depicted in Figure 3.10 (a).

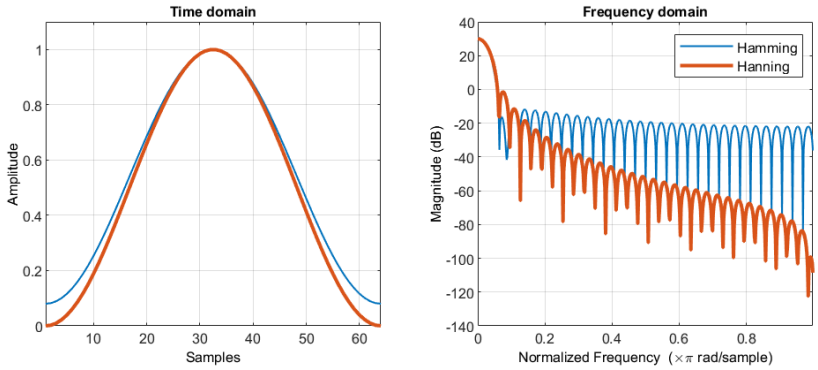


Figure 3.10: Hamming and Hann windowing result in a wide peak but nice low side lobes. Note the dip which occurs next to the main lobe in the Hamming window.

Hamming window is selected over the Hanning window since it is does not quite reach zero and have shown to perform better at cancelling the nearest side lobes —while the Hanning window performs better for cancelling other lobes [National Instruments]. The comparison between these two can be seen depicted in Figure 3.10. The trade off between the use of these windows as opposed to others is that while they yield better frequency resolution, they perform moderately, as compared with other window functions, with side lobes [National Instruments].

3.2.5 Spectrogram

A spectrogram [Cole et al., 1980] is a transform of a signal $x(t)$ to its point wise Fourier transform magnitudes. It can be computed by taking the Fourier spectrum of a short time window as a function of time shift, this is called the short time Fourier transform (STFT). It is useful to capture the changes in frequency with time [Randall, 2011], in other words the non-stationarity of the signal. It is mathematically described as,

$$S(f, \tau) = \int_{-\infty}^{\infty} x(t)w(t - \tau)e^{-i2\pi ft}dt, \tag{3.28}$$

where $w(t)$ is a window which is shifted along the signal. The spectrogram is the amplitude squared,

$$Spec = \mid S(f;\tau) \mid^2, \tag{3.29}$$

which produces a time-frequency image of the signal. In other words a one-dimensional signal is converted into a two-dimensional image. The length of the signal and window size are parameters of the spectrogram representation. If the window size is too short, the spectrogram will fail to capture relevant spectral information; conversely if it is too long, it loses temporal resolution. Hence, the right balance which is application-specific is necessary in its application.

3.2.6 Correlation

Autocorrelation

Time series are correlated by their very nature. However, this dependence will decay over time, and the extent of this decay is reflected in the autocorrelation coefficient. Autocorrelation, simply put, is the correlation of the time series considered, with a copy of itself, sliding along the *x-axis* (i.e. as a function of delay). The peak of the ACF will always occur at lag zero as this is when the signal is an exact replica of itself. It is typically applied to find repeating patterns within a time series signal, such as the presence of periodic signals (harmonics) which may be less easily detected due to the presence of noise. White noise is a term coined for a time series which is independent and identically distributed with a mean of zero.

The autocorrelation function [Stevens, 1950] for a discrete series X_t is defined as,

$$R_{XX}(\tau) = E[X_t\overline{X_{t+\tau}}], \tag{3.30}$$

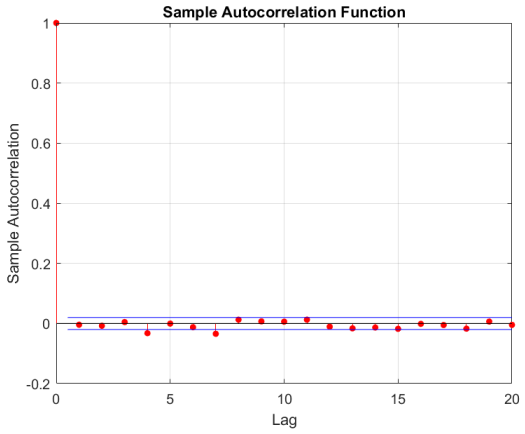
in which $E[\cdot]$ indicates the mathematical expectation, and τ is the lag. This can be expanded to,

$$\rho_{XX}(\tau) = \frac{1}{\sigma_X^2(T-1)} \sum_{t=1}^{T-\tau}[(X_t - \mu_X)(X_{t+\tau} - \mu_Y)] \tag{3.31}$$

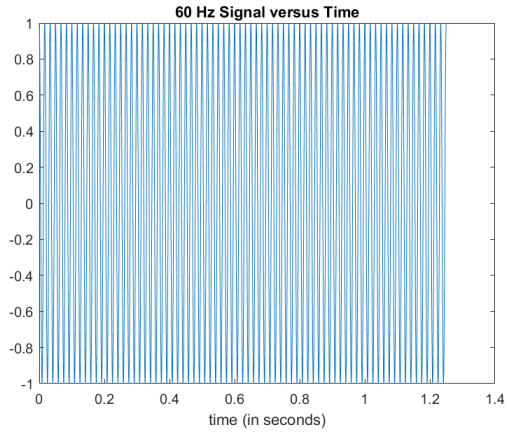
where, σ_X^2 is the sample variance used for normalization.

This can most simply demonstrated with the the ACF of white noise as compared to one with a harmonic present, demonstrated in Figure 3.11. The presence of repeating

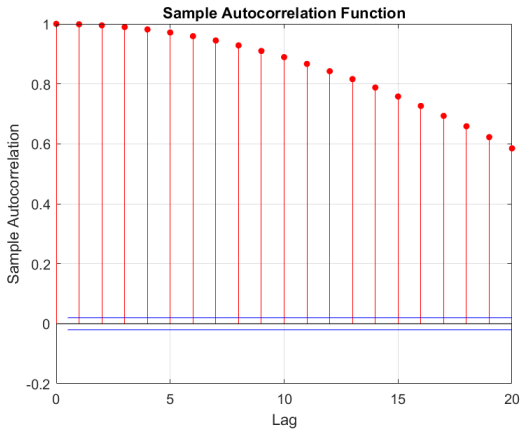
patterns within the time series becomes visible in the ACF as compared with that of the white noise.



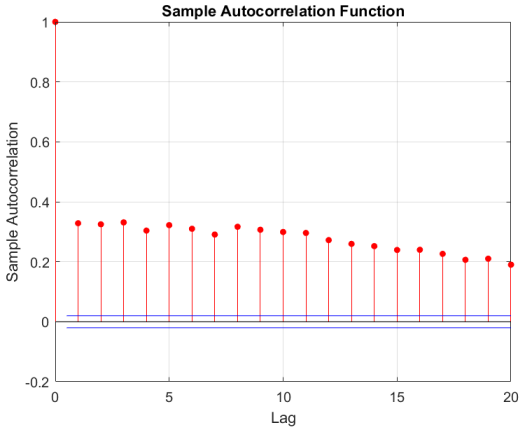
(a) ACF of uniformly distributed white noise.



(b) 60 Hz Signal time series.



(c) ACF of 60 Hz signal time series.



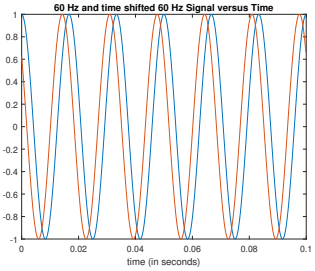
(d) ACF of 60 Hz signal time series with white noise.

Figure 3.11: Autocorrelations of white noise and 60Hz harmonic.

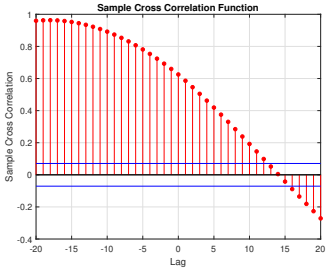
Crosscorrelation

The correlation between two series, X and Y, which represent measurements from two separate sensors recording the data from the same signal source, is referred to as cross correlation. It is the measure of similarity of two time series as a function of displacement of one relative to the other, also termed the sliding dot product.

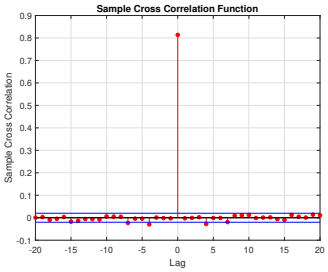
A sample cross correlation function (CCF) can be seen depicted in Figure 3.12. The CCF of Figure 3.12 (a), depicted in Figure 3.12 (b), essentially slides the time shifted 60 Hz signals (Y) along the *x-axis*, calculating the cumulative of the product at each point. When the functions overlap perfectly, this product is maximized. However when the signals are less similar and corrupted by noise, then the CCF, as depicted in Figure 3.12 (c), decreases rapidly.



(a) 60 Hz signal time series, and time shifted 60 Hz signal time series.



(b) CCF of 60 Hz signal time series with time shifted 60Hz signal time series.



(c) CCF of 60 Hz signal time series plus white noise, with 50 Hz signal time series plus white noise.

Figure 3.12: Sample crosscorrelation of two phase-shifted 60Hz sinusoids and with additive noise.

Mathematically, this relationship is represented as follows [Oppenheim et al., 2008],

$$\rho_{XY}(\tau) = \frac{1}{\sigma_X \sigma_Y} \sum_{t=1}^{T-\tau} [(X_t - \mu_X)(Y_{t+\tau} - \mu_Y)], \tag{3.32}$$

where μ_X and σ_X are the mean and standard deviation of the sensor X_t, and similarly for sensor data Y_t, and τ is the lag. τ is useful for determining the time delay between two signals, e.g. the time delays for the propagation of acoustic signals, which is central to the process of leak localization.

The cross correlation can be expressed using the convolution operator $*$ as,

$$\tau_{delay} = \arg\max_{t \in \mathbb{R}}((X * Y)(t)), \tag{3.33}$$

where, the maximum delay can be related to phase delay between two signals. For e.g., in Figure 3.12 the phase delay between the two 60 Hz signals can be seen as 18 sample time instants. In the context of leak localization, this time can be converted to physical distance (location of the leak) through distance-velocity-time relationship, with known speed of sound.

3.2.7 Singular Spectrum Analysis

Singular spectrum analysis (SSA) [Vautard and Ghil, 1989] is a non-parametric and adaptive method, able to decompose a signal into interpretable components without making any normality, linearity or stationarity assumptions. The analysis of a time-series signal using the spectrum of singular values of its *trajectory matrix* is called SSA . A trajectory matrix consists of a time series broken into equal sized segments (a set window length), each segment of which begins at a different lag. SSA is non-parametric, does not assume stationarity or linearity properties in the data and has only one parameter (called the embedding dimension) associated with its application [Golyandina, 2010].

SSA considers a finite length record of a time series $s = \{s_n, n = 1 \dots N\}$ and aims to decompose s as a sum of components, each having a meaningful interpretation. This task is completed through two successive steps detailed below.

Decomposition

Embedding This step consists of mapping the one-dimensional N-samples time series into a sequence of $K = N - L + 1$ lagged column vectors of length L, where L is called the *embedding dimension*. As a result, a trajectory matrix is obtained, expressed as:

$$X = \begin{pmatrix} s_1 & s_2 & \dots & s_K \\ s_2 & s_3 & \dots & s_{K+1} \\ \vdots & \vdots & \vdots & \vdots \\ s_L & s_{L+1} & \dots & s_N \end{pmatrix} \tag{3.34}$$

where each column is a sliding window of length L of data belonging to the time series. Thus, the trajectory matrix X is a Hankel matrix, meaning that it has equal elements

on the secondary diagonals. The only parameter in this step is the window length L, an integer ranging in the interval $[2,\, N-1]$. L should be carefully selected because it directly affects the decomposition. The optimal choice depends on the particularity of the time series and the problem statement.

Singular value decomposition (SVD) The SVD of X (being a real $L \times K$ matrix with rank $R \leq \min(L, K)$), expands this matrix into a sum of weighted orthogonal matrices that are not necessarily Hankel, expressed as:

$$X = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \sum_{i=1}^{R} X_i \quad \text{with } X_i = \sigma_i \boldsymbol{u_i} \boldsymbol{v_i}^T \tag{3.35}$$

where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \cdots, \sigma_R)$, σ_i are the singular values sorted in the descending order, u_i and v_i are, respectively, the associated left and right singular vectors corresponding to the columns of the orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$. The SVD expansion of X can be obtained through the eigendecomposition of the *lag-covariance matrix* $\boldsymbol{C} = XX^T$. This matrix can be factorized as $\boldsymbol{C} = \boldsymbol{U}\,\boldsymbol{\Lambda}\,\boldsymbol{U^T}$ where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. The i-th eigenvalue is equal to σ_i^2. The right singular vectors $\boldsymbol{V} = (\boldsymbol{v_1}, \boldsymbol{...}, \boldsymbol{v_R})$ can be deduced from X and $\boldsymbol{U}$ as $\boldsymbol{v_i} = \boldsymbol{X^T}\,\boldsymbol{u_i}/\sigma_i$. The energy contribution of the i-th eigentriple[1] to the trajectory matrix, given by the ratio $\sigma_i^2 / \sum_{j=1}^{R} \sigma_j^2$, is called the singular spectrum of the time series.

Reconstruction

Grouping This step consists of splitting the set of elementary matrices X_i $(i = 1, ..., R)$ into r disjoint groups and summing the matrices within each group. The result of this process is the expansion of the trajectory matrix X as $X = \sum_{k=1}^{r} X_{I_k}$, where $X_{I_k} = \sum_{i \in I_k} X_i$ is the resulting matrix of group I_k $(k = 1, ..., r)$.

Averaging If the signal components are separable, the resulting matrices after the grouping step are ideally Hankel . Thus, they correspond to the trajectory matrices of some time series. For real-world signals, this seldom happens, thus the resulting matrices X_{I_k} are almost Hankel and the components are approximately separable. The averaging along cross-diagonals of the matrix X_{I_k} aims at solving the problem of finding the time series $x^{(k)}$ for which the trajectory matrix of dimension $(L \times K)$ is the closest to X_{I_k}, in the

[1] The i-th eigentriple is defined by the collection $(\sigma_i, \boldsymbol{u_i}, \boldsymbol{v_i})$.

least-squares sense. In other words, the cross-diagonal averaging of $X_{I_k} = (x_{i,j})$ provides the elements of the time series $\{x_n^{(k)}, n = 1 \dots N\}$ as:

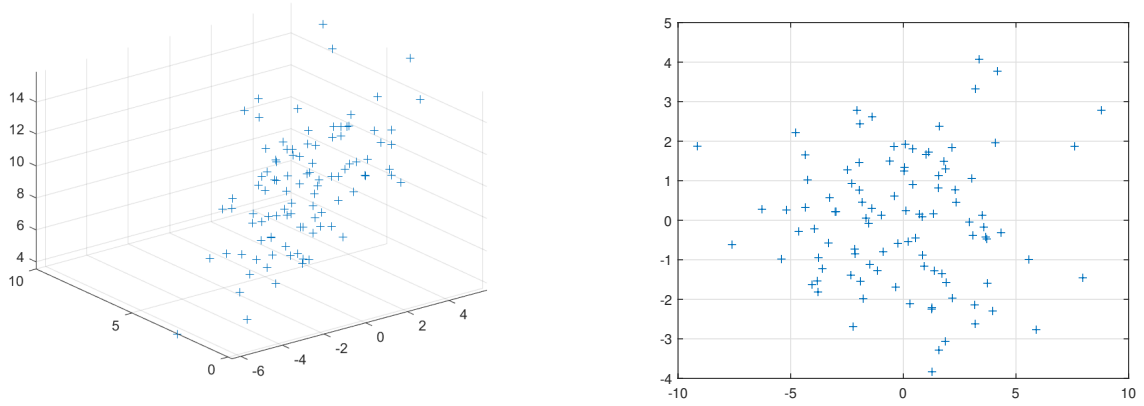
$$x_n^{(k)} = \begin{cases} \frac{1}{n} \sum_{m=1}^{n} x_{m,n-m+1} & \text{for } 1 \leq n < L \\ \frac{1}{L} \sum_{m=1}^{L} x_{m,n-m+1} & \text{for } L \leq n \leq K \\ \frac{1}{N-n+1} \sum_{m=n-K+1}^{L} x_{m,n-m+1} & \text{for } K+1 \leq n \leq N. \end{cases}$$

This cross-diagonal averaging, called Hankelization, can also be applied to each X_i matrix. The resulting time series are referred to as elementary components. This process finally provides an exact expansion of the time series s into L elementary components that satisfies $s_n = \sum_{k=1}^{L} x_n^{(k)}$. The application of SSA to hydro-acoustic signals obtained from the experimental test-bed is described later.

3.3 Statistical and Machine Learning Tools

3.3.1 Dimension Reduction

Discrete feature dimension reduction can be used as a method for feature selection, but it can also be applied for the purpose of feature space reduction for visualization. To represent features in a different coordinate system, dimensionality reduction is often implemented. It is a lower-dimensional representation with as much of the information content as possible about the original data set is preserved, a sample can be seen depicted in Figure 3.13. It involves learning a target function from data where some features are irrelevant. This is applied largely to cope with the *curse of dimensionality*, which encompasses all the problems which arise from working in a higher dimension, as opposed to working in a lower dimension. Fundamentally it states that as the number of features increases, so too must be number of samples and model complexity, as well as an increased possibility of overfitting.



(a) 3-dimensional representation of a random data set.

(b) PCA applied to reduce data set of 2-dimensions.

Figure 3.13: Sample dimension reduction.

This can be applied in both a supervised and unsupervised manner.

- Supervised methods include well-known methods such as neural networks (in which the number of hidden layers is often less than the number of inputs, thus performing dimensionality reduction in which each hidden layer is a logistic function of its inputs), partial least squares [Arenas-García et al., 2007], canonical correlation analysis [Hotelling, 1992], and Fisher's linear discriminant analysis (LDA) method (in which the ratio between between-class variance and within-class variance is maximized for the projected data). While LDA is widely applied and found to be highly effective for visualization purposes, this application focuses mainly on unsupervised cases since the classification is learned, and not known *a priori*.

- Unsupervised methods include the principal component analysis (PCA), factor analysis & principal factor analysis [McDonald, 1970], project pursuit [Safavi and Chang, 2008], multidimensional scaling [Kruskal, 1964], independent component analysis [Bell and Sejnowski, 1995], amongst others. Typically dimension reduction is only applied as a visual aid method for ease of understanding to the user interface portion of this system. However, it can also be applied as a means of feature projection, in order to keep the information contained within the high-dimensional data, while producing a more computationally efficient representation, i.e. lower-dimensional data set.

Principal Component Analysis

The PCA algorithm utilizes the eigenvalues of the covariance matrix [Bishop, 2006]. It can deal with noise very well by discarding lowest eigenvalues and accurately represents the variation in the data while minimizing computational cost. In the case of acoustic data of interest in this dissertation, the data is represented as a selected number of features, k, extracted from a chosen time interval.

PCA performs a linear mapping of any data to a lower-dimensional space by maximizing the variance of the data in the low-dimensional representation. For the implementation of this algorithm, the data sets are represented as an N by k matrix in which each column represents a different feature layer, and$N$ represents the number of time intervals.

The steps to perform the PCA analysis on the matrix A (N by p) is discussed below:

1. The mean of each column of A (μ_l) is subtracted from each value of its respective column, where $l = 1, \ldots, k$ and $i = 1, \ldots, N$:

$$A_{adj} = A_{i,l} - \mu_l \tag{3.36}$$

2. The covariance matrix can then be calculated as,

$$C_A = \frac{1}{N-1} A_{adj} * A_{adj}^T \tag{3.37}$$

   This matrix is symmetric and indicates the spread of the component values around the mean values.

3. By assuming distinct eigenvalues (λ) and finding the solution to the following characteristic equation,

$$|C_A - \lambda I| = 0, \tag{3.38}$$

   the eigenvalues and their corresponding eigen-vectors can be calculated.

4. The eigenvalues are then sorted and the first m ($< k$) points are taken to represent each instance, resulting in an N by m matrix, as opposed to the original N by k input matrix.

3.3.2 Gaussian Mixture models

Based on the central limit theorem, unimodal real world data is typically modeled as a Gaussian distribution. Thus the use of a multimodal Gaussian distribution to model a random variable of complex real-world data makes intuitive sense. GMMs are based on a parametric probability density function. This can be seen depicted in Figure 3.14. Generally when applied, it is used to cluster the feature represented data into k-groups, where k represents each possible state of the system. However, this can also be applied for anomaly detection, where k represents each possible *known* state of the system and the any new instance can therefore be classified as the $k+1$ state.

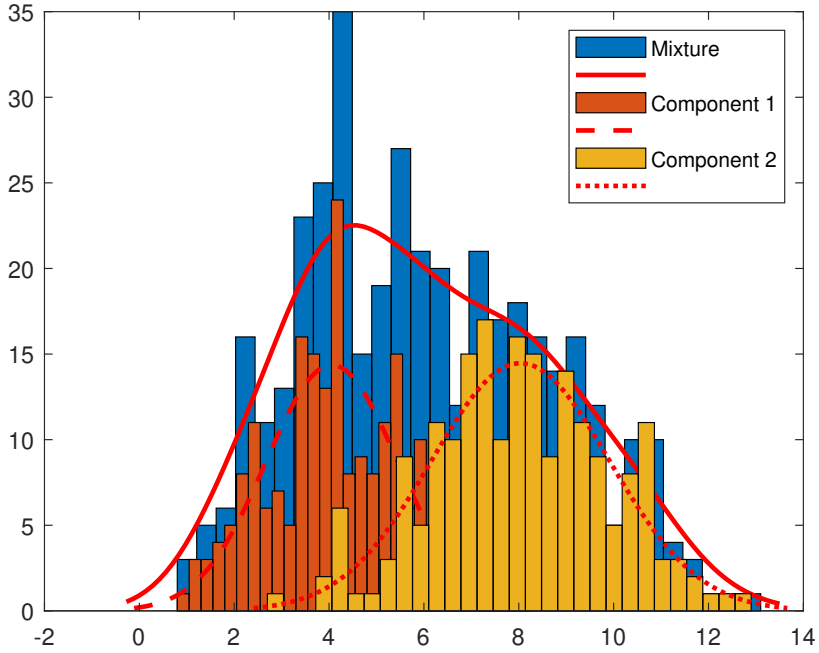


Figure 3.14: Mixture of 1D Gaussians.

The normality model for anomaly detection uses the normal instances of the system i.e., the baseline data set as a weighted sum of Gaussian component densities. For a GMM with K components, this is given by [Bishop, 2006]:

$$p(\mathbf{x}|w,\mu,\Sigma) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}|\mu_k,\Sigma_k), \tag{3.39}$$

where $\mathbf{x}$ is a D dimensional data vector corresponding to the feature set. w_k is the weights associated with for k^{th} mixture satisfying $\sum_{k=1}^{K} w_k = 1$. μ_k and Σ_k are respectively the mean vector and covariance matrix of a D-variate Gaussian density function, $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$, corresponding to k^{th} component of the GMM, which is given by,

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}, \tag{3.40}$$

in which for each component k, μ and Σ are initialized.

The objective is to estimate the GMM parameters, represented as $\Lambda = [\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}]$ with $\boldsymbol{w} = \{w_k, ..w_K\}$, $\boldsymbol{\mu} = \{\mu_k, ..\mu_K\}$ and $\boldsymbol{\Sigma} = \{\Sigma_k, ..\Sigma_K\}$, that best fit the input data set. The optimized set of the parameters are estimated employing the expectation maximization (EM) algorithm [Dempster et al., 1977], which is described below:

Suppose a set of D-dimensional observations $[x_1, \ldots, x_N]$ is used, where N denotes the number of observations. The data matrix $\mathbf{X}$ represents the $N \times D$ data set in which the n^{th} row is given by $\mathbf{x}_n^T$.

1. **Initialization:** Initialize μ_i, Σ_i and w_i using $\mu_i = \frac{\sum^N x_j}{N}$, $\Sigma = \frac{1}{N}\sum_j^N (x_j - \mu_i)(x_j - \mu_i)^T$ and $w_i = \frac{N_i}{N}$, in which N_i is the effective number of instances assigned to component i, and N in the total number of samples in the data set.

2. **Estimation step:** Compute the posterior probability using Bayes rule for each component of the GMM employing,

$$\gamma(z_{nk}) = \frac{w_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)}, \tag{3.41}$$

   in which z_n represents the latent vector corresponding to $\mathbf{x}_n$, for each component k.

3. **Maximization step:** Update the GMM parameters using equations 3.44 to 3.43:

$$\mu_k^* = \frac{\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n}{\sum_{n=1}^{N} \gamma(z_{nk})}, \tag{3.42}$$

$$\Sigma_k^* = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k^*)(\mathbf{x}_n - \mu_k^*)^T}{\sum_{n=1}^{N} \gamma(z_{nk})}, \tag{3.43}$$

$$w_k^* = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N}. \tag{3.44}$$

4. For each iteration (*) w_k, μ_k and Σ_k can be used to evaluate the log likelihood function,

$$\ln P(X|\Lambda) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} w_k \mathcal{N}(x_n|\mu_k, \Sigma_k). \tag{3.45}$$

5. Steps 2 to 4 are repeated until equation (3.45) converges, in which the log likelihood is maximized and result in the optimized set of parameters Λ.

Prior to the application of the EM algorithm, it is necessary to specify the configuration of the GMM as there are several variants on the model represented in equation 3.39. For instance, the covariance matrices, Σ_i can be full or constrained to be diagonal. A full covariance matrix is useful to capture any correlation present within the features for a certain application. However, large dataset and significant number of GMM components to build the model increase the computational complexity with full matrices. In such scenarios, the matrix can be constrained to be diagonal for minimizing the computational time. Hence the choice of full or diagonal matrices depends on the trade off between amount of computational complexity and number of GMM components. As the models described in this study use relatively few GMM components, a full covariance matrix is selected. In addition, GMM parameters can be independent or shared among the components. While independent parameters require individual covariance matrix for each component resulting in increased computational cost for very large dataset, the model can be simplified through shared parameters, i.e., having a single covariance matrix across all components. In the current analysis, GMM parameters are assumed to be independent, as it is expected that different components would be better represented by a different number of Gaussian distributions to better capture the characteristics of the signal. Nevertheless, the choice of these model configurations is often determined by the amount of data available for estimating the GMM parameters and how the GMM is employed in a particular application. For very large data sets with complex structure, the configuration can be simplified with a diagonal covariance matrix and shared GMM parameters to increase the computational efficiency of the model.

Model order selection criteria

The number of Gaussian components (K) is estimated through a sensitivity analysis, where several values of K are tested, and the optimal value is estimated based on model selection criteria. Commonly known selection criteria are Akaike Information Criterion (AIC)

[Akaike, 1974] (derived in Section 3.2.1) and Bayesian Information Criterion (BIC) [Abraham and Box, 1979]. The trade off between these two criteria is that BIC penalizes model complexity more heavily. Typically AIC is at risk of choosing larger model order despite the sample size, while BIC does not run this risk with larger data set [Kuha, 2004]. Moreover, if both positive and negative misclassifications are equally important for a particular application, BIC is chosen. But, AIC is a better option when negative misclassification is more misleading for the application type. Since the data sets used in the current study are sufficiently large and both types of misclassifications are expected to be avoided, BIC is deemed to be a better criterion.

In order to determine the number of mixtures, the BIC is utilized (also known as the Schwarz information criterion). It is derived using a Bayesian model comparison, calculating the posterior probabilities using the full information over the priors [Schwarz et al., 1978]. The evidence for a particular hypothesis is calculated using Laplace's method, as

$$p(\mathcal{D}|M) = \int p(\mathcal{D}|\theta, M)\pi(\theta|M)d\theta, \tag{3.46}$$

where θ are the parameters in the candidate model M, $\mathcal{D}$ represents the training data set, and $\pi(\theta|M)$ is the prior.

$$p(\mathcal{D}|M) = \int p(\mathcal{D}|\theta_{MAP}, M)\pi(\theta_{MAP}|M)\delta\theta, \tag{3.47}$$

where $\hat{\theta}$ are the optimal parameters that are assumed to maximize $\pi(\theta|M)$, $p(\mathcal{D}|\theta_{MAP}, M)$ is the best-fit likelihood, and $\pi(\theta_{MAP}|M)\delta\theta$ is the Occam factor. The BIC is defined as the log-likelihood function and a penalty term as a criterion for model-selection (the Occam factor).

If the assumption is made that the Gaussian prior distribution over parameters is broad, and the Hessian is full rank, [Bishop, 2006] then the BIC can be approximated as,

$$BIC(\mathcal{D}|M) = -2ln(L) + kln(n), \tag{3.48}$$

where L is the maximized value of the likelihood function of the model ($p(\mathcal{D}|\theta_{MAP}, M)$), n is the sample size, and $k = |\theta|$ is the number of parameters estimated by the model. The optimal value for the model order K is associated with the candidate model yielding the minimum value of BIC.

3.3.3 One-class Support Vector Machine

The one-class SVM (OCSVM) developed by Scholkopf et al. [2001] as an advancement to the original two class SVM, converts the traditionally fully-supervised SVM methodology to a semi-supervised classification methodology. The OCSVM training set only requires data from the baseline state(s) of the system and new instances are classified as known or unknown. Basically the baseline state(s) are considered the origin in feature space, and all other data points are separated from the origin using a hyperplane which maximizes the distances between the support vectors (i.e., the subset of points in the known and unknown cases which lie closest to each other). This results in a binary function, returning $+1$ for data which lie within the region recognized by the training data points, and -1 elsewhere.

There are a number of kernels which can be used in the OCSVM decision function. A number of these include, but are not limited to: linear, polynomial, sigmoidal and the Gaussian radial basis functions (RBF). The most popular of these, the RBF kernel, described in equation 3.49:

$$K(x, x') = exp(-\frac{||x - x'||^2}{2\gamma^2}) \tag{3.49}$$

where K is the RBF kernel function, $\gamma \in \mathbb{R}$ is a kernel parameter and $||x - x'||$ is the dissimilarity measure between the training data and the new data.

γ and ν can be tuned to increase accuracy. The γ variable defines the influence of a single training point and its default value is typically set to a value equal to the inverse of the number of features. The ν parameter is found through optimization as described in equation 3.50 as follows:

$$\min_{\omega,\xi,\rho} \frac{1}{2}||\omega||^2 + \frac{1}{\rho N}\sum_i \xi_i - \rho \tag{3.50}$$

subject to $(\omega \cdot \Phi(x_i)) \geq \rho - \xi_i, \xi_i \geq 0$. ν represents an upper bound of the fraction of training errors and a lower bound of the fraction of support vectors. Its default value is typically chosen as 0.5.

3.3.4 Neural Network

Neural networks are approximation functions, often used for binary and multi-class classification Hessel et al. [1999]. In order to perform the classification, the neural network acts

as a data structure storing models of the classes it has been trained to recognize. Certain neural network based algorithms can be used for anomaly detection, identifying input samples as falling outside the scope of the class models it is storing Torok et al. [2013].

Convolutional Neural Network

A convolutional neural network (CNN) is a class of deep, feed-forward ANNs [Le Cun et al., 1990, Yegnanarayana, 2009]. CNNs [Skansi, 2018] consist of an input and an output, with multiple hidden layers. The hidden layers of the CNN typically include a combination of convolutional layers and pooling layers.

During the forward pass, the convolutional layer applies a set of filters (or kernels) to the input in a sliding-window manner, generating a value for each filter at each window location. These values are aggregated in the next layer producing a 2-dimensional activation map for every filter. These activation maps can then be stacked to form the full output volume of the convolution layer. The convolution of each window is meant to emulate the receptive field of an individual neuron to visual stimuli.

Convolution networks usually include max pooling layers which condense a neighbourhood of neurons at one layer into a single neuron in the next layer, summarizing the neighbourhood by its maximum value. This leads to the output layer being smaller than the input layer.

Variational Autoencoder

Autoencoders are a dimensionality reduction tool often used to find efficient data representations [Hinton and Salakhutdinov, 2006]. Spectral anomaly detection techniques try to find a lower dimensional embedding, coined *latent variables*, of the original data set, where anomalies and normal data are expected to separate. These latent variables can be brought back to their original space via *reconstruction*. With this *encoding* and *decoding* it is expected that the network will not be good at reconstructing features outside the statistical gamut of its training set. The difference between the original data and the reconstructed data is referred to as reconstruction error, and can be used as an anomaly score to detect outliers. The encoding and decoding in autoencoders is done in a deterministic way such that the original data map to a single value for each latent variable.

A variational autoencoder (VAE) [Kingma and Welling, 2014] is an autoencoder in which the distributions of each latent variable are approximately unit-variance Gaussian distributions. As such, the original data now maps to a probability distribution for each

latent variable. The encoder process of a VAE depends on a NN and such as a CNN. The loss function in this case is the average of the mean squared error which measures how accurately the network reconstructs the original data, and a latent loss, the Kullback-Leibler (KL) divergence, which measures how closely the latent variables match a unit Gaussian.

3.3.5 Hypothesis testing

Hypothesis tests can be broadly classified as: statistical hypothesis tests, and statistical model hypothesis tests [Peruggia, 2003, Lumley, 2000]. The statistical model based hypothesis test proposes alternative candidate hypothesis which uses a model selection method to choose the appropriate model. The more common of the two methods however, is the statistical hypothesis test, which is a method of statistical inference. It is applied most commonly when two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, as an alternative to an idealized null hypothesis. The comparison is deemed statistically significant if the null hypothesis is proven to be untrue according to a threshold probability (the significance level), i.e. the null hypothesis can be rejected at a determined significance level. Based on the statistical assumptions made about the sample, a number of different tests are available and an appropriate one must be selected. As well the significance level, i.e. threshold under which the null hypothesis will be rejected, must be selected. Tests can either be one-tailed or two-tailed depending on where the region of rejection lies.

There are two conceptual types of errors which are worth considering from a hypothesis test:

- **Type I error** occurs when the null hypothesis is wrongly rejected. The probability of committing a Type I error is called the significance level, often denoted as α.
- **Type II error** occurs when the null hypothesis is wrongly *not* rejected, often denoted as β. The probability of not committing a Type II error is referred to as the Power of the test.

T-test

The one-sample **t-test** is used to compare central values of two independent groups of data, and is most commonly applied when the test statistic would follow a normal distribution.

This test is used for small sample sizes ($n < 30$). In order to test the null hypothesis the statistic t score is calculated as,

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \tag{3.51}$$

where $\bar{x}$ is the sample mean, μ_0 is the population mean, s is the sample standard deviation, and n is the sample size. The degrees of freedom used are $n-1$. While μ_0 need not follow a normal distribution, it is assumed that distribution of $\bar{x}$ is normal.

The standard error of the mean is calculated as $SEM = \frac{s}{\sqrt{n}}$.

Z-test

In contrast to the **t-test**, a **z-test** is used for large sample sizes ($n > 30$), in which the distribution of the test statistic can be approximated by a normal distribution, and for which the population variance is known. In order to test the null hypothesis the statistic z score is calculated as,

$$z = \frac{\bar{x} - \mu_0}{SEM}. \tag{3.52}$$

Wilcoxon Rank-Sum test

The **Wilcoxon Rank-Sum test** is a non-parametric alternative to the two-sample **t-test**, used to compare two independent groups of data. The **rank-sum test** does not make any underlying assumptions about the nature of the two cases, and tests for whether one group tends to produce larger or smaller observations than the second group [Helsel and Hirsch, 1997]. This is in contrast with the use of the **t-test** and **z-test** which are applied based on the assumption that the true distribution of the baseline data is known *a-priori* and are sufficiently large that a normal distribution is observed.

This test is based on ranking the observations of the combined samples, the joint ranks R_k. The test statistic, W_{rs}, is calculated as the sum of the ranks of the group having the smaller sample size [Helsel and Hirsch, 1997]. The null hypothesis,

$$H_0 : Prob[\mu_{x_s} > \mu_{x_p}] = 0.5, \tag{3.53}$$

stated that if μ_{x_s} (of size n) is from μ_{x_p} (of size m), about half of the time, an observation from either group should be expected to be higher than that from the other, so the null hypothesis applies [Helsel and Hirsch, 1997]. While the alternative hypothesis,

$$H_1 : Prob[\mu_{x_s} > \mu_{x_p}] \neq 0.5, \tag{3.54}$$

is a 2-sided test in which μ_{x_s} might be larger or smaller than μ_{x_p}, in which H_0 is rejected if $W_{rs} \leq x*_{\alpha/2,n,m}$ or $W_{rs} \geq x_{\alpha/2,n,m}$, otherwise H_0 is not rejected. When the null hypothesis is rejected, it must be assumed that the alternative hypothesis is true, the two groups differ only in their central values [Helsel and Hirsch, 1997].

Two versions of the one-sided test are as follows:

- The smaller data set has higher values than the larger data set the majority of the time, making the alternate hypothesis $H_1 : Prob[\mu_{x_s} > \mu_{x_p}] > 0.5$, in which H_0 is rejected if $W_{rs} \geq x_{\alpha,n,m}$.
- The smaller data set has lower values than the larger data set the majority of the time, making the alternate hypothesis $H_1 : Prob[\mu_{x_s} > \mu_{x_p}] < 0.5$, in which H_0 is rejected if $W_{rs} \leq x*_{\alpha,n,m}$.

3.3.6 Bhattacharya Distance

The Bhattacharya distance is a statistical measure that has been used as a class separability metric for feature selection [Choi and Lee, 2003]. As well, the coefficient can be extracted to determine the separability of two samples being considered [Khalid et al., 2006].

The Bhattacharya distance is defined as the negative natural logarithm of the Bhattacharya coefficient. For discrete probability distributions p and g this is given by,

$$\mathbf{B_d} = -\ln\left(\rho(p,q)\right), \tag{3.55}$$

where $0 < \mathbf{B_d} < \infty$. The higher the value of $\mathbf{B_d}$, the more dissimilar the distributions and the more separate the corresponding classes. This measure is computed between the histograms of two case sets of feature values in order to quantify the histograms' similarity.

The Bhattacharya coefficient [Bhattacharyya, 1943] is then given by,

$$\rho(p,q) = \left(\sum_{x=1}^{N} \sqrt{p(x)q(x)}\right). \tag{3.56}$$

The Bhattacharya coefficient was originally interpreted geometrically [Derpanis, 2008], as the cosine of the angle between the N-dimensional vectors p and q, i.e., $\rho(p,q) = \cos\theta$. Thus if two populations are identical [Kashyap, 2019],

$$\cos\theta = \sum_{x=1}^{N} \sqrt{p(x)q(x)} = \sum_{x=1}^{N} \sqrt{p(x)p(x)} = \sum_{x=1}^{N} p(x) = 1, \tag{3.57}$$

corresponding to $\theta = 0$.

The value of the limiting case,

$$\rho(p,q) = \sum_{x=1}^{N} q(x)\sqrt{\frac{p(x)}{q(x)}}, \tag{3.58}$$

since $f(x) = \sqrt{x}$ is a concave function, $\sum_{x=1}^{N} q(x) = 1 \Rightarrow q(x) \geq 0$, and therefore based on *Jensen's inequality* [Cover and Thomas, 2012],

$$\rho(p,q) \leq \sqrt{\sum_{x=1}^{N} q(x)\frac{p(x)}{q(x)}}, \tag{3.59}$$

which simplifies to,

$$\rho(p,q) \leq \sqrt{\sum_{x=1}^{N} p(x)}, \tag{3.60}$$

and since $\sum_{x=1}^{N} p(x)$ by construction, $\rho(p,q) \leq 1$. Thus the coefficient lies between 0 and 1 (hence $0 < \mathbf{B_d} < \infty$, since $\ln(0) = -\infty$).

Jensen's inequality

Jensen's inequality generally states that the average value of a convex function is greater or equal to the function of the average. Therefore if f is a concave function, and $\sum_{i=1}^{N} p(i) = 1 \Rightarrow p(i) \geq 0$,

$$\sum_{i=1}^{N} p(i)f(x(i)) \leq f\left(\sum_{i=1}^{N} p(i)x(i)\right). \tag{3.61}$$

3.3.7 KL-divergence

The Kullback–Leibler (KL) divergence [Kullback and Leibler, 1951] is a measure of how one probability distribution is different from a reference probability distribution. KL-divergence has its origins in information theory, in which the higher the probability of an event, the lower its information content. This is the same as saying information is inversely related to the probability of an event. Thus, since $\log p(x)$ is directly related to $p(x)$, it follows that $-\log p(x)$ is inversely related to $p(x)$. Therefore the information content of event x with respect to p can be expressed as,

$$I_p(x) = -\log p(x). \tag{3.62}$$

The value of information if defined as the expected utility of the best action chosen with the new information *minus* the expected utility of the best action chosen without the new information. Therefore the difference of information between $q(x)$ and $p(x)$ is,

$$\Delta I = I_p - I_q = -\log p(x)) + \log q(x) = \log(\frac{q(x)}{p(x)}). \tag{3.63}$$

KL-divergence is the expectation of the difference shown in equation 3.63,

$$E_{\ q}[\Delta I] = \int (\Delta I) q(x) dx = \int q(x) \log(\frac{q(x)}{p(x)}) dx. \tag{3.64}$$

The KL-divergence from P to Q is generally calculated as $D_{KL}[Q(z|X)||P(z|X)]$, where $Q(z|X)$ discrete probability distribution of the projected data X into the latent variable space, and $P(z|X)$ is the true distribution [Kullback, 1997].

$$\begin{aligned} D_{KL}[Q(z|X)||P(z|X)] &= \sum_z Q(z|X) \log \frac{Q(z|X)}{P(z|X)} \\ &= E[\log \frac{Q(z|X)}{P(z|X)}] \\ &= E[\log Q(z|X) - \log P(z|X)]. \end{aligned} \tag{3.65}$$

By Bayes Theorem $P(z|X) = \frac{P(X|z)P(z)}{P(X)}$, therefore equation 3.65 can be expressed as,

$$
\begin{aligned}
D_{KL}[Q(z|X)||P(z|X)] &= E[\log Q(z|X) - \log \frac{P(X|z)P(z)}{P(X)}] \\
&= E[\log Q(z|X) - (\log P(X|z) + \log P(z) - \log P(X))] \\
&= E[\log Q(z|X) - \log P(X|z) - \log P(z) + \log P(X)].
\end{aligned} \tag{3.66}
$$

3.3.8 Performance Measures

In general, the performance of anomaly detection algorithms can be assessed using the following performance measures, which are described next [Bishop, 2006].

- **Accuracy:** It represents the ratio of correctly predicted observations with respect to the total number of observations and can be calculated as:

$$
A_c = \frac{TP + TN}{TP + FP + FN + TN}, \tag{3.67}
$$

  where TP, TN, FP and FN are the true positive, true negative, false positive, and false negative rates, respectively. In the current study, positive and negative are associated with, respectively, the leak free and leak states of the system.

- **Precision:** It represents the ratio of correctly predicted positive or normal observations with respect to the total predicted positive observations. A value closer to 1 is ideal, representing a low false positive rate (FP). It can be calculated as,

$$
P_c = \frac{TP}{TP + FP}. \tag{3.68}
$$

- **Recall:** Also known as the true positive rate (TPR), is estimated as the ratio of correctly predicted positive observations with respect to all positive instances. A value closer to 1 is ideal leading to fewer false alarms. It is calculated using,

$$
R_c = \frac{TP}{TP + FN}. \tag{3.69}
$$

- **F1-score:** It represents the weighted average of precision and recall and is a version of accuracy. It is best used when the number of normal to anomaly instances are not even. F1-score is estimated as,

$$
F_s = \frac{2 * R * P}{R + P}. \tag{3.70}
$$

- **ROC graph and AUC:** The receiver operating characteristic (ROC) graph represents the probability of correctly predicted anomaly event (also called sensitivity) versus the probability of false positive alarms. The area under the curve (AUC) is an indicator of accuracy of the anomaly detection methodology [Güvenir and Kurtcephe, 2013]. AUC values closer to 1 represents an effective predictor, while values closer to 0.5 points towards worthless, or random, predictors.

3.3.9 Time Domain Statistical Features

Some basic time domain statistical features which can be used as traditionally employed features when first performing a technical analysis of time series data. These features can help in reducing the dimension of signals and to gain a more compact representation of the structure and information contained within the data. For a time series represented by x_i Some of these features include [Li et al., 2017],

- **Peak:** The peak can be calculated as,
  $x_{mx} = max(|x_i|)$.
- **Mean:** The mean can be calculated as,
  $x_{me} = \frac{\sum x_i}{n}$.
- **Standard deviation:** The standard deviation can be calculated as,
  $x_{sd} = (\frac{1}{n-1} \sum_{i=1}^{n} (x_i - x_{me})^2)^{1/2}$.
- **R-mean-square:** The RMS can be calculated as,
  $x_{rm} = (\frac{1}{n} \sum_{i=1}^{n} x_i^2)^{1/2}$.
- **Crest factor:** The crest factor refers to the ratio of peak values to the effective value. It can be calculated as,
  $x_{cf} = \frac{x_{mx}}{x_{rm}}$.
- **Energy:** The energy, often referred to as entropy, is used to describe the randomness in the system. It can be calculated as,
  $x_{se} = -\sum p_i * \log_2(p_i)$.

Chapter 4

Details of linear prediction (LP)

This chapter provides the theoretical details of linear prediction, which is central to the approach taken in this dissertation. The context and application of the theory of linear prediction to leaks is described, where the sensitivity of its coefficients to variation is reviewed. Although linear prediction is a well established concept, especially in speech recognition, its relevance to water pipes and leak inducted signals has not been established. LP is presented as a method which can be used to detect and locate small leaks in pressurized water pipes using the cepstral version of LP model coefficients [Ai et al., 2006]. However, unlike speech applications, the short-time spectral information alone is insufficient for both leak detection and localization. This chapter starts with a general overview of the underlying concepts of LP, followed by specific application aspects of the LP principles for the problem of leak detection.

4.1 Linear Prediction

LP has been extensively used to extract the spectral envelope from signals in applications related to speech coding, speech synthesis, speech recognition, speaker recognition and verification, and for speech storage [Fujisaki and Sato, 1973] —LP embodies strong theoretical underpinnings in the field of linear dynamic systems [Makhoul, 1975]. In one of the most interesting and extensively studied applications, the process of generating voiced and unvoiced sounds in the vocal tract has been modeled using LP [Rabiner et al., 2007].

Fundamentally, linear models describe a response variable as a function of predictor variables, as depicted in Figure 4.1. The underlying principle of LP is that the formative

(resonances) response of a linear system can be captured through modeling measurement data and such model parameters contain pertinent information regarding the system properties. This simple central idea is developed further as explained in this chapter.

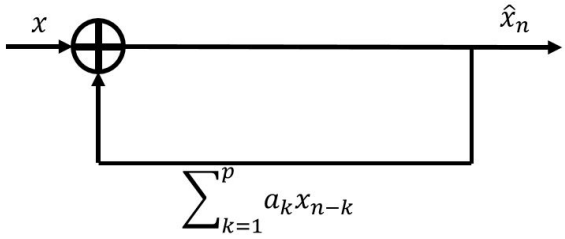


Figure 4.1: Time varying linear predictor p.

The basic idea of LP can be described mathematically as follows [Makhoul, 1975],

$$x(n) = \sum_{k=1}^{p} \alpha_k x(n-k) + G \sum_{l=0}^{q} b_l u(n-l), \tag{4.1}$$

where α_k, b_l and G are all parameters of a hypothesized system; all of which generate a linear combination of past outputs, and present and past inputs which produce a prediction $x(n)$. In this equation, $b_0 = 1$, G is the gain factor. The same mathematical expression can be expressed in the frequency domain by taking $H(z)$ to represent the z-transform (as described in Section 3.2.2), and transforming both sides of equation 4.1,

$$H(z) = \frac{X(z)}{U(z)} = \frac{\sum_{n=-\inf}^{\inf} x(n) z^{-n}}{U(z)} = G \frac{1 + \sum_{l=1}^{q} b_l z^{-l}}{\sum_{k=1}^{p} \alpha_k z^{-k}}, \tag{4.2}$$

where equation 4.2 is the general pole-zero model, or the ARMA model. Two special cases of this model exist: the first of which is the all-zero (MA) model in which $\alpha_k = 0, 1 \leq k \leq p$; the second being the all-pole (AR) model in which $b_l = 0, 1 \leq l \leq q$.

LP models for stationary and ergodic time series processes are also referred to as AR models. This holds approximately true for many applications where only short duration of data are considered, which can individually be assumed to be quasi-stationary. If a system is adequately modeled, this model can capture the underlying eigenstructure of a linear time invariant system. With application to complex systems, Von Storch [1999] discusses an AR(1) model's ability to provide information both of the first order approximation of the second moments and of the linear dynamics using their eigen decomposition for

analysis. They demonstrate that the full spectral and spatial features of a complex system can be described by principal oscillatory patterns and their eigenvalues. This work was continued by Neumaier and Schneider [2001] who further demonstrate the effective use of eigen decomposition for structural analysis of an AR(p) model.

In the problem of probing the eigen structure of the signal, the spectral envelope is of primary interest, i.e. the smoothed envelope curve of the amplitude spectrum preserving the broad structure while not sensitive to the minor local variations. Hence, for compuational simplicity, the discrete-time linear system is described by an all-pole model [Makhoul, 1975]:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}, \tag{4.3}$$

with α_k being the k^{th} coefficient of the p^{th} order linear model and G being the gain factor as described earlier. Here, $H(z)$ represents the Z-transform of the impulse function of the system (also called as transfer function) and z is, in general, a complex number from the z-domain.

4.1.1 Parameters estimation

For the system model in equation 4.3, the signal $x(n)$ is a linear difference model, which is a combination of past samples of the signal ($x(n-k)$) and the excitation $u(n)$, and is given in the form of:

$$x(n) = \sum_{k=1}^{p} \alpha_k x(n-k) + Gu(n), \tag{4.4}$$

The prediction of $x(n)$ with the predictor coefficients a_k, denoted by $\hat{x}(n)$, is given by:

$$\hat{x}(n) = \sum_{k=1}^{p} a_k x(n-k). \tag{4.5}$$

The prediction error, $e(n)$ then becomes:

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=1}^{p} a_k x(n-k). \tag{4.6}$$

The predictor coefficients (a_k) are calculated by minimizing the total squared error (E) estimated using:

$$E = \sum_{n} e_n^2 = \sum_{n} \left[x(n) - \sum_{k=1}^{p} a_k x(n-k) \right]^2, \tag{4.7}$$

in which E is minimized by setting,

$$\frac{\partial E}{\partial a_k} = 0, \quad 1 \leq k \leq p, \tag{4.8}$$

and resulting in:

$$\sum_{k=1}^{p} a_k \sum_{n} x(n-k)x(n-i) = \sum_{n} x(n)x(n-i), \quad 1 \leq i \leq p. \tag{4.9}$$

Equation 4.9 is a set of p equations, which can be solved for p unknowns (a_k, $1 \leq k \leq p$) leading to a minimum E in equation 4.7. The minimum mean squared error can be expressed as,

$$E_n = \sum_{n} (x(n))^2 - \sum_{k=1}^{p} a_k \sum_{n} x(n)x(n-k). \tag{4.10}$$

The parameters a_k are estimated through one of two methods [Makhoul, 1975]:

- The *autocorrelation method*

  The parameters a_k are estimated through the *autocorrelation method* [Makhoul, 1975] by assuming that the error in equation 4.7 is minimized over an infinite duration $-\infty < n < \infty$ so that equation 4.9 and 4.10 reduce, respectively to:

  $$\sum_{k=1}^{p} R(i-k)a_k = R(i), \quad 1 \leq i \leq p, \tag{4.11}$$

  and,

  $$E_n = R(0) - \sum_{k=1}^{p} a_k R(k), \tag{4.12}$$

  where,

  $$R(i) = \sum_{n=-\infty}^{\infty} x(n)x(n-i). \tag{4.13}$$

  In equation 4.11, the auto-correlation matrix is a symmetric Toeplitz matrix, which can be solved efficiently through the Levinson-Durbin algorithm [Durbin, 1960].

Levinson-Durbin algorithm

For a matrix that is a symmetrix positive definite Toeplitz matrix, i.e. all the elements on a given diagonal in the matrix are equal, it can be solved effectively using the Levinson-Durbin algorithm [Levinson, 1946, Durbin, 1960]. To solve for the prediction coefficients $a = [a_1, a_2, \ldots, a_p]^T$, equation 4.11 can be expanded in matrix form as,

$$\begin{bmatrix} R_0 & R_1 & \cdots & R_{p-1} \\ R_1 & R_0 & \cdots & R_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & \cdots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_p \end{bmatrix}$$

which can be expanded to a system of equations (in which the order is denoted by a superscript), from which the last expression can be subtracted resulting in the following for the first $p-1$ rows,

$$\begin{aligned} R_0 a_1^{(p)} + R_1 a_2^{(p)} + \cdots + R_{p-2} a_{p-1}^{(p)} &= R_1 - R_{p-1} a_p^{(p)} \\ &\vdots \\ R_{p-2} a_1^{(p)} + R_{p-3} a_2^{(p)} + \cdots + R_0 a_{p-1}^{(p)} &= R_{p-1} - R_1 a_p^{(p)}, \end{aligned} \tag{4.14}$$

and an equation containing the last row,

$$R_{p-1} a_1^{(p)} + R_{p-2} a_2^{(p)} + \cdots + R_1 a_{p-1}^{(p)} = R_p - R_0 a_p^{(p)}. \tag{4.15}$$

By defining $\tilde{R} = [R(p), R(p-1), \cdots, R(1)]^T$, and $\tilde{a}^{(p)} = [a_p^{(p)}, a_{p-1}^{(p)}, \cdots, a_1^{(p)}]^T$, and exploiting the symmetry in equation 4.11, equations 4.14 can be rewritten as,

$$\begin{bmatrix} a_1^{(p)} \\ a_2^{(p)} \\ \vdots \\ a_{p-1}^{(p)} \end{bmatrix} = a^{p-1} - a_p^p \tilde{a}^{(p-1)}. \tag{4.16}$$

This can be written to calculate for the single coefficients as,

$$a_i^{(p)} = a_i^{(p-1)} - a_p^{(p)} a_{p-i}^{(p-1)}, \quad i = 1, \cdots, p-1. \tag{4.17}$$

The one missing coefficient $a_p^{(p)}$, referred to as the partial correlation coefficient (PARCOR), can be calculated using equations 4.16 and 4.15. By isolating for $a_p^{(p)}$ the

PARCOR can be calculated as,

$$a_p^{(p)} = \frac{R_p - (\tilde{R}^{(p-1)})^T a^{(p-1)}}{R_0 - (\tilde{R}^{(p-1)})^T \tilde{a}^{(p-1)}}. \tag{4.18}$$

- The *covariance method*

  The parameters a_k are estimated through the *covariance method* [Makhoul, 1975] by assuming that the error in equation 4.7 is minimized over a finite interval $0 \leq n \leq N-1$ so that equation 4.9 and 4.10 reduce, respectively to:

$$\sum_{k=1}^{p} a_k \varphi_{ki} = \varphi_{0i}, \quad 1 \leq i \leq p, \tag{4.19}$$

  and,

$$E_n = \varphi_{00} - \sum_{k=1}^{p} a_k \varphi_{0k}, \tag{4.20}$$

  where the covariance of the signal $x(n)$ in the given intervals is,

$$\varphi_{ki} = \sum_{n=0}^{N-1} x(n-i)x(n-k). \tag{4.21}$$

  Since the covariance matrix is a symmetric positive-semidefinite matrix, it can be solved efficiently through the Cholesky decomposition (as described in Appendix B) of the covariance matrix [Benoıt, 1924, Higham, 2009].

  If a matrix is a Hermitian —a complex square matrix that is equal to its own conjugate transpose —positive-definite, it can be effectively decomposed into the product of a lower triangular matrix, and its conjugate transpose, using the Cholesky decomposition [Parker, 2017]. However, the covariance matrix is Hermitian, positive semi-definite and thus the diagonal entries of the triangular matrix are not equal and allowed to be zero [Parker, 2017].

If signal $x(n)$ obeys the model described by equation 4.4 exactly, i.e., $a_k = \alpha_k$, then $e(n) = Gu(n)$. This implies that the input signal is proportional to the error signal. Since the filter $H(z)$ is fixed, the total energy in the input signal ($Gu(n)$) must equal the total energy in the error signal and thus the Gain factor G is estimated as,

$$G^2 = E_n. \tag{4.22}$$

Once G is obtained, the LP spectrum can be estimated using,

$$H(e^{j\omega}) = \frac{G}{1 - \sum_{k=1}^{p} a_k e^{-j\omega k}} = \frac{G}{A(e^{-j\omega})}, \tag{4.23}$$

where, j is the complex number with value as $\sqrt{-1}$ and ω is the frequency of the system.

4.2 Leak characterization

Acoustic signatures caused by leaks can be assumed to follow the assumptions of plane wave theory sufficiently away from the source and this makes linear prediction a powerful tool to capture the primary resonant responses of the fluid-pipe coupled linear system. The source/system model for linear predictive analysis of hydro-acoustic signals in water filled pipes is illustrated in Figure 4.2, where the acoustic signal $x(n)$ is modeled as the output of a linear, slowly time-varying system excited by $u(n)$.

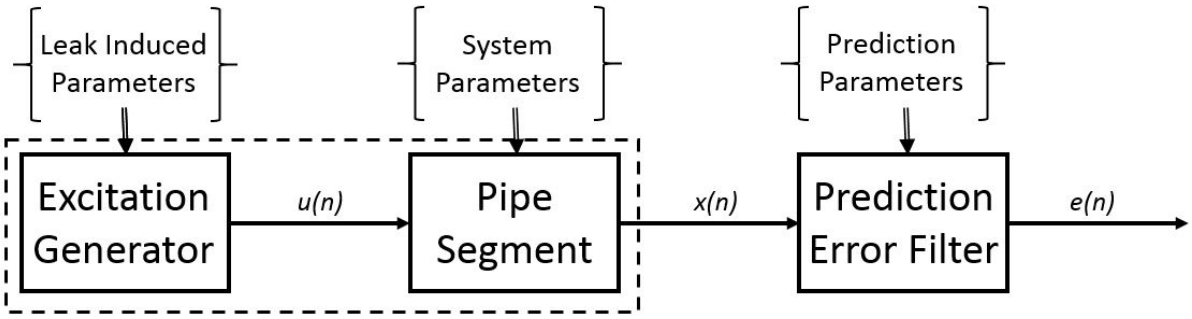


Figure 4.2: Model for linear predictive analysis of leak signals.

With reference to Figure 4.2, the acoustic leak signature $x(n)$ is assumed to be the output of a discrete-time linear system, which is time-invariant within a relatively short time segment. A leak is assumed to introduce excitation characterized by a narrow-band spectrum which is convolved with the impulse response of the fluid-filled pressurized pipe. Such a model, represented by the dashed bounding box in Figure 4.2, mimics the composite spectrum effects of radiation, pipe system, and leak-induced excitation pulse shape over a relatively short measurement period. In the absence of a leak, this system represents the composite effects assuming that the ambient (leak-free) fluid excitation is of broad-band nature (flat spectrum).

Given these assumptions, the auto-correlation function obtained from a short segment of an acoustic signal contains crucial information regarding the presence of a leak. Assume that the system impulse response of a linear system is given by $h(n)$, then the output is obtained through the convolution,

$$x(n) = u(n) * h(n). \tag{4.24}$$

It can then be shown that the autocorrelation follows,

$$R_x(\tau) = R_u(\tau) * R_h(\tau), \tag{4.25}$$

where $R_x(\tau)$ is the autocorrelation of $x(n)$, which can be calculated as the convolution of the autocorrelation functions $R_u(\tau)$ and $R_h(\tau)$ corresponding to $u(n)$ and $h(n)$ respectively.

When a leak is characterized by a narrow-band spectrum, the autocorrelation of the response is also narrow-banded, but is shaped by the composite effects of the leak-induced spectrum as well as the impulse response of the fluid-filled pipe system. On the other hand, the auto-correlation for the case when $u(n)$ is white (seen as an extreme case for the broad-band case) is zero everywhere except for $\tau = 0$. Hence (for the leak-free case),

$$R_x(\tau) = R_h(\tau). \tag{4.26}$$

These results show that the presence of a leak significantly changes the nature of the short-term Fourier spectrum (short-term autocorrelation and short-term Fourier spectrum contain the same information) the spectral peaks are shaped by its presence. The autocorrelation functions taken over relatively short finite time segments could potentially reveal distinct system characteristics in leak versus no-leak scenarios, even for finite-length segments, as shown later. It is interesting to note that a similar idea has been proposed in the context of leak detection earlier [Yang et al., 2013], however their methodology involved directly extracting features from the autocorrelation rather than being model-based.

In this LP model according to equation 4.3, the excitation $u(n)$, which is unknown, but can be assumed as being broad-band for the leak-free case and quasi-periodic pulses for the leak-case [Ferrante and Elghobashi, 2004], is treated in an indirect way; the excitation is whatever is needed to produce $x(n)$, which is the acoustic signal. Linear predictive analysis allows for the excitation gain, G, and the filter coefficients α_k, to be estimated in a very computationally efficient manner, as described in Section 4.1.1.

4.2.1 LP application to leak-induced signals

The basic premise of this work is that a leak introduces quasi-periodic pulse excitation to the system, which means the time lapse between two consecutive pulses is time-varying

(jitter), but within a fixed interval. This assumption is supported by previous observations of non-random coherent structures present in turbulent boundary layers [Ferrante and Elghobashi, 2004]. Such a spectrum can be characterized by a distribution of energy across a narrow band of frequencies and low-pass, which is convolved with the pipe system. The absence of a leak renders the leak-induced spectrum flat, i.e., broad-band with respect to the pipe system dynamics.

To test these assumptions, acoustic signals are acquired from a straight polyvinyl chloride (PVC) water-filled pipe, 12.4 m in length and 15.2 cm in diameter, using hydrophones with $-175dB$ sensitivity and an LMS SCADAS data acquisition system. The pipe is pressurized to approximately 345 kPa using a service inlet and a simulated leak of 0.6 cm diameter is located at 5.3 m from the service inlet end of the pipe section, while the hydrophone is located 7.3 m away from the same end. A schematic diagram of the test bed is presented in Figure 4.3. Data frames of 500 ms duration are selected for analysis. A hamming window is applied to the data frames prior to estimating the auto-correlation functions and the Fourier spectra.

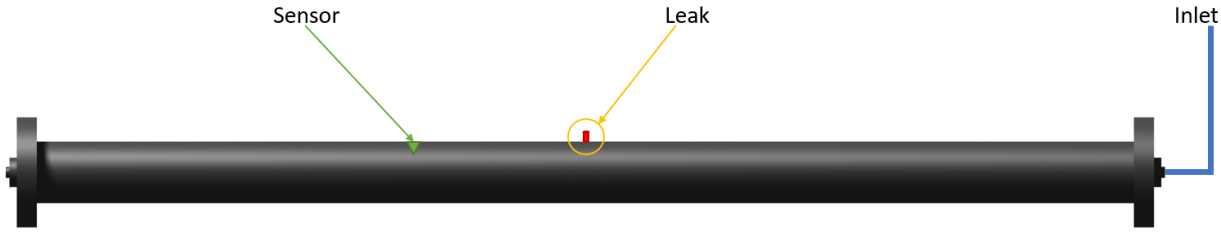


Figure 4.3: The single pipe laboratory setup showing components used in the study (not to scale).

Figure 4.4 shows the short-term autocorrelation function (STACF) and the short-time Fourier spectra (STFS) for the leak-free and leak cases. Subplot (a) and subplot (b) show the STACF and STFS for the leak-free case. The peaks in the STFS correspond to the impulse responses and other disturbances (common to the two cases) of the system. Subplot (c) and (d) show the same for the case when a leak is present and the introduction of leak dynamics into the overall system is evident in the shape of the spectral peaks and the STACF. For instance, the STACF for the leak case in subplot (c) is narrow band in nature with distinct periodicity as compared to the broad band nature of the STACF for the leak free case in subplot (a). The change in the spectral shape is clearly evident in Figure 4.5, which shows the LP spectra generated using four model orders, $p = 50,\ 100,\ 200,\ 500$.

As seen in Figure 4.5, the leak energy is contained in the low-frequency region, which is consistent with previous studies [Muggleton and Brennan, 2004]. As shown in the Figure 4.4 (b) and (d), model order $p = 50$ matches the general shape of the STFS, but does not represent all its local peaks and valleys, while as the model order increases (for example, $p = 500$) the LP spectrum converges towards the true spectrum. For the purposes of leak detection, finer details of the true spectrum is not required to be captured and a lower model order representing the gross spectrum is sufficient as shown in Figure 4.5. While there is no constraint in choosing a higher model order for the current application, too large of the model is not advantageous with respect to computationally efficiency. With reference to the literature on commonly adopted model orders, for speech signals, this is around 20 [Oirere et al., 2015]. An exploratory analysis revealed a model order of 50 as sufficient for this application.

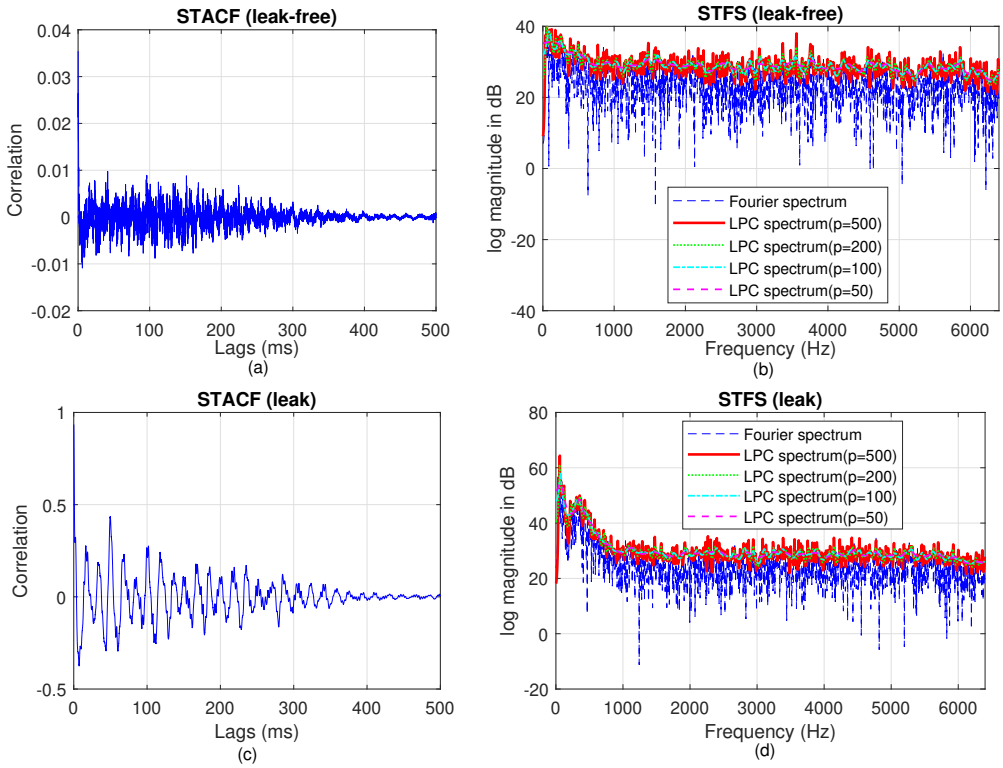


Figure 4.4: STACF and STFS and corresponding LP spectrum for hydro-acoustic signals in cases of normal or leak-free and leak events

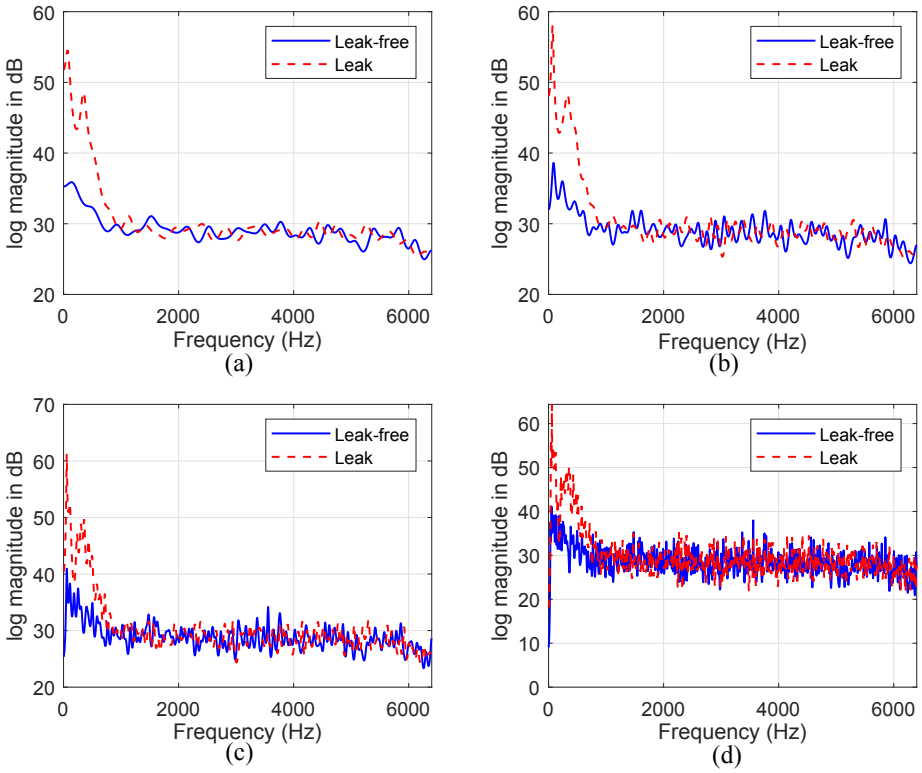


Figure 4.5: LP spectrum for leak-free and leak cases, for LP order: (a) $p = 50$, (b) $p = 100$, (c) $p = 200$ and (d) $p = 500$

4.3 Leak sensitive features

The coefficients representing predictive models for an acoustic signal are meant to approximate closely the key spectral information in the original signal. Such coefficients can be regarded as distinctive features for the data set, on the condition that they provide reasonable representation of the acoustic signals. In the current study, LP coefficients are used as features from acoustic signals as they capture the underlying eigenstructure of a linear time invariant system. This has been shown in the context of linear models of the LP type with applications to both simple (univariate) and complex (multi-variate) systems [von Storch and Zwiers, 2002]. While in the univariate case (single sensor measurement) the eigen values are of primary interest, for the multi-variate case (when multiple sensor measurements are used simultaneously), the eigen vector contains additional spatial infor-

mation regarding the underlying oscillatory patterns of the time series. In the case of an uni-variate model of order p, it can be shown that the eigen value decomposition of the $p \times p$ transition matrix results in an inference of the eigen structure of the underlying time series. The transition matrix can be constructed through representing the linear model in equation 4.4 in the state space.

By introducing the p-dimensional state vector $y(n) = [x(n),\ x(n-1), ... x(n-p+1)]'$ for all n, the LP model can be re-expressed as [West and Harrison, 2006],

$$x(n) = A'y(n) \tag{4.27}$$

$$y(n) = By(n-1) + Ae(n) \tag{4.28}$$

where, $A = \begin{bmatrix} 1 \\ 0 \\ 0 \\ . \\ . \\ 0 \end{bmatrix}$ and $B = \begin{bmatrix} \alpha_1 & \alpha_2 & . & . & \alpha_{p-1} & \alpha_p \\ 1 & 0 & . & . & 0 & 0 \\ 0 & 1 & . & . & 0 & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & . & . & 0 & 0 \\ 0 & 0 & . & . & 1 & 0 \end{bmatrix}$.

The state evolution or transition matrix B has an eigen decomposition as $B = D\Lambda D^{-1}$. The eigen vector matrix D of size $p \times p$ has columns that are the eigen vectors of the corresponding eigen values in the diagonal of the matrix Λ. Since B is real valued, either the eigen values are real or complex, which may occur in conjugate pairs. While real eigen values represent only scalar scaling of eigen vectors, complex eigen values represent scaling as well as rotation of the eigen vectors. In general, the eigenvalues are precisely the characteristic roots of the predictive model and hence a stationary LP model is characterized by the transition matrix B, which has all eigenvalues of less than unit modulus, whether real or complex.

Using the LP coefficients ($p = 50$) estimated from the data set collected from the simplified single pipe system described in the previous section, the transition matrix B is constructed for both leak free and leak cases. All the eigen values estimated for B matrix are shown in Figure 4.6 along with the 6 largest eigen values. These results show that the overall eigen structure of the transition matrix for leak free and leak signals are distinctively different. It should be noted that the leak free and leak signals are distinguishable as long as the leak signatures are not dissipated before reaching the sensor location.

In order to project the important characteristics of the time series into a lower dimension, principal component analysis (PCA) [Bishop, 2006] is performed on the p values of a_k and the three fundamental PCA are taken as the representative feature set. The steps

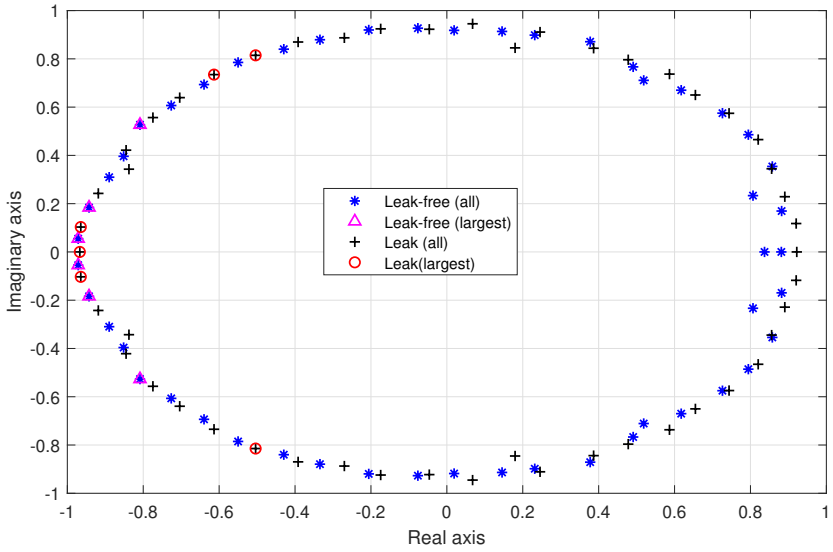


Figure 4.6: All eigen values for leak-free and leak cases along with the six largest values

to perform the PCA analysis on the N by p matrix A is described in 3.3.1. The features extracted from the data set by estimating the LP coefficients and subsequently applying PCA on the coefficients are denoted as LP-PCA features. The application of PCA on the LP coefficients not only helps in projecting the key characteristics of the acoustic data onto three principal components, but also allows for a dimensionality reduction of the feature vector.

Chapter 5

Lab Results

5.1 Introduction

This chapter presents a laboratory case study for the application of both model free and model based methods for semi-supervised leak detection in WDNs. The model free methods employed in this chapter are SSA—it was used as a key step in processing the data—followed by the application of a OCSVM classification methodology; and, a NN utilizing a deep autoencoder. The model based method employed later in the chapter utilizes LP, coupled with a GMM classification methodology, and is extended for localization.

In order to identify an ideal method for the field case study described in Chapter 6, three methods were considered. First, typical time domain features were reviewed, an in an effort to increase the detection accuracy of these common statistical features, SSA was selected as a pre-processing methodology. While this method yielded promising results, the amount of tuning required made wide scale deployment difficult. Following this, a NN approach was assessed. Once the results of simple statistical features were reviewed, it was evident that without SSA pre-processing detection accuracy was poor, while including SSA as a pre-processing step required more parameter tuning than can be effectively deployed for wide scale field studies. As such an effort to address the need for feature engineering was made. In utilizing spectrograms of the leak free data, the proposed NN extracted features and good classification was obtained. However this method was far too computationally intensive for a wide scale deployment. Finally an LP methodology was reviewed which focused on computational efficiency, as well as expanding the proposed methodology to localization.

This chapter is organized as follows: first, the laboratory test bed and it's associated data sets are reviewed; then, the results for the proposed SSA methodology for leak detection are reported; followed by a deep autoencoder based methodology for leak detection ; next, the results for the proposed LP methodology are reported, including leak detection, and leak localization results; and finally a comparative summary of these results is discussed. The results of this laboratory exercise is intended to guide the selection of the method employed for field implementation subsequently.

5.2 Laboratory Test Bed

The laboratory experimental set up simulates a small portion of a typical full-scale water distribution system in North America. This test-bed is by no means meant to fully mimic actual distribution systems, but intended to capture the geometry of the pipes, typical noise and the operational pressure and the associated variations. While a simple single pipe network, depicted in Figure 4.3 was used for a review of leak characteristics, in order to experimentally validate the proposed methodologies, the system required added complexity. Two iterations of the laboratory network exist as this dissertation progressed and more complexity was added. First the single pipe was augmented to a tee network including a service valve to simulate noisy conditions. Once the first proposed methodology was proven on this set up, an attempt to incorporate added complexity was made. The network was then augmented to include more tees as well as multiple loops. These two laboratory networks are described in detail next.

5.2.1 First Iteration

The first iteration of the test bed consists of a relatively simple roughly orthogonal layout of PVC pipes, tees and a fire hydrant. The pipes are made of 'Grey Scale 80 PVC' pipes with a 15.24 cm inner diameter, typically used in Canadian and US full-scale networks. A total length of approximately 20 m of pipes comprised the system along with three simulated leaks created at different locations, one service connection valve and a fire hydrant monitoring station where the sensor is mounted. The fire hydrant location was specifically chosen as it mimics actual implementation in the field for this dissertation. One end of the pipe is directly connected to the building's water supply distribution system, one is a retrofitted fire hydrant, the remaining are terminated with end caps (one of them has a valve which opens to simulate the flow case or a larger leak). The three small leaks are simulated using a 6.35 mm inner diameter valve, which simulated 0.25 L/sec flow,

while the service line is simulated using a 2.54 cm valve. Figure 5.1 illustrates the system layout at scale. Leak 1 is located approximately 3 m from the sensor, while leaks 2 and 3 are located approximately 2 m from the sensor. The acoustic data was collected using a hydrophone located at the base of the fire hydrant, mounted using a specially designed hydrant valve stem. This type of arrangement where the sensors are placed at the base of the hydrant allows for continuous operation of the hydrant without the need to flood the hydrant during data collection.

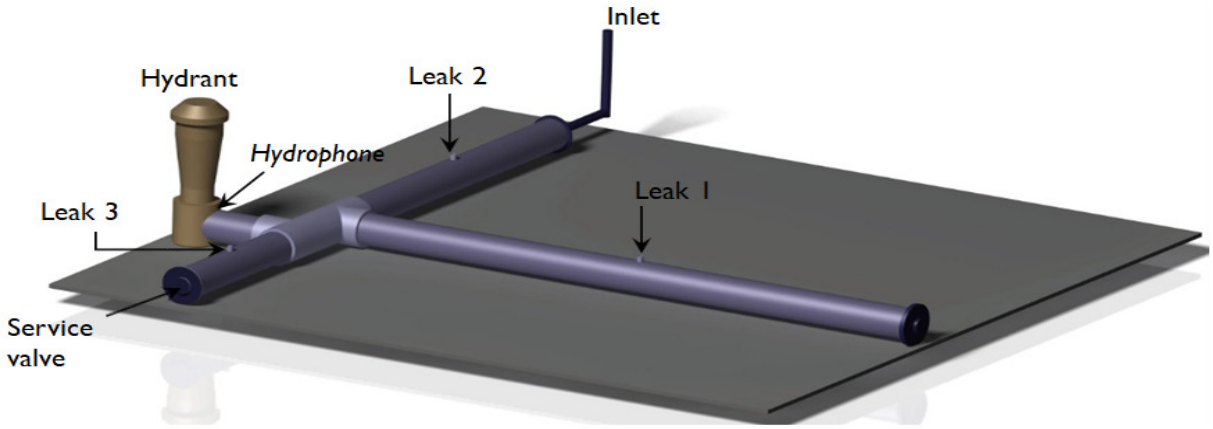


Figure 5.1: First iteration of the laboratory pipe network showing key components used in the experimental study (not to scale).

5.2.2 Second Iteration

The second iteration of the test bed consists of a series of grey scale 80 PVC pipes with 15.24 cm inner diameter, two tees, a fire hydrant, one service connection valve, four simulated leaks at different locations and a city line inlet. The total length of the pipe system is approximately 30 m. A schematic diagram of the test bed is presented in Figure 5.2. The hydrophone used for the laboratory monitoring system was the Teledyne Reson TC4013, which was selected for its high sensitivity in order to determine the region of the spectrum required for analysis.

Leaks are simulated by opening a 0.64 cm valves at four locations, each of which results in flows ranging from $18 - 20$ L/min when fully opened. Two hydrophones, represented by *Sensor 1* and *Sensor 2*, are used to measure the acoustic characteristics. The first leak (*Leak 1*) is located 496.6 cm from *Sensor 1* (taking the most direct path). The second and

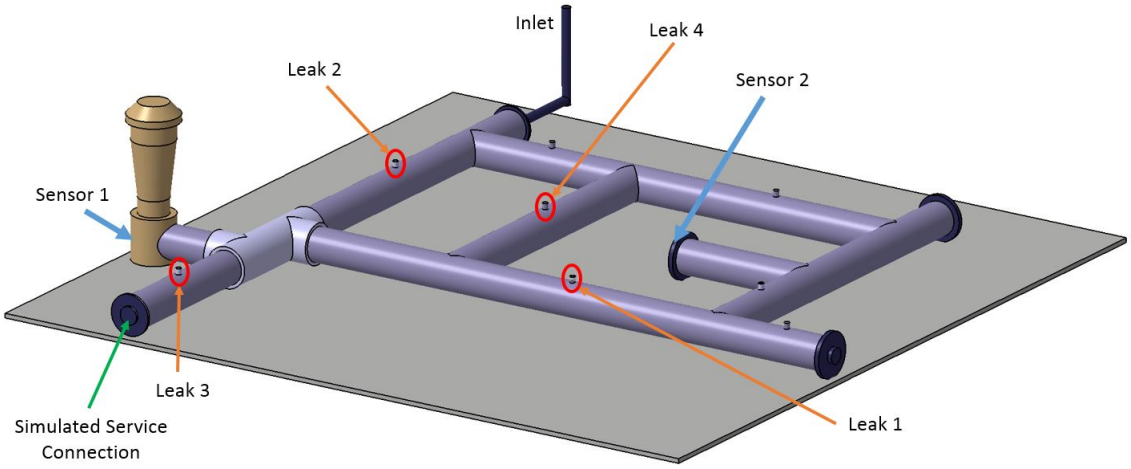


Figure 5.2: Second iteration of the laboratory pipe network showing key components used in the experimental study (not to scale).

third leaks, i.e., *Leak 2* and *Leak 3*, are situated at 360.7 *cm* and 292.1 *cm* from *Sensor 1*, respectively. The fourth leak (*Leak 4*) is located at 562.6 *cm* from *Sensor 1* and 911.9 *cm* from *Sensor 2*. The simulated leaks create different acoustic signatures at various locations due to the different impedance changes in the acoustic wave propagation path resulting from bends and tees in the system and this configuration yields a rich set of test cases to evaluate the proposed approach.

A service line is incorporated in the test bed by installing a 2.54 *cm* valve located 288.3 *cm* from the branch which leads to the fire hydrant, which simulates typical water usage demand from the system. During the experiments, *Flow* and *No flow* cases are generated in the pressurized water pipe system by opening and closing this service line. The *No flow* case represents a relatively less noisy environment, while the *Flow* case is associated with the presence of very high background noise in the pipe system caused by this service line. The inlet (Figure 5.2) is connected to the City of Waterloo's main distribution system. The system is pressurized to an average value of 345 KPa, but was observed to fluctuate between 310 KPa to 380 KPa, mainly depending on the building usage. This fluctuation is representative of head tank pressures, as detailed by a brief study of the head tank pressure found in the city of Guelph, as outlined in Appendix C. *Sensor 1* is installed at the base of the hydrant through a specially designed hydrant valve stem as shown in Figure 6.1a, while *Sensor 2* is inserted into the end cap.

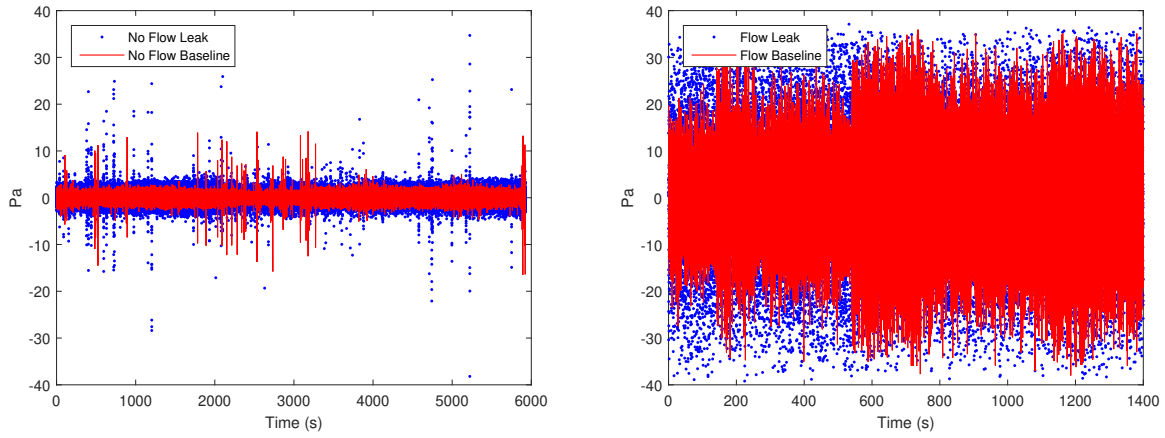
The laboratory test bed in this research has characteristics in the range of past studies discussed in Section 2.4. These laboratory test-beds consisted of total pipe network lengths between 10 m - 100 m with leak sizes ranging from a quarter inch to four inches, in a variety of typical and atypical pipe diameters for WDNs [Jia et al., 2015, Lazhar et al., 2013, Ferrante et al., 2013, Khulief et al., 2011, Khalifa et al., 2010, Soares et al., 2008, Covas et al., 2006, Lee et al., 2005, Mpesha et al., 2001]. These studies reviewed different methodologies on vastly different test-beds, as such there is a large diversity in performance, depending on the specific laboratory test-bed's characteristics (i.e. material, diameter, pressurization method). The performance measure is only meaningful in the context of the specific laboratory test bed's characteristics. This lack of standardization makes direct comparison to other methods difficult.

5.2.3 Data Collection

From a data-driven algorithmic stand-point, leak detection in noisy conditions is a challenging problem, especially when the leak occurs concurrently within the variability associated with service conditions, such as opening of a neighboring valve. For this reason two baseline scenarios are considered. Experiments and data acquisition were conducted under a total of four scenarios: (1) ambient, which means there is no leak and the valve is closed; (2) under leak condition, where a leak is present and the valve is closed; (3) valve, where the network is leak-free and the valve is open, (4) valve & leak, where a leak is present and the valve is open. For each scenario, the acoustic signals were acquired at different times over a month with a sampling frequency of 1.35 kHz. This sampling frequency corresponds to the maximum frequency that can be achieved by the 16—bit A/D custom DAQ system used. This DAQ system consists of a custom micro-controller board, local storage and a communications module supporting wireless communications. The power was provided using an external battery. This DAQ was consciously chosen in order to mimic the system to be deployed in the field in full-scale tests. A newer generation supporting 24-bit A/D and a more powerful micro-controller was developed and used in the field tests as described in Chapter 6.

The data acquisition rate used is deemed adequate for the purpose of this study since the leak information is typically found towards the lower end of frequency spectrum, under 300 Hz [Hunaidi et al., 2000]. The data was collected for up to two minutes in 10 to 30 second intervals at different times throughout a given day and repeated for nearly a month. This was done in order to ensure variability in the data as well as minimize the effects of overfitting to the system conditions. Since the pipes are connected directly to the building water distribution system, the experimental test bed also reflects the supply and

demand patterns typical of full-scale WDNs. Acoustic pressure data was collected using a hydrophone (SensorTech SQ26-13) of sensitivity $-193\ dB$ re $1V/\mu Pa$ with the raw voltage signal having an added pre-amplification of 20 dB gain prior to data-acquisition. Sample time histories of the acquired signals for the second iteration of the laboratory setup are shown in Figure 5.3 for the four scenarios previously described.



(a) No flow (b) Flow

Figure 5.3: Hydrophone measurements without and with the presence of a leak when the service line valve is (a) closed, and (b) opened.

Leak signals from different leak locations (L1, L2, L3 shown in Figure 5.2) were collected to form one dataset. The frequency spectra associated to signals in noisy and quiet conditions are illustrated in Figure 5.4. The variability in the leak versus leak free cases can be seen in Figure 5.2. For certain segments of the collected data set the baseline, or leak-free, and leak events differ visibly, e.g., time stamp 300 to 500 seconds in sub-figure (b) (having an RMS value of 14 Pa for the leak case, as compared to 8 Pa for the baseline case). Conversely, for other segments the baseline and leak events signals appear similar in the time series (as well as yielding nearly identical RMS values, e.g., time stamp 600 to 800 seconds in sub-figure (b) yield and RMS value of 13.01 Pa for the leak case and 12.90 Pa for the baseline case). This variability is likely caused by the different leak locations contained throughout the leak case data set, how similar the leak case data is to the baseline is likely influenced by the proximity and impedances present between the sensor and the leak. The different locations of the leak allow for variability in the travel time and exposure to different impedance changes in the system, thus inducing different levels of energy attenuation and noise. While one would assume that the leak located furthest

from the sensor would produce the most difficult to detect leak, experimentally the data has shown that this is not the case and is a function of the number and type of elements which cause impedance changes. This observation is also consistent with findings by others [Jia et al., 2015]. The effect of bends or tees on dissipating leak signatures is equal or more significant than straight line distance for low frequencies, as the low frequency signals have been shown to travel relatively large distances in fluid filled pipes with very little attenuation [Aristegui et al., 2001]. Thus, the incorporation of bends and tees in the pipe network was an important aspect in this test bed in order to produce a data set with adequate variability and complexity mimicking field conditions.

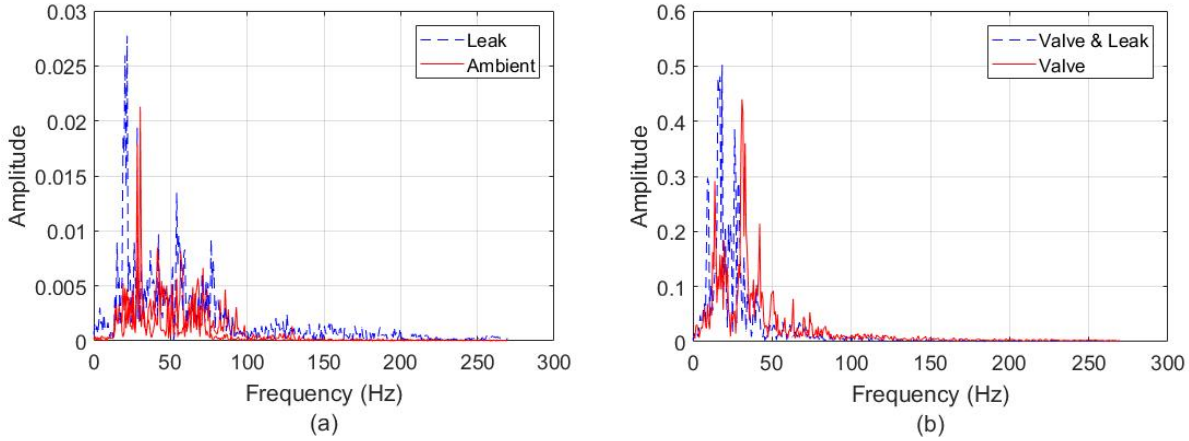


Figure 5.4: Spectra of acoustic signals in the absence and presence of a leak; (a) valve closed, (b) valve open

The leaks located in different areas of the pipe create different leak signatures, leaks located along the length of the pipe produce leak signatures which are much more easily detected, however, leaks located perpendicular, or after a Tee, produce leak signatures which are more difficult to see visibly in the frequency domain, as well as more difficult to classify. Thus while analysis was done for all leak cases, this chapter presents the results for the worse case of leak, which is leak 3 (Figure 5.2). It can be seen in Figure 5.4 that most of the energy in the spectrum—for both leak and ambient cases—are limited to under 250 Hz. This is consistent with previous reports, for e.g., [Hunaidi and Chu, 1999]. Furthermore, there is a visible lifting of the spectral energy in Figure 5.4(a) across the aforementioned frequencies, which means that delineating leak versus ambient cases should be fairly straight-forward, even using basic feature sets derived directly from the

time series. On the other hand, opening the valve increases the amplitude of the signal considerably (nearly 5-10 times across key regions of the spectrum) and changes the inherent structure and frequency content of the acoustic signals. From a classification standpoint, this increases the overlap of features obtained from the two scenarios and obfuscates leak detectability.

5.3 SSA Results

This section proposes the use of a non-parametric method, called the Singular Spectrum Analysis (SSA) Vautard and Ghil [1989], Harmouche et al. [2018] as a tool for pre-processing hydroacoustic signatures. SSA is chosen for its simplicity, as it has only one tunable parameter, does not require stationarity, linearity or normality assumptions about the processed data and has seen widespread use in other fields, including biology Yufeng and Saniie [2015], Celka and Colditz [2002], Sanei and Hassani [2015]. While time-frequency methods such as wavelet transforms have been used in the context of burst event detection and localization Srirangarajan et al. [2013], SSA is more directly related to the eigen structure of the signal and will form excellent comparison with model based method (LP) described later.

SSA on its own can only decompose the signal into its constituent parts and is a pre-processing technique to extract leak components from complex measurements. For leak detection, the decomposed signals are combined with an ensemble one-class support vector machine Scholkopf et al. [2001] in an unsupervised approach. As will be shown in this study, leaks buried in high non-stationary background noise results in a change of the frequency structure of the signal, which are readily discernible once pre-processed using SSA. Significant improvement in the detection performance can achieved compared to using raw features alone.

5.3.1 Data Processing

Enhancing the ability of descriptive features to discriminate between leak and non-leak data is essential for the purposes of leak detection. Utilizing the first iteration of the laboratory set up the experimental hydro-acoustic signals are partitioned into consecutive segments of 5000 samples each. Entropy, effective value and spectral peak are computed for each segment.

SSA with an embedding dimension $L = 500$ is then applied to each frame. In general, L is empirically chosen according to the particularity of the time series [Golyandina, 2010]. As a rule of thumb, in order to extract a given oscillation, L is recommended to be larger than its fundamental period. More importantly, a large L ($L < N/2$) is recommended to decompose an arbitrary signal of length N. Here, it is selected large enough so that the window embeds the entire signal spectrum ($3Hz$-$300Hz$). Each frame of 5000 samples is hence decomposed into 500 independent elementary components.

The use of SSA to increase the discrimination capability of the features is validated using a OCSVM methodology [Yin et al., 2014] described in Section 3.3.3 , which is used to model the features computed from leak-free SSA-processed data. The procedure is as follows:

1. Leak-free hydro-acoustic signals are collected under different conditions.
2. The signals are segmented into frames and SSA is applied. The SSA parameter L is selected based on the prior analysis of the frequency content of the signals.
3. Each frame gives two components which are reconstructed from the pre-selected SSA elementary components, one corresponds to group I_1 and the second corresponds to group I_2.
4. Features are then computed from the obtained components. Thirty percent of the computed feature values are randomly selected and used for testing. A OCSVM model with Gaussian kernel is trained using the remaining 70% of feature values.

Leak detection accuracy is evaluated by estimating the ROC curve (described in Section 3) using leak and leak-free data.

5.3.2 Features analysis

Using the proposed SSA based approach, a group of elementary components is assumed to be carrying the signatures related to the leak, while the remaining components form the background variation which are insensitive to the leak. The leak sensitive group is identified and used to compute the features. The identification of the sensitive elementary components is straightforward by referring to the singular spectrum. Figure 5.5 shows the box plot of the first 10 singular values computed for the data frames that correspond to the open valve case. The intervals of singular values for leak (filled box plot) and non-leak data are illustrated. For a particular component of order I (where order follows the

decreasing sequence of singular values), a high overlap in the box plot means the component is not sensitive to the leak presence, since the overlap indicates that the distribution of the singular value in question is not perfectly separated. This is the case of the first two singular values for example. However, the fifth and sixth singular values show minimal intersection between the box plots for the leak and leak-free data, and therefore are more sensitive to the presence of leaks.

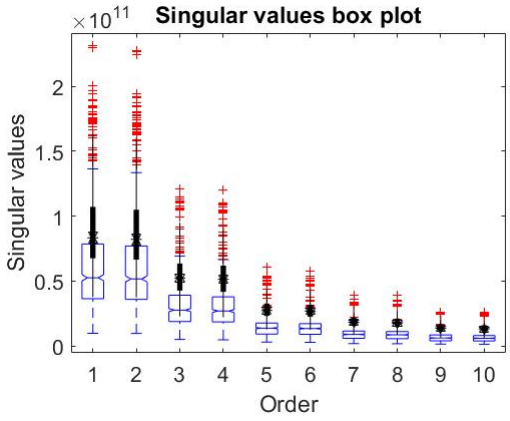


Figure 5.5: Box plot of the first 10 singular values computed for the data frames of open valve case, the filled box plot refers to leak data

This procedure, applied to the data corresponding to the quiescent water system (valve closed), leads to the selection of elementary components of orders in $I_1 = \{21, ..., L\}$ to reconstruct the signal carrying the leak signature. When applied to the data corresponding to the noisy water system (valve open), it leads to the selection of elementary components of orders in $I_2 = \{5, ..., 20\}$ to reconstruct the leak signal. The leak signal is the sum of those chosen elementary components; features are computed using the obtained SSA components, instead of the raw signals.

Figure 5.6 shows the features histograms computed on the reconstructed leak signals. In the left-hand column (valve closed), the leak signals are reconstructed using group I_1, while group I_2 is used to reconstruct leak signals in the right-hand column (valve open) of Figure 5.6. It is clear from the histograms that the Bhattacharya distance of the leak signals obtained through SSA decomposition is large and hence should enhance discrimination capability.

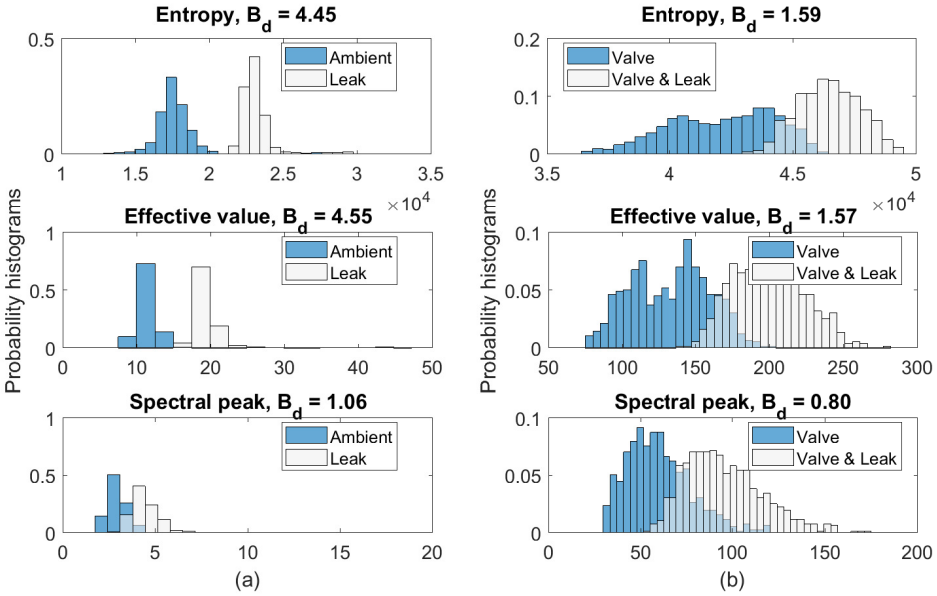


Figure 5.6: Histograms and Bhattacharya distance corresponding to the entropy, the effective value and the spectral peak, computed on the SSA components of leak and non-leak time-series (a) valve closed, (b) valve open

The AUC is evaluated for different parameters (ν, γ) of the SVM model, and several runs are performed, where at each iteration 70% of leak-free data are randomly selected for training and 30% for testing. Figure 5.7 shows the AUC averaged over multiple runs, for different values of the SVM model parameters. It is clear that the SVM parameters have a large impact on the detection performance and large values of both ν and γ are favorable in the data set used in this study. In practice, however, due to the possible lack of historical and leak data, the model is trained using the current set of data in order to maximize a given metric of the detection quality. With the availability of new data the model can be readjusted to achieve better accuracy. In the current case, it is shown that for a given value of ν, the AUC increases with γ and reaches a maximum which is in the range [0.85, 0.92]. As a comparison, the maximum AUC values range from [0.67, 0.83] if SSA is not used and the same features are computed from raw time series data.

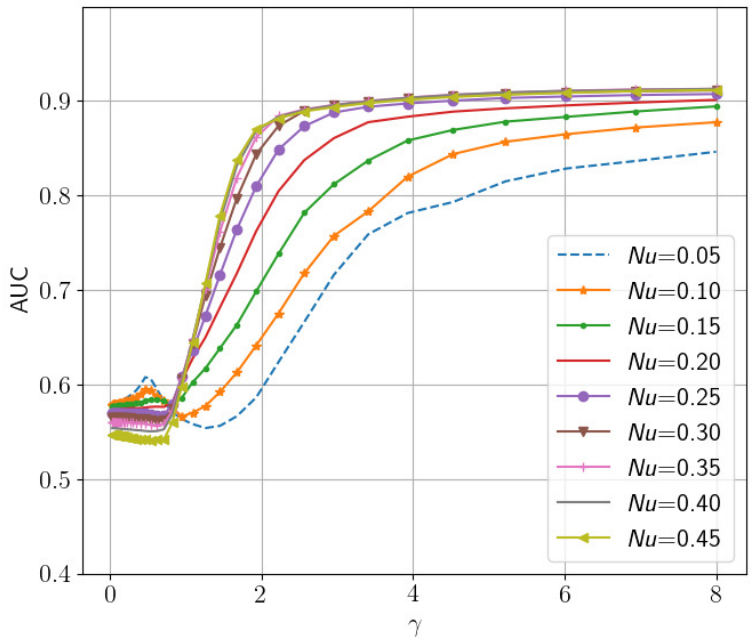


Figure 5.7: Evaluation of leak detection accuracy using the AUC of a OCSVM model based on SSA components.

5.3.3 Leak detection

Where leak data are absent or minimal, the selection of components based on the analysis of their sensitivity to leaks is impractical. However, under the assumption that the set of elementary components can be divided into two disjoint subsets of sensitive and insensitive components, an OCSVM can be trained for each elementary component and L separate models can be obtained. Over time, poor models will be recognized by their randomly varying predictions, while efficient models based on the highly sensitive components are likely to make consistent predictions. Training and continuously tuning L models can be costly especially when L is large. This number can be reduced based on the following characteristics of SSA decomposition.

The elementary components can be separated into signal and noise components. The signal components are the first few ones and their number, denoted by d, can be determined using information criteria [Kumar et al., 2006]. On the one hand, after the signal dimension

is selected, $(L - d)$ residual elementary components can be combined in order to form one noise component. On the other hand, generally speaking, the signal components are expected to represent trends and oscillations. In the case of the hydro-acoustic signals, they mainly consist of oscillations. In this case, each pair of elementary components corresponds to an oscillation and those can be combined to form one component. This reduces d into $d/2$ components, thus leading to $d/2 + 1$ components in total including the residual one.

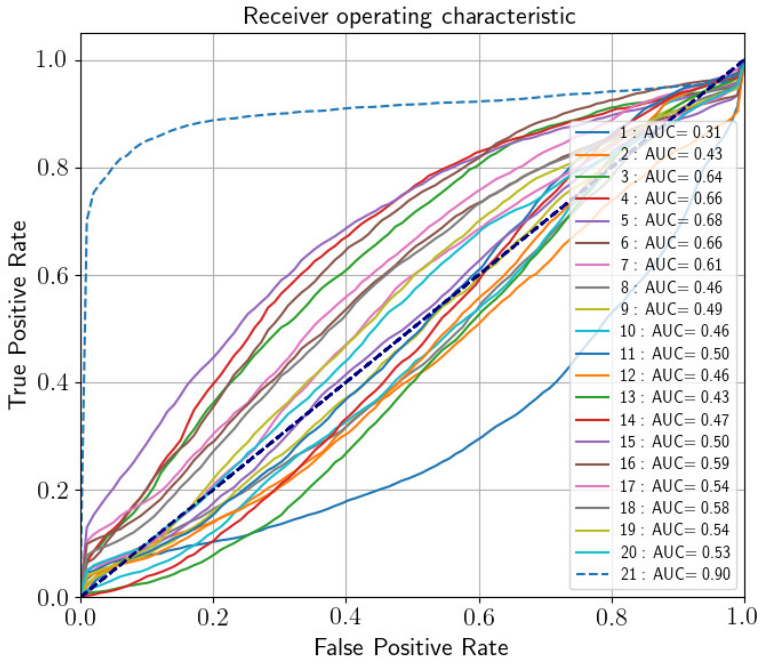


Figure 5.8: Evaluation of leak detection accuracy using the AUC of a OCSVM model when the valve is closed

This approach is based on the assumption that SSA components can be separated into a set of sensitive and less sensitive (to leak) components. For experimental validation, with $L = 500$, the signal dimension used is 40. 21 components are extracted according to the previously described procedure, and a OCSVM is learned for each component using the leak-free data. All models are treated equally with the same parameter values, which are inferred from the previous performance results. In order to validate the stated assumption, the models are tested against leak and leak-free data. ROC curves, along with AUC values, are estimated for each model and the results are shown in Figure 5.8-5.9. When the valve

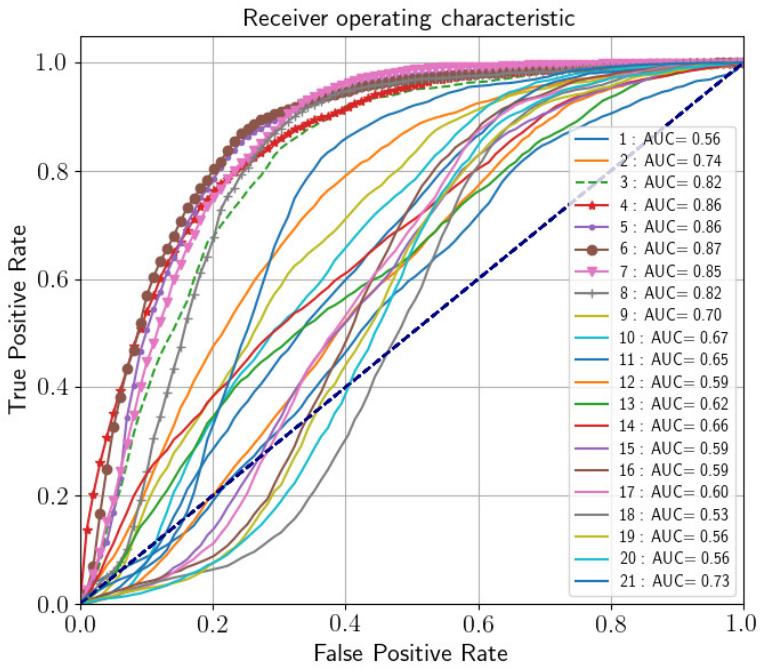


Figure 5.9: Evaluation of leak detection accuracy using the AUC of a OCSVM model when the valve is open

is closed and the network is relatively quiet, the 21st component results in an AUC of 0.90 (dashed line in Figure 5.8), meaning a high detection accuracy, while the remaining components lead to AUC values around 0.5, which mean they perform almost random predictions. The models are also tested against leaks that occur when the valve is open. It is shown that many components are expected to perform poorly (AUC < 0.6) while few components (2, 9, 21) result in good predictions. However, the components 3, 4, 5, 6, 7, 8 (marked lines in 5.9) lead to AUC > 0.8, with 4, 5, 6 have AUC > 0.85. Once a leak occurs, those latter components (4, 5, 6) give deterministic and constant alarms leading to a robust decision making process.

The experimental results from this proposed nonparametric algorithm based on SSA decomposition of raw measurements has many advantages. The approach is fully data-driven and only requires the availability of sensor data. It has the potential to detect small leaks, while signature are hidden in background noise, thanks to the decomposition of the raw signals into elementary components. While this method allows for significant insight

into the data, exploratory analysis is required to identify the leak sensitive components. As well, the use of some leak data is required for parameter tuning. The tuning of these parameters can greatly affect the classification accuracy of the results. These not only make the proposed methodology difficult to deploy on a wide scale, but it is also important to note that this study is limited to the task of leak detection and not localization, which is a requirement for useful field deployment.

5.4 Results of using NN

In addition to being problematic in situations where there is significant variability in baseline conditions, hydroacoustic-based leak detection studies require some feature engineering. Feature engineering is the application of domain knowledge in order to identify and select a subset of case sensitive features from a data set (e.g., mean, variance) to be used as input for machine learning algorithms. While feature engineering can identify appropriate features in the context of laboratory-based leak detection studies [Cody et al., 2017, 2018, Khulief et al., 2011], it is not clear if the set of relevant features will be the same or even constant across an actual WDN. Repeating the feature selection step prior to applying the classification algorithm for every new monitoring location in a WDN would be time-consuming and potentially problematic. One deep learning technique that is free from the need for feature engineering and new in the context of leak detection is the autoencoder. Deep autoencoders have typical application in image classification [Geng et al., 2015, Chen et al., 2014] and speech recognition [Principi et al., 2017], as well as in fraud detection [Vartouni et al., 2018]. Since hydro-acoustic data can also be represented as a spectrogram, autoencoders could be applied as another novelty detection approach for leak detection.

This section proposes a semi-supervised method using autoencoders for leak detection in water distribution systems based on spectrograms of hydro-acoustic data. The proposed spectogram-based novelty detection method combines a deep 2D convolutional neural network (CNN) within a variational autoencoder (VAE). The deep VAE will include 2D CNN layers for pre-processing the spectrograms, followed by a variational autoencoder layer to reach the latent layer. The proposed data-driven methodology is, to the best of the authors knowledge, the first application of autoencoders to hydro-acoustic leak monitoring of WDNs. This contribution is significant because the approach does not require training with leak data sets and thus eliminates the need for the feature engineering/selection process, overcoming a key limitation of current data driven approaches applied to the problem of leak detection in pipe networks. Results from this implementation will serve as a good reference for the model based LP method to be described later.

5.4.1 Data Processing

The data utilized for the proposed methodology are spectrograms of hydro-acoustic measurements collected using the first iteration of the laboratory test bed. Spectrogram images are used to train the deep learning tool so typical frequencies expected within the system can be recognized. The presence of a leak is expected to cause a change in the energy distribution in the transform. While for smaller leaks this difference may be difficult to discern visually, it is expected that the proposed methodology will be able to detect the change. Preprocessing was applied to the data prior to the spectrogram calculation. In order to minimize aliasing caused by the Fourier Transform a Hanning window was applied to each data subset prior to spectrogram calculation.

A normalized sample of the baseline data is shown in the example spectrogram in Figure 5.10 (with mathematical description derived in Section 3.2.5) to provide readers an idea of each observation that is generated. However it should be noted that, for analysis, the data was not normalized.

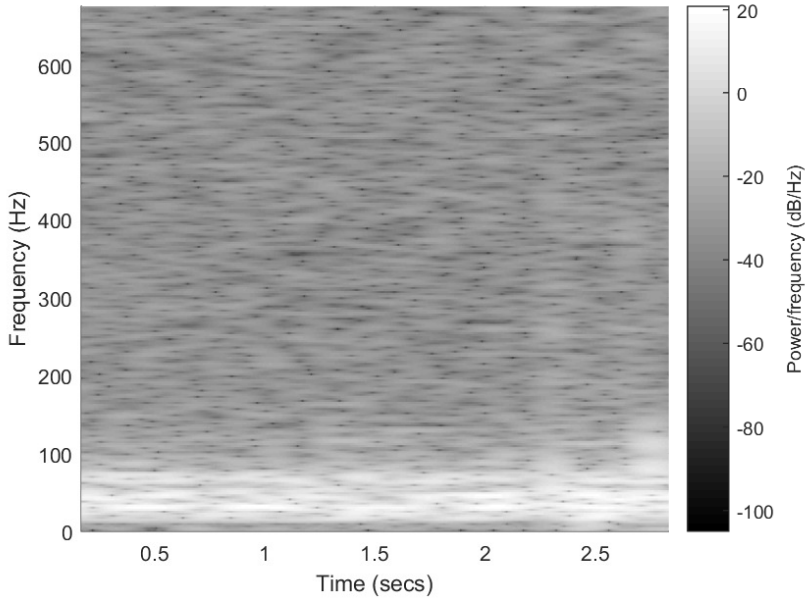


Figure 5.10: An example normalized spectrogram of baseline —leak-free case. Visualization of the input training data. With sampling rate = 1350Hz, window = Hann, window size = 4050 and overlap length of 50%.

The hydrophone signals are partitioned using a 4,050-sample window segment with

50% overlap into adjacent frames, that is 3 seconds of data is used for each instance. The subjective choice of a three second window was based on an inspection of the raw data collected. This is done to both the leak-free and leak scenarios. As well a parametric study was conducted to understand the effect of various window lengths on classification accuracy using simple statistical features as a OCSVM methodology, and it was concluded that between 1 *second* and 4 *seconds* the change in results were negligible, however when segment lengths exceeding 5 *seconds* the accuracy was reduced significantly.

5.4.2 Implementation of proposed ANN

Tensorflow [Abadi et al., 2016] is an open source software library for high performance numerical computation. It is used extensively for developing ANNs. Keras is a high-level ANN application program interface, written in Python and runs on top of Tensorflow. An ANN was constructed using a CNN and VAE combination using Keras as a wrapper for TensorFlow. For this, batch sizes of 30 were selected and the neural network was trained for 30 epochs. The training set consisted of only baseline data, this included 990 spectrograms for training with a validation set of 630 spectrograms. The test sets of 1000 spectrograms consisted of equal parts baseline and leak cases. The general classification methodology of the proposed work is summarized in Figure 5.11 and the variational autoencoder/decoder framework is visually depicted in Figure 5.12.

Implementation of CNN

A spectrogram is passed through a CNN consisting of convolutional layers and max pooling layers. The output of the CNN is then fed into the variational autoencoder, described in Section 5.4.2. The CNN is then repeated in reverse to reconstruct the original image. The loss function is taken as the mean of the loss using a Mean Squared Error (MSE) between the original and reconstructed image, as seen in equation (5.1), and KL divergence of the input image and latent layer, as seen in equation (5.5) described in Section 5.4.2.

The MSE is the performance measure most widely selected for regression applications. The standard form of the MSE loss function is,

$$\mathcal{L}(\hat{y}, y) = \frac{1}{N}\sum_{i}^{N}[y_i - \hat{y}_i]^2, \tag{5.1}$$

where $(y_i - \hat{y}_i)$ is the residual which the MSE loss function targets to minimize, where y represents the element-wise L2 loss.

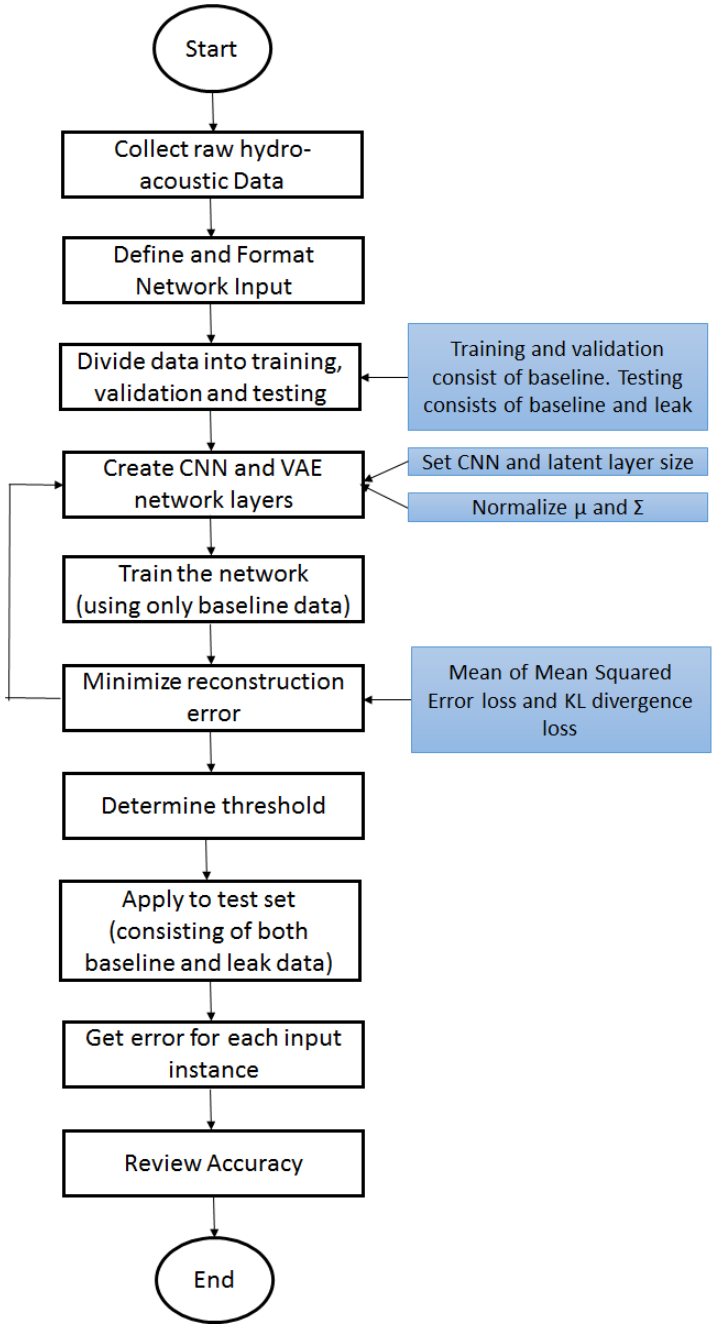


Figure 5.11: Overall structure of the proposed novelty detection methodology.

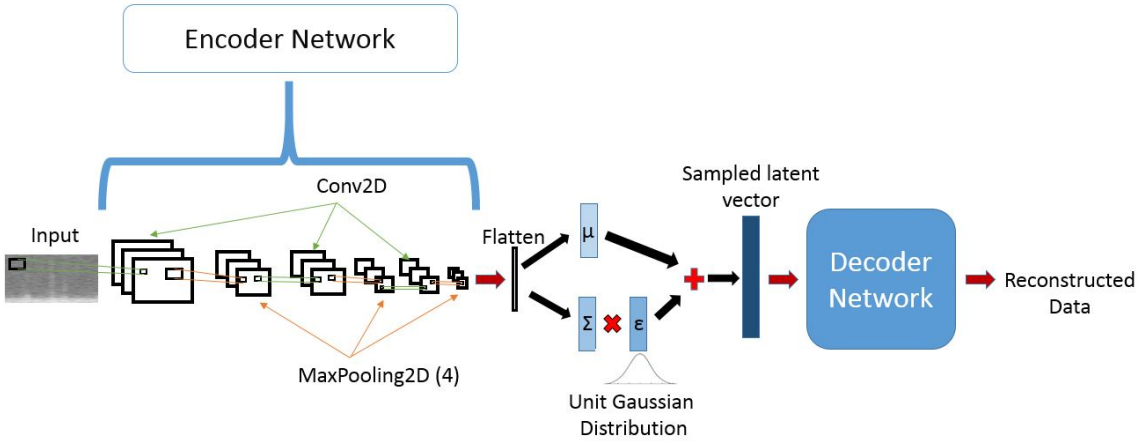


Figure 5.12: Convolutional Neural Network structure —the encoder segment applied in this dissertation includes a sequence of convolutional and max pooling layers. Flattening is applied, in which the elements are reordered from a multi-dimensional array into a 1-D array. The Conv2D layers use a 3 x 3 window size with a rectified linear unit (ReLU) activation function. The output layer uses a Sigmoid (logistic) activation function. The MaxPooling2D layers use a 2 x 2 window size (indicated by the 4 in the figure). The framework for variational autoencoder/decoder is also depicted. The latent space has an assigned dimension of 2, and is then passed to a *dense* (fully connected) layer with a linear activation function, outputting a vector. The decoder network is an exact inverse of the encoder network.

During the reconstruction portion of the CNN, up sampling is applied in the place of max pooling. Up sampling is the process of duplicating the original sample to increase the sampling rate, or, in this case, window size. This is referred to as the nearest-neighbour interpolation method for up sampling. Three two dimensional convolutional (Conv2D) layers were used with a 3 x 3 window size, padding applied to the original image, a rectified linear unit (ReLU) activation function within the network and a Sigmoid (logistic) activation function of the output layer. The first layer created 16 filters while the following two produce 8 filters. Between each of these layers max pooling (and subsequently up sampling) is applied [Yang et al., 2018], as previously described, using a 2 x 2 window size.

Implementation of VAE

The encoder takes input X and outputs $\mu(X)$ and $\Sigma(X)$, which are used as parameters of a Gaussian of the projection of X into the latent variable space [Kingma and Welling, 2014]. The latent variable space can thus follow a unit Gaussian distribution. This is done using the reparameterization trick, by assuming the original data is Gaussian with some mean and standard deviation and then restandardizing it so that the mean is now 0 with a standard distribution of 1. With this assumption in mind, when sampling from this standard normal distribution in which the mean and standard deviation is known, the sampling operation can be implemented as,

$$z = \mu(X) + \Sigma(X)\epsilon, \tag{5.2}$$

where, $\epsilon \sim N(0, 1)$. A latent variable dimension of 2 was selected for the proposed methodology.

The second half of the loss function is the KL loss (described in Section 3.3.7) from each data point in the mini-batch.

Since $P(z) = N(0, 1)$, and $Q(z|X)$ is Gaussian with parameters $\mu(X)$ and $\Sigma(X)$, the KL-divergence between these two can be computed in closed form,

$$D_{KL}[N(\mu(X), \Sigma(X))||N(0, 1)] = \frac{1}{2}(tr(\Sigma(X)) + \mu(X)^T\mu(X) - k - \log det(\Sigma(X))), \tag{5.3}$$

in which k is the dimension of the Gaussian, $tr(X)$ is the trace function (the sum of the diagonal of matrix X), and $det(\Sigma(X))$ is the determinant of the standard deviation matrix of X. The determinant of a diagonal matrix can be computed as the product of its diagonal. Since $\Sigma(X)$ is a diagonal matrix it can be considered a vector. Thus equation (5.3) can be simplified to,

$$D_{KL} = \frac{1}{2}\sum_k(\Sigma(X) + \mu^2(X) - 1 - \log\Sigma(X)). \tag{5.4}$$

While mathematically equation (5.4) is correct, in practice $\Sigma(X)$ is modeled as $\log\Sigma(X)$ as it is more numerically stable. Hence the second half of the loss function is calculated using,

$$D_{KL} = \frac{1}{2}\sum_k(\exp(\Sigma(X)) + \mu^2(X) - 1 - \Sigma(X)). \tag{5.5}$$

The overall loss function, or reconstruction error, used for the described model takes into account both equation 5.5 and equation 5.1, and is defined as,

$$Loss = \frac{D_{KL} + \mathcal{L}(\hat{y}, y)}{2}. \tag{5.6}$$

This loss is first minimized during the training phase to calibrate the network weights. The training phase reconstruction errors are also analyzed to determine the reconstruction error threshold given an application specific allowable type I error (false positive rate). The loss equation is then used in the testing phase where the error values are used for classification.

5.4.3 Leak detection

For this section once the data is prepared, the evaluation is performed through the following steps:

1. The overall data set is divided into 3 second windows with 50% overlap. Spectrograms of each of the 3 second windows are then created. This creates an image of size 256 x 3600
2. The set of spectrograms of the leak-free baseline case is divided into three parts: training, validation and test sets.
3. The CNN-VAE network is built using the training and validation sets from the baseline, leak-free scenario, thus creating a semi-supervised system.
4. The classification error and accuracy of the obtained model is evaluated using the test set of the leak-free case, as well as the 0.25 L/sec leak scenario. The accuracy is contingent on selecting a threshold reconstruction error related to the baseline case.

The training, validation and test sets consisting of 60%, 20% and 20% of the available 2121 leak-free spectograms, respectively; while the test set also included an equal number of leak cases to the 20% leak-free data. The test set utilized insured an equal number of ambient and leak case data thus the accuracy reported is a weighted accuracy. The data used during training and validation were not used for testing, as well no leaky data samples were used to select a threshold, the threshold was determined statistically based on the training data assuming an allowable Type 1 error percentage. Only the test set

is used to determine the classification accuracy. Using the training and validation sets, the connection weights of our network are adjusted. During the model fitting, a function was added that saved the weights yielding the lowest loss value. These weights were only updated if the loss value was smaller than the previously saved loss (associated with the weights saved).

The CNN-VAE spectogram reconstruction model is then used to generate the reconstructed spectrograms using the test set described. Leak detection performance is evaluated by estimating the receiver operating characteristics (ROC) curve. An ROC curve using various thresholds was constructed, and can be seen in Figure 5.13. The area under ROC curve (AUC), is an indicator for the accuracy of leak detection [Domingues et al., 2018, Güvenir and Kurtcephe, 2013]. The AUC value ranges from [0, 1], in which values closer to 0.5 indicate performance comparable to chance and values closer to 1 OR 0 indicate an almost perfect predictor. The AUC score for the ROC curve in Figure 5.13 is 0.974.

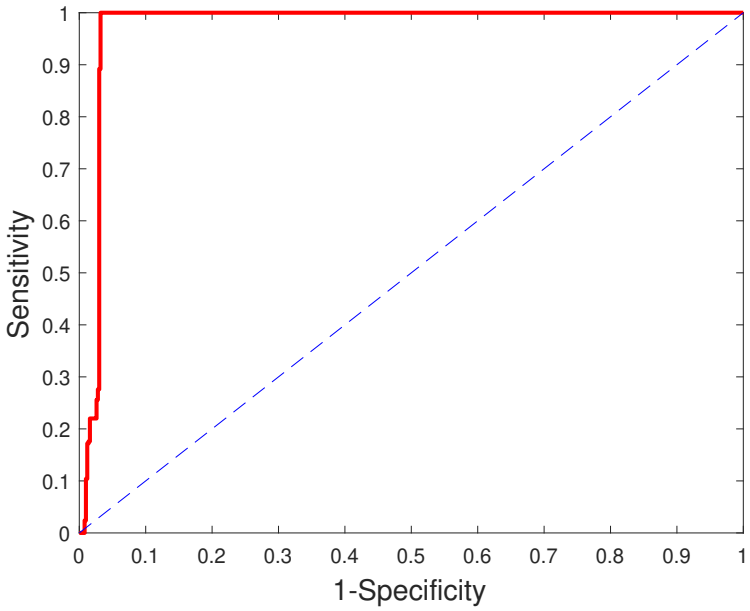


Figure 5.13: ROC curve of 0.25 L/sec leak (6.35 mm valve). Showing the Sensitivity (True Positive Rate) versus 1-Specificity (False Positive Rate).

The mean and standard deviation of an assumed normal distribution of the reconstruction error of the training data is computed and a threshold is set at two standard deviations from the mean, thus encompassing 97% of the training data. That is, an as-

sumed false detection rate of 3% is permitted. A reconstruction error threshold of 299.85 was selected, which represents two standard deviations above the mean of the training data reconstruction error distribution.

Using the test set and model described, an accuracy of 97.2% was observed. Furthermore, a precision of 92% and F1-score of 96% were observed. The overall classification sensitivity, specificity, etc. can be found in the confusion matrix outlined in Table 5.1. The low type I and type II errors are indicative of the effectiveness of the proposed methodology for leak detection.

Table 5.1: Confusion matrix of proposed network classification rates for 0.25 L/sec leak (6.35 mm valve).

True Class	**Predicted Class**	
	Leak	Non-leak
Leak	100%	0%
Non-leak	4%	96%

Figure 5.14 demonstrates the responsiveness of the classification accuracy with leak size and selected threshold (e.g., one through four standard deviations higher than the mean reconstruction error of the training data). The second, larger leak case, produced with a valve diameter approximately four times the size of the smaller leak. Results demonstrate that using a threshold of two standard deviations is a robust choice.

The selected threshold lies on the upper end of the training data reconstruction error distribution and thus calibration using leak scenario data is not necessarily needed if the reconstruction error is modeled as a statistical distribution and the threshold taken as two standard deviation above the mean, thus encompassing 97% of the baseline class reconstruction error. The threshold can be varied depending on the importance of type I vs. type II error associated with the desired application.

While the proposed autoencoder methodology is a useful in determining leak sensitive features within the spectrograms, with results achieving high accuracy (97% classification accuracy), it is important to note that this study is limited to the task of leak detection for environments with noise levels comparable to the test bed. If the baseline system is sufficiently noisy, such as during high demand hours, detection of minimal changes to this scenario would not be distinguishable with the proposed framework. As well, this method is very computationally taxing and significant processing and time is required for large scale

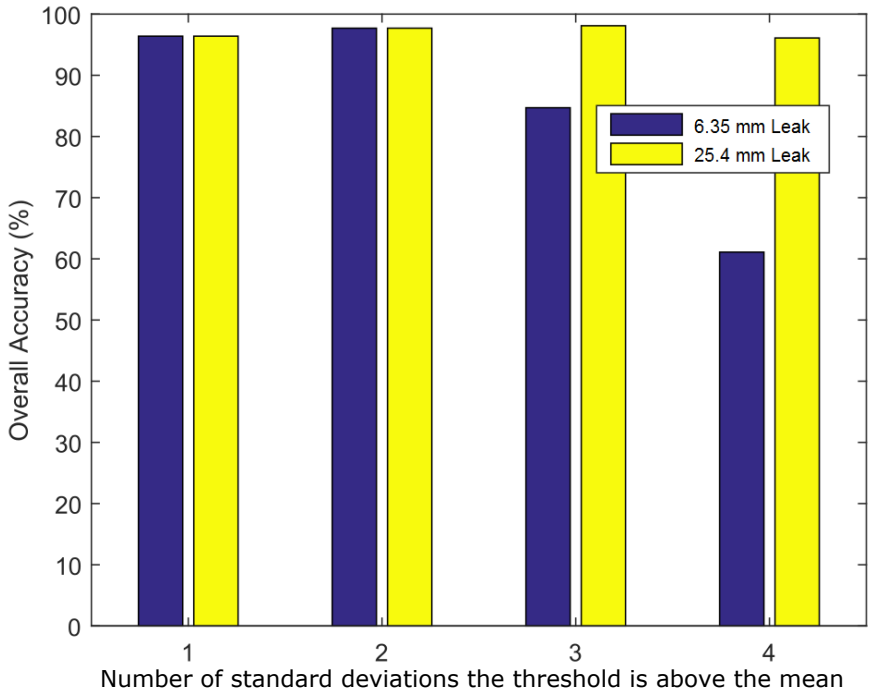


Figure 5.14: Accuracy of ANN model in detecting anomalies for different detection thresholds for two leak sizes. Overall accuracy represents the percentage of correctly labeled test instances (since the test set is equally weighted with leak and leak-free data this value equates a weighted accuracy score).

deployment. Finally this method is also limited to detection and can not be expanded to localization, which is a requirement for useful field deployment.

5.5 LP Results

While the previously described methodologies yielded adequate results, when field deployment is considered, the need for significant parameter tuning in the proposed SSA method, and the computational inefficiency of the NN method must be addressed. Long-term continuous monitoring applications require tools which balance noise robustness and computational complexity so that the process of generating leak sensitive features is relatively simple, while providing good robustness to background noise and good detectability.

In Chapter 4 LP was presented as a parametric modeling technique to detect and locate small leaks in pressurized water pipes. This section proposes the use of LP as a semi-supervised data-driven anomaly detection approach utilizes the features extracted from the LP coefficients representing the underlying acoustic signals. LP was selected because of its computational efficiency, as well as it's ability to be expanded to localization. In terms of leak localization, correlation techniques generally rely on having access to relatively long time segments. However, transmitting large data samples for online monitoring applications is expensive and not feasible for wide scale field deployment. Compared to correlation techniques employing raw signals, it is shown that shorter segments of LP reconstructed signals can achieve similar levels of accuracy as those employing longer segments of raw time series, which is a key advantage in long term online implementation applications.

5.5.1 Data Processing

The acoustic signals collected using the second iteration of the laboratory test bed are first separated into a series of individual time-segments, of approximately 3.5 seconds duration, with 50% overlap between adjacent frames. Each segment of the measurement signal generates one sample in the feature matrix. Hence, the number of samples used for the analysis is equal to the number of frames. To avoid aliasing, a hamming window is applied to each time frame prior to extracting the $p = 50$ order LP-PCA features.

The overall analysis consists of two steps. In the first step, a feature analysis is performed employing features from both normal and leak data samples. The second step involves anomaly detection where the model is trained based only on the normal data and the trained model is validated from both normal and leak data samples. Upon extracting features from each frame, the feature samples are thus divided into different subsets of samples accordingly. For the feature analysis, it is assumed that sample of leak data are available; hence, 20% of both the normal and leak data are randomly chosen for feature analysis and the remaining samples are retained for anomaly detection. The anomaly

detection methodology is performed in two phases, training and testing, through a bootstrapping approach with 100 iterations, in which the data samples for each phase are chosen randomly. For the training phase, the normal data is employed, in which 70% of the remaining data samples (after feature analysis) are used, and the remaining 30% of the samples are employed for testing. For an unbiased testing of the trained model, an equivalent number of leak samples are added to the normal samples. The current study investigates the *No flow* and *Flow* cases separately, thus creating two scenarios for leak detection: i) *No flow*-normal and *No flow*-leak; and ii) *Flow*-normal and *Flow*-leak. For the purposes of comparison, peak, mean, standard deviation, root mean square, crest factor and energy [Li et al., 2017] (collectively referred to as time-domain (TD) features hereafter) are also extracted from the time segments.

Leak localization in WDNs has been widely reviewed in literature [Candelieri et al., 2014, Farley et al., 2013, De Silva et al., 2011, Bracken and Johnston, 2009, Osama Hunaidi, 2006, Hunaidi et al., 2004, Bond et al., 2004, Ozevin and Yalcinkaya, 2013]. One major limitation of these methods is that significant data transmission is required for processing. Leak localization is achieved by applying the LP method to filter the signals from two sensor locations followed by application of the correlation method. This method allows for a reduction in the transmission requirements by approximating the signal with a lower dimensional model. The LP filter corresponding to the prediction model in equation 4.5 is represented as,

$$P(z) = \sum_{k=1}^{p} a_k z^{-k} \tag{5.7}$$

where the output and the input of the filter is respectively, $\hat{x}(n)$ and $x(n)$ as shown in Figure 5.15. First, the model coefficients a_k for a given signal $x(n)$ are estimated employing the auto-correlation method, followed by estimating the reconstructed signal $\hat{x}(n)$ through the application of the LP filter as shown in equation 5.7.

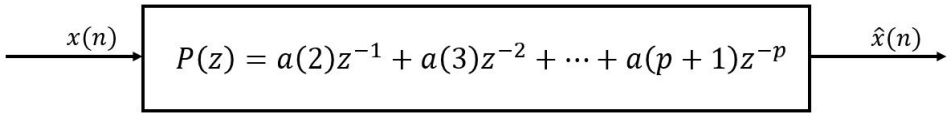


Figure 5.15: Application of LP filter in reconstruction of signal

The cross correlation (described in Section 3.2.6) of two signals (i.e., the two sensor locations S1 and S2) is applied to the two filtered measurements to determine the the time lag (τ).

The location of a leak, D_1 from a reference sensor, is calculated using Gao et al. [2004]

$$D_1 = \frac{D - c\tau}{2}, \tag{5.8}$$

where, c is the sound propagation velocity derived in Section 3.1.3 in the pipe and D is the distance between the two sensor locations, as depicted in Figure 5.16.

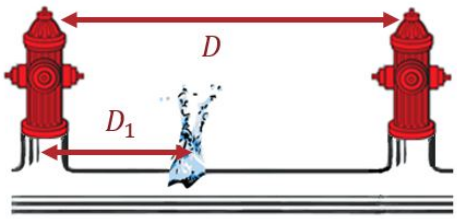


Figure 5.16: Localization parameters for cross-correlation based distance from a reference sensor.

5.5.2 Feature analysis

Prior to detecting anomalies, a feature analysis is performed in order to investigate the performance of the LP-PCA features in separating the baseline and leak data sets. Feature analysis plays an important role before building an anomaly detection model, as the selection of strong features alleviates the need for any preprocessing of the data, thus reducing the computational complexity of the problem. In general, if the separating capacity of the features is weak (i.e. the distributions are not sufficiently separable), the detection model would have poor performance and vice versa. It should be noted that for detection data from *Sensor 1* is used.

The histogram plots of the first LP-PCA features for baseline and leak events are presented in Figure 5.17 with respect to *No flow* and *Flow* scenarios. A general observation on the LP-PCA features in the figure is that the LP-PCA features perform well in separating the two events for both scenarios, though it is clearly more separable for the *Flow* scenario as compared to the *NO flow* scenario. This is a distinct advantage when dealing with field data, which is typically associated with high background noise and more closely simulated by the *Flow* case. The measure of similarity of the two probability distributions is also clearly shown from the estimated Bhattacharyya distances (B_d), described in Section 3.3.6, on the histogram plots. The values of B_d ranges from 0 to ∞, where 0 represents no separation between two distributions and ∞ points towards no similar instances between the

samples. It can be seen that in case of *No flow* scenario, two of the three LP-PCA features possess relatively higher values of B_d (-6.86 and -1.82) showing the capability of these features for separating the leak event from the baseline or leak free event. In case of *Flow* scenario, all the features perform very well, specifically the two features with B_d value as $-\infty$ (subplots (b) and (f)) implying no overlap between the two events.

It should be noted that with different data sets, the observations may be different; that is to say the above observation for the *No flow* and *Flow* scenarios is dependent on the sample sizes for feature analysis and the particular data sets available for analysis. The results are expected to vary based on the noise level and structure of the data sets. Nevertheless, the current study pursues the performance of the feature sets coupled with the GMM in detecting leaks for both scenarios independently, rather than comparing them against one another.

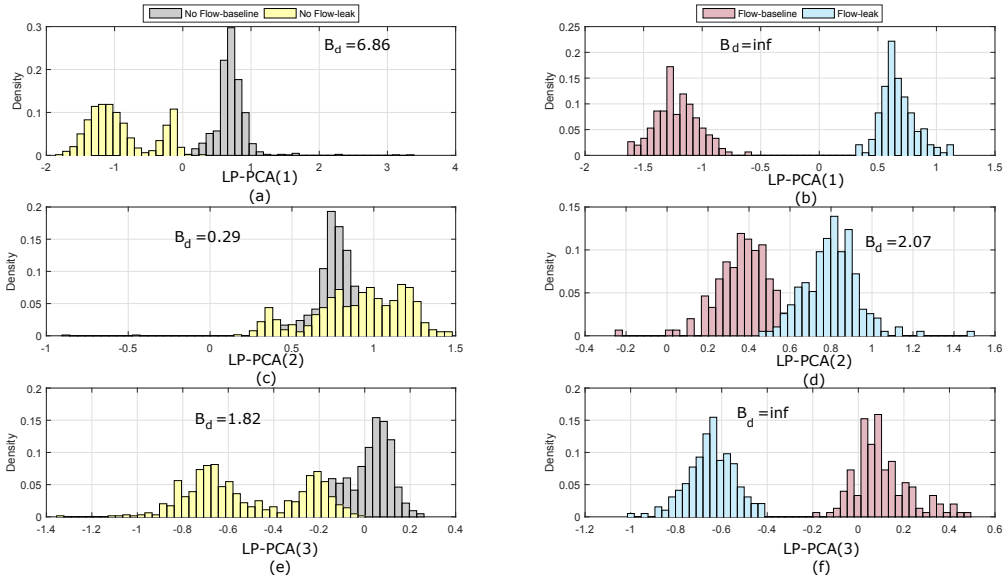


Figure 5.17: Histograms for the first three principal components of LP coefficients i.e., LP-PCA(i) with component $i = 1,\ 2,\ 3$ for *No flow* and *Flow* scenarios

5.5.3 Leak detection

For this stage of analysis, training and test sets of data samples are prepared as discussed in Section 5.5.1. Both the training and testing sets are selected randomly for the 100

bootstrapping iterations and the average of performance measures in predicting leak events are estimated. Prior to the construction of the predictive model of leak detection, the optimal number of components are selected based on the training data set. As shown in Figure 5.18, the K values corresponding to the minimum BIC values are 2 and 1 for the LP-PCA features in case of *No flow* and *Flow* scenarios, respectively. The number of components K depends on the data structure and complexity. With more complex and non-normal nature of the data set, the number of GMM components to best fit the data increases. Since the data structures corresponding to the two scenarios studied here are not similar, they require different K values to best fit the data. Following this, the parameters of the K Gaussian mixture components ($\Lambda = [\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}]$) are obtained using EM algorithm and the GMM model is constructed for detection.

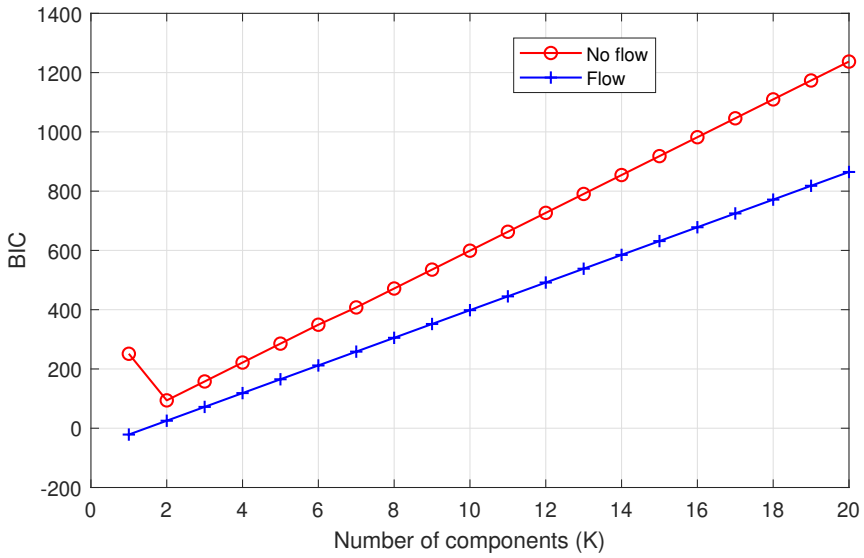


Figure 5.18: Sensitivity analysis of number of components based on BIC for LP coefficients as features

Once the GMM model is constructed from the training data set, the probability density of the test data set ($p_{test}(\mathbf{x}^{test}|\Lambda)$) is determined using equation 3.39 with the estimated number of components and GMM parameters from the training data set. In order to determine if a test sample is *normal* (i.e., corresponding to baseline), p_{test} is compared with a threshold value, which is established based on different percentile values of the probability density function (pdf) of the training data set (p_{train} estimated through equation 3.39). The 2.5, 5.0, 7.5, 10.0, 12.5 and 15.0 percentile values are adopted in the current study and

the detection accuracies are investigated based on different threshold values. Nevertheless, the detection threshold can be adjusted according to the users' preferences and based on the availability of training data. Upon establishing the threshold value (d_t), each test set instance is then flagged as leak or an anomaly if $p_{test} < d_t$.

First, both the *No flow* and *Flow* scenarios are studied by adopting a 5.0 percentile value for the detection threshold. Then the accuracy of the detection methodology is reviewed for different thresholds. For demonstration, the histogram of probability densities of samples being *normal* (i.e., p_{train}) and the estimated threshold value is presented in Figure 5.19 for LP-PCA features in case of *No flow* scenario. The overall accuracy (A_c), precision (P_c), recall (R_c) and f1-score (F_s) are estimated, and reported in Table 5.2 for the *No flow* and *Flow* scenarios. A general observation from the detection results is that the GMM with LP-PCA features perform very well in detecting leaks with accuracies of 97.62% and 97.32% for the *No flow* and *Flow* scenarios, respectively.

An in-depth discussion of the performance measures under these scenarios is presented below along with comparing the performance of LP-PCA features to the traditionally employed TD features, which include peak, mean, standard deviation, root-mean-square, crest factor and energy [Li et al., 2017], as defined in Section 3.3.9. The normality model for the TD features are built following a similar methodology as described for the LP-PCA features (Section 5.5.1), in which the number of components are calculated to be 6 and 3 respectively for the *No flow* and *Flow* scenarios.

In the case of the *No flow* scenario, the LP-PCA based features result in higher accuracy of 97.62% (Table 5.2) as compared to the TD features with accuracy of 81.89%. Similar observations are also made for the other performance measures. For example, the precision, which implies what proportion of positive identifications are actually correct, is 100% for LP-PCA features as compared to the value of 75.54% for TD features. However, the recall estimates return close values for both feature sets. This can be explained through the confusion matrix as reported in Table 5.3, where the first row represents values of TP and FN and second row has the values of FP and TN. As defined in Section 3.3.8, precision depends on TP and FP. While the TP reported for both feature sets are almost the same, FP is very high for TD features leading to lower estimates of precision. On the other hand, recall depends on TP and FN (Section 3.3.8), which are very similar for both the feature sets yielding close recall values (95.46% and 93.23% respectively). Another way of evaluating the performance of the detection framework is through the ROC curves as shown in Figure 5.20(a). By comparing the two ROC curves, it can be concluded that the proposed approach of anomaly detection based on integrating LP-PCA features with GMM significantly enhances the leak detection performance resulting in AUC value of 0.99 as compared to 0.93 for TD features in *No flow* scenario.

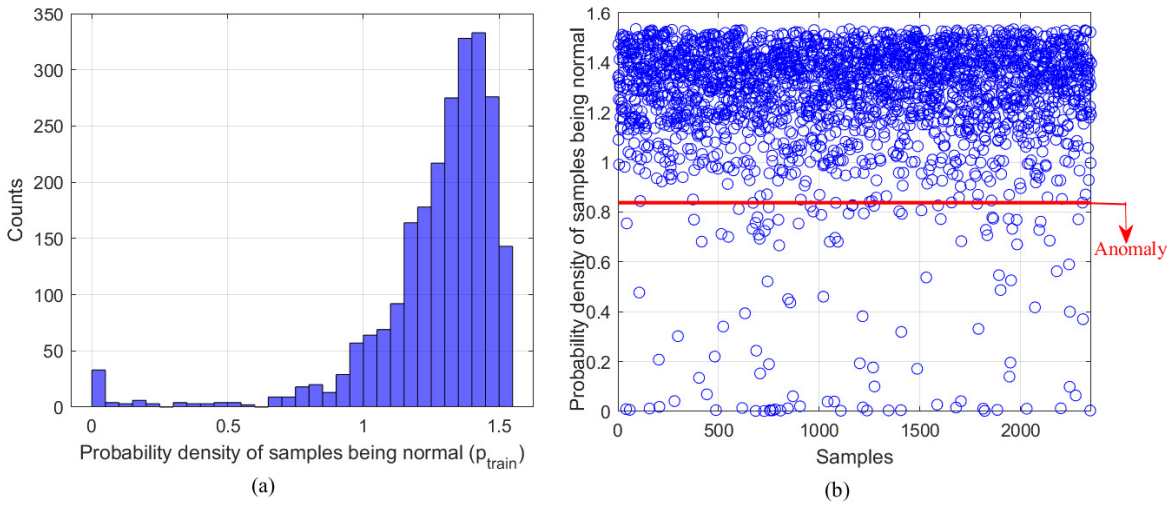


Figure 5.19: (a) Histogram of p_{train} and (b) threshold of samples being *normal* in case of LP-PCA features under *No flow* scenario

Similar observations are also made for the *Flow* scenario from the performance measures as reported in Tables 5.2 and 5.4. In the case of the Flow scenario, the TD features perform very poorly with an accuracy of 55.12% as compared to 97.38% for LP-PCA features. In specific, TD features yield a very high FP (83.63%) as shown in Table 5.4, which is reflected through very low precision measure (52.91%). On the other hand, the LP-PCA features significantly increases performance of the leak detection methodology resulting in 0% FP which leads to precision measure of 100%. The ROC curves for both feature sets in Figure 5.20(b) and subsequent AUC values also point towards the high performance of LP-PCA features as compared to the TD features in this scenario.

Table 5.2: Performance of GMM in detecting leaks

Performance measure (%)	**No flow**		**Flow**	
	LP-PCA	TD	LP-PCA	TD
Accuracy	97.62	81.89	97.38	55.12
Precision	100	75.54	100	52.91
Recall	95.46	93.23	95.06	72.94
F1-score	97.56	84.09	97.3	67.65

Table 5.3: Confusion matrix for *No flow* scenario

True Class	**Predicted Class**			
	LP-PCA features		TD features	
	Normal	Anomaly	Normal	Anomaly
Normal	95.24%	4.76%	95.04%	4.96%
Anomaly	0%	100%	31.26%	68.74%

Table 5.4: Confusion matrix for *Flow* scenario

True Class	**Predicted Class**			
	LP-PCA features		TD features	
	Normal	Anomaly	Normal	Anomaly
Normal	94.76%	5.24%	93.86%	6.14%
Anomaly	0%	100%	83.63%	16.37%

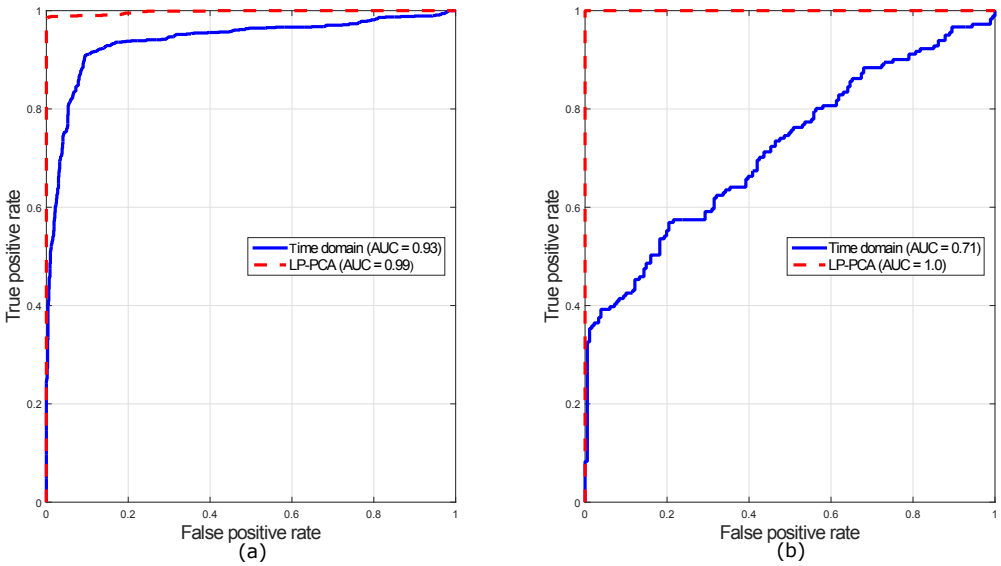


Figure 5.20: ROC curves for TD and LP-PCA based features in cases of (a) No flow and (b) Flow scenarios

It should be stressed that the above observations and conclusions are based on the choice of detection threshold and the performance of the GMM in detecting leaks is likely to be sensitive to this threshold value. A sensitivity study is conducted based on threshold

values in accordance with different percentiles of the PDF of samples being *normal* and the overall accuracy of the GMM is shown in Figure 5.21. As shown, the accuracy of the GMM in detecting anomaly depends on the choice of the threshold level. For example, the highest accuracy for TD features are associated with higher percentile values, such as 10.0 and 25.0 for the *No flow* and *Flow* scenarios, respectively. In case of these features, the highest accuracy primarily depends on the relative decrease and increase in TP and TN values based on the threshold levels. On the other hand, LP-PCA features requires very low percentile values such as 2.5 for both scenarios. Nevertheless, the LP-PCA features performs very well in detecting leaks with accuracy ranging from 92% to 99% while the accuracy of TD features range from 51% to 88% depending on the detection threshold value and background noise. These differences in the performances of the TD and LP-PCA features are mainly due to the poor separation between the two events by the TD features leading to significant sensitivity of the model accuracy to the threshold levels as compared to the LP-PCA features. Based on large amount of baseline data and small amount of leak data in the overall data set, the threshold can be calibrated in the future to yield the maximum accuracy for the detection algorithm.

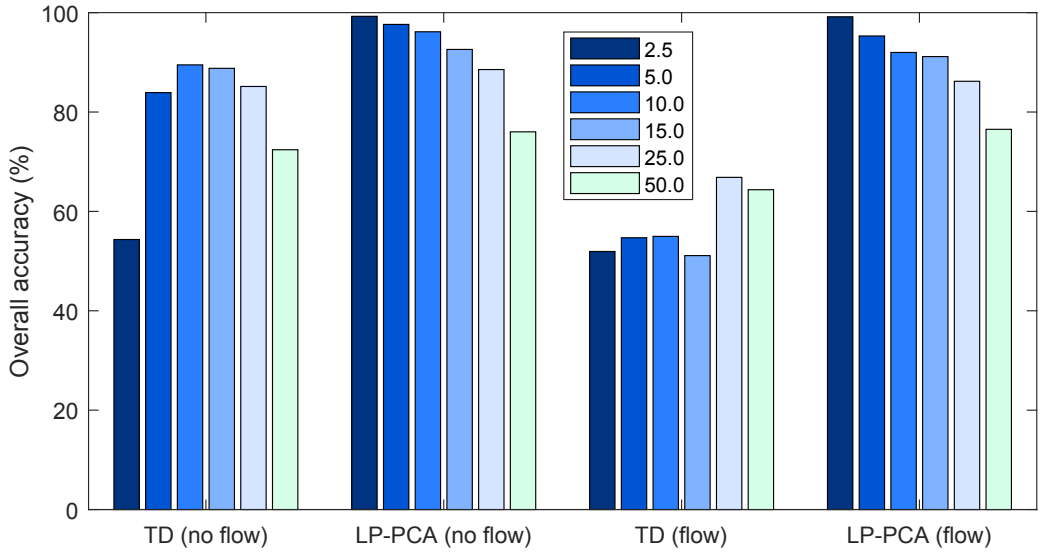


Figure 5.21: Accuracy of GMM model in detecting anomalies for different detection thresholds based on different percentiles of the PDF of the normal samples

5.5.4 Leak localization

Leak localization based on correlation techniques generally rely on having access to relatively long time segments. However, transmitting large data samples for online monitoring applications is expensive and has often been viewed as one of the main hurdles in cost-effective leak monitoring applications. The computational simplicity as well as the ability of LP method to be deployed locally to reduce the overall volume of data transmitted can be advantageous for long-term applications. The current study proposes the LP method as a filtering tool, where the reduced filtered data is used to localize leaks. One of the significant advantages of this approach is that reasonably good level of accuracy in terms of localization can be achieved with relatively short segments of LP filtered data, as described next.

The cross correlation method is applied, as described in Section 5.5.1, on both raw and LP filtered signals of time lengths ranging from 5 s to 30 s. Using parameters specific to the laboratory test bed, the theoretical value of the sound propagation velocity (c) in water-filled PVC pipe is first calculated following Section 3.1.3, as 458 m/s [Neutrium, 2014]. This is consistent with the experimentally observed velocity range of 450 m/s to 520 m/s for PVC pipes by Hunaidi and Chu [1999]. It is important to note that acoustic velocity, which plays a central role in localization, is affected by variabilities in both physical, environmental and sensor elements. Hence, errors in localization are to be expected as a result of this stochasticity. For this study, the theoretically determined value of c is employed with the the experimentally determined τ from the cross correlation between the two sensor measurements for localization. Results shown are from ten trials, where D_1 is estimated using equation 5.8 descibed in Section 5.5.1, with $D = 1474.5\ cm$, and the deviation of the estimated leak location from the true location (i.e. percent error) for several cases is reported in Table 5.5. The true D_1 value is 911.9 cm.

As expected, the estimated D_1 in Table 5.5 shows that there is a general trend where longer duration samples produce more accurate results, reflected in a reduction of localization errors. Using raw data alone, the average errors are relatively high for shorter time segments, especially for 5 sec. However, the average localization errors are reduced for shorter time segments (5 sec) with the use of even low order LP filters. Most notably, the error in estimating D_1 is decreased from 31.43% to 13.45% after applying an LP filter of order 10. The application of higher order LP filters however does not improve the localization estimates in terms of average localization errors. Also, importantly, the correlation technique is shown here to produce acceptable results even in more complex geometries, especially when long-term monitoring techniques are viewed as a first step yielding coarse localization information prior to deploying more accurate inspection methods.

Table 5.5: Comparison of average estimated D_1 (in cm, without parenthesis) and percentage error (within parenthesis) for raw and LP filtered signals for different signal duration

Time length (s)	Raw signal	Filtered signal using LP of order p		
		$p = 10$	$p = 25$	$p = 50$
5	1198.4 (31.4%)	1030.4 (13.5%)	1026.7 (13.3%)	1026.7 (13.3%)
10	1011.2 (12.9%)	1021.3 (12.7%)	1020.1 (12.5%)	1015.7 (12.4%)
15	1012.3 (11.8%)	1019.0 (12.1%)	1022.4 (12.4%)	1013.4 (11.7%)

This section presents a computationally efficient method of semi-supervised leak detection as well as localization of water distribution systems using the concept of LP. Results from this study show that LP can be used to extract leak-sensitive features to facilitate reliable leak detection, while at the same time can be used effectively for leak localization. Application of a computationally simple GMM model for leak detection shows that this method can be employed for leak detection in long-term monitoring field applications. In addition to using it as an effective leak detection tool, results show that LP is able to achieve good signal approximation based on the underlying model coefficients, thereby reducing the length of time traces required for correlation. Laboratory experimental studies show that the resulting LP spectral envelope captures the overall behaviour of system, while being able to differentiate between various leak and no-leak cases in both low and high background noise cases.

5.6 Summary

This chapter presents a laboratory case study for the application of both model free and model based methods for semi-supervised leak detection in WDNs. While both methods of data collection provide important insights and results pertaining to leak detection, there are a few key differences between the two. Model free methods in general can be helpful in providing insight into the underlying structure and information contained within the data itself. The first method employed that is model free in this chapter are SSA, which is used as a key step in pre-processing the data, following by the application of a OCSVM

classification methodology. If only the best results are reported for this method, i.e. the optimal parameters are chosen, the results from this method are reasonably good ($0.8 < AUC < 0.9$). However, if these parameters are not appropriatelyy selected the results are quite poor, approaching random ($0.3 < AUC < 0.6$). This coupled with the need for adequate selection of SSA components, and the fact that this methodology was tested on only the first iteration of the laboratory set up, not incorporating the added complexity of the updated system, demonstrates the difficulties this method would face in a large scale field deployment.

The other model free method reviewed is a deep autoencoder methodology. While this yielded extremely high accuracy (97% classification accuracy), and effectively eliminated the need for feature engineering, it is associated with significant computational overhead. As well, this high accuracy was obtained using only low noise data tested on the simplified first iteration of the laboratory test bed. While this method is promising, the large computaitonal overhead also demonstrates significant limitations with regards to field deployment.

One of the main limitation of these two methods is that they are only applicable for leak detection and cannot extended to leak localization within the same methodological framework. In contrast, the model based LP method coupled with a GMM classification methodology can be a computationally efficient alternative for leak detection, while being able to extend effectively to leak localization as well. The results demonstrate the benefits of the proposed LP based methodology for wide scale autonomous field deployment over the model-free methods, while providing similar accuracies in detection and being able to localize as well. As well, emphasis should be made on the application of LP for leak localization, for which these model-free based methods could not be extended.

While the results from this study are promising, it is important to acknowledge the limitations of the study. First, despite all the efforts taken to ensure data acquired from the system is adequately representative of data in an actual field setting, given the laboratory constraints it is not possible to simulate actual field conditions. This study focuses on detecting and localizing the presence of relatively small and single leak cases (one at a time) within a given section of the distribution system. It does not consider the effects of different leak opening sizes and multiple breaks within a given section, nor the effects of surrounding soil medium.

Chapter 6

Field Results

6.1 Introduction

Building off the conclusions drawn from the previous chapter, this chapter shows that LP performs well for leak detection in a field test-bed case study of a WDN. As well, a new multi-step localization method is able to achieve reliable localization results. This therefore shows that the resulting LP spectral envelope captures the overall behaviour of the system, while being able to differentiate between various leak and no-leak cases. The data analyzed int his chapter is obtained through an extensive field instrumentation program, which took over an year to complete; starting the summer of 2017 through fall of 2018. The field test bed consists of a subsection of a DMA which is unisolated and for whom the variables which may be affecting the baseline of the system are unknown. The data used in this chapter is acquired from hydrophones located at various hydrant locations, using retrofitted hydrant stems, thus eliminating the need to flood the hydrant; this enables the proposed method to be viable for long-term monitoring. The data is then processed and analyzed using the LP method described in the previous chapters.

The proposed retrofitted hydrant system aims to overcome many challenges and limitations associated with long-term passive monitoring of WDN systems. It offers a convenient and affordable solution for event-detection in WDNs, while maintaining minimal installation cost. The hydrant retrofitting aspect of the proposed system allows it to be installed easily in many locations without incurring the expensive installation costs associated with excavating pipes for installation and more invasive monitoring approaches. By simply retrofitting a fire hydrant, the installation can be completed within hours and water distribution in that area would not be affected. The computational efficiency of the proposed

methodology allows for long-term monitoring in field applications. In addition, the localization methodology allows for two granularities of localization, while using relatively short time signal lengths. The short time signal lengths minimizes the data-transmission requirements, which is one of the main impediments in other full-scale implementations of leak-detection technology.

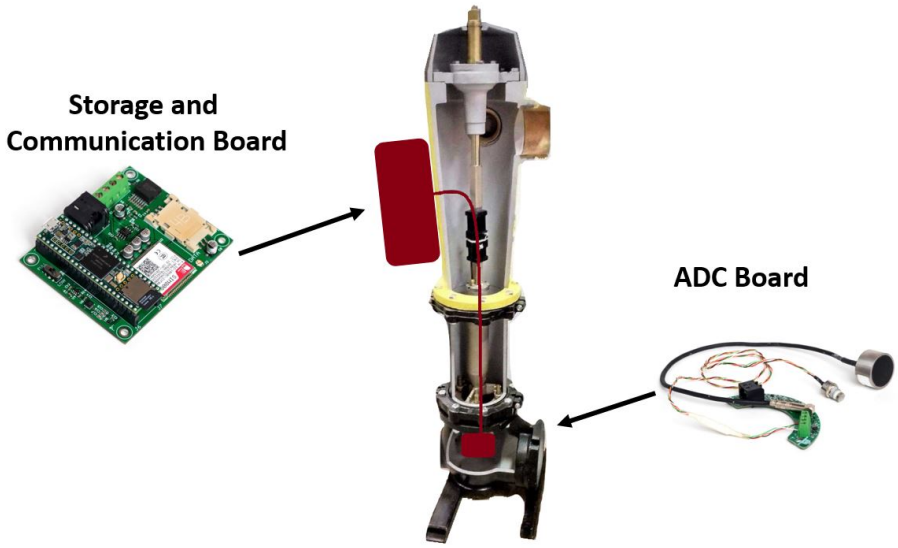
This chapter is organized as follows: first, the hardware system used in the field data collection is described, which includes the both the hardware used to retrofit hydrants and the data collection system; next the field test bed and it's associated data set are outlined; the data is then analyzed in order to better characterize the baseline and determine the need for advanced classification methods; next, the overall methodology used for the field study is described, including signal pre-processing, the algorithmic steps involved, and the results for the leak-detection and localization tasks using LP; finally a summary of the main conclusions and limitations are reported.

6.2 Sensors and Data-Acquisition System

The proposed sensors and the full data-acquisition system were developed during the course of this project; a commercial version of this system is currently being offered as an off-the-shelf product, partly resulting from the activities undertaken as a part of this research.[1] The state of WDN is monitored using the following four sensors: (i) hydrophone, (ii) pressure, (iii) accelerometer, and (iv) temperature. The data is acquired, stored, and transmitted using a custom-designed data-acquisition system. For this study, only those results obtained using the hydrophone are reported as this sensor type was specifically designed for this end application and has been incorporated into the commercial product, informed by the activities conducted during the course of this dissertation.

Access to the water column within the WDN is often an intrusive exercise, as are not many locations where access is readily available. Previous studies have resorted to inserting hydrophones into the main lines, using access points such as valves, or drilling into the pipe wall directly [Whittle et al., 2010, 2013]. In cold climates, such access is often restricted and avoided, mainly due to risk of freezing, and hence a specially-designed retrofit which can be inserted into existing fire hydrants was used. This is depicted in Figure 6.1a, where the retrofitted sensor housing can be installed easily in existing hydrant locations without the need to excavate pipes or the need to access valve stems.

[1] Digital Water Solutions - https://digitalwater.solutions/



(a) Cross-sectional view of hydrophone mounting on the hydrant stem.

(b) ADC board housing and sensor mount.

Figure 6.1: Hydrophone mounting unit

The data-acquisition system was designed in house to enable analog-to-digital conversion, processing, storage and wireless communication. Basic signal-processing is also enabled at the sensor locations, in order to minimize data transfer overhead and to maximize battery life. Data-acquisition is achieved using a custom analog-to-digital Conversion (ADC) board, with 24-bit resolution. The four sensors depicted in Figure 6.1 are located within a stainless steel unit at the base of the hydrant, with the hydrophone and pressure sensors exposed to the water column. A 24-bit resolution was chosen compared to less complex 16 or 12 bit primarily because there of a large uncertainty and potentially significant range in the acoustic noise levels within WDNs. Power (battery), data storage, and communication modules are located at the top of the hydrant (street level) and the data is transferred along the height of the hydrant using a single Cat5e power-over-ethernet (PoE) cable to the ADC. The processing, storage and communication modules include a Teensy 3.6 micro-processor board located on a custom designed printed circuit board (PCB), a flash memory storage module, a GPS chip receiver with an antenna, and a 2G cellular modem. The firmware was written in C++ with serial peripheral interface (SPI) communications between the storage, communications, and the ADC modules. The software architecture was designed such that diagnostic data is transmitted at set intervals during low demand hours, primarily during the night. Raw data values are stored in the flash

storage module for post-processing and analysis, as required. Data visualization and user alerts are made possible using a web-interface, however this aspect is not described here as it is considered not relevant for this dissertation.

6.3 Field Test Bed

The field test-bed represents a small section of the WDN within a city's distribution system. The test-bed is part of a residential area in south-western Ontario (Guelph), Canada and consists of approximately $1,500\ m$ of grey scale 80 PVC pipe. The full test-bed is illustrated in Figure 6.2.

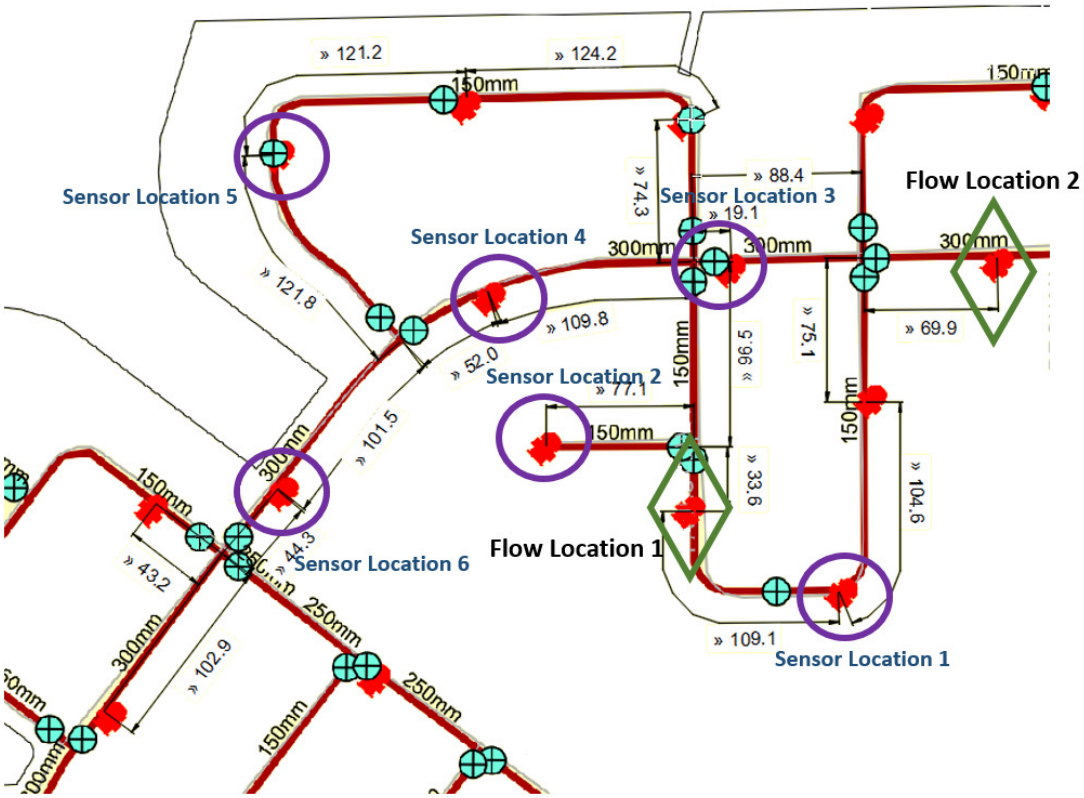


Figure 6.2: Flow and sensor locations in the test-bed; circles indicate sensor locations while the diamonds indicate flow locations.

Approximately $930\ m$ of the test-bed consists of pipes with $15.24\ cm$ diameter, while the remaining $570\ m$ of the pipes have a diameter of $30.48\ cm$. The test region includes two full

hoops, three intersections, and two tees. A total of five fire hydrants were retrofitted with a specially-designed data-acquisition system, while two additional hydrants were selected to simulate leaks. Typical system pressures range from $360 - 380\ kPa$ within the test area, depending on various factors such as demand and pump statuses. The system was not controlled in any way and the typical usage characteristics were not modified during the course of testing. Furthermore, all the hydrants remained fully operational throughout the testing period, despite them being retrofitted with new sensor attachments and data-acquisition systems.

The locations of the simulated leaks were selected to simulate different acoustic impedance characteristics and distances from the sensor locations (*SLs*). Various combinations of intersections and bends between the flow locations (*FLs*) and SLs encapsulate different possible configurations the sensors may face with possible leaks in the system.

6.3.1 Data Collection

Controlled distribution flow events were created by attaching a 2.5 *inch* valve to a hydrant flow location, at which the flow rate is measured. Three dates were selected in the Fall of 2018 for leak simulation tests: October 13th, October 21st, and November 3rd. On each of these test-dates, flow locations, described in Figure 6.2, were simulated between midnight and 4-AM local time. At each of these flow locations, four flow amounts were simulated: 200 L/min, 100 L/min, 50 L/min, and 25 L/min. This test procedure spanned the full three-day period, where, on each of these test days, the two hydrants were flowed in sequence spanning nearly three hours, starting at midnight through to 3AM. These times were selected based on historical information of low-demand hours. Figure 6.3 shows a graph model of the water distribution system, with sensor locations and *flow locations 1* and *2* (hereafter referred to as *FL 1* and *FL 2*, respectively) identified.

On the first field test-day data collected at *SL 1, 2, 4* and *6*; on the second test day, data collected at *SL 1, 2, 4, 5* and *6*; and finally on the third day data collected at *SL 1, 2, 4, 3, 5* and *6*. For each hydrant location, the sequence in which data was collected was as follows: (1) leak-free data, (2) 200 L/min leak, (3) 100 L/min leak, (4) 50 L/min leak, (5) 25 L/min leak, (6) leak-free data. The collection of leak-free data, before and after the leak cases were simulated, allowed for the underlying variability in the system to be sufficiently encapsulated and thus the baseline models could be adequately calibrated.

During the course of all the field tests, spanning a period of nearly two months, between four and six sensor locations (hydrants) were instrumented. Leaks were simulated at the two locations described earlier and shown in Figure 6.3, wherein time was synchronized

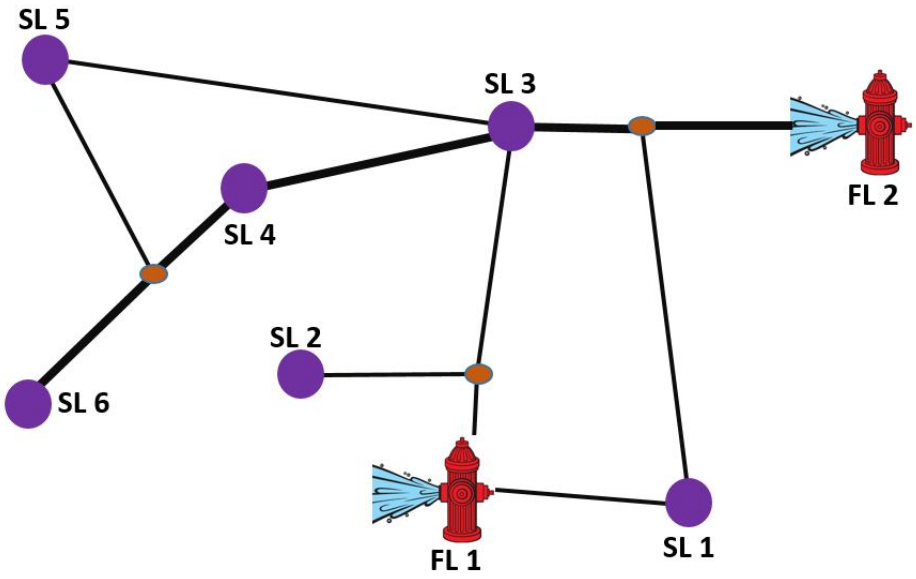


Figure 6.3: Graph model for portion of pipe network layout of the WDN in Figure 6.2. The thicker lines indicate 300 mm diameter lines, while the thinner lines indicate 150 mm diameter lines.

using a FONA 808 MiniGSM + GPS module via cellular network. In order to ensure sufficient variability in the data set, the acoustic signals were acquired on different days and with different flow sizes. Data was collected at a sampling frequency of 4 kHz, and the voltage signals were pre-amplified with 20 dB gain, prior to storage and transmission. The data-acquisition rate was set based on laboratory tests and preliminary signal processing was conducted.

6.4 Baseline Characterization

Before a complete analysis is undertaken, it is informative to understand the variability that exists in the acoustic fields within a live WDN. There are several such factors that cause such variability and are mostly uncontrollable from a monitoring program standpoint: automatic pumps, usage and repairs to name a few. The exercise to be described next attempts to study the variability in the acoustic pressure measured by the hydrophones from the retrofitted hydrant systems. The objective is to enquire whether relatively well

established statistical hypothesis tests are adequate to determine the presence of the leak within the natural variability which exists in the acoustic field environment.

The short-time auto-correlation function (STACF) and its counterpart, the short-time Fourier spectrum (STFS), capture the leak-induced resonances through the estimated LP coefficients. This is evident, e.g., in the STACF corresponding to *FL 1* for two flow cases, 200 and 50 L/min, at all sensor locations (see Figure 6.4).

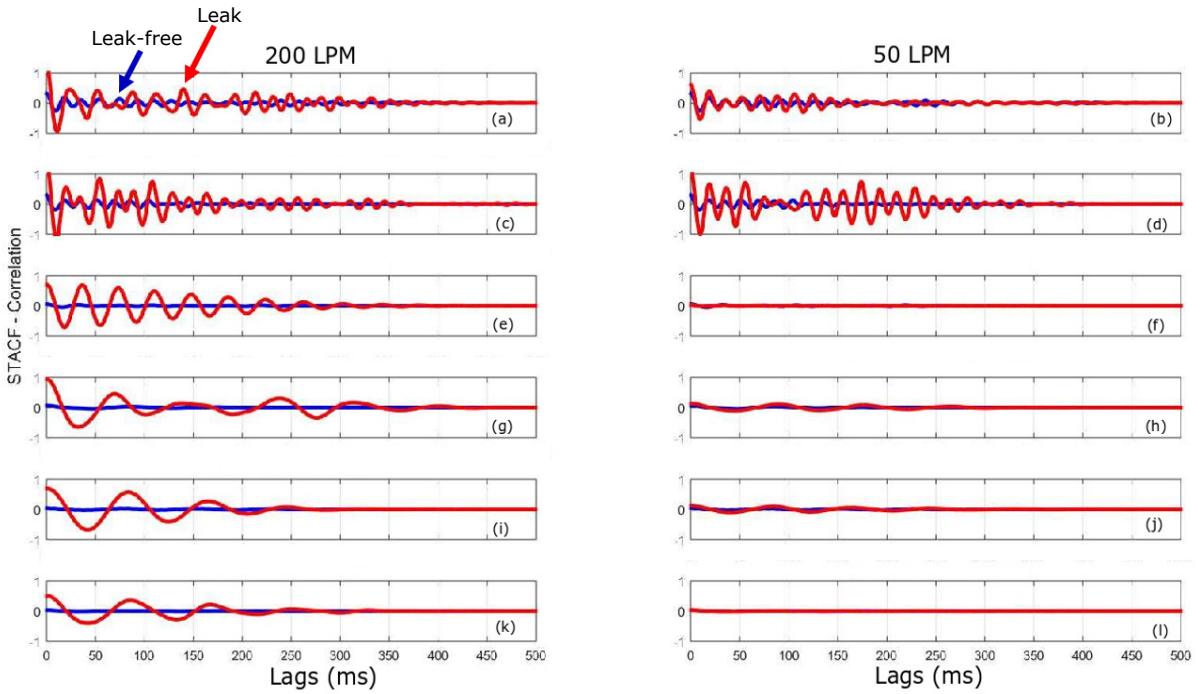


Figure 6.4: STACF for all sensor locations for *200 L/min and 50 L/min*. In which (a), (b) represent SL 1; (c), (d) represent SL 2; (e), (f) represent SL 3; (g), (h) represent SL 4; (i), (j) represent SL 5; and (k), (l) represent SL 6.

A general trend can be observed in the STACFs; a distinct periodicity is introduced into the system when the leak is present and this periodicity is prominent at larger flow amounts. The difference in the magnitude of the leak energy is predominantly visible in the low-frequency region of the spectrum, which is consistent with laboratory findings and previous studies from [Muggleton and Brennan, 2004]. The periodicity in the correlation functions is more evident for leak cases involving larger flow amounts (the 200 L/min case versus the 50 L/min case) and for sensors located closer to the leak location, compared to those further away. It can be seen that for the smaller leak case of 50 L/min flow, when the

sensor is located further from the leak source (such as Figure 6.4 (f), (h), (l)), the STACF of the *leak-free* and *leak* cases are practically indistinguishable. In contrast, sensors very close to the leak source produce clearly distinguishable STACFs, even for smaller leak sizes (such as Figure 6.4 (b), (d)).

While the strength of STACFs reduce further from the source and such differences not visually distinguishable, the Bhattacharya distance (which shows the divergence of two statistical distributions obtained from LP-PCA features) indicted good separability for the same data set (*FL 1*), wherein five out of the six sensor locations have a Bhattacharya distance that is indicative of little to no overlap for at least one of the LP-PCA features. This is the case for the 200 L/min flow case, as well as for all six sensor locations for the 50 L/min flow case. This shows that LP-PCA features are capable of being able to separate the *leak* event from the *leak-free* case for both large (200 L/min) and smaller (50 L/min) flow cases.

6.4.1 Hypothesis testing

Prior to describing the LP method for leak-detection, it is instructive to understand the issues in applying standard statistical tests as a means of leak-detection. Although simple statistical tests would be the most ideal, they are prone to significant amounts of Type I error. As such more sophisticated methods must be considered.

T-test

Assuming that the historical baseline condition for the system is available *a priori*, upon measuring new test samples, a hypothesis test can be conducted to determine the presence of a leak. The null hypothesis is posed as follows:

$$H_0 : \mu_{x_p} = \mu_{x_s}, \tag{6.1}$$

where $\mu_x(n) = \frac{\sum^N \sqrt{\frac{1}{n}(x_1^2+x_2^2+\cdots+x_n^2)}}{N}$ is the average of N signal root mean squares (RMS), each computed based on n observations, produced in the same test area. Given a large historical sample set, μ_{x_p} can be regarded as the population mean; μ_{x_s} constitutes the sample mean. Generally, it is easy to obtain μ_{x_p} under baseline (leak-free) conditions. By simulating leak and leak-free cases, the corresponding sample averages can be used to check against the null hypothesis to assess Type I (falsely rejecting H_0 when there is no leak) and Type II (failing to reject H_0 when there is a leak) errors.

A test statistic, $T(\mu_{x_s})$, is a function of the desired confidence interval. Therefore, the critical region (CR) is defined as,

$$CR = [|T(\mu_{x_s}) - \mu_{x_p}| > z_\alpha * \frac{\sigma_{x_p}}{\sqrt{n_{x_s}}}] \tag{6.2}$$

The null hypothesis is rejected when equation 6.2 does not hold true. The threshold of $\mu \pm z_\alpha * \frac{\sigma}{n}$ can be adjusted by changing z_α, therefore changing the probability of false alarm, i.e., the Type I error.

The hypothesis tests, as formulated above, are prone to Type I errors. In order to demonstrate this, 1,500 5-second average RMS samples were randomly drawn from nearly 13,000 samples to estimate μ_{x_p}, which constitutes the baseline acoustic data that is spread over two months of a measurement campaign. This yielded $\mu_{x_p} = 1.63\ Pa$ to be used in the hypothesis test. On three separate dates, outside the period used to create the baseline statistics, μ_{x_s} was calculated by randomly drawing 10 samples corresponding to each day. The results of the null hypothesis tests are as follows:

$$\textbf{Day1} : t(9) = 20.109, SEM = 0.102, p < .00001$$
$$\textbf{Day2} : t(9) = 1.045, SEM = 0.389, p = .323$$
$$\textbf{Day3} : t(9) = -4.872, SEM = 0.102, p < .001$$

It is clear that these results are associated with Type I errors, for two out of three days tested. Since the leak-free data is more often being classified as falsely rejecting H_0 despite there being no leak.

Z-test

Similarly, these three dates can be considered a single data set, wherein 40 of the 150 samples are randomly selected. A **z-test** is applied, which yielded $z = 1.772, SEM = 0.244, p = .0767$. This means, given that H_0 is true, the probability of observing something less likely than what was observed is 7.7%. This is just around the threshold ($p \leq 0.05$), which is consistent with the results produced by the t-test. Given that one of the three dates was not rejected, when all three test dates are randomized, it is reasonable that these results would lie close to α. If the confidence intervals on the null hypothesis are widened, Type II error increases significantly before the Type I error can be reduced.

Wilcoxon Rank-Sum test

The use of the **t-test** and **z-test** are based on the assumption that the true distribution of the baseline data is known *a-priori* and are sufficiently large, so that a normal distribution is observed, however the lack of years of data make this simply an assumption. Therefore, if the leak and leak-free data sets are simply taken as two independent data sets (i.e. assuming the underlying distribution of the two data sets is unknown and there are no pairings between the observation groups) the **Wilcoxon rank-sum test** can be applied.

The 2-sided Wilcoxon Rank-Sum test, when applied to the hydro-acoustic data, results in either Type I, Type II error, or both. In order to demonstrate this, 3-second RMS samples were generated for the data set associated with each date and case (baseline acoustic data and leak acoustic data). Two versions of this null hypothesis test were run, and the results at the 5% significance level are as follows:

1. The baseline data on each date was compared with the leak data from that same date:

$$\begin{aligned}
&\textbf{Day1} : W_{rs} = 64091, z = -17.9622, p < .00001 \\
&\textbf{Day2} : W_{rs} = 145435, z = 0.2530, p = .8003 \\
&\textbf{Day3} : W_{rs} = 360571, z = -0.3595, p = .7192
\end{aligned}$$

2. Two subsets of baseline data taken approximately a half hour apart are compared for each date:

$$\begin{aligned}
&\textbf{Day1} : W_{rs} = 19691, z = 4.8137, p < .00001 \\
&\textbf{Day2} : W_{rs} = 7818, z = -13.9417, p < .00001 \\
&\textbf{Day3} : W_{rs} = 43090, z = -11.4859, p < .00001
\end{aligned}$$

Based on the results of the first set of Wilcoxon Rank-Sum tests run it is clear that these results are associated with Type II errors, for two out of three days tested. Conversely, based on the results of the second set of tests run, for all three days tested the results are associated with Type I errors. These results indicate that for two of the three test dates both Type I and Type II errors occur, while on the remaining test date Type I error is observed.

These tests were repeated using baseline data from the day prior as well as three days prior leading up to the leak test date and the similar trends were observed. Specifically when the data from the day prior to the leak date, but during the same time frame, was used the results were clearly associated with Type I errors. As well, when multiple days prior, but at the same time of day, was used the results were also clearly associated with Type I errors. The significant variability of the systems baseline day to day leads to this significant association with Type I errors. Similarly, it is expected that as the baseline is increased to include more days the Type II errors will also become more prominent.

Due to the inevitably high Type I error, simple statistical tests are inadequate. To address this issue, more sophisticated signal-processing methods should be considered, further substantiating the proposed approach.

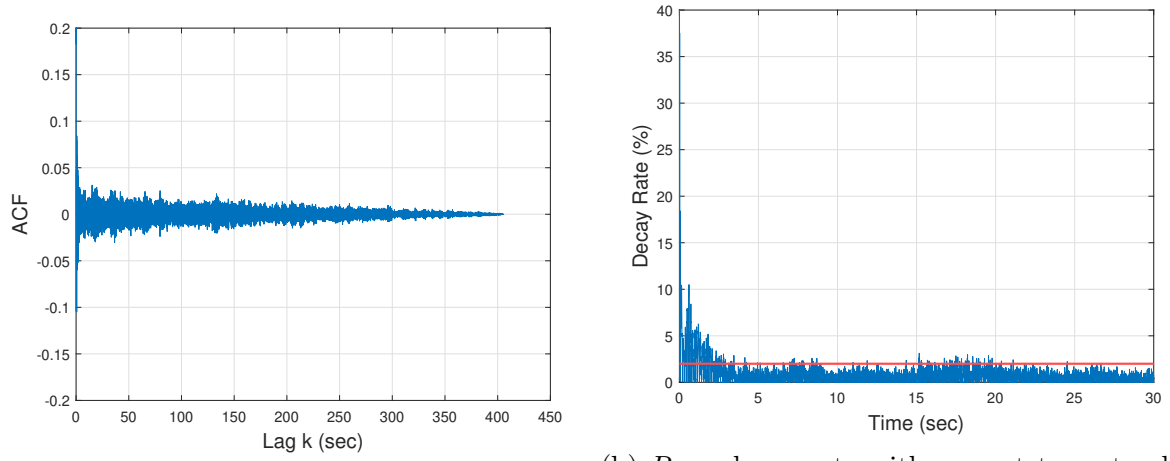
6.5 Data Processing

Prior to performing leak-detection, key data-preparation steps are discussed. In order to determine the length of the time segments used, a brief study of the autocorrelation of a time segment is done to ensure each segment is adequately uncorrelated, as is necessary for the application of the leak detection methodology. This if then followed by the steps outlined to pre-process the data for classification.

6.5.1 Autocorrelation

A general assumption of GMM is that the data set must be independent and identically distributed (IID) [Bishop, 2006]. When time series data is considered, the property considered to be representative of this is the system's stationarity. For an ergodic Gaussian process, the underlying mean and auto-correlation does not change over time. Ensuring IID for classification is not possible in the context of time-series data, however weak ergodicity can be assumed if the data is mean-centered (to zero) and the time windows are sufficiently spaced apart. The latter is due to the rapidly decaying correlation over time.

The autocorrelation for a sample ambient case for measurements given by $X(t)$ described in equation (3.31), is shown in Figure 6.5a. The correlation in Figure 6.5a decays rather quickly and this decay can be better represented compared to a standard reference. The standard reference used here is the largest value of the autocorrelation, which occurs at the lag, $\tau = 0$. The decay rate graph can be seen in Figure 6.5b.



1

(a) Normalized autocorrelation of Time Series (b) R_{XX} decay rate with respect to a standard reference

Figure 6.5: Autocorrelation analysis of sample hydro-acoustic time series.

It can be seen that the auto-correlation decays by nearly 90% in under a second. The red line shown is a 97% decay, which occurs after 3 seconds. Several tests were done with different time frames and the results showed negligible differences. Thus the LP features calculated are taken from 1 *second* time frames with a 50% overlap between adjacent frames.

6.5.2 Leak-detection

For each flow trial and sensor location, the training data sets are created for baseline comparisons. They consist of data acquired immediately prior and following each test, as they are assumed to contain the most relevant baseline spectral characteristics of the system near the measurement period. Details regarding the data-preparation process, the estimation of LP coefficients, and the detection of anomalous events consist of the following steps:

1. The acoustic signals for both the *leak-free* and *leak* case are first segmented into approximately 1 *second* time frames, with a 50% overlap between adjacent frames. Each frame generates one sample of the feature matrix. Hence, the number of samples for the analysis is equal to the number of frames.

2. LP features are estimated using a model order of $p = 50$ for both *leak-free* and *leak* cases. A Hamming window is applied to each frame prior to extracting the LP coefficients. The three principal components of these 50 coefficients are taken for each sample.

3. Anomaly-detection is undertaken in two phases: training and testing, where 40% of the baseline (simulated leak-free) samples are used for the training phase and the remaining 60% are used for the testing phase. The unconventional use of a smaller training set in comparison with the test set was done in order to better mimic the typically large amount of baseline instances the system will likely encounter. Thus by not over fitting to the training data, the results can better reflect how the model will classify baseline instances. A 10-fold cross-validation procedure is applied to validate the stability of the model and, thus, accurately estimate its performance. Average performance measures are reported in all cases.

4. The overall anomaly-detection results are then reviewed using the performance measure described previously (A_c) and reported and discussed.

Training and test sets were randomly selected for the 10-fold cross validation iterations. The training sets were used to determine the optimal number of mixtures (K). For the majority of cases, values of K ranged from 1 to 3. Different K values were required to best fit the baseline training data, because each sensor location has distinctly different characteristics and the system has significant day-to-day usage variability. The calculated K values, along with the EM algorithm, were then used to determine the Gaussian mixture parameters, θ.

After the GMM is constructed, a threshold was selected based on an assumed allowable Type II error of 5% (two standard deviations away from the mean). This threshold value can be adjusted based on preferences of Type I versus Type II error. Finally, the overall accuracy of the detection methodology was determined using the test samples.

6.5.3 Leak localization

Two-step field localization

Correlation-based localization techniques require the transmission of high resolution raw time series data, usually collected over relatively long periods of time, which is of order of several seconds followed by the calculation of statistical averages. As the number of sensor

nodes increase (typical city-scale installations could potentially have dozens or hundreds of such nodes), transmitting data from all the nodes would not be cost-effective. Filtering methods have shown to reduce transmission overhead in laboratory settings [Cody et al., 2019] and could alleviate this burden to some extent; however, this does not eliminate the need to transmit data from all the nodes to a central location completely. In this chapter, a new approach is described, where localization is performed in two steps. The first step (Step 1) localizes at a coarse resolution, where the main result is to determine the general area of the leak and the most pertinent sensor streams required for correlation analysis. In the subsequent step (Step 2), localization is performed using cross-correlation, using filtered sensor time-series data identified in Step 1. Such a two-step process eliminates the need for transmitting all the data streams, thereby reducing transmission volume and cost in full-scale applications. These two steps are explained next in detail.

Step 1: Coarse localization

Using statistical features such as root-mean-square (RMS) for this specific application is associated with large Type II error [Cody et al., 2019]. However, as shown in this chapter, RMS could provide a simple means to localize the leak region to a subset of sensors. Features such as RMS involve relatively simple mathematical operations and can easily be implemented at the sensor node level, while achieving substantial data compression (lossy) when used as a screening step. Figure 6.6 shows the distribution of RMS values corresponding to flow generated at *FL 1* and sensor data at *SL 2*. It is clear that the distributions are clearly separable for the larger flow amount (200L/min) (the Bhattacharya distance [Cody et al., 2018] for this case $B_d = 1.33$. However, RMS appears less useful as a direct means to delineate smaller leaks from the baseline cases, such as those around 100 L/min or less (B_d values are substantially less, $B_d = 0.07 - 0.3$).

While RMS is not a reliable leak-detection indicator on its own, it is found to be a good screening tool for leak-localization. It is found that while RMS does not perform well in terms of inter-distribution separability, i.e. between the baseline and leak distributions, RMS provides an effective measure of proximity when sorted by magnitude (intra-distribution of RMS); the true order of the magnitude of RMS at the sensor locations can be sorted and typically follows the order of distance from the simulated leak location. It is important to note here that the leak detection step precedes the leak localization and is not triggered when a leak is not detected.

The following pre-processing steps summarize the overall process of determining the most pertinent sensor locations to be selected for correlation analysis:

1. The average baseline RMS is calculated using leak-free data collected prior to simulated leaks.

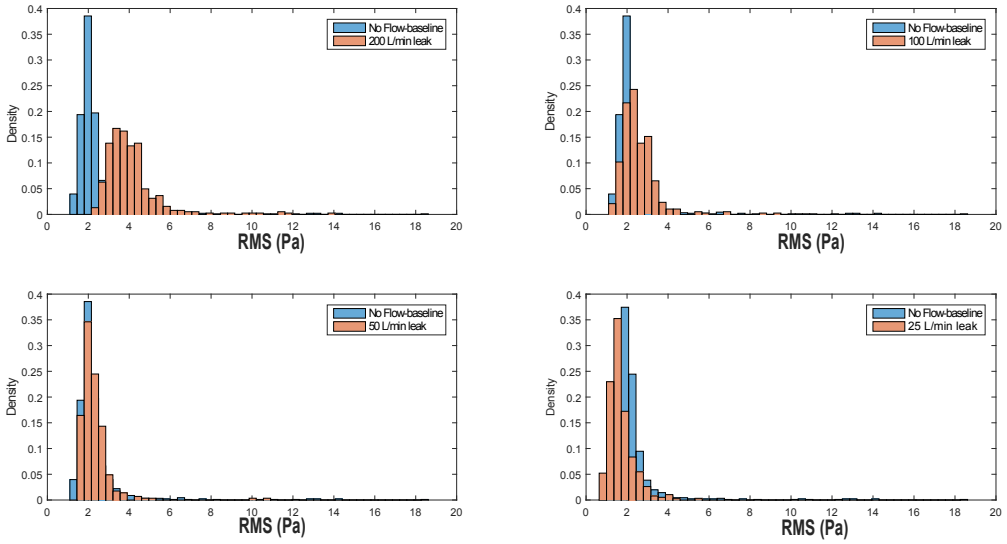


Figure 6.6: Histogram of RMS (Pa) for different flow amounts using Field Trial 2 data at *SL 2* for *FL 2*. RMS shows good inter-distribution separability for the large flow case (200 L/min), however this is not the case for lower flow amounts.

2. A filter consisting of three 2nd-order Butterworth notch filters applied at 30, 60 and 120 Hz using a 2 Hz stopband, in order to remove electrical noise noted in the system, likely caused by the presence of large pumps in the system; is applied for each 3 *second* sample.

3. At every instance, the intensity of the acoustic signal for each sensor location, as described by its RMS value, is calculated and normalized by subtracting the baseline RMS associated with the SL.

4. These values are then ranked from largest to smallest.

Step 2: Fine localization

The conclusion of Step 1 results in the two closest locations to the leak to be identified for further processing. Once the sub-set of sensor locations closest to the leak location have been identified, fine leak localization is achieved by applying the LP method to filter the signals, followed by applying the correlation method. The LP filter corresponding to the predictor model is shown in equation 4.5. First, the model coefficients a_k are estimated

by employing the auto-correlation method, followed by estimating the reconstructed signal $\hat{x}(n)$ through the application of the LP filter of model order $p = 50$ on signals that are ≈ 2 *second* in length. Followed by the use of the same filter described in *Step 1*.

The location of a leak, at a distance D_1 from a reference sensor, is calculated using equation 5.8. The sound propagation velocity for the 15.24 cm and 30.48 cm diameter pipes (with a pipe wall thickness of 1.1 cm and 1.75 cm, respectively) are calculated using well-known relationships, as seen in equation 3.13 [Gao et al., 2004, Pinnington and Briscoe, 1994], in which a pipe is considered thin-walled if $D/e > 10$, which is the case for both the pipe diameters considered, thus the current case is considered thin-walled and free to expand throughout, assigning a value of $\psi = 1$. This results in sound propagation speeds of 458 m/s and 403 m/s, respectively.

6.6 LP Results

The general detection and localization methodology of the proposed work is summarized in Figure 6.7.

6.6.1 Leak-detection

The performance measures for all sensor locations produced by all leak amounts at *FL 1* and *2* are visually represented in Figure 6.8, with a detailed account of these results found in Tables 6.1 and 6.2. Due to the nature of data collection and establishing a baseline, the data set is equally weighted. While the occurrence of a leak versus a leak free case in reality would never approach equal weight within a data set, the information required to accurately weight the data set was unavailable. Since the conclusions that would have been drawn by assessing the uneven data set would be based on assumptions without adequate information, this assessment was omitted. For this reason, and due to its popularity as the most commonly used performance metric for classifiers [García et al., 2009], *Accuracy* (A_c) as the measure of performance in Figure 6.8 is selected.

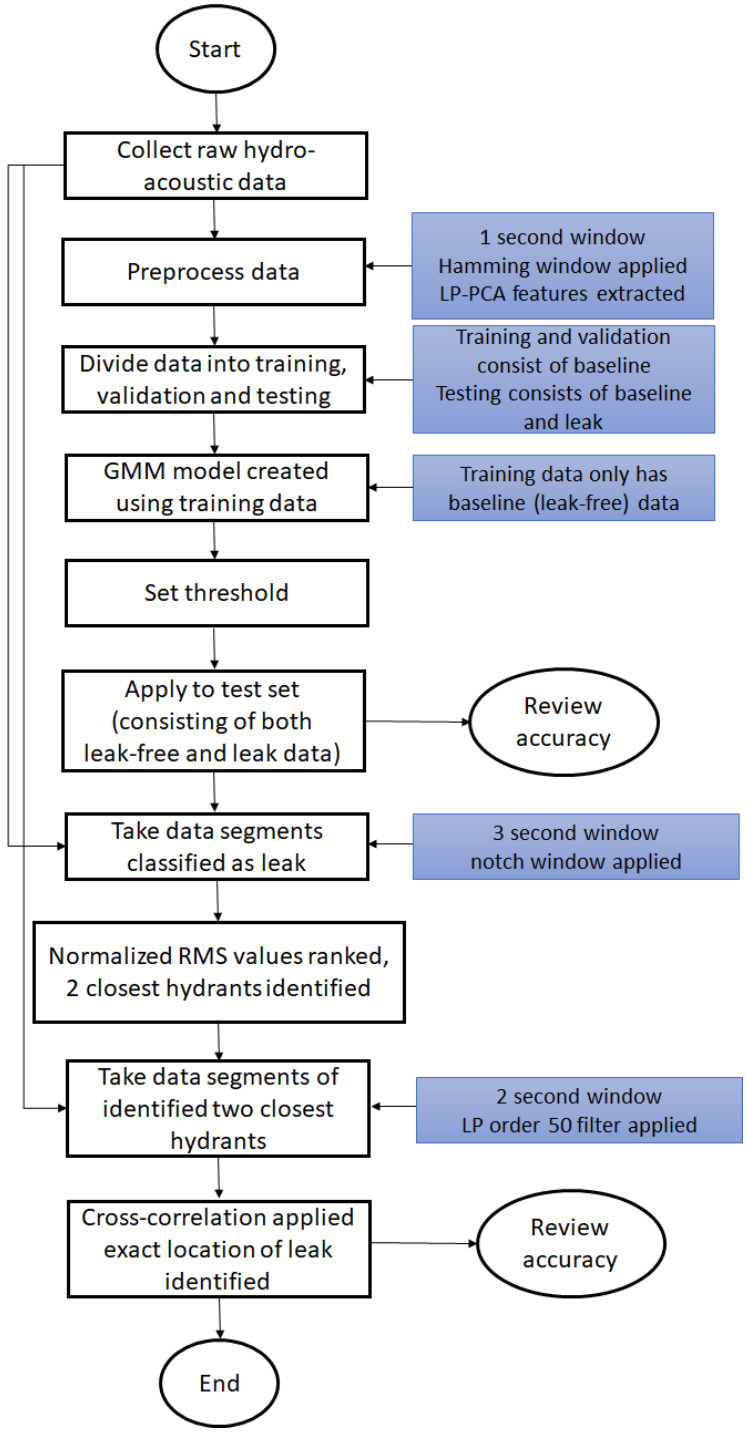
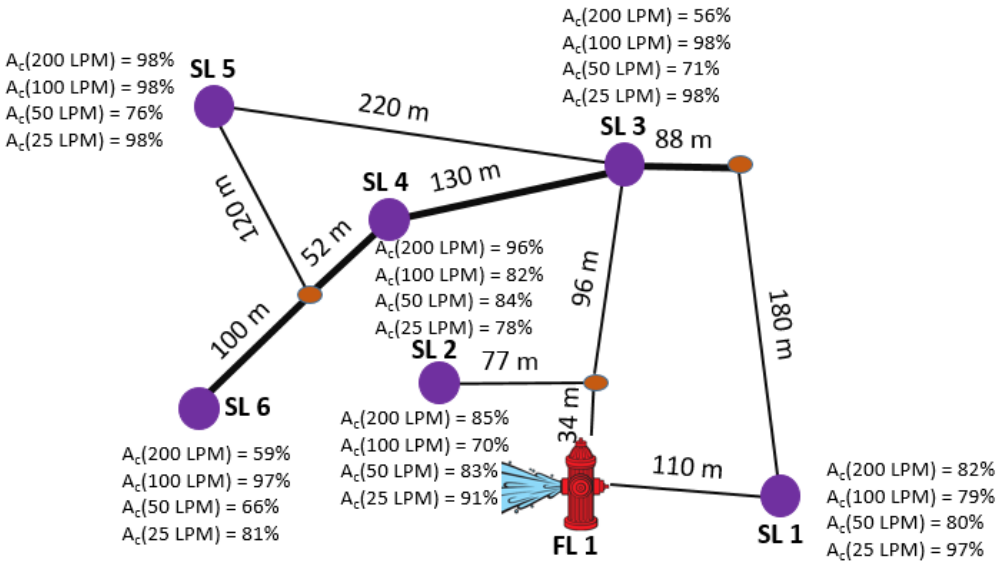
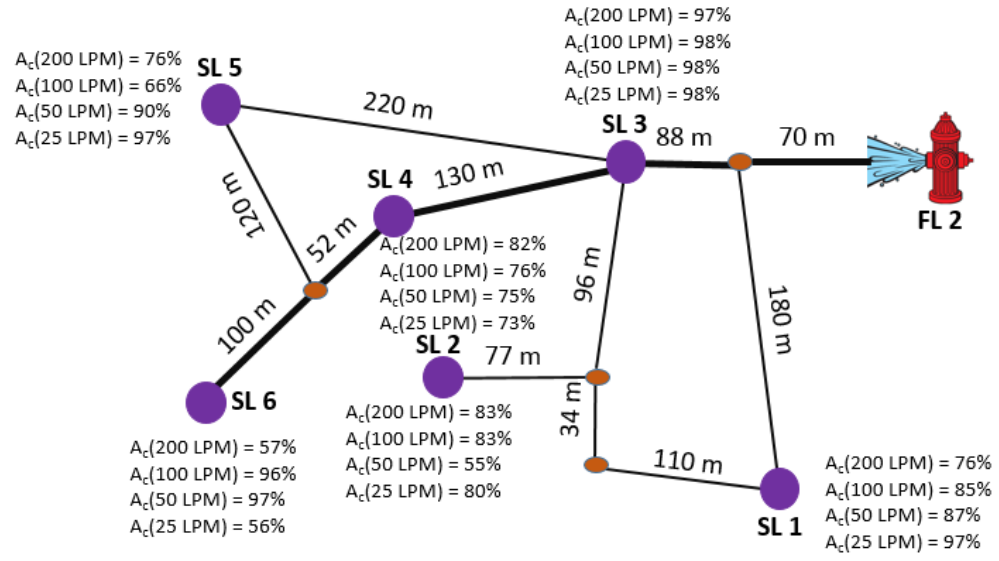


Figure 6.7: Overall structure of the detection and localization methodology implemented on the field data.



(a) Flow Location 1



(b) Flow Location 2

Figure 6.8: Accuracy of detection for all flow cases at (a) *FL 1* and (b) *FL 2* for all sensor locations. The locations can be seen in Figure 6.2. Thick solid lines represent larger diameter pipes and thin lines represent the smaller diameter ones.

Table 6.1: Detailed results for leak detection at flow location 1.

200 *L/min* ***Flow Case***						
Sensor Location	*1*	*2*	*3*	*4*	*5*	*6*
Accuracy	0.76	0.83	0.97	0.82	0.76	0.57
Precision	0.78	0.85	0.99	0.83	0.79	0.60
Recall	0.72	0.84	0.95	0.90	0.82	0.61
AUC	0.81	0.82	0.98	0.79	0.97	0.53
100 *L/min* ***Flow Case***						
Sensor Location	*1*	*2*	*3*	*4*	*5*	*6*
Accuracy	0.85	0.83	0.98	0.76	0.66	0.96
Precision	0.87	0.87	1.00	0.80	0.66	0.98
Recall	0.89	0.81	0.95	0.73	0.81	0.93
AUC	0.85	0.82	0.98	0.78	0.77	0.94
50 *L/min* ***Flow Case***						
Sensor Location	*1*	*2*	*3*	*4*	*5*	*6*
Accuracy	0.97	0.55	0.98	0.75	0.90	0.97
Precision	1.00	0.60	1.00	0.78	0.90	1.00
Recall	0.92	0.57	0.95	0.83	0.92	0.93
AUC	0.97	0.63	0.98	0.80	0.98	0.69
25 *L/min* ***Flow Case***						
Sensor Location	*1*	*2*	*3*	*4*	*5*	*6*
Accuracy	0.97	0.55	0.98	0.75	0.90	0.97
Precision	1.00	0.60	1.00	0.78	0.90	1.00
Recall	0.92	0.57	0.95	0.83	0.92	0.93
AUC	0.97	0.63	0.98	0.80	0.98	0.69

Figure 6.8a details the detection accuracy across the system for all considered flow amounts due to the simulated leak at *FL 1*. As *SLs 1* and *2* are nearly equidistant to *FL 1*, their detection accuracy is similar across all flow amounts. It is likely the inconsistent low detection accuracy at *SL 3* for the 200 *L/min* case is the result of a large flow occurring at the junction of two different size pipes—mimicking typical flow direction variation—and thus making detection more challenging. It should be noted that *SL 3* was only installed for one of the three test dates and unlike the other SLs, the test set at this location was not repeated.

Table 6.2: Detailed results for leak detection at flow location 2.

200 L/min ***Flow Case***						
Sensor Location	*1*	*2*	*3*	*4*	*5*	*6*
Accuracy	0.82	0.85	0.56	0.96	0.98	0.59
Precision	0.85	0.87	0.58	1.00	1.00	0.61
Recall	0.81	0.88	0.73	0.92	0.95	0.76
AUC	0.82	0.73	0.58	0.96	0.96	0.56
100 L/min ***Flow Case***						
Sensor Location	*1*	*2*	*3*	*4*	*5*	*6*
Accuracy	0.79	0.70	0.98	0.82	0.98	0.97
Precision	0.82	0.74	1.00	0.88	1.00	1.00
Recall	0.76	0.77	0.94	0.81	0.95	0.94
AUC	0.65	0.67	0.83	0.83	0.96	0.97
50 L/min ***Flow Case***						
Sensor Location	*1*	*2*	*3*	*4*	*5*	*6*
Accuracy	0.80	0.83	0.71	0.84	0.76	0.66
Precision	0.84	0.85	0.69	0.87	0.84	0.68
Recall	0.70	0.85	0.89	0.85	0.76	0.58
AUC	0.80	0.84	0.67	0.96	0.76	0.67
25 L/min ***Flow Case***						
Sensor Location	*1*	*2*	*3*	*4*	*5*	*6*
Accuracy	0.97	0.91	0.98	0.78	0.98	0.81
Precision	1.00	0.92	1.00	0.84	1.00	0.86
Recall	0.93	0.92	0.94	0.68	0.95	0.76
AUC	0.93	0.89	0.83	0.79	0.98	0.74

Overall, with the exception of the 200 L/min at *SL 3, SL6* and 50 L/min at *SL 6*, over 70% detection accuracy is achieved at all sensor locations for the case of simulated leak at *FL1*. In terms of the maximum and minimum distances, *SL 6* is located over 412 m away and both *SL 1* and *SL 2* are at least 110 m from the simulated leak location (*FL 1*). Furthermore, it is interesting to observe that the lowest flow amount of 25 L/min is associated with a minimum accuracy of 78%, which shows that the flow amounts in the range tested does not seem to play a large role in terms of detection accuracy. Accuracy results for *FL 2* at all sensor locations are shown in Figure 6.8b. The results are as expected

with respect to *SL 3*, yielding the highest accuracy, of over 97% in all flow cases. As with the previous case, the detection accuracy in most cases is over 70%. As with the previous flow case, the detection rate for the lowest flow amount is higher than 73%, with one exception (*SL 6*), which is located the furthest from *FL 2*.

6.6.2 Leak localization

Leak localization follows leak detection and utilizes the two-step procedure described previously. This section describes the results obtained using this two-step procedure for the case of two simulated leaks at locations *FLs 1* and *2*.

Flow Location 1

Step 1: Results for the proposed localization methodology are first presented for *FL 1*, where the graph representation consisting of six measurement points, eleven pipe lengths, and three pipe junctions, is depicted in Figure 6.3. The summary of the localization results for this step for all *FL 1* tests are shown in Table 6.3. The results presented in this table were created by aggregating concurrent data from all sensor locations and converting the raw time series into RMS of 3 *second* samples.At every instance, the intensity of the acoustic signal for each sensor location, as described by its RMS value, is ranked from the largest to the smallest. The underlying hypothesis is that the highest intensity values occur closest to the simulated leak location and the least value for the location farthest. To account for the statistical uncertainty, this experiment is repeated for all the instances ($2,500$ instances per test date) for each case of simulated leak. A polling scheme is employed to rank the sensor locations according to their intensity and then converted into a percentage. This process is repeated for all the three test dates and results are then averaged. Normalization across each sensor location is performed in order to account for the unavailability of certain sensor locations on all the three test dates.

Results in Table 6.3 show that the detection results in terms of identifying the most proximal leak location(s) is correct for the four closest SLs. There is larger uncertainty in the ranking (e.g. *SL 4* and *SL 6* are both ranked in the fourth position), which could be a result of several factors, including the number of impedance changes and the local acoustic environment near the source. It is important to underscore that for the purposes of the correlation exercise in the next step, it is only sufficient to rank at a minimum two locations correctly.

Table 6.3: Average location rank as a percentage for *FL 1* (true order: SL1/SL2, SL3, SL4, SL6, SL5) of all flow cases. The selected rank for each sensor location is in **bold**.

Rank	**SL 1**	**SL 2**	**SL 3**	**SL 4**	**SL 5**	**SL 6**
1st	18.4	**45.1**	19.1	1.1	6.7	6.8
2nd	**42.2**	14.0	8.0	5.4	5.4	16.1
3rd	1.9	3.7	**28.8**	24.6	23.1	29.8
4th	1.7	3.5	16.3	**25.2**	25.8	**32.1**
5th	19.5	7.5	9.0	18.9	**28.9**	*15.2*
6th	16.3	26.3	18.8	24.7	*10.0*	0.0

Step 2: Localization results from the first step show that *SLs 1* and *2* are the two closest locations to the simulated leak, with *SL 3* also being within the general region. The previously described pre-processing was applied to ≈ 2 *second* concurrent time series data from the three sensor locations. Cross-correlation, using a theoretically obtained sound propagation velocity of 458 m/s, was performed and the lag corresponding to the maximum positive correlation was used to calculate D_1. For all cases, the average D_1 calculated for each flow amount for the three test dates, yielding locations for the simulated leaks as summarized in Table 6.4. The practical challenges associated with determining the true D_1 are the unknown lengths of laterals in the data available and the lack of experimentally determined value of the speed of sound in the system. Hence, the true D_1 used is as an approximate value and the results obtained from using SL 1 and SL 3 show relatively large error (between 7.5-24.8%) compared to SL 1 and SL 2 (1-11.9%). It is not surprising that the former results in larger errors as $SL-3$ is farther from the leak location and at a junction of two pipe sizes. However, using SL 1 and SL 2, the correlation method yields results within 12% for the worst case. As well, better localization accuracy is achieved for the lower flow amounts compared to the highest flow amount (200 L/min).

Flow Location 2

Step 1: The graph model for this case is shown in Figure 6.3. The localization results from the first step for all tests (dates and amounts) are shown in Table 6.5. Unlike *FL 1*, *FL 2* is located at the extremity of the test area and is located along the larger diameter pipe. As seen from the results presented in Table 6.5, both *SL 3* and *SL 1* can be considered equally probable as candidates in terms of the closest from the simulated leak, with *SL 3* being the

Table 6.4: Average D_1 (in meters), μ_{D_1}, for each flow amount using *SLs 1* and *2*, and *SLs 1* and *3*. In both cases D_1 is taken as the distance from *SL 1* to the leak location, thus for both cases, the true distance is approximately $D_1 = 109\ m$. Percentage errors are reported in parenthesis.

Flow Amount	$\mu_{D_1(SL1-2)}\ (m)$	$\mu_{D_1(SL1-3)}\ m$
200 L/min	95.99 (11.9%)	136.03 (24.8%)
100 L/min	107.91 (1.0%)	131.19 (20.4%)
50 L/min	100.40 (7.9%)	117.26 (7.5%)
25 L/min	99.59 (8.6%)	100.62 (7.7%)

true closest location. *SL 4* is correctly classified as the next closest, and *SL 5* is correctly classified as the furthest SL. *SL 2* and *SL 6*, are mis-classified, as they are relatively far from the flow location. However, this does not affect the subsequent step as the only two closest locations have to be correctly ranked for the ensuing step.

Table 6.5: Average location rank as a percentage for *FL 2* (true order: SL3, SL1, SL4, SL2, SL6, SL5). The selected rank for each sensor location is in **bold**.

Rank	**SL 1**	**SL 2**	**SL 3**	**SL 4**	**SL 5**	**SL 6**
1st	**33.1**	22.6	**52.7**	1.1	2.5	10.8
2nd	*27.4*	18.4	17.8	9.3	2.0	**27.3**
3rd	11.1	8.2	8.6	**34.4**	5.7	26.2
4th	8.7	*8.7*	7.6	24.1	20.7	24.2
5th	13.6	**33.0**	3.2	11.3	16.7	*11.4*
6th	6.1	9.1	0.0	19.7	**52.3**	0.0

Step 2: Based on the results of the first step presented in Table 6.5, the proximity ranking of sensor locations show both *SLs 1* and *3* as being the most proximal sensor locations, while *SL 4* ranked as the next closest. Due to this, it can be inferred that the leak is located closer to, if not along, the main section of the pipe. Localization on the main section of the pipe is done most effectively by using *SLs 3*, and *4* as their locations are associated with the least impedance changes. The localization method is once again applied to ≈ 2 *second* samples; cross-correlation using a sound propagation velocity of 403 m/s, calculated for larger diameter pipe segment along which the leak and SLs under review lie, was performed

and the location of the maximum lag in the correlation graph was used to calculate the average D_1 for each flow amount.

Table 6.6: Average D_1 (in meters), μ_{D_1}, for each flow amount using *SLs 3* and *4*. D_1 is taken as the distance from *SL 4* to the leak location, thus for both true $D_1 = 268\ m$. Percentage errors are shown in the parenthesis.

Flow Amount	$\mu_{D_1(SL3-4)}$
200 L/min	261.88 (2.3%)
100 L/min	230.85 (13.9%)
50 L/min	222.47 (17.0%)
25 L/min	247.90 (7.5%)

The cross-correlation of *SLs 3* and *4* was reviewed for the different flow amounts collected: the D_1 values estimated from this algorithm are summarized in Table 6.6 together with the errors. As mentioned previously, due to the presence of laterals, the true D_1 is not known exactly; however the 268 m based on the information provided (this is believed to be within a few meters of actual) was used for comparison purposes. The results obtained for $FL-2$ are similar in terms of errors compared to $FL-1$, despite the fact that this layout can be viewed as more challenging compared to the former.

6.7 Summary

This chapter presents a field case study for the application of LP for semi-supervised leak-detection in WDNs. A novel two-part localization methodology is presented, which first isolates the general region of the leak, then pin-points a more exact location. The system developed in this study overcomes many challenges and limitations that were previously associated with long-term passive monitoring in WDN systems. This system offers a convenient and affordable solution for event detection in WDNs, while maintaining minimal installation cost and energy-consumption.

Results from this study outline the use of LP features, coupled with a GMM, and show that the proposed methodology is able to achieve leak-detection. Furthermore, the localization results using correlation method shows that it is possible to achieve localization, which can then inform more local inspections and intervention strategies for pinpointing leaks. The computational efficiency of the proposed methodology allows for long-term

monitoring in field applications. In addition, the localization methodology allows for two granularities of localization, while using short time signal lengths. The short time signal lengths minimizes the data-transmission requirements, which is one of the main impediments in other full-scale implementations of leak-detection technology. These results are presented using data obtained from a section of a WDN, using a custom hydrant-mounted data-acquisition system with unique hardware and software, designed specifically for the case study described here.

While the results from this study are promising, it is important to acknowledge the limitations of this study. The effectiveness of the proposed methodology was proven in one field test-bed; there is significant additional validation tests required before this method can be generalized or uniformly applied to a range of conditions. Moreover, while it is expected that this methodology will perform robustly with different pipe materials, especially since PVC is typically considered most challenging for leak detection and leak localization, other pipe materials have not been evaluated in this dissertation. This study assumes that there is only one leak at a time in a given section and does not make any claims in terms of detection when multiple leaks occur in a given test area. Finally, the effect of different soil materials surrounding the pipe system has also not been studied.

Chapter 7

Concluding Remarks

7.1 Summary of contributions

In this dissertation, the problem of long term passive monitoring of leak detection and determining their corresponding locations is addressed. Basic principles of LP are extended and applied to the problem of leak detection and localization. Results from two test beds are considered: a laboratory test bed for proof of concept and a corresponding field test bed for validation of the developed event detection and localization algorithms. The field test bed was selected for its similarity to the laboratory test bed, as they are the same material and the majority of the field test bed consisted of the same diameter of pipe as that found in the laboratory test bed. The purpose of the field test was to validate the test results obtained from the laboratory tests. Specific challenges faced with long term data collection and the use of data under unknown conditions were discussed, and semi-supervised algorithms facilitating leak detection and localization under these conditions were developed. The following are the main contributions of this dissertation:

1. A sensor hardware and software platform with specifications for the retrofitted hydrant mounted monitoring systems was developed, specifically for long-term passive monitoring systems which balance the density (of sensors), granularity of scale and the reliability of event detection in a live WDN, for year-round monitoring. The retrofitted fire hydrant system allows for a low-cost implementation with low granularity in order to isolate general regions in the system to direct the tedious high-resolution, high man-hour detection processes. The low-maintenance and low granularity aspects of this passive monitoring system allows for cities with different

budgets to deploy this system with whatever initial cost they can accommodate, changing the granularity and accuracy of the system. The major tasks accomplished for the development of this system are as follows:

   (a) The development of hardware with adequate capabilities for data sampling and collection.

   (b) In practical implementation, the system facilitates autonomous deployment. The developed software enables passive data collection during selected times of interest.

2. A LP based framework is established for the analysis of acoustic signatures and was found to be a powerful tool in capturing the primary resonant responses of the fluid-pipe coupled linear system, and thus a representative feature as to the state of the system. The main highlights from the development of this framework are as follows:

   (a) The sensitivity of linear prediction coefficients to leak induced signals was validated for a single pipe segment in the laboratory.

   (b) The LP framework allows for an unified treatment of both leak detection and localization problems, which is a significant advantage.

   (c) The proposed framework allows for a semi-supervised anomaly detection implementation which is sufficiently robust and computationally efficient for long term field applications.

3. Laboratory and field case studies were used to validate the LP based framework and test the developed decision support systems for autonomous implementation of event detection for the hydrant mounted system. The main highlights from the experimental case studies are as follows:

   (a) A laboratory test bed for proof of concept validation of the proposed methodology, which is relatively representative of field conditions, was developed. While it is not completely representative of a field system, it involves an increased number of bends and laterals, as well as actual hydrants mounted to the system to more accurately represent field conditions and features.

   (b) By using advanced statistical and machine learning methods coupled with the proposed LP based framework, the described system has the capacity to identify anomalies accurately and with sufficient computational efficiency that it can be deployed in field monitoring situations. However, as it is a purely data-driven

approach, this is contingent on a good representation of the baseline of the system.

(c) The proposed system was experimentally deployed and verified in a subset of water distribution network under live conditions. A comprehensive database containing months of hydro-acoustic data of typical water distribution system operating at various hours and months, across a DMA at a hydrant level, was developed. This database can be used to:

   i. Better understand hydraulic conditions as they occur in the field.
   ii. Better the development of accurate baselines throughout the year and algorithmic improvements.

7.2 Limitations

While the results from this dissertation are promising, it is important to acknowledge the limitations. The effectiveness of the proposed methodology was proven in one field test-bed; there is significant additional validation tests required before this method can be generalized or uniformly applied to a range of conditions. Data collected at low demand hours are used for the pilot study discussed herein, a review of the effectiveness of the proposed methodology on data collected throughout the day should also be considered. As well, an assumption is made that the detection of anomalies is limited to the presence of leaks. No attempt to identify other sources of anomalies is made within this dissertation. Moreover, while it is expected that this methodology will perform robustly with different pipe materials, especially since PVC is typically considered most challenging for leak detection and leak localization, other pipe materials have not been evaluated in this dissertation. This dissertation assumes that there is only one leak at a time in a given section and does not make any claims in terms of detection when multiple leaks occur in a given test area or any attempt at identifying the size of the leak that is detected. The effect of the surrounding soil properties on the detection accuracy is also not addressed.

7.3 Directions for future study

Based on the research work proposed in this dissertation, the following research directions can be pursued for extending the methodology:

1. While this study extended over multiple years, actionable data was only obtained for time spanning a few months. Various factors including the need for field support staff (operating hydrants can only be undertaken by licensed staff), budget and scheduling limited the ability to collect extensive data. Hence, there is a tremendous scope to not only extend the duration of data collection, but also to increase the sensor density.

2. From an anomaly detection standpoint, a fairly limited set of tools have been employed on acoustic data; there is definitely a lot of room to test and deploy powerful machine learning algorithms in the future. The employment of a probabilistic Bayesian inference approach, for example, would enrich the methodology by incorporating the epistemic uncertainty around the LP and GMM parameters.

3. There is also potential to add concurrent data from other sensor types such as pressure and accelerometers into the training and validation processes. Utilizing the strengths from different sensor types would strengthen the methodology and likely improve performance.

4. Although PVC is considered the most difficult pipe material for detection and localization due to the flexible nature of the materials and thus the increased signal attenuation, the review of the proposed framework on other pipe materials would be a logical next step to the work summarized within this dissertation. It is expected that the rigid nature of cast iron and concrete will improve the detection and localization accuracies when this framework is implemented in those systems.

5. In addition to the pipe material, the effect of various types of surrounding back-fill material on the quality of results need to be studied. This will likely have significant effects on the rigidity and in turn the attenuation of the signal.

6. The described framework can be extended to include other important variables such as leak size, presence and number of multiple leaks and frequency of pressure transients in the system.

7. In the proposed framework, data from each sensor unit is modeled independently as a uni-variate time series. A multivariate modelling approach can also be taken in the future, where data from multiple units can be taken simultaneously and modeled as as vector time series, which could potentially yield additional insight into the leak characteristics.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

Bovas Abraham and George EP Box. Bayesian analysis of some outlier problems in time series. *Biometrika*, 66(2):229–236, 1979.

ADS. Ads® eureka digital system user manual, 2009. URL https://www.adsenv.com/sites/default/files/manuals/adseurekadigitalsystemusermanuala0.pdf.

Changsheng Ai, Honghua Zhao, Rujian Ma, and Xueren Dong. Pipeline damage and leak detection based on sound spectrum lpcc and hmm. In *Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on*, volume 1, pages 829–833. IEEE, 2006.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

K Aksela, M Aksela, and & R Vahala. Leakage detection in a real distribution network using a SOM. *Urban Water Journal*, 64(January):279–289, 2009. ISSN 1573-062X. doi: 10.1080/15730620802673079.

Taha Al-Washali, Saroj Sharma, Fadhl Al-Nozaily, Mansour Haidera, and Maria Kennedy. Modelling the leakage rate and reduction using minimum night flow analysis in an intermittent supply system. *Water*, 11(1):48, 2019.

Michael Allen, Ami Preis, Mudasser Iqbal, Seshan Srirangarajan, Hock Beng Lim, Lewis Girod, and Andrew J. Whittle. Real-time in-network distribution system monitoring to improve operational efficiency. *Journal - American Water Works Association*, 103(7): 63–75, 2011. ISSN 0003150X.

Fabrício Almeida, Michael Brennan, Phillip Joseph, Stuart Whitfield, Simon Dray, and Amarildo Paschoalini. On the acoustic filtering of the pipe and sensor in a buried plastic water pipe and its effect on leak detection: an experimental investigation. *Sensors*, 14 (3):5595–5610, 2014.

Gary Anguiano, Stuart Strum, Victor Medina, Scott Waisner, Wendy Condit, John Matthews, and Ryan Stowe. Innovative acoustic sensor technologies for leak detection in challenging pipe types. Technical report, Naval Facilities Engineering Command Port Hueneme United States, 2016.

Jerónimo Arenas-García, Kaare B Petersen, and Lars K Hansen. Sparse kernel orthonormalized pls for feature extraction in large data sets. In *Advances in Neural Information Processing Systems*, pages 33–40, 2007.

C Aristegui, MJS Lowe, and P Cawley. Guided waves in fluid-filled pipes surrounded by different fluids. *Ultrasonics*, 39(5):367–375, 2001.

Michele Basseville and Igor V. Nikiforov. *Detection of abrupt changes: theory and application*. Englewood Cliffs: Prentice Hall, 1993.

M Batté, BMR Appenzeller, D Grandjean, S Fass, V Gauthier, F Jorand, Laurence Mathieu, M Boualam, S Saby, and JC Block. Biofilms in drinking water distribution systems. *Reviews in Environmental Science and Biotechnology*, 2(2-4):147–168, 2003.

Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

Commandant Benoıt. Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrésa un systeme d'équations linéaires en nombre inférieura celui des inconnues. application de la méthodea la résolution d'un systeme défini d'équations linéaires (procédé du commandant cholesky). *Bulletin géodésique*, 2(1):67–77, 1924.

Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science & Business Media, 2006.

Anthony Bond, Brian Mergelas, and Cliff Jones. Pinpointing leaks in water transmission mains. *Proceedings of ASCE Pipeline 2004*, pages 1–10, 2004. doi: 10.1061/40745(146)91.

Alexandre Bonton, Christian Bouchard, Benoit Barbeau, and Stephane Jedrzejak. Comparative life cycle assessment of water treatment plants. *Desalination*, 284:42–54, jan 2012. ISSN 0011-9164. doi: 10.1016/j.desal.2011.08.035.

George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

Marc Bracken and Bill Cain. Transmission main and plastic pipe leak detection using advanced correlation technology: Case studies. In *Pipelines 2012: Innovations in Design, Construction, Operations, and Maintenance, Doing More with Less*, pages 147–157. ASCE, 2012.

Marc Bracken and Dave Johnston. Advances in acoustic methods for locating leaks in plastic pipe and trunk mains. In *Pipelines 2009: Infrastructure's Hidden Assets*, pages 508–523. ASCE, 2009.

Anders Brandt. Spectrum and Correlation Estimates Using the DFT. In *Noise and vibration analysis: signal analysis and experimental procedures*, chapter 10, pages 205–243. John Wiley & Sons, 2011a.

Anders Brandt. Experimental Frequency Analysis. In *Noise and vibration analysis: signal analysis and experimental procedures*, chapter 9, pages 177–204. John Wiley & Sons, 2011b.

Bruno Brunone. Transient test-based technique for leak detection in outfall pipes. *Journal of water resources planning and management*, 125(5):302–306, 1999.

Bruno Brunone. Detecting leaks in pressurised pipes by means of transients. *Journal of Hydraulic Research*, 39(4):1–9, 2001. ISSN 0022-1686. doi: 10.1080/00221686.2004.9641189.

Allan R. Budris. Damage control: Avoiding destructive water hammer conditions, Apr 2014. URL https://www.waterworld.com/municipal/technologies/pumps/article/16192896/damage-control-avoiding-destructive-water-hammer-conditions.

The Ontario Building Code. Buildingcode.online, 2018. URL http://www.buildingcode.online/271.html.

Environment Canada. 2011 municipal water use report – municipal water use 2009 statistics, 2011.

Statistics Canada. Human Activity and the Environment: Section 3: The demand for water in Canada, dec 2013. URL http://www.statcan.gc.ca/pub/16-201-x/2010000/part-partie3-eng.htm.

A Candelieri, D Conti, and F Archetti. A graph based analysis of leak localization in urban water networks. *Procedia Engineering*, 70:228–237, 2014.

Antonio Caputo and Pacifico Pelagagge. Using Neural Networks to Monitor Piping Systems. *Process Safety Progress*, 2(2):119–127, 2003.

Myrna V Casillas Ponce, Luis E Garza Castañón, and Vicenç Puig Cayuela. Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities. *Journal of Hydroinformatics*, 16(3):649–670, 2013.

P. Celka and P. Colditz. A computer-aided detection of EEG seizures in infants: a singular-spectrum approach and performance comparison. *IEEE Transactions on Biomedical Engineering*, 49(5):455–462, 2002.

Yushi Chen, Zhouhan Lin, Xing Zhao, Gang Wang, and Yanfeng Gu. Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected topics in applied earth observations and remote sensing*, 7(6):2094–2107, 2014.

C. Choi and C. Lee. Feature extraction based on the bhattacharyya distance. *Pattern Recognition*, 36:1703–1709, 2003.

R. Cody, J. Harmouche, and S. Narasimhan. Leak detection in water distribution pipes using singular spectrum analysis. *Urban Water Journal*, 15(7):636–644, 2018.

R. Cody, P. Dey, and S. Narasimhan. Linear prediction for leak detection in water distribution networks. *Journal of Pipeline Systems - Engineering and Practice*, 2019.

Roya Cody, Sriram Narasimhan, and Bryan Tolson. Ccwi2017: F47 'one-class svm – leak detection in water distribution systems', Sep 2017.

Ronald A Cole, Alexander I Rudnicky, Victor W Zue, and D Raj Reddy. Speech as patterns on paper. *Perception and production of fluent speech*, pages 3–50, 1980.

A.F. Colombo, P. Lee, and B.W. Karney. A selective literature review of transient-based leak detection methods. *Journal of Hydro-Environment Research*, 2(4):212–227, 2009. ISSN 15706443. doi: 10.1016/j.jher.2009.02.003.

James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

D Covas, H Ramos, N Lopes, and AB Almeida. Water pipe system diagnosis by transient pressure signals. In *8th Annual Water Distribution Systems Analysis Symposium, Cincinnati, OH, Aug*, pages 27–30, 2006.

Didia Covas and Helena Ramos. Case Studies of Leak Detection and Location in Water Pipe Systems by Inverse Transient Analysis. *Journal of Water Resources Planning and Management*, 136(2):248–257, 2010. ISSN 0733-9496. doi: 10.1061/(ASCE)0733-9496(2010)136:2(248).

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Dhammika De Silva, John Mashford, and Stewart Burn. Computer aided leak location and sizing in pipe networks, 2011.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Yong Deng, Wen Jiang, and Rehan Sadiq. Modeling contaminant intrusion in water distribution networks: A new similarity-based dst method. *Expert Systems with Applications*, 38(1):571–578, 2011.

Konstantinos G Derpanis. The bhattacharyya measure. *Mendeley Computer*, 1(4):1990–1992, 2008.

M Deyi, J van Zyl, and M Shepherd. Applying the favad concept and leakage number to real networks: A case study in kwadabeka, south africa. *Procedia Engineering*, 89:1537–1544, 2014.

M. Domingues, R. Filippone, P. Michiardi, and J. Zouaoui. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, 74:406–421, Feb 2018.

James Durbin. The fitting of time-series models. *Revue de l'Institut International de Statistique*, pages 233–244, 1960.

Echologics. Echowave® acoustic water leak detector, 2020. URL https://www.echologics.com/services/large-diameter-leak-detection/echowave/.

Mohammed S. El-Abbasy, Fadi Mosleh, Ahmed Senouci, Tarek Zayed, and Hassan Al-Derham. Locating Leaks in Water Mains Using Noise Loggers. *Infrastructure System*, 22(3):1–1, 2016.

D G Eliades and M M Polycarpou. Leakage fault detection in district metered areas of water distribution systems. *Journal of Hydroinformatics*, 14(4):992, 2012. ISSN 1464-7141. doi: 10.2166/hydro.2012.109.

EPA. Statistics and Facts - WaterSense, 2015. URL https://www.epa.gov/watersense/statistics-and-facts.

European Commission. EU Reference document Good Practices on Leakage Management WFD CIS WG PoM. Technical report, European Commission, 2015.

Harry B Evans. *Water distribution in ancient Rome: the evidence of Frontinus*. University of Michigan Press, 1997.

Sami Eyuboglu, Hanan Mahdi, and Haydar Al-shukri. Detection of water leaks using ground penetrating radar. *Proceedings of the Third International Conference on Applied Geophysics*, 2003.

Mohamed Fahmy and Osama Moselhi. Automated Detection and Location of Leaks in Water Mains Using Infrared Photography. *Journal of Performance of Constructed Facilities*, 24(3):242–248, 2010. ISSN 0887-3828. doi: 10.1061/(ASCE)CF.1943-5509.0000094.

Frank J Fahy and Paolo Gardonio. *Sound and structural vibration: radiation, transmission and response*. Elsevier, 2007.

P. Fanner, J. Thornton, and R. Liemberger. Leakage Management Technologies. *AWWA Research Foundation Denver*, 2007.

B Farley, S R Mounce, and J B Boxall. Development and Field Validation of a Burst Localization Methodology. *Journal of Water Resources Planning and Management*, 139(December):604–613, 2013. ISSN 07339496. doi: 10.1061/(ASCE)WR.1943-5452.0000290.

Antonino Ferrante and Said Elghobashi. On the physical mechanisms of drag reduction in a spatially developing turbulent boundary layer laden with microbubbles. *Journal of Fluid Mechanics*, 503:345–355, 2004.

M. Ferrante, C. Massari, B. Brunone, and S. Meniconi. Leak behaviour in pressurized PVC pipes. *Water Science and Technology: Water Supply*, 13(4):987–992, 2013. ISSN 16069749. doi: 10.2166/ws.2013.047.

Marco Ferrante, Bruno Brunone, and Silvia Meniconi. Wavelets for the Analysis of Transient Pressure Signals for Leak Detection. *Journal of Hydraulic Engineering*, 133(11): 1–8, 207.

Sam Fox, Will Shepherd, Richard Collins, and Joby Boxall. Experimental quantification of contaminant ingress into a buried leaking pipe during transient events. *Journal of Hydraulic Engineering*, 142(1):04015036, 2015.

H. Fujisaki and Y. Sato. Evaluation and comparison of features in speech recognition. *Annu. Rep. Eng. Res. Inst.*, 32:213–218, 1973.

Sadaoki Furui. *Digital speech processing: synthesis, and recognition*. CRC Press, 2000.

Y. Gao, M. J. Brennan, P. F. Joseph, J. M. Muggleton, and O. Hunaidi. A model of the correlation function of leak noise in buried plastic pipes. *Journal of Sound and Vibration*, 277(1-2):133–148, 2004. ISSN 0022460X. doi: 10.1016/j.jsv.2003.08.045.

Y. Gao, M.J. Brennan, P.F. Joseph, J.M. Muggleton, and O. Hunaidi. On the selection of acoustic/vibration sensors for leak detection in plastic water pipes. *Journal of Sound and Vibration*, 283(3-5):927–941, 2005. ISSN 0022460X. doi: 10.1016/j.jsv.2004.05.004.

Yan Gao, Yuyou Liu, Yifan Ma, Xiaobin Cheng, and Jun Yang. Application of the differentiation process into the correlation-based leak detection in urban pipeline networks. *Mechanical Systems and Signal Processing*, 112:251–264, 2018.

Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera. A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*, 13(10):959, 2009.

Jie Geng, Jianchao Fan, Hongyu Wang, Xiaorui Ma, Baoming Li, and Fuliang Chen. High-resolution sar image classification via deep convolutional autoencoders. *IEEE Geoscience and Remote Sensing Letters*, 12(11):2351–2355, 2015.

Nina E. Golyandina. On the choice of parameters in singular spectrum analysis and related subspace-based methods. *Statistics and Its Interface*, 2010.

Jinzhe Gong, Martin Lambert, Aaron Zecchin, Angus Simpson, Nicole Arbon, and Young-il Kim. Field study on non-invasive and non-destructive condition assessment for asbestos cement pipelines by time-domain fluid transient analysis. *Structural Health Monitoring*, 15(1):113–124, 2016. ISSN 1475-9217. doi: 10.1177/1475921715624505.

James-A Goulet, Sylvain Coutu, and Ian FC Smith. Model falsification diagnosis and sensor placement for leak detection in pressurized pipe networks. *Advanced Engineering Informatics*, 27(2):261–269, 2013.

Environment Government of Canada and Climate Change Canada. Environment and Climate Change Canada - Environmental Indicators - Residential Water Use, mar 2012. URL https://www.ec.gc.ca/indicateurs-indicators/default.asp?lang=en{&}n=7E808512-1.

Marcos Guillen, Jean-Francois Dulhoste, Gildas Besancon, and Rafael Santos. Study of a flow model for detection and location of leaks and obstructions in pipes. *9th International Conference on Modeling, Optimization & SIMulation, Bordeaux, France*, 06 2012.

Suman Gupta. Canadas freshwater - alive, 2013. URL http://www.alive.com/lifestyle/canadas-freshwater/.

H. A. Güvenir and M. Kurtcephe. Ranking instances by maximizing the area under roc curve. *IEEE Transactions on Knowledge and Data Engineering*, 25(10):2356–2366, Oct 2013.

J. Harmouche, D. Fourer, F. Auger, P. Borgnat, and P. Flandrin. The sliding singular spectrum analysis: a data-driven nonstationary signal decomposition tool. *IEEE Transactions on Signal Processing*, 66(1):251–263, 2018.

Dennis R. Helsel and Robert M. Hirsch. *Statistical methods in water resources Book 4, Chapter A3*. U.S. Geological Survey, 1997.

G. Wayne Hennigar. Water leakage control and sonic detection. *Canadian Water Resources Journal*, 9(3):51–57, Jan 2013. doi: 10.4296/cwrj0903051.

G Hessel, W Schmitt, K Van der Vorst, and F-P Weiss. A neutral network approach for acoustic leak monitoring in the vver-440 pressure vessel head. *Progress in Nuclear Energy*, 34(3):173–183, 1999.

Nicholas J Higham. Cholesky factorization. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(2):251–254, 2009.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Keith W Hipel and A Ian McLeod. *Time series modelling of water resources and environmental systems*. Elsevier, 1994.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992.

David M Hughes and Chandan Venkatesh. Reduction of non-revenue water through continuous acoustic monitoring. Technical report, Champaign, IL: Illinois Sustainable Technology Center, 2016.

Osama Hunaidi. Detecting Leaks in Water-Distribution Pipes. *Institute for Research in Construction*, 92(October):1–6, 2000.

Osama Hunaidi and Wing T Chu. Acoustical characteristics of leak signals in plastic water distribution pipes. *Applied Acoustics*, 58:235–254, 1999. ISSN 0003682X. doi: 10.1016/S0003-682X(99)00013-4.

Osama Hunaidi and Peter Giamou. Ground-penetrating radar for detection of leaks in buried plastic water distribution pipes. In *International Conference on Ground Penetrating Radar*, pages 783–786, 1998.

Osama Hunaidi, Wing Chu, Alex Wang, and Wei Guan. Detecting leaks in plastic pipes. *Journal / American Water Works Association*, 92(2):82–94, 2000. ISSN 0003150X.

Osama Hunaidi, Alex Wang, M. Bracken, T Gambino, and C Fricke. Acoustic methods for locating leaks in municipal water pipe networks. *International Water Demand Management Conference*, pages 1–14, 2004.

Andrew C Jackson, James P Butler, Emil J Millet, FREDERIC G Hoppin Jr, and STANLEY V Dawson. Airway geometry by analysis of acoustic pulse response measurements. *Journal of Applied Physiology*, 43(3):523–536, 1977.

Mohammadamin Jahanpour. *Pressure Sensor Placement for Leak Diagnosis under Demand Uncertainty in Water Distribution Systems*. PhD thesis, University of Waterloo, 2019.

Alan Jeffrey. *Advanced engineering mathematics*. Elsevier, 2001.

Zi-guang Jia, Liang Ren, Hong-nan Li, Siu-Chun Ho, and Gang-bing Song. Experimental study of pipeline leak detection based on hoop strain measurement. *Structural Control and Health Monitoring*, 22(5):799–812, 2015.

Y. Jin, W. Yumei, and L. Ping. Approximate entropy-based leak detection using artificial neural network in water distribution pipelines. In *11th International Conference on Control Automation Robotics & Vision (ICARCV)*, pages 1029–1034. IEEE, dec 2010.

Ravi Kashyap. The perfect marriage and much more: Combining dimension reduction, distance measures and covariance. *Physica A: Statistical Mechanics and its Applications*, 536:120938, 2019.

M Sohail Khalid, M Umer Ilyas, M Saquib Sarfaraz, and M Asim Ajaz. Bhattacharyya coefficient in correlation of gray-scale objects. *Journal of Multimedia*, 1(1):56–61, 2006.

Atia E Khalifa, Dimitris M Chatzigeorgiou, Kamal Youcef-Toumi, Yehia A Khulief, and Rached Ben-Mansour. Quantifying acoustic and pressure sensing for in-pipe leak detection. In *International Mechanical Engineering Congress and Exposition*, pages 489–495. American Society of Mechanical Engineers, 2010. ISBN 978-0-7918-4450-2. doi: 10.1115/IMECE2010-40056.

YA Khulief, A Khalifa, R Ben Mansour, and MA Habib. Acoustic detection of leaks in water pipelines using measurements inside pipe. *Journal of Pipeline Systems Engineering and Practice*, 3(2):47–54, 2011.

DP. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *stat*, volume 1050, page 10, 2014.

Lawrence E Kinsler, Austin R Frey, Alan B Coppens, and James V Sanders. *Fundamentals of acoustics*. John Wiley & Sons, Inc., 1999.

Frank Kirchner and Joachim Hertzberg. A prototype study of an autonomous robot platform for sewerage system maintenance. *Autonomous robots*, 4(4):319–331, 1997.

Robin Kirkham, Patrick D Kearney, Kevin J Rogers, and John Mashford. Pirat—a system for quantitative sewer pipe assessment. *The International Journal of Robotics Research*, 19(11):1033–1053, 2000.

Gregory J Kirmeyer and Katherine Martel. *Pathogen intrusion into the distribution system*. American Water Works Association, 2001.

David Knight. Components and materials: Part 6, 2007. URL http://g3ynh.info/zdocs/comps/part_6.html.

Judice LY Koh, Mong Li Lee, Wynne Hsu, and Kai Tak Lam. Correlation-based detection of attribute outliers. In *International Conference on Database Systems for Advanced Applications*, pages 164–175. Springer, 2007.

Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

Jouni Kuha. Aic and bic: Comparisons of assumptions and performance. *Sociological methods & research*, 33(2):188–229, 2004.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

V. Kumar, J. Heikkonen, J. Rissanen, and K. Kaski. Minimum description length denoising with histogram models. *IEEE Transactions on Signal Processing*, 54(8):2922–2928, 2006.

H-B Kuntze and H Haffner. Experiences with the development of a robot for smart multisensoric pipe inspection. In *Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No. 98CH36146)*, volume 2, pages 1773–1778. IEEE, 1998.

D W Kurtz. Developments in Free-Swimming Acoustic Leak Detection System For Transmission Pipelines. *The Conference and Exposition for AWWA DSS*, 2006. doi: 10.1061/40854(211)25.

David W Kurtz. Case studies for a free-swimming acoustic leak detection system used in large diameter transmission pipelines. In *Pipelines 2007: Advances and Experiences with Trenchless Pipeline Projects*, pages 1–4. ASCE, 2007.

A Lambert. Accounting for losses: The bursts and background estimates (BABE) concept. *Water Environ.*, 8(2):205—-214, 2004.

AO Lambert. Accounting for losses: The bursts and background concept. *Water and Environment Journal*, 8(2):205—-214, 1994.

AO Lambert, TG Brown, M Takizawa, and D Weimer. A review of performance indicators for real losses from water supply systems. *Journal of Water Supply: Research and Technology—AQUA*, 48(6):227–237, 1999.

Kevin E Lansey, W El-Shorbagy, I Ahmed, J Araujo, and CT Haan. Calibration assessment and data collection for water distribution networks. *Journal of Hydraulic Engineering*, 127(4):270–279, 2001.

Kevin Laven, Xiangjie Kong, and Rasko Ojdrovic. Condition assessment of in-service ferous mains. In *Pipelines 2010: Climbing New Peaks to Infrastructure Reliability: Renew, Rehab, and Reinvest*, pages 899–908. ASCE Online Library, 2010.

Ayed Lazhar, Lamjed Hadj-Taïeb, and Ezzeddine Hadj-Taïeb. Two leaks detection in viscoelastic pipeline systems by means of transient. *Journal of Loss Prevention in the Process Industries*, 26(6):1341–1351, 2013. ISSN 09504230. doi: 10.1016/j.jlp.2013.08.007.

Y Le Cun, B Boser, J S Denker, D Henderson, R E Howard, W Hubbard, and L D Jackel. Handwritten Digit Recognition with a Back-Propagation Network. In *NIPS 1990*, pages 396–404, 1990.

Mark W LeChevallier, Richard W Gullick, Mohammad R Karim, Melinda Friedman, and James E Funk. The potential for health risks from intrusion of contaminants into the distribution system from pressure transients. *Journal of Water and Health*, 1(1):3–14, 2003.

Pedro Lee, M F Lambert, A R Simpson, J P Vítkovsky, and D Misiunas. Leak location in single pipelines using transient reflections. *Australian Journal of Water Resources*, 11 (1):53–65, 2007.

Pedro J Lee, John P Vítkovskỳ, Martin F Lambert, Angus R Simpson, and James A Liggett. Frequency domain analysis for detecting pipeline leaks. *Journal of Hydraulic Engineering*, 131(7):596–604, 2005.

Della J Leslie-Milbourne, Alan E Vardy, and Arris S Tijsseling. Transient fsi in a pipe system with elbow and tee junction. In *Proc. of the 8th Int. Conf. on Flow-Induced Vibration, FIV2004, Paris, France*, volume 1, pages 355–360, 2004.

S.-S. Leu and Q.-N. Bui. Leak Prediction Model for Water Distribution Networks Created Using a Bayesian Network Learning Approach. *Water Resources Management*, 30(8): 2719–2733, 2016. ISSN 15731650. doi: 10.1007/s11269-016-1316-8.

Norman Levinson. The wiener (root mean square) error criterion in filter design and prediction. *Journal of Mathematics and Physics*, 25(1-4):261–278, 1946.

Rui Li, K. Huang, H., Xin, and Tao Tao. A Review of Methods for Burst/Leakage Detection and Location in Water Distribution Systems. *Water Science & Technology: Water Supply*, 15(3):429–441, 2015. ISSN 1606-9749. doi: 10.2166/ws.2014.131.

Zhenlin Li, Haifeng Zhang, Dongjie Tan, Xin Chen, and Hongxiang Lei. A novel acoustic emission detection module for leakage recognition in a gas pipeline valve. *Process Safety and Environmental Protection*, 105:32–40, 2017.

Thomas Lumley. Kendall's advanced theory of statistics. volume 2a: classical inference and the linear model. alan stuart, keith ord and steven arnold, arnold, london, 1998, no. of pages: xiv+ 885. price:£ 85.00. isbn 0-340-66230-1. *Statistics in Medicine*, 19(22): 3139–3140, 2000.

Yifan Ma, Yan Gao, Xiwang Cui, Michael J Brennan, Fabricio CL Almeida, and Jun Yang. Adaptive phase transform method for pipeline leakage detection. *Sensors*, 19(2): 310, 2019.

John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4): 561–580, 1975.

Alberto Martini, Marco Troncossi, and Alessandro Rivola. Vibroacoustic measurements for detecting water leaks in buried small-diameter plastic pipes. *Journal of Pipeline Systems Engineering and Practice*, 8(4):04017022, 2017.

John Mashford, Dhammika De Silva, Stewart Burn, and Donavan Marney. Leak Detection in Simulated Water Pipe Networks Using Svm. *Applied Artificial Intelligence*, 26(5): 429–444, 2012. ISSN 0883-9514. doi: 10.1080/08839514.2012.670974.

J May. Leakage, pressure and control. In *BICS International Conf. on Leakage Control, London*, 1994a.

John May. Pressure dependent leakage. *World water and environmental engineering*, 17 (8):10, 1994b.

L. Mays, editor. *Ancient Water Technologies*. Springer Netherlands, Dordrecht, 2010. ISBN 978-90-481-8631-0 978-90-481-8632-7. URL http://link.springer.com/10.1007/978-90-481-8632-7.

Roderick P McDonald. The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23(1):1–21, 1970.

Ministry of the Solicitor General. Ofm - tg-03-1999: Fire protection water supply guideline for part 3 of the ontario building code, Mar 2016. URL http://www.mcscs.jus.gov.on.ca/english/FireMarshal/Legislation/TechnicalGuidelinesandReports/TG-1999-03.html.

Dalius Misiunas, Martin F. Lambert, Angus R. Simpson, and Gustaf Olsson. Burst detection and location in water distribution networks. *Water Science & Technology*, 5(3-4): 71–80, 2005a.

Dalius Misiunas, John Vitkovsky, Gustaf Olsson, Angus Simpson, and Martin Lambert. Pipeline break detection using pressure transient monitoring. *Journal of Water Resources Planning and Management*, 131(4):316–325, 2005b.

Dalius Misiunas, Martin Lambert, and Angus Simpson. Transient-based periodical pipeline leak diagnosis. In *Water Distribution Systems Analysis Symposium 2006*, pages 1–19, 2008. doi: 10.1061/40941(247)55.

J. Morrison, S. Tooms, and D. Rogers. District metered areas: Guidance notes, 2007.

Gaudenz Moser, Stephanie German Paal, and Ian FC Smith. Performance comparison of reduced models for leak detection in water distribution networks. *Advanced Engineering Informatics*, 29(3):714–726, 2015.

SR. Mounce, J. Machell, and JB. Boxall. Development of artificial intelligence systems for analysis of water supply system data. In *Water Distribution Systems Analysis Symposium*, pages 1–15, 2006.

SR Mounce, JB Boxall, and J. Machell. Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows. *Water Resources Planning and Management*, 136(3):309–318, 2010. ISSN 0733-9496. doi: 10.1061/(ASCE)WR.1943-5452.0000030.

Stephen R Mounce and John Machell. Burst detection using hydraulic data from water distribution systems with artificial neural networks. *Urban Water Journal*, 3(1):21–31, 2006.

Stephen R. Mounce, John Machell, and Joby B. Boxall. Development of artificial intelligence systems for analysis of water supply system data. *Water Distribution Systems Analysis Symposium 2006*, pages 1–15, 2008. doi: 10.1061/40941(247)91.

Stephen R. Mounce, Richard B. Mounce, and Joby B. Boxall. Novelty detection for time series data analysis in water distribution systems using support vector machines. *Journal of Hydroinformatics*, 13(4):672, 2011. ISSN 1464-7141. doi: 10.2166/hydro.2010.144.

Witness Mpesha, Sarah L Gassman, and M Hanif Chaudhry. Leak detection in pipes by frequency response method. *Journal of Hydraulic Engineering*, 127(2):134–147, 2001.

JM Muggleton and MJ Brennan. Leak noise propagation and attenuation in submerged plastic water pipes. *Journal of Sound and Vibration*, 278(3):527–537, 2004.

JM Muggleton, MJ Brennan, and RJ Pinnington. Wavenumber prediction of waves in buried pipes for water leak detection. *Journal of Sound and Vibration*, 249(5):939–954, 2002.

JM Muggleton, MJ Brennan, and PW Linford. Axisymmetric wave propagation in fluid-filled pipes: wavenumber measurements in in vacuo and buried pipes. *Journal of Sound and Vibration*, 270(1-2):171–190, 2004.

Mukkamala, Srinivas, Guadalupe Janoski, and Andrew Sung. Intrusion detection using neural networks and support vector machines. In *IJCNN'02. Proceedings of the 2002 International Joint Conference on Neural Networks*, 2002.

Harrison E. Mutikanga. *Water Loss Management: Tools and Methods for Developing Countries*. PhD thesis, Delft University of Technology, 2012.

Harrison E Mutikanga, Saroj K Sharma, and Kalanithy Vairavamoorthy. Methods and Tools for Managing Losses in Water Distribution Systems. *JOURNAL OF WATER RESOURCES PLANNING AND MANAGEMENT*, 139(April):166–174, 2013. doi: 10.1061/(ASCE)WR.1943-5452.0000245.

Amir AF Nassiraei, Yoshinori Kawamura, Alireza Ahrary, Yoshikazu Mikuriya, and Kazuo Ishii. A new approach to the sewer pipe inspection: Fully autonomous mobile robot" kantaro". In *IECON 2006-32nd Annual Conference on IEEE Industrial Electronics*, pages 4088–4093. IEEE, 2006.

National Fire Protection Association. Fire Code, 2015.

National Instruments. Understanding ffts and windowing. URL https://learn.ni.com/badges/resources/1022.

Arnold Neumaier and Tapio Schneider. Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS)*, 27(1):27–57, 2001.

Neutrium. Speed of sound in fluids and fluids in pipes, 2014. URL https://neutrium.net/fluid{_}flow/speed-of-sound-in-fluids-and-fluid-in-pipes/.

Si Tran Nguyen Nguyen, Jinzhe Gong, Martin F Lambert, Aaron C Zecchin, and Angus R Simpson. Least squares deconvolution for leak detection with a pseudo random binary sequence excitation. *Mechanical Systems and Signal Processing*, 99:846–858, 2018.

Lei Ni, Juncheng Jiang, Yong Pan, and Zhirong Wang. Leak location of pipelines based on characteristic entropy. *Journal of Loss Prevention in the Process Industries*, 30(1): 24–36, 2014. ISSN 09504230. doi: 10.1016/j.jlp.2014.04.004.

NWRI and Meteorological Service of Canada. Threats To Water Availability in Canada, 2004.

Aaron M Oirere, Ganesh B Janvale, and Ratnadeep R Deshmukh. Automatic speech recognition and verification using lpc, mfcc and svm. *Muranga University of Technology Institutional Repository — School of Computing and IT*, 2015.

A. V. Oppenheim, A. S. Willsky, and S.H. Nawab. *Signals and systems*. Prentice Hall, Upper Saddle River, NJ, 2nd ed. edition, 2008.

Alex Wang Osama Hunaidi. A new system for locating leaks in urban water distribution pipes. *Management of Environmental Quality*, 17(4), 2006. ISSN 1477-7835. doi: 10.1108/14777830610700928.

Didem Ozevin and Hazim Yalcinkaya. New Leak Localization Approach in Pipelines Using Single-Point Measurement. *Journal of Pipeline Systems Engineering and Practice*, 04013020(8):1–8, 2013. ISSN 1949-1190. doi: 10.1061/(ASCE)PS.1949-1204.0000163.

Anubha Panchal, Ketakee Dagade, Shubhangi Tamhane, Kiran Pawar, and Pradnya Ghadge. Automated Water Supply System and Water Theft Identification Using PLC and SCADA. *Journal of Engineering Research and Applications*, 4(4):67–69, 2014.

KA Papadopoulou, MN Shamout, B Lennox, D Mackay, AR Taylor, JT Turner, and X Wang. An evaluation of acoustic reflectometry for leakage and blockage detection. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 222(6):959–966, 2008.

Michael Parker. *Digital Signal Processing 101: Everything you need to know to get started*. Newnes, 2017.

Rohit Patel, Er Mukesh Kumar, AK Jaiswal, and Romini Saxena. Design technique of bandpass fir filter using various window function. *IOSR journal of electronics and communication engineering*, 6:52–57, 2013.

R Perez, V Puig, J Pascual, A Peralta, E Landeros, and Ll Jordanas. Pressure sensor distribution for leak detection in barcelona water distribution network. *Water science and technology: water supply*, 9(6):715–721, 2009.

Ramon Perez, Gerard Sanz, Vicenc Puig, Joseba Quevedo, Miquel Angel Cuguero Escofet, Fatiha Nejjari, Jordi Meseguer, Gabriela Cembrano, Josep M Mirats Tur, and Ramon Sarrate. Leak localization in water networks: a model-based methodology using pressure sensors applied to a real network in barcelona [applications of control]. *IEEE Control Systems*, 34(4):24–36, 2014.

Mario Peruggia. Model selection and multimodel inference: A practical information-theoretic approach (2n ed.). *Journal of the American Statistical Association*, 98(463): 778, 2003.

RJ Pinnington and AR Briscoe. Externally applied sensor for axisymmetric waves in a fluid filled pipe. *Journal of Sound and vibration*, 173(4):503–516, 1994.

Boaz Porat. *A course in digital signal processing*, volume 1. Wiley New York, 1997.

Z. Poulakis, D. Valougeorgis, and C. Papadimitriou. Leakage detection in water pipe networks using a Bayesian probabilistic framework. *Probabilistic Engineering Mechanics*, 18(4):315–327, 2003. ISSN 02668920. doi: 10.1016/S0266-8920(03)00045-6.

Emanuele Principi, Fabio Vesperini, Stefano Squartini, and Francesco Piazza. Acoustic novelty detection with adversarial autoencoders. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 3324–3330. IEEE, 2017.

Ranko S Pudar and James A Liggett. Leaks in pipe networks. *Journal of Hydraulic Engineering*, 118(7):1031–1046, 1992.

Raido Puust, Zoran Kapelan, Dragan Savic, and Tiit Koppel. Probabilistic leak detection in pipe networks using the scem-ua algorithm. In *Water Distribution Systems Analysis Symposium 2006*, pages 1–12, 2008.

Zhexian Qi, Feifei Zheng, Danlu Guo, Holger R Maier, Tuqiao Zhang, Tingchao Yu, and Yu Shao. Better understanding of the capacity of pressure sensor systems to detect pipe burst within water distribution networks. *Journal of Water Resources Planning and Management*, 144(7):04018035, 2018.

Lawrence R Rabiner, Ronald W Schafer, et al. Introduction to digital speech processing. *Foundations and Trends® in Signal Processing*, 1(1–2):1–194, 2007.

Alina Racoviceanu, Bryan Karney, Christopher Kennedy, and Andrew F. Colombo. Life-Cycle Energy Use and Greenhouse Gas Emissions Inventory for Water Treatment Systems. *Journal of Infrastructure Systems*, 13(4):261–270, 2007. ISSN 1076-0342. doi: 10.1061/(ASCE)1076-0342(2007)13:4(261).

Robert Bond Randall. *Vibration-based condition monitoring: industrial, aerospace and automotive applications*. John Wiley & Sons, 2011.

Sidra Rashid, Usman Akram, and Shoab A Khan. Wml: wireless sensor network based machine learning for leakage detection and size estimation. *Procedia Computer Science*, 63:171–176, 2015.

M Romano, Z Kapelan, and D Savić. Real-time leak detection in water distribution systems. *Water Distribution Systems Analysis 2010*, pages 1074–1082, 2011. doi: 10.1061/41203(425)97.

Erich Rome, Joachim Hertzberg, Frank Kirchner, Ulrich Licht, and Thomas Christaller. Towards autonomous sewer robots: the makro project. *Urban Water*, 1(1):57–70, 1999.

Ali Sadeghioon, Nicole Metje, David Chapman, and Carl Anthony. Smartpipes: smart wireless sensor networks for leak detection in water pipelines. *Journal of sensor and Actuator Networks*, 3(1):64–78, 2014.

Haleh Safavi and Chein-I Chang. Projection pursuit-based dimensionality reduction. In *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XIV*, volume 6966, page 69661H. International Society for Optics and Photonics, 2008.

A Ejah Umraeni Salam, Muh Tola, Mary Selintung, and Farouk Maricar. A leakage detection system on the water pipe network through support vector machine method.

In *2014 Makassar International Conference on Electrical Engineering and Informatics (MICEEI)*, pages 161–165. IEEE, 2014.

S. Sanei and H. Hassani. *Singular spectrum analysis of biomedical signals*. Springer, 2015.

T. Sato and A. Mitra. Leak detection using the pattern of sound signals in water supply systems. *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems*, 6529:1–9, 2007.

B. Scholkopf, JC Platt, J Shawe-Taylor, AJ Smola, and RC Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2): 461–464, 1978.

sensoft. pulseekko®: Deep ground penetrating radar, 2020. URL https://www.sensoft.ca/products/pulseekkopro/overview-pulseekko/.

Yong Woo Shin, Min Soo Kim, and Sang Kwon Lee. Identification of acoustic wave propagation in a duct line and its application to detection of impact source location based on signal processing. *Journal of mechanical science and technology*, 24(12):2401–2411, 2010.

A Shulemovich. Flow-induced vibrations caused by roughness in pipes conveying fluid. *Journal of Applied Mechanics*, 53(1):181–186, 1986.

Reinaldo a. Silva, Claudio M. Buiatti, Sandra L. Cruz, and João a.F.R. Pereira. Pressure wave behaviour and leak detection in pipelines. *Computers & Chemical Engineering*, 20 (96):S491–S496, 1996. ISSN 00981354. doi: 10.1016/0098-1354(96)00091-9.

Sandro Skansi. Convolutional neural networks. In *Introduction to Deep Learning*, pages 121–133. Springer, 2018.

L. A. Smith, K. A. Fields, A. S. C. Chen, and A. N. Tafuri. Leak and Break Detection and Repair of Drinking Water Systems, 2000.

Alexandre Kepler Soares, Dídia I Covas, and Luisa Fernanda Reis. Analysis of pvc pipe-wall viscoelasticity during water hammer. *Journal of Hydraulic Engineering*, 134(9): 1389–1394, 2008. ISSN 0733-9429. doi: 10.1061/(ASCE)0733-9429(2008)134:9(1389).

Alexandre Kepler Soares, Didia I. C. Covas, and Luisa Fernanda R. Reis. Leak detection by inverse transient analysis in an experimental PVC pipe system. *Journal of Hydroinformatics*, 13(2):153, 2011. ISSN 1464-7141. doi: 10.2166/hydro.2010.012.

S. Srirangarajan, M. Allen, A. Preis, M. Iqbal, H. B. Lim, and A. J. Whittle. Wavelet-based burst event detection and localization in water distribution systems. *Journal of Signal Processing Systems*, 72(1):1–16, 2013.

Pavol Stajanca, Sebastian Chruscicki, Tobias Homann, Stefan Seifert, Dirk Schmidt, and Abdelkarim Habib. Detection of leak-induced pipeline vibrations using fiber—optic distributed acoustic sensing. *Sensors*, 18(9):2841, 2018.

KN Stevens. Autocorrelation analysis of speech sounds. *The Journal of the Acoustical Society of America*, 22(6):769–771, 1950.

I. Stoianov, L. Nachman, S. Madden, T. Tokmouline, and M. Csail. PIPENET: A Wireless Sensor Network for Pipeline Monitoring. In *6th International Symposium on Information Processing in Sensor Networks*, pages 264–273, apr 2007. doi: 10.1109/IPSN.2007.4379686.

Ivan Stoianov, Lama Nachman, Andrew Whittle, Sam Madden, and Ralph Kling. Sensor Networks for Monitoring Water Supply and Sewer Systems: Lessons from Boston. *Water Distribution Systems Analysis Symposium 2006*, pages 1–17, 2006. doi: 10.1061/40941(247)100.

SubSurface Leak Detection. Zcorr digital correlating loggers, 2019. URL http://www.subsurfaceleak.com/zcorr_logger_prod.html.

AL-Washali Taha, Saroj Sharma, and Maria Kennedy. Methods of assessment of water losses in water supply systems: a review. *Water Resources Management*, 30(14):4985–5001, 2016.

Lamjed Hadj Taieb, Lazhar Ayed, and Ezzeddine Hadj Taieb. Leak detection in viscoelastic pipe by transient analysis. In *Condition monitoring of machinery in non-stationary operations*, pages 69–79. Springer, 2012.

Lizhe Tan and Jean Jiang. *Digital signal processing: fundamentals and applications*. Academic Press, 2018.

Yuriko Terao and Akira Mita. Robust water leakage detection approach using the sound signals and pattern recognition. *Sensors and Smart Structures Technologies for Civil,*

Mechanical, and Aerospace Systems 2008, Pts 1 and 2, 6932:69322D–69322D–9, 2008. ISSN 0277-786X. doi: 10.1117/12.775968.

IWA The World Bank. The world bank and the international water association to establish a partnership to reduce water losses, Sep 2016. URL https://www.worldbank.org/en/news/press-release/2016/09/01/the-world-bank-and-the-international-water-association-to-establish-a-partnership-to

The World Counts. Water, water everywhere... but not a drop to drink, 2020. URL https://www.theworldcounts.com/stories/average-daily-water-usage.

ARD Thorley. Pressure transients in hydraulic pipelines. In *ASME*, 1968.

J Thornton. *Water Loss Control Manual*. New York: McGraw-Hill, 2002.

Matthew M Torok, Mani Golparvar-Fard, and Kevin B Kochersberger. Image-based automated 3d crack detection for post-disaster building assessment. *Journal of Computing in Civil Engineering*, 28(5):A4014004, 2013.

University of Sheffield. Leaky pipes can allow contaminants into our drinking water, jun 2015. URL http://phys.org/news/2015-06-leaky-pipes-contaminants.html.

Melissa Valentine, editor. *Water Audits and Loss Control Programs - Manual of Water Supply Practices, M36*. American Water Works Association, 3rd edition, 2009.

Jakobus E van Zyl and AM Cassa. Modeling elastically deforming leaks in water distribution pipes. *Journal of Hydraulic Engineering*, 140(2):182–189, 2014.

Ali Moradi Vartouni, Saeed Sedighian Kashi, and Mohammad Teshnehlab. An anomaly detection method to detect web attacks using stacked auto-encoder. In *2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, pages 131–134. IEEE, 2018.

A Vassiljev, T Koppel, R Puust, D Savic, King G Walters, ST Khu, et al. Calibration of the model of an operational water distribution system. In *Proceedings of the fifth international conference on engineering computational technology, Stirlingshire, United Kingdom: Civil-Comp Press*, 2005.

R. Vautard and M. Ghil. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D: Nonlinear Phenomena*, 35(3):395–424, 1989.

Jorge Vento. Leak Detection and Isolation in Pressurized Water Pipe Networks using Interval LPV Models. *Networks*, pages 36–41, 2009.

D. J. Vicente, L. Garrote, R. Sanchez, and D. Santillan. Pressure management in water distribution systems: Current status, proposals, and future trends. *Journal of Water Resources Planning and Management*, 142(2), 2016.

François Vince, Emmanuelle Aoustin, Philippe Bréant, and François Marechal. LCA tool for the environmental evaluation of potable water production. *Desalination*, 220(1–3): 37–56, mar 2008. ISSN 0011-9164. doi: 10.1016/j.desal.2007.01.021.

Hans Von Storch. Misuses of statistical analysis in climate research. In *Analysis of Climate Variability*, chapter 15, pages 11–26. Springer, 1999.

Hans von Storch and Francis W Zwiers. Statistical analysis in climate research, 2002.

Lynn P. Wallace. *Water and Revenue Losses: Unaccounted for Water*. American Water Works Foundation, 1987. ISBN 0898674174.

L Wang, A Chen, SA Flamberg, JB Nestleroth, M Royer, and AF Williams. Field demonstration of innovative leak detection/location technologies in conjunction with pipe wall thickness testing for water mains. In *Pipelines 2010: Climbing New Peaks to Infrastructure Reliability: Renew, Rehab, and Reinvest*, pages 1188–1198. American Society of Civil Engineers, 2010.

Mike West and Jeff Harrison. *Bayesian forecasting and dynamic models*. Springer Science & Business Media, 2006.

Andrew J. Whittle, Lewis Girod, Ami Preis, Michael Allen, Hock Beng Lim, Mudasser Iqbal, Seshan Srirangarajan, Cheng Fu, Kai Juan Wong, and Daniel Goldsmith. WaterWiSe@SG: A Testbed for Continuous Monitoring of the Water Distribution System in Singapore. *Water Distribution Systems Analysis 2010*, pages 1362–1378, 2010. doi: 10.1061/41203(425)122.

Andrew J Whittle, Michael Allen, Ami Preis, and Mudasser Iqbal. Sensor networks for monitoring and control of water distribution systems. In *Pipelines 2012: Innovations in Design, Construction, Operations, and Maintenance, Doing More with Less*. International Society for Structural Health Monitoring of Intelligent Infrastructure, 2013.

Pierre Wickramarachi. Effects of windowing on the spectral content of a signal. *Sound and vibration*, 37(1):10–13, 2003.

G. Williams and G. Kuczera. Analyzing SCADA to understand the contribution of hydraulic pressures to trunk-main failure. *Procedia Engineering*, 89:1452–1459, 2014. ISSN 18777058. doi: 10.1016/j.proeng.2014.11.472.

Worldometer. Current world population, 2019. URL https://www.worldometers.info/world-population/.

Yipeng Wu and Shuming Liu. A review of data-driven approaches for burst detection in water distribution systems. *Urban Water Journal*, 14(9):972–983, 2017.

Zheng Yi Wu and Paul Sage. Water loss detection via genetic algorithm optimization-based model calibration. In *Water Distribution Systems Analysis Symposium 2006*, pages 1–11. ASCE, 2008.

Zheng Yi Wu, Paul Sage, and David Turtle. Pressure-Dependent Leak Detection Model and Its Application to a District Water System. *Journal of Water Resources Planning and Management*, 136(1):116–128, 2010. ISSN 0733-9496. doi: 10.1061/(ASCE)0733-9496(2010)136:1(116).

Jin Yang, Yumei Wen, and Ping Li. Leak location using blind system identification in water distribution pipelines. *Journal of Sound and Vibration*, 310(1-2):134–148, 2008. ISSN 10958568. doi: 10.1016/j.jsv.2007.07.067.

Jin Yang, Yumei Wen, Ping Li, and Xingke Wang. Study on an improved acoustic leak detection method for water distribution systems. *Urban Water Journal*, 10(2):71–84, 2013.

Xincong Yang, Heng Li, Yantao Yu, Xiaochun Luo, Ting Huang, and Xu Yang. Automatic pixel-level crack detection and measurement using fully convolutional network. *Computer-Aided Civil and Infrastructure Engineering*, 33(12):1090–1109, 2018.

Guoliang Ye and Richard Andrew Fenner. Kalman Filtering of Hydraulic Measurements for Burst Detection in Water Distribution Systems. *Journal of Pipeline Systems Engineering and Practice*, 2(1):14–22, 2011. ISSN 1949-1190. doi: 10.1061/(ASCE)PS.1949-1204.0000070.

B Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

Shen Yin, Xiangping Zhu, and Chen Jing. Fault detection based on a robust one class support vector machine. *Neurocomputing*, 145:263–268, 05 2014.

L. Yufeng and J. Saniie. Singular spectrum analysis for trend extraction in ultrasonic backscattered echoes. *Proc. IEEE Int. Ultrasonics Symposium*, pages 1–4, 2015.

Yves Filion, MacLean, H., and Karney, B. Life-Cycle Energy Analysis of a Water Distribution System. *Journal of Infrastructure Systems*, 10(3):120–130, 2004. ISSN 1076-0342. doi: 10.1061/(ASCE)1076-0342(2004)10:3(119).

Min-Ling Zhang and Zhi-Hua Zhou. A k-nearest neighbor based algorithm for multi-label classification. In *Granular Computing, 2005 IEEE International Conference on*, volume 2, pages 718–721. IEEE, 2005.

Qingzhou Zhang, Zheng Yi Wu, Ming Zhao, Jingyao Qi, Yuan Huang, and Hongbin Zhao. Leakage zone identification in large-scale water distribution systems using multiclass support vector machines. *Journal of Water Resources Planning and Management*, 142 (11):04016042, 2016.

APPENDICES

Appendix A

List of Publications

The following is a list of journal publications and conference papers resulting from the work contained in this dissertation:

Publications

1. **Cody, R.**, Harmouche, J., and Narasimhan, S. (2018). Leak detection in water distribution pipes using singular spectrum analysis. *Urban Water Journal, 15(7)*, 636-644.

2. **Cody, R. A.**, Dey, P., and Narasimhan, S. (2020). Linear prediction for leak detection in water distribution networks. *Journal of Pipeline Systems Engineering and Practice, 11(1)*, 04019043.

3. **Cody, R. A.**, Tolson, B. A., and Orchard, J. (2020). Detecting leaks in water distribution pipes using a deep autoencoder and hydroacoustic spectrograms. *Journal of Computing in Civil Engineering, 34(2)*, 04020001.

4. **Cody, R.** and Narasimhan, S.. A Field implementation of linear prediction for leak-monitoring in water distribution networks. *Advanced Engineering Informatics* (undergoing revisions).

Conference Proceedings - Full paper

1. **Cody, R.**, Narasimhan, S., and Tolson, B. (2017). One-class SVM–leak detection in water distribution systems. *Proc., Computing and Control for the Water Industry*, CCWI 2017.

Appendix B

Cholesky Decomposition

The Cholesky decomposition (or Cholesky factorization) factors a Hermitian, positive-definite matrix into the produce of a lower triangular matrix and its conjugate transpose, such that $A = L^T L$. It can be computed by a form of Gaussian elimination that takes advantage of the symmetry and definiteness, such that each element can be computed as [Benoıt, 1924],

$$A(i,j) = \begin{cases} a_{i,i} = \sum_{k=1}^{i} l_{k,i}^2 & : j = i, \\ a_{i,j} = \sum_{k=1}^{i} l_{k,i} l_{k,j} & : j > i. \end{cases} \tag{B.1}$$

This can be expressed in terms of L as,

$$L(i,j) = \begin{cases} l_{i,i} = \sqrt{a_{i,i} - \sum_{k=1}^{i-1} l_{k,i}^2}, \\ l_{i,j} = \frac{1}{l_{j,j}} \left(a_{i,j} - \sum_{k=1}^{j-1} l_{i,k} l_{j,k} \right). \end{cases} \tag{B.2}$$

Appendix C

Head Tank Pressure

This summary attempts to statistically characterize the variability that exists in the overall system pressure within a live WDN. This can be done by studying the variability in the measured pressure in the head tank, located in the City of Guelph, Ontario where the field tests have been conducted. A histogram of the head tank pressures for the dates 10/13/2018, 10/21/2018 and11/03/2018 are generated and shown in Figure C.1.

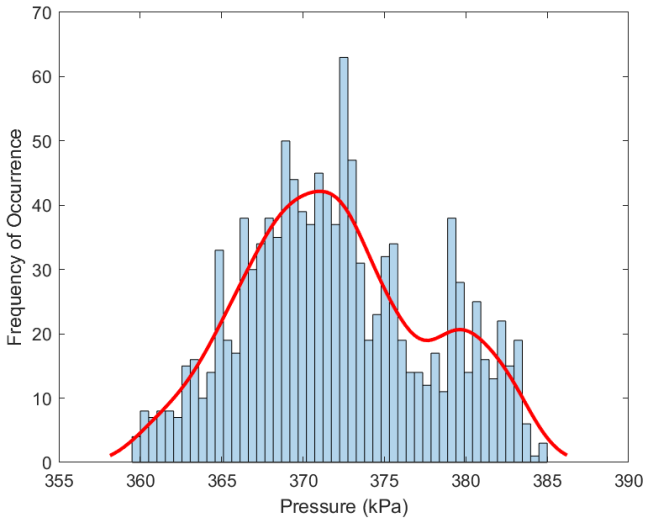


Figure C.1: Distribution of leak and leak free pressure data.

Appendix D

Equation of motion of a membrane

The following derivation for the equation of motion of a membrane is taken from Kinsler et al. [1999]. It is included within this dissertation for the sake of completeness.

If the tension per unit length, T is uniform at all points and the deflections $u(x, y, t)$ during motion are small, then according to Figure D.1, the net vertical forces in the x-axis and y-axis, respectively, are,

$$\begin{aligned} T\Delta y(sin\beta - sin\alpha) &\approx T\Delta y(tan\beta - tan\alpha) \\ &= T\Delta y(\frac{\partial u}{\partial x}|_{x+\Delta x, y_1} - \frac{\partial u}{\partial x})|_{x,y_2}) \\ T\Delta x(sin\beta - sin\alpha) &\approx T\Delta x(tan\beta - tan\alpha) \\ &= T\Delta x(\frac{\partial u}{\partial y}|_{x_1, y+\Delta y} - \frac{\partial u}{\partial y})|_{x_2,y}). \end{aligned} \tag{D.1}$$

By applying Newton's law and summing the forces in equation D.1 for equilibrium,

$$\frac{\partial^2 u}{\partial t^2} = \frac{T}{\rho}[\frac{\frac{\partial u}{\partial x}|_{x+\Delta x,y_1} - \frac{\partial u}{\partial x})|_{x,y_2}}{\Delta x} + \frac{(\frac{\partial u}{\partial y}|_{x_1,y+\Delta y} - \frac{\partial u}{\partial y})|_{x_2,y})}{\Delta y}] \tag{D.2}$$

$$= c^2(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}) \tag{D.3}$$

$$= c^2 \nabla^2 u \tag{D.4}$$

where $c^2 = \frac{T}{\rho}$ and ∇^2 is the Laplacian operator.

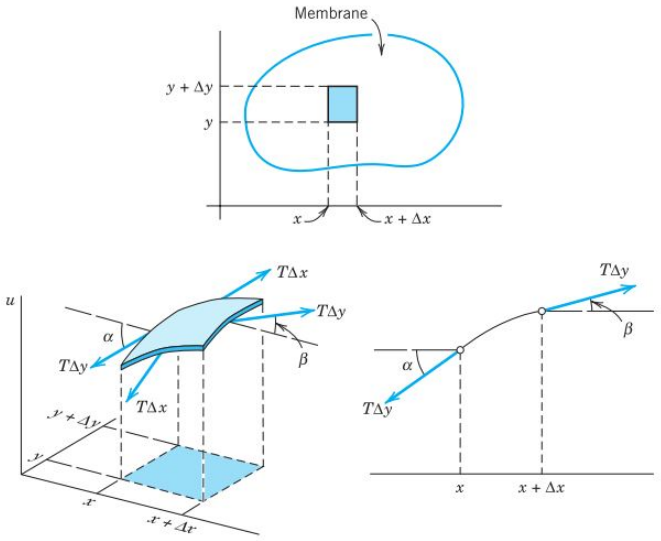


Figure D.1: Forces acting on the stretched membrane [Jeffrey, 2001].

In order to derive the equation for normal modes, the equation of motion shown in equation D.4 is assumed to have solutions of the form,

$$\mathbf{u} = \Psi e^{i\omega t}, \tag{D.5}$$

where Ψ is a function of only the position.

By substituting equation D.5 into equation D.4, and introducing wave number, (k), as $k = \frac{\omega}{c}$ gives the Helmholtz equation,

$$\nabla^2 \Psi + k^2 \Psi = 0, \tag{D.6}$$

which can be expressed in cylindrical coordinates as,

$$\frac{\partial^2 \Psi}{\partial r^2} + \frac{1}{r}\frac{\partial \Psi}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \Psi}{\partial \theta^2} + k^2 \Psi = 0. \tag{D.7}$$

By assuming $\Psi = \mathbf{R}(r)\mathbf{\Theta}(\theta)$, applying separation of variable and multiplying equation D.7 by $r^2/\mathbf{\Theta}$ and re-arranging gives,

$$\frac{r^2}{\mathbf{R}}\left(\frac{\partial^2 \mathbf{R}}{\partial r^2} + \frac{1}{r}\frac{\partial \mathbf{R}}{\partial r}\right) + k^2 r^2 = -\frac{1}{\mathbf{\Theta}}\frac{\partial^2 \mathbf{\Theta}}{\partial \theta^2}. \tag{D.8}$$

If $\frac{\partial^2 \mathbf{\Theta}}{\partial \theta^2} = -m^2 \mathbf{\Theta}$ has the harmonic solutions $\mathbf{\Theta}(\theta) = cos(m\theta + y_m)$, then, with m fixed in value, equation D.8 becomes *Bessel's equation*,

$$\frac{\partial^2 \mathbf{R}}{\partial r^2} + \frac{1}{r}\frac{\partial \mathbf{R}}{\partial r} + (k^2 - \frac{m^2}{r^2})\mathbf{R} = 0. \tag{D.9}$$

The solutions to this equation take the form of *Bessel's functions* of order m of the first kind $J_m(kr)$ and second kind $Y_m(kr)$,

$$\mathbf{R}(r) = \mathbf{A}J_m(kr) + \mathbf{B}Y_m(kr). \tag{D.10}$$

Equation D.10 are oscillatory functions of kr whose amplitude reduces roughly as $1/\sqrt{kr}$. As kr tends to 0, the function $\mathbf{B}Y_m(kr)$ becomes unbounded. However, the membrane that extends across origin should have finite displacement at $r = 0$. Thus it requires $\mathbf{B}$ in Eq. D.10 to be zero which reduces to,

$$\mathbf{R}(r) = \mathbf{A}J_m(kr). \tag{D.11}$$

At $r = a$, the boundary condition $\mathbf{R}(a) = 0$ requires $J_m(ka) = 0$. If the values of the function for J_m that cause it to be zero, are represented by j_{mn} and k_{mn}, in which $k_{mn} = j_{mn}/a$, then solutions of circular membrane with fixed rim becomes,

$$\mathbf{u}_{r,\theta,t} = \mathbf{A}_{mn}J_m(k_{mn}r)cos(m\theta + \gamma_{mn})e^{i\omega_{mn}t}, \tag{D.12}$$

where, $\mathbf{A}_{mn} = A_{mn}e^{j\phi_{mn}}$.

The fundamental frequency of the system can be obtained using zeros of the first kind $(J_m(kr))$ of Bessel's function, the speed of sound c in a pipe, and the pipe radius a,

$$f_{mn} = \frac{1}{2\pi}\frac{j_{mn}c}{a}. \tag{D.13}$$

The real part of equation D.12 gives the physical motion of (m, n)th modes,

$$u_{r,\theta,t} = A_{mn}J_m(k_{mn}r)cos(m\theta + \gamma_{mn})cos(\omega_{mn}t + \phi_{mn}). \tag{D.14}$$

Sample mode shapes can be seen in Figure 3.1.