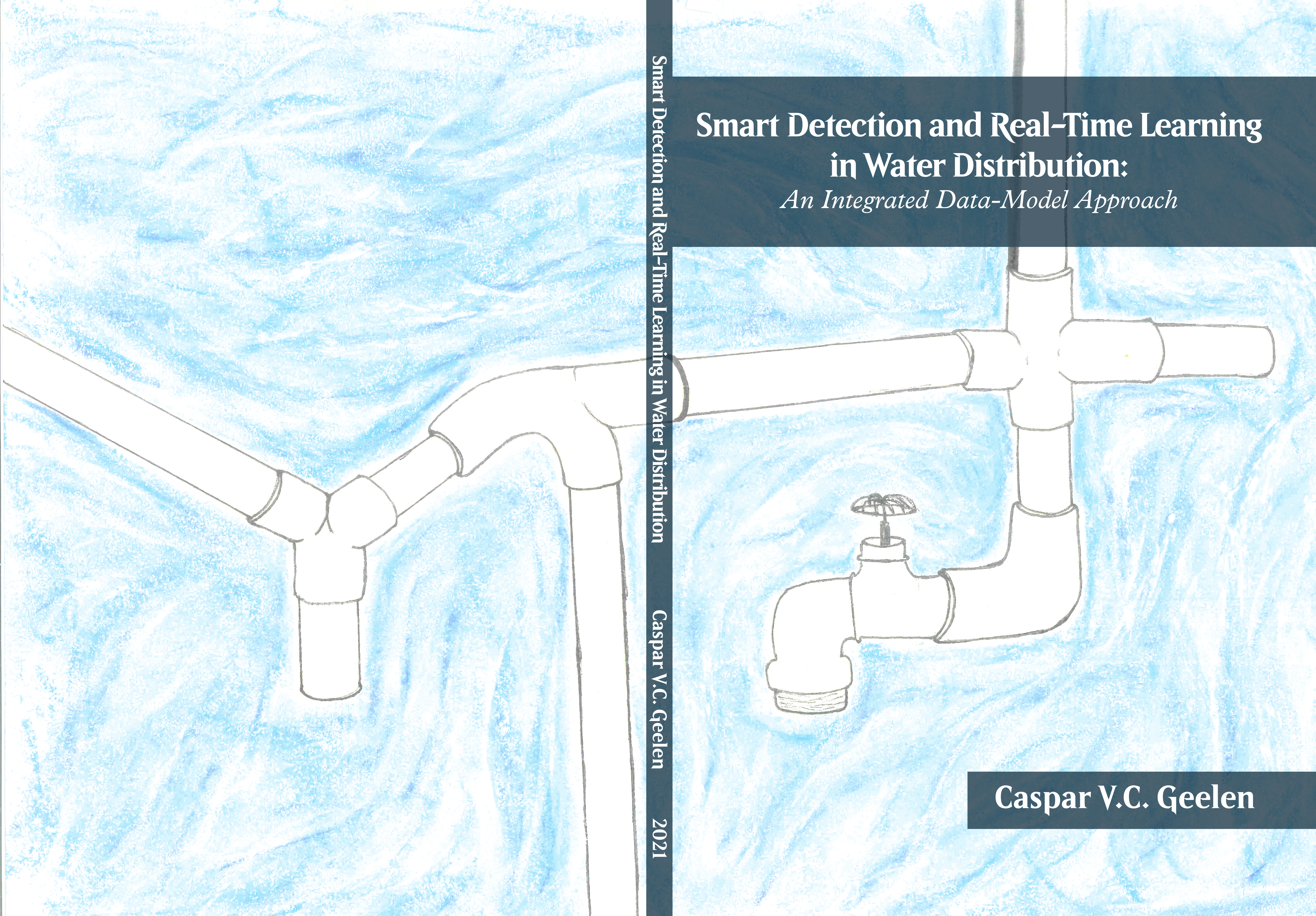


Smart Detection and Real-Time Learning in Water Distribution:

An Integrated Data-Model Approach

Caspar V.C. Geelen

Smart Detection and Real-Time Learning in Water Distribution

Caspar V.C. Geelen

2021

Propositions

1. In water distribution networks, non-burst anomalies present in high resolution pressure and flow sensor signals are underused data sources.
   (this thesis)

2. For water companies, placing sensors with the aim to improve hydraulic models is more important than optimizing burst detectability.
   (this thesis)

3. Dutch drinking water companies should merge into one single company.

4. Water companies may considerably improve their water management by using microbial communities as biosensors in addition to conventional water quality sensors.

5. For prediction purposes, data should not be shuffled before splitting in training and validation sets.

6. Jargon is the enemy of communication.

7. Low coffee quality is detrimental to work efficiency.

Propositions belonging to the thesis, entitled

Smart Detection and Real-time Learning in Water Distribution:
An Integrated Data-Model Approach

Caspar V.C. Geelen
Wageningen, 25 January 2022

Smart Detection and Real-Time Learning in Water Distribution:

An Integrated Data-Model Approach

Caspar V.C. Geelen

Thesis Committee

Promotors
Prof. Dr Karel J. Keesman
Personal chair, Applied Mathematics
Wageningen University & Research

Prof. Dr Jaap Molenaar
Professor of Applied Mathematics
Wageningen University & Research

Co-promotor
Dr Doekle R. Yntema
Scientific Project Manager
Wetsus, European Centre of Excellence for Sustainable Water Technology, Leeuwarden

Other Members
Prof. Dr Bedir Tekinerdogan, Wageningen University & Research
Prof. Dr Hans J. Zwart, University of Twente, Enschede
Prof. Dr Jan A.H. Hofman, University of Bath, United Kingdom
Dr Yvonne Hassink-Mulder, Vitens, Zwolle

This research was conducted under the auspices of the Graduate School for Socio-Economic and Natural Sciences of the Environment

Smart Detection and Real-Time Learning in Water Distribution:

An Integrated Data-Model Approach

Caspar V.C. Geelen

Thesis

Submitted in fulfilment of the requirements for the degree of doctor

at the Wageningen University

by the authority of the Rector Magnificus,

Prof. Dr. A.P.J. Mol,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Tuesday 25 January 2022

at 1:30 p.m. in the Aula.

Caspar V.C. Geelen
Smart Detection and Real-Time Learning in Water Distribution:
An Integrated Data-Model Approach
146 pages.

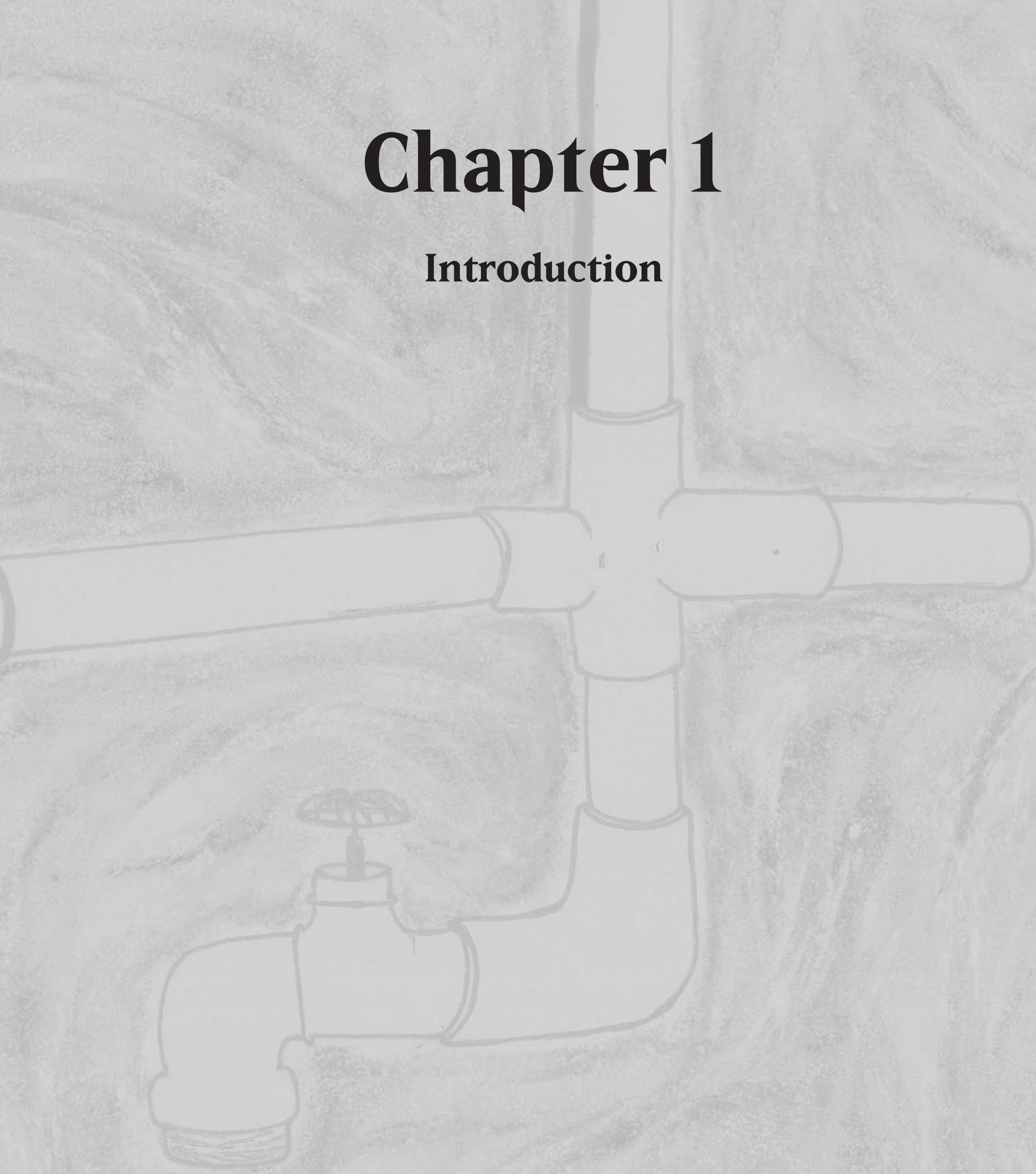
PhD Thesis, Wageningen University, Wageningen, the Netherlands (2022)
With references and summaries in English and Dutch

ISBN:	978-94-6395-994-0
DOI:	https://doi.org/10.18174/555533

Table of Contents

Chapter 1

Introduction



1. Introduction

1.1. Drinking Water in the Netherlands

Although invisible and not at the forefront of everyone's mind, there is a vast network of underground infrastructure below our feet, consisting of conduits and cables transporting, amongst others, gas, water, sewage, internet, and electricity. For every one km of visible, aboveground infrastructure in the form of car and bicycle roads, ten km of unseen infrastructure in the form of cables and pipes is present belowground (CBS; COB 2018). One of the many underground conduit networks is the drinking water distribution network (WDN).

The Netherlands is a country with a relatively high availability of fresh water sources in addition to a steadily increasing mean yearly downpour (KNMI 2020). However, this does not mean that there is sufficient access to fresh water. Although yearly downpour is increasing, only the winters are getting wetter, while recent summers are plagued with dry spells of increasing duration. Due to the drier summers and limited water storage capacity, drinking water winning from groundwater more and more often exceeds permitted levels in recent summer periods (Vitens). Therefore, even for a country with an abundance of fresh water, drinking water production and distribution remain a top priority.

A redeeming feature of the Dutch drinking water is the excellent state of the conduit networks, with a non-revenue water fraction of only 5.5%, or 1.5 $m^3km^{-1}day^{-1}$, compared to the Western European median of 7.2 $m^3km^{-1}day^{-1}$ (Vewin 2017). This non-revenue water basically covers water losses as well as unbilled water, including water from fire hydrants. Part of this success can be explained by the relatively new and well maintained conduit network, with an average conduit age of 35 years, compared to 45 and 75 years for US and UK, respectively (Vewin 2016; Folkman 2018). Although the water losses of Dutch drinking water infrastructure are relatively low, maintaining the existing infrastructure is costly and presents on average 59.2% of the investments of Dutch drinking water companies in 2016, which is the fastest growing investment category (Vewin 2017). Reducing water loss by optimizing management of the distribution network therefore presents one of the most important responsibilities of the water companies.

1.2. Water Losses

Prevention of water loss is not only important in reducing the loss of a valuable resource, bursts in drinking water conduits can also lead to contamination of the drinking water and damage to surrounding infrastructure (Scozzari et al. 2021). Conduit bursts are often caused by a combination of various factors, of which not all are easily identified. Although new conduits of plastic materials are often more durable than older iron or cement conduits, pipe bursts still occur regularly, often without a well understood cause. Although cement and iron pipes get weaker with age, ageing is almost never the sole reason of a burst. Often a final trigger will result in failure of an already degraded conduit. Temperature, traffic, vegetation, ground deposition, groundwater levels, and ground acidity, amongst others, can all play a role in conduit degradation, and can all be the final straw that breaks the camel's back.

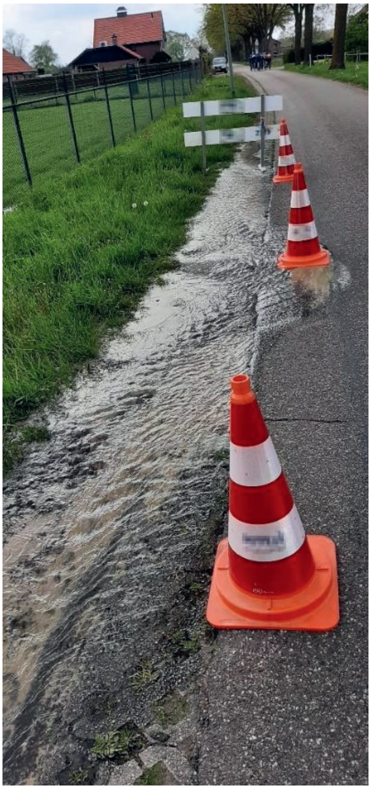


Figure 1.1 – Burst water pipe floods adjacent deer pasture.

Although conduit bursts are complex processes, action must be taken to limit water losses. In general two types of action can be distinguished: proactive and reactive leakage control strategies, based on either equipment-based or model-based approaches. Reactive leakage control entails remedying pipe bursts in a timely fashion, while proactive leakage control concerns the broader and more general field of preventing water loss by optimizing grid operation and management, using techniques such as timely replacement of ageing and damaged pipes, identification of weaker assets, optimizing network pressures and valve operations.

Due to continuous operational and environmental stress, conduits degrade with time. The condition of the conduit material is a major factor in pipe bursts, but pipe location and internal conditions also play a role in pipe bursts (Barton et al. 2019). Since internal conditions play a role, optimal management of water flow, pressure, and quality may reduce risk of bursts. Water hammers can be detrimental to already weakened pipes. Water hammers are waves of surging pressure moving with up to 1200 ms^{-1}, which arise if pressurized water is suddenly forced to stop or change direction (Chaudhry 2014). Common causes of water hammers are sudden valve movements, or sudden changes in the amount of water that is being subtracted from the network at a specific location. The many processes influencing pipe degradation and eventual breaking, of which most are not actively measured, make it very difficult to correctly predict asset conditions. Equipment-based inspection techniques are therefore important for both asset condition verification and improvement of distribution network models. Although advances are being made, pipe bursts are still at the order of the day, and therefore reactive leakage control strategies also remain of high importance.

1.3. Equipment-based Inspection

One form of proactive leakage control is timely replacement of assets, before pipe bursts can occur. In order to determine when a pipe should be replaced, the condition of the pipes needs to be assessed. Asset life expectancy of plastic pipes exceeds 100 years, compared to 50 years for cement pipes (Breen 2006; Slaats 2015). Since most cement pipes were placed in the 1970's, a significant fraction of the Dutch water infrastructure consists of aged pipes (Vewin 2017). However, short term replacement of all cement pipes is not economically viable and would be wasteful for pipes that are still in good condition. Solely replacing degraded pipes, before leakages occur, is the preferred strategy.

1.3.1. Ultrasonic Inspection

Equipment-based inspection techniques can be used to measure the condition of underground assets and can thus be used to assess asset condition. A wide selection of camera, acoustic, and ultrasonic inspection methods are available (Liu and Kleiner 2013).

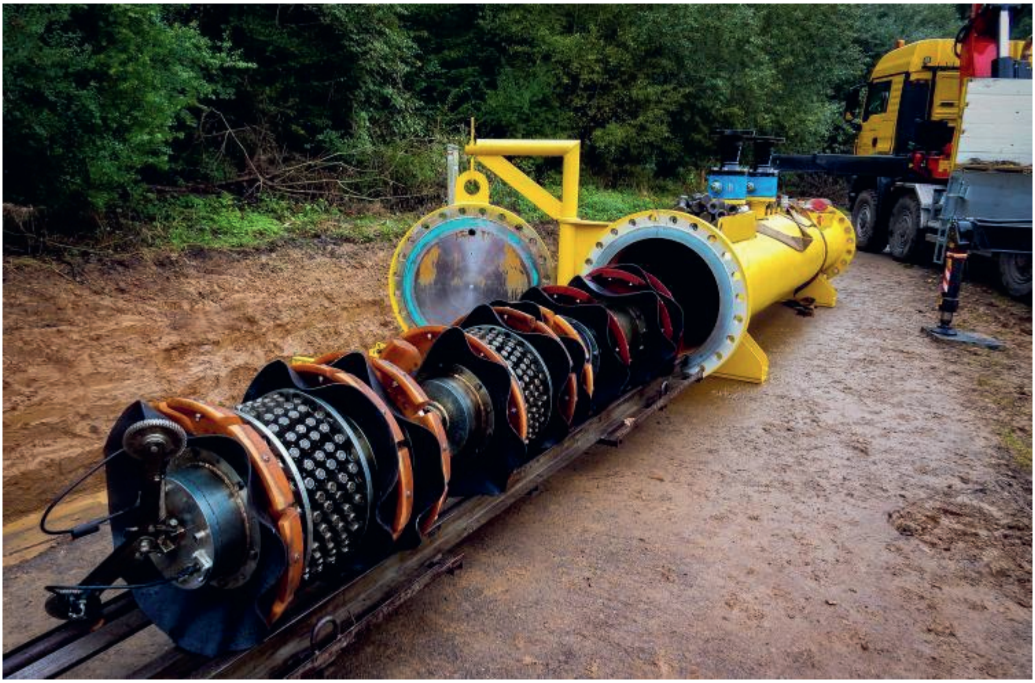


Figure 1.2 – Pipeline inspection gauge equipped with various sensors used to assess condition of water mains (Brynych 2018).

A non-destructive inspection technology currently used in practice is ultrasonic inline inspection, where a robot is sent through cement pipes in order to measure pipe condition using reflecting ultrasonic beams (Delgadillo et al. 2016, 2020). This tool can be used to create a detailed 3D-image of the condition of assets, and allows for measurements on a centimeter scale or even smaller. This equipment-based method can thus be used as a form of proactive leakage control, where damaged and degraded pipes are replaced before a burst can occur. Although not yet used for PVC water mains, promising research is being conducted about the application to PVC mains.

1.3.2. Electromagnetic Inspection

In addition to acoustic and ultrasonic methods, various electromagnetic methods are available as well. One such method employed in the field is ground penetrating radar, which is used to investigate subsurface assets in a non-invasive and non-destructive way. Radio wave electromagnetic radiation is transmitted from the device and the signals reflected from the subsurface pipe are recorded again by the radar. These reflected signals can then be used to determine presence of cracks, tears or holes in the pipes. These electromagnetic techniques can be used for both proactive and reactive leakage control, and thus be used either to assess asset conditions or to localize existing leaks. Ground penetrating radar provides lower resolution information about the condition of the pipes than ultrasonic inspection (Amran et al. 2018; De Coster et al. 2019; Al Qahtani et al. 2020).

1.3.3. SMART Pipes

Instead of using mobile tools equipped with acoustic and electromagnetic sensors, sensors can also be embedded inside of pipe walls during casting of pipes. These so-called SMART pipes allow for proactive leakage control, since they allow continuous monitoring of the material condition and early detection of crack formation in PVC. Since SMART pipes boast embedded sensors, this technique is most promising for new pipes. Large-scale deployment remains costly, but ongoing research reveals the potential of this technology, as smart pipes provide useful information regarding asset condition and decrease the need for on-site inspections (Liu and Kleiner 2013; Tran et al. 2021).

1.4. Model-Based Techniques

Equipment-based inspection technologies are high accuracy methods for condition assessment and leakage detection in underground pipes (Puust et al. 2010). However, due to the complexity and novelty of these ultrasonic and electromagnetic technologies, these methods are still cost, labor and expertise intensive, and therefore not yet suitable for regular deployment on large sections of the water distribution network. Due to these concerns, there is a need for machine learning algorithms or hydraulic models capable of timely leak detection, allowing fast response to burst events and thus minimal water loss (Scozzari et al. 2021).

1.4.1. Pipe Lifetime Models

Water companies therefore rely on computer algorithms, such as pipe lifetime models, in order to predict pipe condition and chance of failure (Malm et al. 2012; Barton et al. 2019; Snider and McBean 2020). Hence, pipe lifetime models are an indispensable tool for realizing proactive leakage control. An estimate of remaining asset lifetime can be made based on pipe properties, pressure regime, and surroundings. Those assets that are predicted to be at the end of their lifecycle can be inspected in order to validate the model and determine actual condition of the conduits. In practice, assets subjected to equipment-based inspection are always pre-selected using model-based techniques. The inspection results can also be used to validate and update the lifetime expectancy models.

1.4.2. Machine Learning

Any model-based approach used in optimizing water distribution, has the goal to provide additional information and insight into the ongoing processes. The resulting information is highly valued by water companies, although the required data is as of yet not always correctly collected, stored or categorized. The information contained in the currently collected raw big data sets can be compared with an iceberg, where we only see the tip visible above water, while not yet aware of the vast mass of information still hiding from us underwater. Consequently, the water industry has been referred to as "data-rich, but information poor" (Scozzari et al. 2021). Rapid advances in machine learning provide ample tools to expose more of the iceberg, although attention should be paid to the underlying assumptions and biases of these techniques. For a brief overview of big data and machine learning techniques relevant in the context of this thesis are detailed in Appendix I – Big Data and Machine Learning.

1.4.3. Water Balance Models

Current methods of reactive leakage control still rely heavily on customers as substitutes for sensors. Customer response to water quantity and quality issues are used to detect and localize bursts. Although these responses are an important factor in leakage control, it does not provide additional insight in the causes of bursts and is therefore not sufficient in order to facilitate reactive leakage control.

Most drinking water networks are complex cyber-physical systems, where the prevalence of water quantity and quality sensors are used to monitor the network in real-time. In order to facilitate reactive leakage control in the form of timely detection and localization of conduit bursts, water companies increasingly rely on algorithms based on real-time data from flow and pressure sensors. Traditionally, water distribution networks are divided into District Metering Areas (DMA's), hydraulically isolated sections of a company's network, where each inflow into the DMA is metered (Figure 1.3). Since all inflows and outflows of each DMA are precisely measured, the water demand of each DMA can be accurately monitored (Hutton and Kapelan 2015a; Bui et al. 2020). Using real-time algorithms, the measured water demand per DMA can then be compared with a prediction of the expected water demand in order to identify disturbances in the water distribution process. Since medium to large bursts result in an abnormally high water "demand" compared to predicted demand, these differences between measured and predicted demand can serve as burst detection tool. Real-time burst alerts enable timely localization and solving of the detected leakage. Water demand forecasting thus represents a form of reactive leakage control.

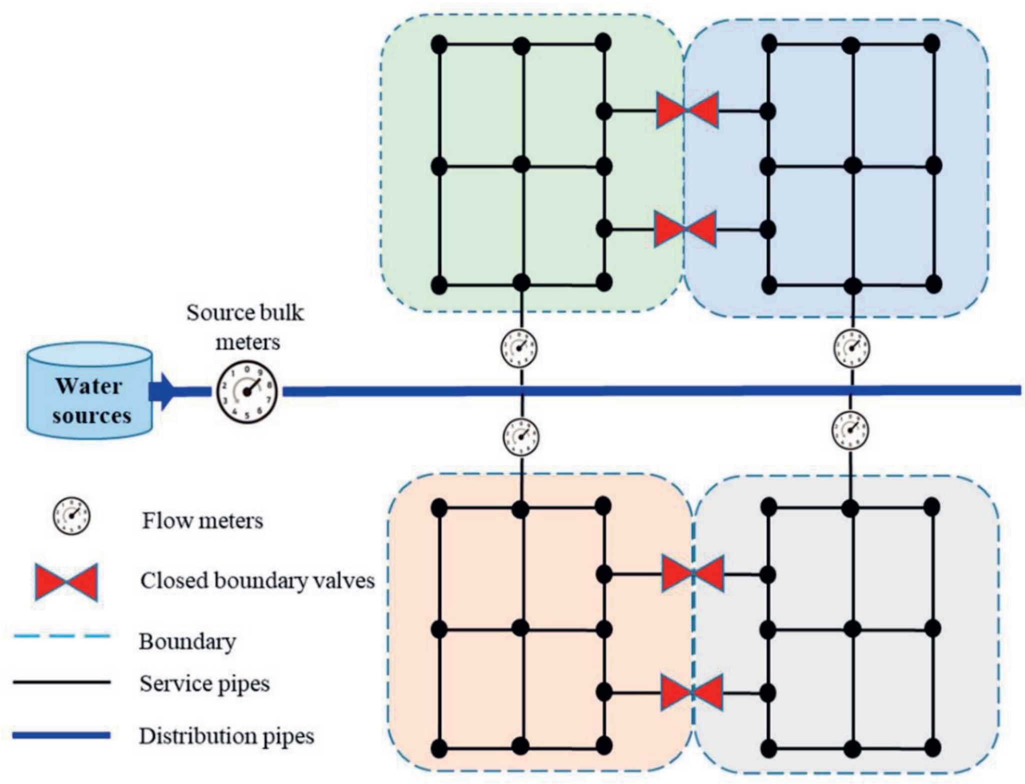


Figure 1.3 - Schematic Diagram of District Metering Areas (DMA's) and their boundaries (Bui et al. 2020).

Most burst detection methods rely on classification and/or statistics using several weeks of historical water balance measurements in order to capture the seasonality of the water demand per DMA (Wu and Liu 2017). Besides traditional regression methods for water demand forecasting (Billings and Jones 2008; Adamowski et al. 2012; Hutton and Kapelan 2015a; Papageorgiou et al. 2015; Froelich 2016; Candelieri 2017), recent methods often make use of burst event databases and employ neural networks or other supervised machine learning methods, optionally combined with a regression approach in order to facilitate burst detection (Babel and Shinde 2011; Bai et al. 2014; Xu et al. 2018; Pacchin et al. 2019).

Burst detection algorithms also benefit from inclusion of forecasts of exogenous factors that influence water demand, such as weather forecasts or a holiday schedule. During hot summer days the water demand may be higher, due to filling of swimming pools, car washing, or garden sprinkling. During holidays, people tend to wake up later and thus also shower later in the morning. If accurate forecasts of these exogenous processes are available, these can be used to improve demand forecasts (Wu et al. 2018; Zubaidi et al. 2018).

1.4.4. Hydraulic Models

Due to DMA-based network monitoring, the exact flow and pressure at the boundaries of each DMA and at each water production location are accurately known and closely monitored in real-time. Although burst detections on DMA level are accurate, this does not help to determine the exact location of a burst within the DMA in question. The flows and pressures in the network assets within a DMA are far less known compared to the DMA boundaries. In order to close this gap, hydraulic models of each DMA can be used to elucidate these flows and pressures inside of the network. Based on network topology and asset properties, such as diameter, length, slope, and roughness, an accurate virtual copy of the water distribution network can be built. Together with water demand estimates per consumption point and the metered flows leaving water production locations, this model can be used to simulate the flows through each pipe, the pressure in each junction, and the water quality throughout the network, in real-time, creating a so-called “Digital Twin” of the water distribution network (Conejos Fuertes et al. 2020; Scozzari et al. 2021). Besides real-time virtual monitoring of the entire network, hydraulic models allow simulation of hydraulic scenarios, such as periods of high water consumption or large bursts. This, in turn, also helps water companies to prepare for these extreme scenarios and to optimize network operation and design.

Where machine learning algorithms are essential for timely detection of bursts, hydraulic models may help to localize these bursts. For example, a hydraulic model can be used to simulate virtual bursts in various network locations and then investigate what the predicted flows and pressures are at those points in the network where a flow or pressure sensor is placed in the real network (Nagar and Powell 2004; Sarrate et al. 2014; Díaz et al. 2016; Boatwright et al. 2018; Marchi et al. 2018). This way a database of burst signatures may be built up, that can be compared with the signature of sensor measurements at the moment of a real burst (Farley et al. 2012; Sophocleous et al. 2019).

Machine learning algorithms and hydraulic models are two sides of the same coin: different techniques that complement each other and together provide full network insight, minimize water loss and optimize network durability. Additional sensor placement will improve burst detectability, and accuracy of hydraulic models of the network. Accurate models can, in turn, be used to determine optimal placement of extra sensors if required. These new sensors then will improve burst detectability and hydraulic model accuracy. This circular process will thus help elucidate the exact processes happening in the water network and improve network management.

1.5. Scientific Gap

1.5.1. Non-Burst Anomalies

Current time series analysis methods are mainly used for burst detection and aim to reduce the false positive burst alarms and to minimize detection of non-burst anomalies (Babel and Shinde 2011; Bai et al. 2014; Xu et al. 2018; Pacchin et al. 2019). However, burst events are not the only abnormalities in the time series measurements collected by sensors throughout the water distribution systems. Although cause and effect of these non-burst abnormalities is not always clear, these events do have an impact on the system (Martínez-Codina et al. 2015). Unexpected and fast changes in flow and pressure can cause additional stress on the system or indicate that valve or pump settings are not functioning as intended. If the underlying causes of these abnormalities are not identified, the cumulative stress of these events can have potentially devastating consequences. There is a need for more detailed analyses of these recurring anomalies, as they show potential in enabling sensor-based proactive leakage control (Millán-Roures et al. 2018).

1.5.2. Burst Detection

Dividing the water distribution networks into DMA’s provides more insight in customer water demand and allows for water demand forecasting based on the water mass balance per DMA (Laucelli et al. 2017). These DMA-wide short term water demand forecasts allow for a high accuracy of burst detection. However, false positive bursts alarms remain a problem. If too many false alarms are issued, operators will no longer trust the software or respond with the same intensity (Scozzari et al. 2021). Accurate incorporation of water demand seasonality based on prior measurements may help to reduce false positives (Benítez et al. 2019). However, exogenous factors, such as weather, holidays, or festivities, are often also responsible for deviating water demand. In order to minimize false burst alarms as a consequence of these exogenous influences, there is a need for novel and robust burst detection methods.

1.5.3. Sensor Placement

Measurements from additional sensors increase the accuracy of various leakage control algorithms, such as burst detection methods (Giustolisi et al. 2008; Casillas et al. 2013; Hagos et al. 2016; Steffelbauer and Fuchs-Hanusch 2016; Cugueró-Escofet et al. 2017; Boatwright et al. 2018; Qi et al. 2018; Quiñones-Grueiro et al. 2018; Xu et al. 2020). However, the position of a sensor within a network is critical in order to maximize the additional information provided by the sensor. Although sensor placement and operation costs have decreased in recent years, there remains a tradeoff between information gain and sensor costs (Scozzari et al. 2021). Optimal placement of network sensors is therefore an important process to consider.

Traditional sensor placement methodologies focus on burst detectability as the criteria for optimal placement of additional pressure sensors and utilize a sensitivity-based approach to determine those placements (Steffelbauer and Fuchs-Hanusch 2016; Cugueró-Escofet et al. 2017; Boatwright et al. 2018; Qi et al. 2018; Quiñones-Grueiro et al. 2018). Junctions most sensitive to burst-induced pressure changes are most suitable for a sensor. As of yet, not all water utility companies make use of pressure sensors in burst detection, but instead rely on DMA-wide water demand forecasting to detect bursts. Most current methodologies do not incorporate pressure dynamics into the calculations for optimal sensor placement (Giustolisi et al. 2008; Boatwright et al. 2018). However, pressure sensors with a measurement frequency of at least once per few seconds are capable of detecting pressure transients. These pressure transients, or water hammers, are often caused by pipe bursting and are an important indicator of bursts. It has been shown that detection of these transients plays a crucial role in burst detection (Srirangarajan et al. 2012; Lee et al. 2016). Therefore, there is a need for optimal sensor placement methodologies capable of including very fast pressure dynamics.

1.6. Research Aim

Nowadays, the value of data has become more and more apparent to water distribution companies. How to expose the information contained within this supply of big data remains a challenge. Both machine learning algorithms and hydraulic models offer innovative opportunities to extract concrete information from the real-time data.

One of these crucial data sources are the plethora of flow and pressure sensors deployed in the network. The real-time data recorded by these sensors is known to be valuable for burst detection, although suppression of false burst alarms needs permanent attention. Abnormalities in the time series data of water demand could indicate a burst. However, not all these abnormalities correspond to bursts, where the cause and effect of many is as of yet not completely understood. This begs the questions:

- How can recurring anomalies in sensor time series help improve proactive leakage control?
- How can water bursts be detected with high precision?

Knowing that there is a wealth of information contained within the time series data gathered in practice, the added value of real-time flow and pressure sensors for water distribution management is undeniable. Placement of additional sensors will strengthen decision support tools and early warning systems, such as burst detection and localization algorithms. However, the location of sensors placement greatly influences usefulness of measured data. Since sensor placement remains costly, it is of vital importance to optimize the placement of new sensors. Key challenges in this regard are:

- What is an appropriate framework for optimal sensor placement, such that the procedure is robust with respect to accuracy of burst detection algorithms and hydraulic models?
- How can placement of multiple sensors be made suitable for large-scale water distribution networks?

1.7. Thesis Outline

After this introduction in Chapter 1, Chapters 2 and 3 will describe two machine learning algorithms that can be used to obtain more information about leakage control from time series data of existing flow and pressure sensors.

- Chapter 2 introduces a method to detect recurring non-burst abnormalities present in the time series data from flow or pressure sensors or DMA water balances. By timely detection of recurring abnormal patterns, early warnings can be issued to ensure timely identification and mitigation of the causes of these recurring anomalies, thus facilitating proactive leakage control.

- Chapter 3 presents a novel and robust burst detection method. The water demand nowcasting method can be used to predict in real-time the DMA water demand or expected flow at a flow sensor, using data from other flow and pressure sensors within the same DMA to reduce false positive burst alarms.

Chapters 4 and 5 describe a model-based approach from systems theory to investigate where best to place additional flow or pressure sensors in water distribution networks.

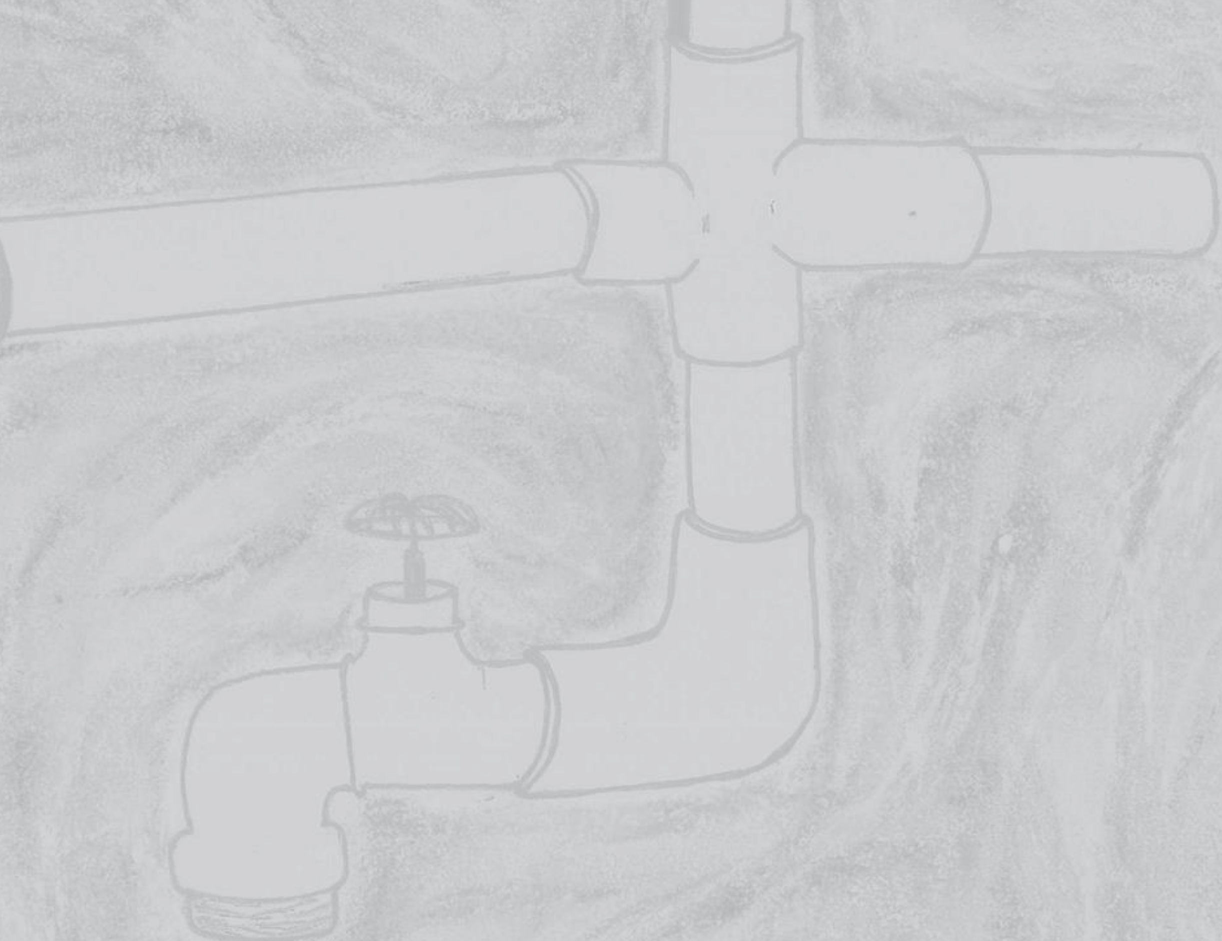
- Chapter 4 deals with a state-space hydraulic model of pressurized fluid transport. This model is used for optimal placement of additional flow and pressure sensors, where optimal is defined as sensor placement that maximizes network observability, a metric describing how well the flow and pressure of each location in the network can be calculated based on data from chosen sensor placement alone.

- Chapter 5 expands on this methodology for optimal sensor placement, by investigating placement of multiple sensors sequentially and simultaneously in real-scale networks. In order to adapt the sensor placement methodology for larger networks, a network skeletonization method as well as a more robust optimality criteria will be introduced.

Finally, Chapter 6 presents a general discussion, as well as an outlook for further research concerning machine learning and hydraulic modeling methods to be used to obtain more insight in water distribution networks.

Chapter 2

Monitoring Support for Water Distribution Systems Based on Pressure Sensor Data



A version of this chapter has been published as:
Geelen, C.V.C., Yntema, D.R., Molenaar, J., Keesman, K.J., 2019. Monitoring Support for Water Distribution Systems based on Pressure Sensor Data. Water Resource Management 33, 3339–3353. https://doi.org/10.1007/s11269-019-02245-4

2. Monitoring Support for Water Distribution Systems Based on Pressure Sensor Data

Abstract

The increasing age and deterioration of drinking water mains is causing an increasing frequency of pipe bursts. Not only are pipe repairs costly, bursts can also lead to contamination of the Dutch non-chlorinated drinking water, as well as damage to other above- and underground infrastructure. Detection and localization of pipe bursts have long been priorities for water distribution companies. Here we present a method for proactive leakage control, referred to as Monitoring Support. Contrary to most leak prevention methods, our method is based on real-time pressure sensor measurements and focuses on detection of recurring pressure anomalies, which are assumed to be indicative of misuse or malfunctioning of the water distribution network. The method visualizes and warns for both recurring and one-time anomalous events and offers monitoring experts an unsupervised decision support tool that requires no training data or manual labeling. Additionally, our method supports any time series data source and can be applied to other types of distribution networks, such as those for gas, electricity and oil. The performance of our method, including both instance-based and feature based clustering, was validated on two pressure sensor data sets. Results indicate that feature based clustering is the best method for detection of recurring pressure anomalies, with accuracy F1-scores of 92% and 94% for a 2013 and 2017 data set, respectively.

2.1. Introduction

The Netherlands has an excellent drinking water distribution system (WDS), with water losses of only 6%, compared to 25% and 16% for the US and UK, respectively (Rosario-Ortiz et al. 2016). The relatively good state of the Dutch drinking water infrastructure is in part caused by the replacement of at least half of the distribution network since 1970, resulting in an average pipe age of 33 to 37 years, compared to an estimated 75 to 80 years in the UK. Although the pipes are relatively new, the actual state of the water mains is largely unknown. Pipe bursts regularly occur, causing damage to other above- and underground infrastructure as well as requiring costly repairs. The Dutch drinking water is not chlorinated, which means that contamination as a consequence of bursts will not be neutralized by chlorine, therefore introducing more risk to consumers. In order to ensure proper functioning, water companies need to assess the probability of failure and apply leakage control.

Currently, the probability of pipe failure is estimated based on pipe properties, historical (failure) data and external conditions, with emphasis on reactive leakage control in the form of leak detection and localization (Mounce et al. 2003; Puust et al. 2010; Bakker et al. 2014; Gelazanskas and Gamage 2014; Okeya et al. 2015; Wu et al. 2016). However, to deal with the unknown state and continuous degradation of pipes, a proactive strategy, with a focused on leak prevention, is required. The objective of this study is therefore to present and evaluate a method for proactive leakage control.

Although various leak detection methods have been developed and tested, leak prevention methods have only recently been published (Wang et al. 2012; Xu et al. 2013; Kabir et al. 2015; Leu and Bui 2016; Kakoudakis et al. 2017). Although powerful, these methods frequently rely on supervised machine learning, requiring extensive data on pipe properties and external conditions. However, these methods often do not incorporate available real-time pressure and flow sensor data. Moreover, internal pipe conditions and grid management can also play a role in asset failure. In addition to extensive data sets, for the training of supervised models, classification labels are also required. Lastly, since these methods mostly use historical data, real-time implementation was not considered.

Our method focuses on proactive leakage control and offers an early warning and decision support system for proactive management of the WDS, which helps to prevent future bursts and malfunctioning. Contrary to the previously mentioned leak prevention studies, our method is based on real-time sensor data only, detecting recurring pressure anomalies which are indicative of misuse or malfunctioning within the WDS. Additionally, our method provides monitoring experts with an unsupervised decision support tool that requires no training data or manual labeling. Unsupervised learning is particularly suited for recurring pattern detection due to its robustness regarding detection of novel recurring patterns (Kotsiantis and Pintelas 2004). Clustering of anomalies allows detection of clusters containing a common recurring pattern. In this paper, both instance-based and feature-based clustering is applied to two pressure data sets from the Dutch drinking water company Vitens. Lastly, our method supports any time series data and can be applied to other distribution networks, such as those for gas, electricity or oil.

2.2. Materials & Methods

The detection of anomalous and recurring pressure patterns is divided into three steps: detection of anomalous events (Figure 2.1a), clustering of events (Figure 2.1b) and visualization of recurrence history (Figure 2.1c).

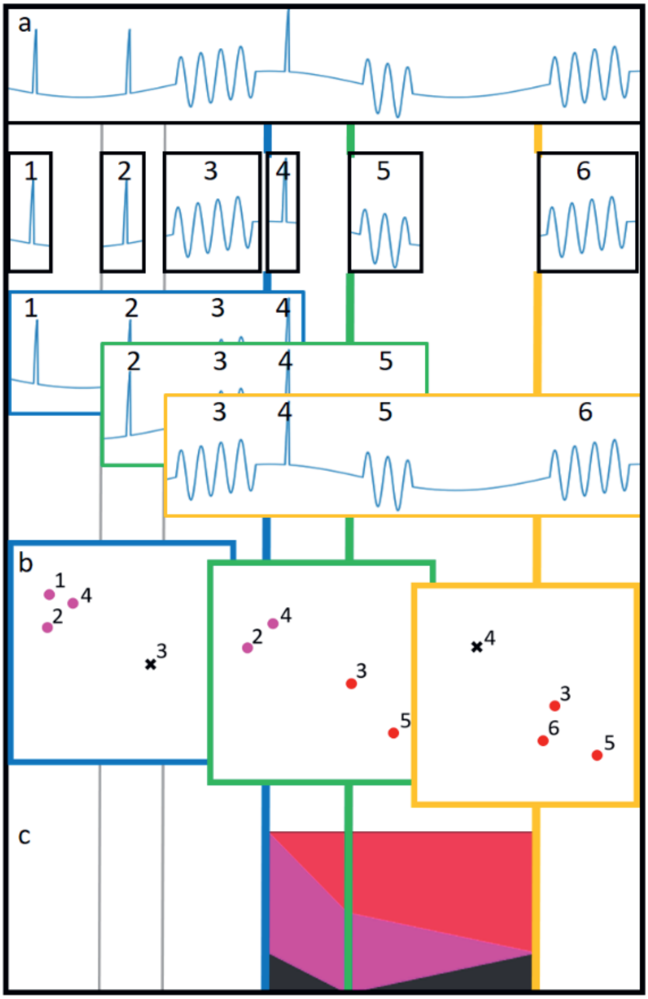


Figure 2.1 - Flowchart of Monitoring Support. a) Measured time series subjected to anomaly detection (six anomalous events). Whenever a new anomalous event is detected, windows of a preset number of preceding events are created (solid, dashed and dotted windows with four events per window in this example). These windows are then subjected to b) moving window clustering, resulting in two clusters (star and circle) and outliers (black cross). c) Clustering results are then summarized in fingerprint graphs, stacked area plots with time of event detection on the horizontal axis and frequency of pattern occurrence per cluster on the vertical axis.

2.2.1. Data Sets

Access to actual and historical pressure sensor data was provided by Vitens, a Dutch drinking water company. A known case of recurring anomalous pressure patterns followed by a pipe burst was investigated from 1/6/2012 to 1/6/2013, hereafter referred to as the 2013 data set. In addition, a recent data set from another pressure sensor is used, with measurements from 18/5/2017 to 17/11/2017, hereafter referred to as the 2017 data set. Both pressure sensors were situated close to water reservoirs. As a preprocessing step, erratic measurements were removed. Resampling and linear interpolation in time were used to obtain a constant sampling interval of one second.

2.2.2. Event Detection

Anomalous events were detected using a moving window range statistic, defined as the difference between maximum and minimum values of every ten-seconds moving window, divided by the window size of ten seconds. A ten-seconds window range statistic was used instead of the derivative, so as to avoid problems associated with noise present in the pressure measurements. Measurements with a range statistic of more than two kPa/s were flagged as anomalous (Figure 2.1a), since rapid pressure changes of this magnitude are most often caused by events that are relevant for the purposes of this study. Although quite simple, the range statistic and absolute range threshold were found to be able to detect all relevant anomalous events. Since anomaly detection is an important and complicated process, a more extensive definition of anomaly detection will most likely improve performance (Branisavljević et al. 2011; Mounce et al. 2014; Scozzari and Brozzo 2017). However, for illustration of our method on the aforementioned data sets, the current metric is sufficient and suitable. The anomalies were combined into events, where anomalous measurements within a 15 min duration were considered to be part of one event (Figure 2.1a). Next, each event was extended with two minutes of preceding and two minutes of succeeding measurements to ensure the entire anomaly and context were captured as a single event.

2.2.3. Event Clustering

Recurrence of anomalous pressure patterns was defined as the repetition of similar anomalous events. Events were clustered in order to detect which events are similar and probably have the same cause. Clustering is an unsupervised method for grouping of similar events based on the distances between events. For this, events were represented by vectors, after which the distance between these vectors can be calculated. Events with a low distance between them are deemed similar and were included in the same cluster. Each cluster corresponds to a specific recurring and anomalous pattern (Figure 2.1b). The vectors assigned to each event were based either on event measurements (instance-based) or on each event's characteristic features (feature-based) (Fulcher and Jones 2014). In this study, clustering was performed using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (McInnes et al. 2017), which clusters events based on their density within a vector space. Unlike similar clustering methods, such as DBSCAN (Ester et al. 1996) and Mean Shift (Ray and Benammar 2002), HDBSCAN uses a hierarchical minimum density threshold and is better in detecting varying cluster shapes. HDBSCAN also allows clustering with a precomputed distance matrix and has the capacity to distinguish core samples from outliers. Since the presented method is intended for real-time application, clustering needs to be performed anew when novel events are detected. Clustering was performed over the most recent 150 events using a moving window of these events whenever a new anomalous event was detected (Figure 2.1b). This moving window approach ensures real-time applicability and detection of distinct clusters for the different recurring patterns present in the investigated data. The window size can be adjusted if requested. However, larger window sizes potentially result in merging of clusters due to a higher overall density of events, making the distinguishing of local denser areas more difficult. Smaller window sizes can result in failure to detect recurring patterns with a low frequency of occurrence.

2.2.4. Distances for Instance-Based Clustering

In order to calculate the distance between two event vectors of different lengths, the vectors are clipped to equal lengths. Clipping was done based on the maximum cross-correlation between both events (Figure 2.2b). For every pair of event time series, the lag related to the maximum cross-correlation was removed (Figure 2.2a), followed by clipping of the non-overlapping tails of both events (Figure 2.2a) to obtain events of equal length (Figure 2.2c). Optionally, Dynamic Time Warping (DTW) can then be applied in order to correct for temporal drift, which increases the accuracy of the succeeding distance calculations (Figure 2.2d) (Aghabozorgi et al. 2015). In this study, DTW was limited to warping of up to 5% of the total event duration in both directions. After clipping and DTW, the Euclidean distance between events was calculated and corrected by dividing by the length of the events before being subjected to clustering.

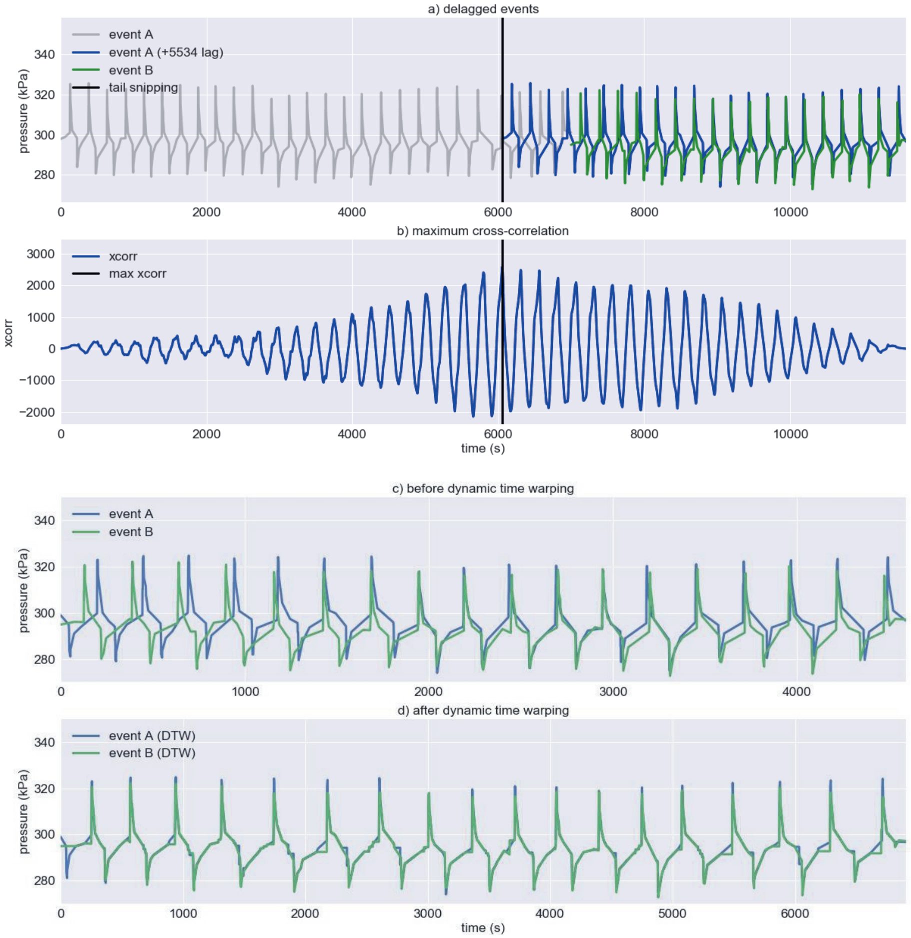


Figure 2.2 – from top to bottom: a) Time series clipping based on maximum cross correlation of standardized events. b) Cross correlation between both events. c) Events A and B after clipping and before DTW d) Events A and B after DTW

2.2.5. Distances for Feature-Based Clustering

For each event, 43 features were calculated (Appendix 2A). In each clustering window of 150 events, the features of these events were scaled by median subtraction followed by interquartile range division, ensuring that scaling was robust for outliers. The features were chosen so as to be robust for distinguishing between a limited number of recurring patterns. After scaling, the distances between each event pair's feature vectors were calculated and the resulting distance matrix was subjected to the clustering method.

2.2.6. Fingerprint Graphs

Fingerprint graphs (Figure 2.3) present an effective overview of the periods of recurrence for different type of patterns and their respective frequency of occurrence. When a new anomalous event is detected, the clustering results of the corresponding 150-event window is added to the fingerprint graph as a vertical white slice. Each colored area depicts a recurring pattern, where each pattern's height depicts its frequency of occurrence within the 150-event window and its length corresponds to the duration of the pattern recurrence (Figure 2.1c).

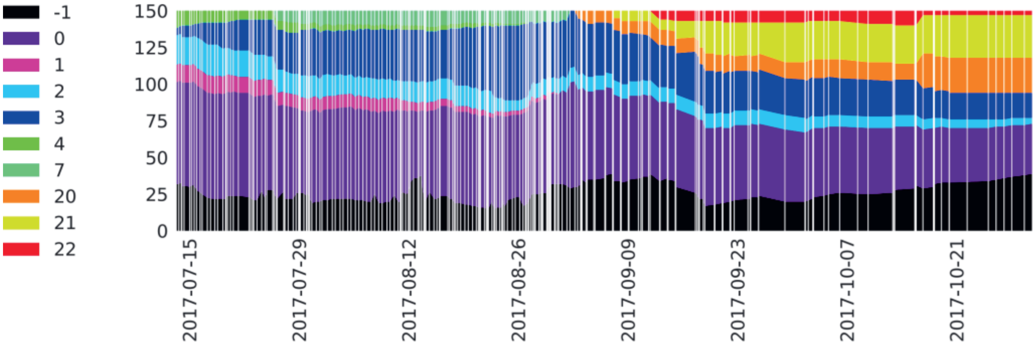


Figure 2.3 - Example fingerprint graph, showing the recurring anomalous pressure patterns as different areas in the stacked area graph. Each legend entry number matches a separate cluster. The bottom, black area with legend entry −1 represents outlier events, which are deemed non-recurring

2.2.7. Validation Report

The validation report depicts the precision (fraction true positives among detected positives), recall (fraction of true positives among actual positives) and F$_1$-score ($F_1 = 2 * precision * recall * (precision + recall)^{-1}$) for each true recurring pattern present in the manually labeled validation data (van Rijsbergen 1979). In order to calculate these scores, cluster ID numbers were mapped to the validation labels. Clusters mapping to the same pattern were deemed a single cluster for the sake of accuracy scores calculation only.

2.3. Results

The method was applied to pressure data of the WDS of Vitens. In order to validate the method, a known case of pressure pattern recursion leading to a pipe burst was investigated, as well as a more recent data set from 2017. The data set from 2013 contains a rapid crack propagation event at 2013/03/12 18:03 (Figure 2.4). The pipe in question was already under strain due to angular displacement and sub-zero temperatures. However, afterwards it was concluded that the burst probably occurred due to pressure oscillations caused by the interaction of two upstream pumps connected in parallel. Repeated activation and deactivation of these pumps led to these recurring oscillations, which had been occurring for over two months before the coincidence with sub-zero temperatures and additional pipe stress caused by traffic led to a burst.

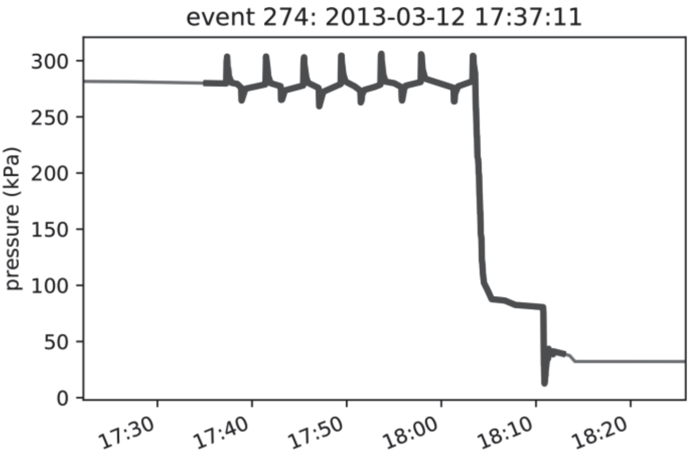


Figure 2.4 – Pressure sensor data from the 2013 data set containing the pipe burst at 12-03-2013 18:03

To prevent future malfunctioning and to obtain more insight into the behavior inside the pipes, we developed a method functioning as a real-time decision support and early warning system for recurring unwanted pressure patterns. By timely detection of recurring anomalous pressure patterns, the 2013 pump malfunction could have been identified earlier and thus been prevented alltogether. As a proof of concept, our method has been applied to the 2013 (Figure 2.5) and 2017 data sets (Appendix 2B) using instance-based clustering with and without DTW and feature-based clustering. In order to assess the real-time performance of the method, it was applied to the 2013 and 2017 data sets with moving windows, as a stand-in for real-time application.

When a novel anomalous event was detected in the pressure time series data of a sensor, the most recent 150 events time window was again clustered. Events that belong to the same cluster were assumed to be part of the same recurring anomalous pressure pattern. Based on the manually labeled validation data (Figure 2.5d), there are five main types of recurring patterns present in the 2013 data set (labeled as fast oscillation, oscillation, slope, spike and valley) (Figure 2.6).

As mentioned before, the 2013 burst (Figure 2.4) probably happened because of recurring pressure oscillations (Figure 2.6 Oscillation), which in turn were caused by erroneous behavior of two pumps upstream of the sensor. Without having this prior knowledge, our method detects these oscillations and so would have provided an early warning of the problem months in advance of the eventual burst.

Besides the oscillations, four other recurring patterns are detected. The fast oscillation events most probably occurred as a consequence of rapid pump activation and deactivation. The slope pattern (Figure 2.6 Slope) consists of rapid pressure increases due to increased pumping activity. The slope events occur mostly in the early morning, where rapid pump activations cause the pressure to rise to a higher pressure than is necessary, before gradually decreasing again. The spike pattern (Figure 2.6 Spike) consists of pressure transients, caused by rapid pump, valve or water consumption changes. Pressure transients may cause (gradual) degradation and deformation of pipes, connections or valves (National Research Council 2006). Lastly, the valley patterns (Figure 2.6 Valley) consists of short pressure drops where for a short period of time water diversion or increased water consumption causes temporary but considerable pressure drops.

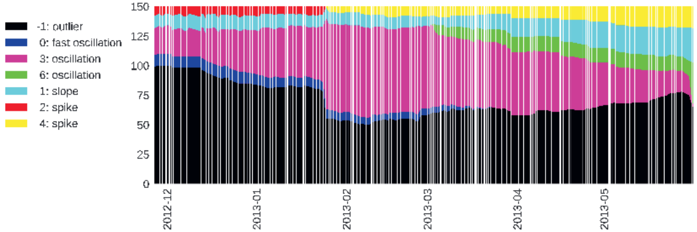
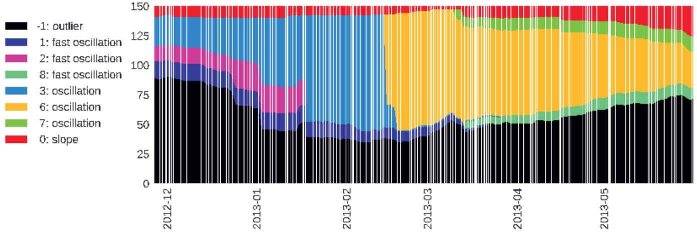
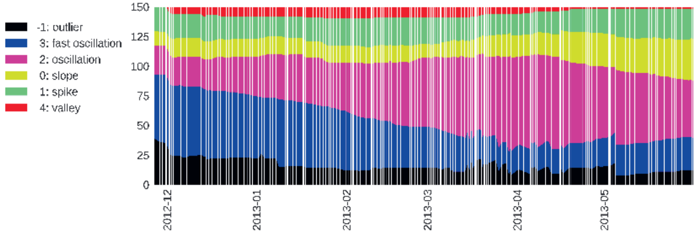
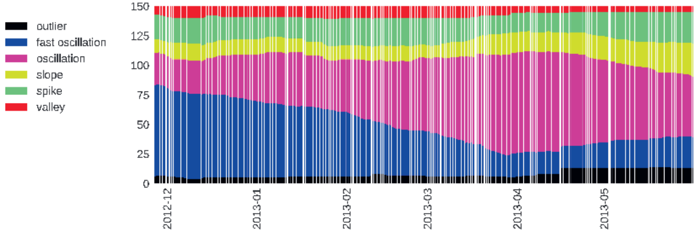
<table>
<tr><th>Method</th><th>Fingerprint Graphs</th><th>Validation Report</th></tr>
<tr><td>(A) Instance-Based Clustering</td><td>
</td><td>
<table>
<tr><th></th><th>Precision</th><th>Recall</th><th>F₁-Score</th><th>Support</th></tr>
<tr><td>Fast Oscillation</td><td>1.00</td><td>0.10</td><td>0.18</td><td>108</td></tr>
<tr><td>Oscillation</td><td>0.75</td><td>0.73</td><td>0.74</td><td>113</td></tr>
<tr><td>Slope</td><td>0.71</td><td>0.53</td><td>0.61</td><td>45</td></tr>
<tr><td>Spike</td><td>0.89</td><td>0.47</td><td>0.62</td><td>51</td></tr>
<tr><td>Valley</td><td>0.00</td><td>0.00</td><td>0.00</td><td>17</td></tr>
<tr><td>Average / Total</td><td>0.81</td><td>0.42</td><td>0.49</td><td>334</td></tr>
</table></td></tr>
<tr><td>(B) Instance-Based Clustering with DTW</td><td>
</td><td>
<table>
<tr><th></th><th>Precision</th><th>Recall</th><th>F₁-Score</th><th>Support</th></tr>
<tr><td>Fast Oscillation</td><td>1.00</td><td>0.38</td><td>0.55</td><td>108</td></tr>
<tr><td>Oscillation</td><td>0.71</td><td>0.83</td><td>0.76</td><td>113</td></tr>
<tr><td>Slope</td><td>0.75</td><td>0.53</td><td>0.62</td><td>45</td></tr>
<tr><td>Spike</td><td>0.00</td><td>0.00</td><td>0.00</td><td>51</td></tr>
<tr><td>Valley</td><td>0.00</td><td>0.00</td><td>0.00</td><td>17</td></tr>
<tr><td>Average / Total</td><td>0.66</td><td>0.48</td><td>0.52</td><td>334</td></tr>
</table></td></tr>
<tr><td>(C) Feature-Based Clustering</td><td>
</td><td>
<table>
<tr><th></th><th>Precision</th><th>Recall</th><th>F₁-Score</th><th>Support</th></tr>
<tr><td>Fast Oscillation</td><td>0.99</td><td>0.90</td><td>0.94</td><td>108</td></tr>
<tr><td>Oscillation</td><td>1.00</td><td>0.91</td><td>0.95</td><td>113</td></tr>
<tr><td>Slope</td><td>0.95</td><td>0.89</td><td>0.92</td><td>45</td></tr>
<tr><td>Spike</td><td>0.94</td><td>0.92</td><td>0.93</td><td>51</td></tr>
<tr><td>Valley</td><td>1.00</td><td>0.53</td><td>0.69</td><td>17</td></tr>
<tr><td>Average / Total</td><td>0.98</td><td>0.89</td><td>0.93</td><td>334</td></tr>
</table></td></tr>
<tr><td>(D) Manual Labeling, Validation Reference</td><td>
</td><td></td></tr>
</table>

Figure 2.5 - Results of the 2013 data set. The legend lists the cluster ID number and which pattern matches best with that clusters ID, as derived from the manually labeled validation data. Clusters numbers mapping to the same recurring pattern, were deemed a single cluster for the sake of accuracy scores calculation only

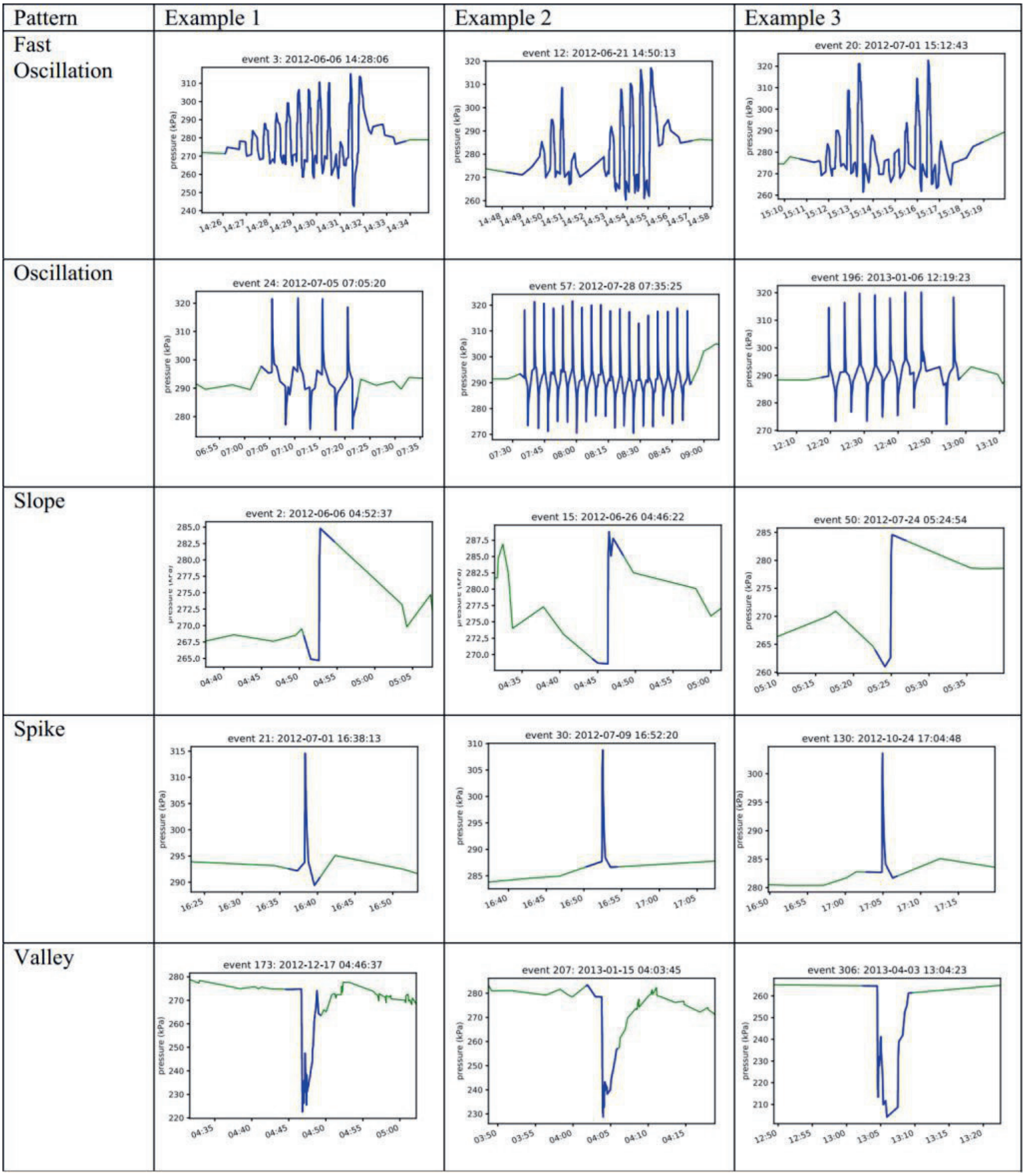


Figure 2.6 - Examples of the recurring patterns present in the 2013 data set

2.4. Discussion and conclusions

The fast oscillations events show a large variation between them (Figure 2.6). Consequently, both instance-based methods and the feature-based method show a lower recall for fast oscillations compared to other patterns (Figure 2.5abc, validation reports). Besides a lower recall, our method often detects multiple clusters matching the fast oscillation recurring pattern, due to the large variation between various fast oscillation events. (Figure 2.5b fingerprint graph clusters 1,2 and 8 all correspond to the fast oscillation pattern. The same is true for Figure 2.5c clusters 3 and 6.)

2.4.1. Method Comparison

Most spike events closely resemble half a period of an oscillation event, resulting in a small instance-based clustering distance between these events, especially after event clipping. This phenomenon is reflected in the low accuracy of spike detection for instance-based clustering without and with DTW (Figure 2.5ab: F_1-scores of 0.62 and 0.00, respectively), as opposed to the high accuracy using feature-based clustering (Figure 2.5c: F_1-score of 0.930). To some extent, the same occurs for slope events resembling parts of valley events (Figure 2.5abc: slope F_1-scores of 0.61 and 0.62 for instance-based with and without DTW, respectively, versus 0.92 for feature-based clustering). Because of this low distance between parts of both patterns, instance-based clustering is less suitable for distinguishing oscillation and spike events compared to feature-based clustering, which does not rely on the distances between events as calculated for instance-based clustering.

Like fast oscillations, there is a large variation between the valley events. Additionally, only 17 out of the 334 events in the 2013 data set represent valleys. As a result, instance-based clustering is unable to detect the valley recurring pattern (Figure 2.5ab: F_1-scores 0.00 and 0.00 for with DTW and without DTW instance-based clustering) and feature-based clustering shows a low recall of 0.69 for valley detection (Figure 2.5c).

Since an unsupervised approach was taken in this study, novel patterns that did not occur in the past could still be detected successfully, such as the oscillation pattern seen in the 2017 data set (Figure II-1, Appendix 2B). Not only do new patterns occur as time progresses, the types of patterns detected also differ widely between sensors, as can be seen when comparing the 2013 and 2017 data set results. Consequently, an unsupervised method is considered the most suitable approach for detecting pattern recurrence in sensor data.

Feature-based clustering requires a suitable selection of features capable of distinguishing recurring patterns. The unsupervised approach means it is not possible to automatically choose a set of features most suited for grouping pressure anomalies or to weigh features based on suitability. Therefore, additional care is required for initial feature selection. However, even though the 2013 and 2017 data sets differ widely in recurring patterns present, the currently selected features show high accuracies detecting and distinguishing between recurring patterns (Figure 2.5, Appendix 2B). Feature-based clustering also outperforms instance-based clustering, as can be seen from the F1-scores of 0.93 and 0.94 for feature-based clustering of the 2013 and 2017 data sets, compared to 0.49/0.82 and

0.52/0.80 for the no DTW/DTW instance-based clustering of 2013 and 2017 data sets, respectively (Figure 2.5, Appendix 2B). This indicates that the currently chosen set of features are robust for clustering 150 event windows (Appendix 2A).

2.4.2. Method Performance

Our method fills the gap for real-time sensor-based and proactive leakage control methods. Besides recurrence detection, the method offers an easy framework for monitoring pressure measurements. Our method finds all anomalous pressure events and detects which contain a recurring pattern. The method can isolate, visualize and summarize both recurring and onetime events and so helps to determine the cause and potential consequences of the aberrant pressure events. Combined with an unsupervised approach, our method represents a powerful tool that alleviates the grid monitoring workload of monitoring experts.

Overall, our method shows promising results regarding recurrence detection and visualization. Although only the performance with time series data from pressure sensors was investigated, flow data or data from other distribution systems can also be used. By choosing a suitable anomaly detection method, our method can be applied to any time series data where recurrence of unwanted or artificial patterns are likely to occur.

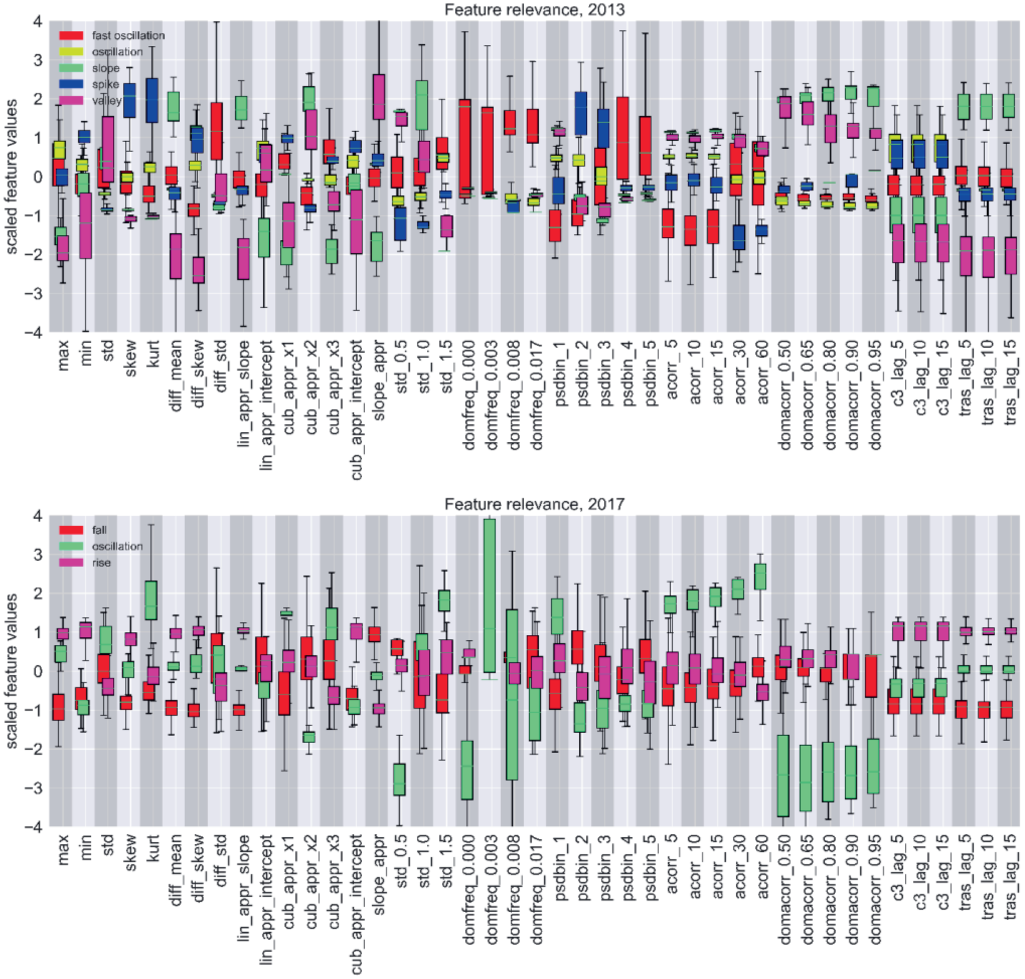
Our application to real data shows that feature-based clustering is the preferred method for detecting recurring pressure anomalies. This implies that selection of these features is a crucial ingredient of this approach. Implementation of our method and/or testing more data sets will allow reevaluation of chosen features over time, if required. However, since an average accuracy F1-score of 93.5% was achieved with the current feature-based unsupervised method, current features show robustness for clustering of 150 event windows.

Appendix 2A

Features used for feature-based clustering (Schreiber and Schmitz 1997; Fulcher and Jones 2014; Christ et al. 2018)

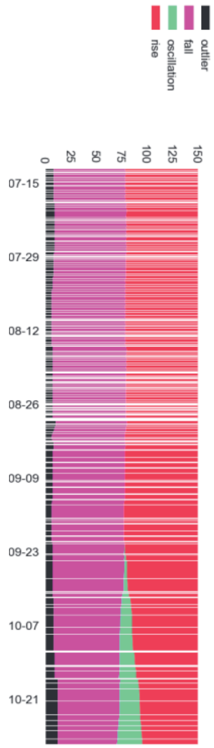
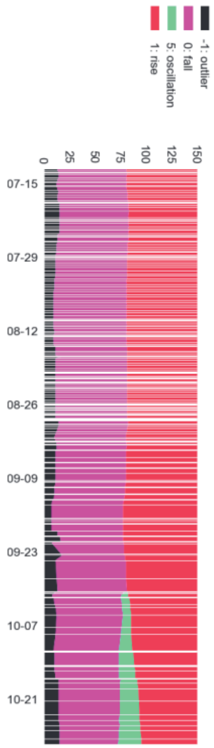
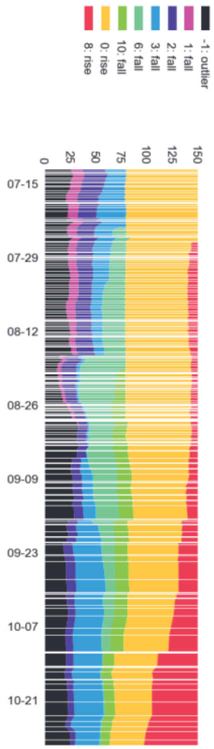
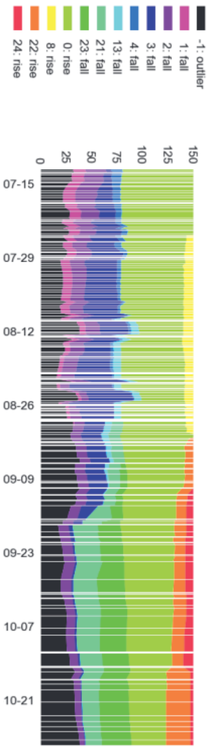
Features	***Description***
max	Maximum of the event measurements
min	Minimum of the event measurements
std	Standard deviation of the event measurements
skew	Unbiased skewness, normalized by N-1
kurt	Unbiased kurtosis using Fisher's definition of kurtosis, normalized by N-1.
diff_mean	Mean of the first order derivative of the event measurements
diff_skew	Unbiased skewness, normalized by N-1, of the first order derivative of the event measurements
diff_std	Standard deviation of the first order derivative of the event measurements
lin_appr_slope	Slope estimate p_1 from a linear least squares fit ($y = p_1x + p_2$)
lin_appr_intercept	Intercept estimate p_2 from a linear least squares fit ($y = p_1x + p_2$)
cub_appr_x1	Polynomial coefficient estimate p_1 from a cubic least squares fit through the sorted data ($y = p_1x^3 + p_2x^2 + p_3x + p_4$)
cub_appr_x2	Polynomial coefficient estimate p_2 from a cubic least squares fit through the sorted data ($y = p_1x^3 + p_2x^2 + p_3x + p_4$)
cub_appr_x3	Polynomial coefficient estimate p_3 from a cubic least squares fit through the sorted data ($y = p_1x^3 + p_2x^2 + p_3x + p_4$)
cub_appr_intercept	Intercept estimate p_4 from a cubic least squares fit through the sorted data ($y = p_1x^3 + p_2x^2 + p_3x + p_4$)
slope_appr	Difference between initial and final event value divided by the number of seconds between
std_0.5	Fraction of event measurements larger than 0.5 times the standard deviation
std_1.0	Fraction of event measurements larger than the standard deviation
std_1.5	Fraction of event measurements larger than 1.5 times the standard deviation
domfreq_0.000	Dominant Power Spectral Density frequency
domfreq_0.003	Dominant Power Spectral Density frequency above 1/300 s^{-1}
domfreq_0.008	Dominant Power Spectral Density frequency above 1/120 s^{-1}
domfreq_0.017	Dominant Power Spectral Density frequency above 1/60 s^{-1}
psdbin_1	Power Spectral Density fraction between frequencies 1/600 and 1/103 s^{-1}
psdbin_2	Power Spectral Density fraction between frequencies 1/103 and 1/56 s^{-1}
psdbin_3	Power Spectral Density fraction between frequencies 1/57 and 1/39 s^{-1}
psdbin_4	Power Spectral Density fraction between frequencies 1/39 and 1/30 s^{-1}
psdbin_5	Power Spectral Density fraction between frequencies 1/30 and 1/24 s^{-1}
acorr_5	Autocorrelation with a lag of 5 seconds
acorr_10	Autocorrelation with a lag of 10 seconds
acorr_15	Autocorrelation with a lag of 15 seconds
acorr_30	Autocorrelation with a lag of 30 seconds
acorr_60	Autocorrelation with a lag of 60 seconds
domacorr_0.50	Fraction of autocorrelation function above 50% correlation
domacorr_0.65	Fraction of autocorrelation function above 65% correlation
domacorr_0.80	Fraction of autocorrelation function above 80% correlation
domacorr_0.90	Fraction of autocorrelation function above 90% correlation
domacorr_0.95	Fraction of autocorrelation function above 95% correlation
c3_lag_5	Time series non-linearity measure using a lag operator of 5 seconds
c3_lag_10	Time series non-linearity measure using a lag operator of 10 seconds
c3_lag_15	Time series non-linearity measure using a lag operator of 15 seconds
tras_lag_5	Time reversal asymmetry statistic using a lag operator of 5 seconds
tras_lag_10	Time reversal asymmetry statistic using a lag operator of 10 seconds
tras_lag_15	Time reversal asymmetry statistic using a lag operator of 15 seconds

Feature performance evaluation for the 2013 (top) & 2017 (bottom) data set. After feature calculation for all events within a single data set, each feature is standardized. For each pattern present in the manually labeled validation, a box plot of feature values is made per pattern. Features showing well-separated pattern-specific box plots with low within pattern variation are most suitable for separating the patterns present in the investigated data set. As can be derived from both the 2013 and 2017 data set box plots and validation reports (Figure 2.5), current features are deemed suitable and robust for the intended goal of detecting recurrence of anomalous patterns



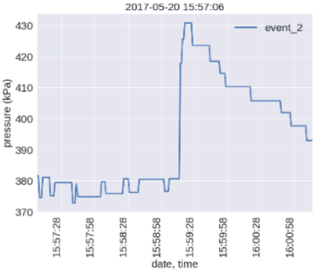
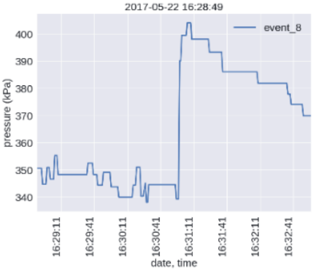
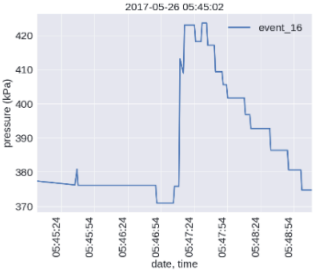
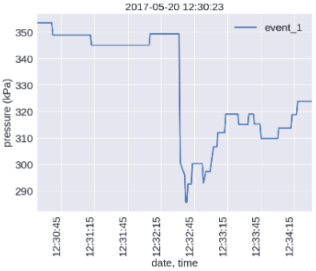
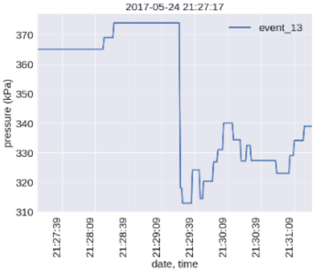
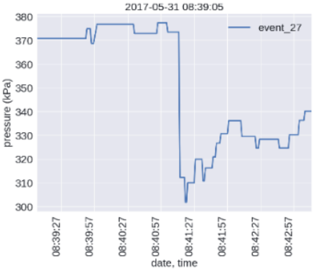
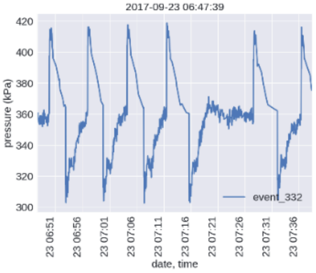
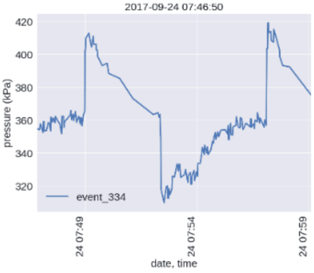
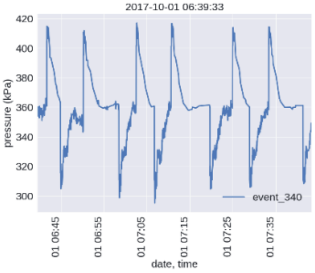
Appendix 2B

Results of the 2017 data set for instance-based clustering without DTW (A), with DTW (B), feature-based clustering (C) and validation using manual labeling (D). As can be seen from the validation labeling (D), three different recurring patterns are found in the data set, referred to as rise, fall and oscillation. The legend lists the cluster ID number and which pattern matches best with that clusters ID, as derived from the manually labeled validation data. Clusters numbers mapping to the same recurring pattern, were deemed a single cluster for the sake of accuracy scores calculation only.

	D) Manual Validation	C) Feature-Based	B) I.-B. With DTW	A) Instance-Based
Fingerprint Graph				

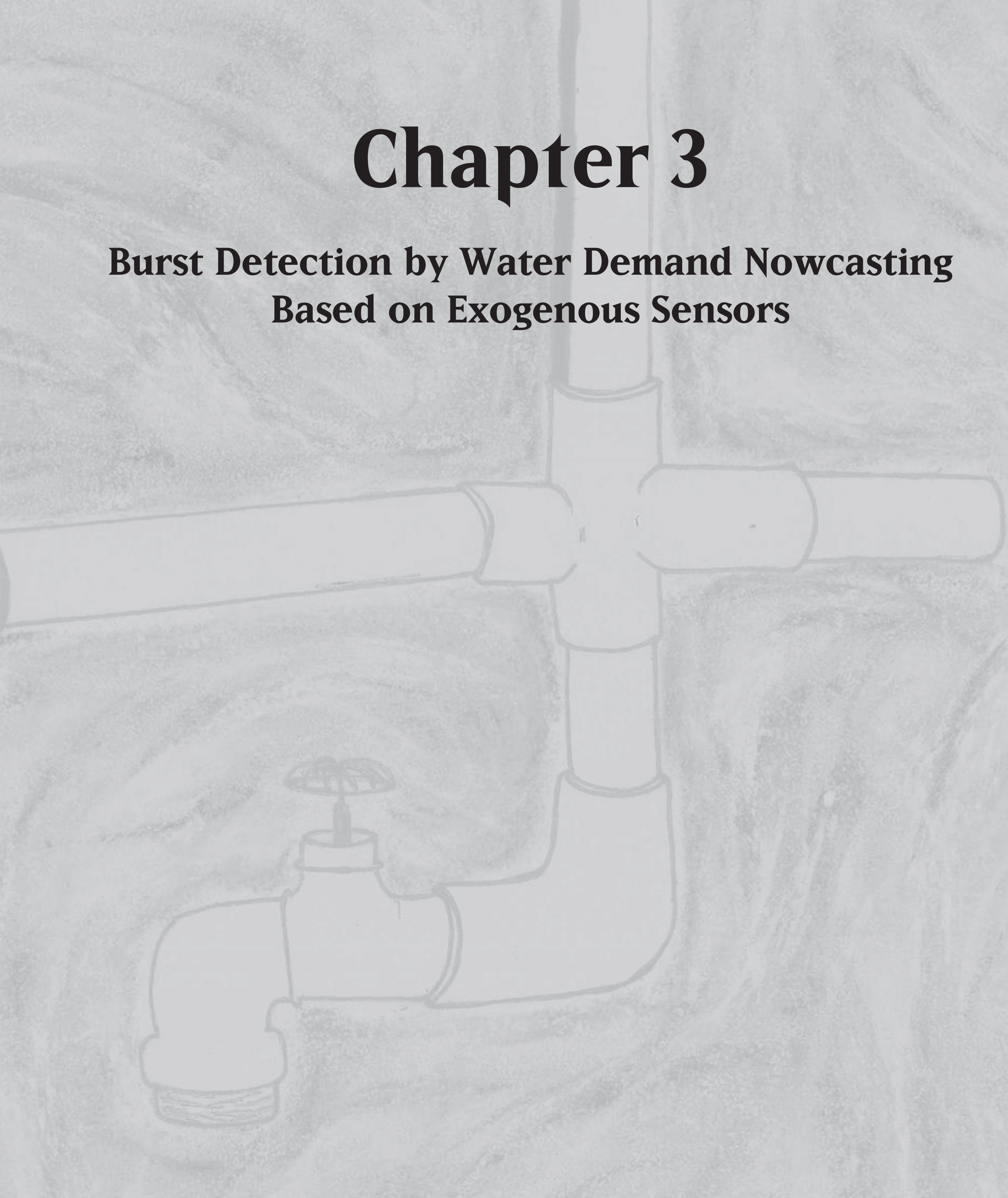
Validation Report	C) Feature-Based				B) I.-B. With DTW				A) Instance-Based			
	Precision	***Recall***	***F_1***	***Support***	***Precision***	***Recall***	***F_1***	***Support***	***Precision***	***Recall***	***F_1***	***Support***
Fall	0.96	0.93	0.94	164	0.94	0.78	0.85	164	0.93	0.77	0.84	164
Oscillation	1.00	0.80	0.89	25	0.00	0.00	0.00	25	0.00	0.00	0.00	25
Rise	0.95	0.93	0.94	160	0.83	0.94	0.88	160	0.91	0.92	0.91	160
Average	**0.96**	**0.92**	**0.94**	**349**	**0.82**	**0.80**	**0.80**	**349**	**0.85**	**0.79**	**0.82**	**349**

Examples of the recurring patterns present in the 2017 data set

<table>
<tr><th>Pattern</th><th>Example 1</th><th>Example 2</th><th>Example 3</th></tr>
<tr><td>Rise</td><td> </td><td> </td><td> </td></tr>
<tr><td>Fall</td><td> </td><td> </td><td> </td></tr>
<tr><td>Oscillation</td><td> </td><td> </td><td> </td></tr>
</table>

Chapter 3

Burst Detection by Water Demand Nowcasting Based on Exogenous Sensors



A version of this chapter has been published as:

Geelen, C.V.C., Yntema, D.R., Molenaar, J., Keesman, K.J., 2021. Burst Detection by Water Demand Nowcasting based on Exogenous Sensors. Water Resour. Manag. 35, 1183–1196. https://doi.org/https://doi.org/10.1007/s11269-021-02768-9

3. Burst Detection by Water Demand Nowcasting Based on Exogenous Sensors

Abstract

Bursts of drinking water pipes not only cause loss of drinking water, but also damage below and above ground infrastructure. Short-term water demand forecasting is a valuable tool in burst detection, as deviations between the forecast and actual water demand may indicate a new burst. Many of burst detection methods struggle with false positives due to non-seasonal water consumption as a result of e.g. environmental, economic or demographic exogenous influences, such as weather, holidays, festivities or pandemics. Finding a robust alternative that reduces the false positive rate of burst detection and does not rely on data from exogenous processes is essential. We present such a burst detection method, based on Bayesian ridge regression and Random Sample Consensus. Our exogenous nowcasting method relies on signals of all nearby flow and pressure sensors in the distribution net with the aim to reduce the false positive rate. The method requires neither data of exogenous processes, nor extensive historical data, but only requires one week of historical data per flow/pressure sensor. The exogenous nowcasting method is compared with a common water demand forecasting method for burst detection and shows sufficiently higher Nash-Sutcliffe model efficiencies of 82.7% - 90.6% compared to 57.9% - 77.7%, respectively. These efficiency ranges indicate a more accurate water demand prediction, resulting in more precise burst detection.

3.1. Introduction

Water distribution networks form an extensive and complex underground infrastructure, coping with a water demand that changes over time and per location. Due to this complexity, optimal management and operation of the distribution network is a challenging task. Suboptimal management has wide ramifications, such as faster deterioration of pipes, insufficient water pressure, increased burst frequency and higher operational costs (Billings and Jones 2008; Kozłowski et al. 2018). Forecasting water demand will help optimize network management and facilitates fault detection (Brentan et al. 2017). Water demand forecasting is challenging, since water consumption depends on many environmental, economic, and demographic factors with temporal and spatial variation (Hutton and Kapelan 2015b). One high priority use of short-term water demand forecasting is burst detection. Burst detection methods are typically based on detecting significant deviations between measured and predicted water demand. Conventionally, the measured water consumption in a District Metering Area (DMA) is compared to a forecast based on historical measurements of water consumption on e.g. the same day in the week and the same time on that day. Significant deviations between the forecasted and the current water consumption indicate a burst, if a suitable and accurate forecasting method is used (Hutton and Kapelan 2015a). The most frequently used methods for water demand forecasting are based on univariate time series models, such as autoregressive moving average (ARMA) models (Hutton and Kapelan 2015b). ARMA models are suitable for short-term forecasts of water demand, as these models are strong in capturing the specific periodic patterns of water consumption. However, water demand is not only a function of periodic water consumption, but is also influenced by exogenous processes, such as holidays, festivals, the weather, pandemics, or other non-periodic deviations water consumption. Ordinary ARMA models do not take into account these exogenous processes, resulting in an increased false positive rate of burst detection (Billings and Jones 2008). In order to take into account exogenous processes, (multiple) (non-)linear regression or exogenous ARMA models were used, under the condition that extensive data on each of these exogenous processes are available (Adamowski et al. 2012; Papageorgiou et al. 2015; Froelich 2016; Candelieri 2017).

Recent methods make use of neural networks (NN) or other supervised machine learning methods, or hybrid methods that combine NN with univariate/regression forecasting models (Babel and Shinde 2011; Bai et al. 2014; Xu et al. 2018; Pacchin et al. 2019). Similar to exogenous ARMA models, these methods are capable of incorporating exogenous data and boast reliable forecasts, but require extensive historical data for training and are accompanied by large forecast uncertainties, which cannot always be quantified (Hutton and Kapelan 2015b; Anele et al. 2017). Although powerful, even these NN and hybrid models still require identification of all relevant exogenous processes with corresponding data. Identifying the many environmental, economic and demographic exogenous processes that influence drinking water demand as well as collecting all the corresponding data, is not realistic or feasible for most water distribution companies. A method that does not depend on data of exogenous processes would be invaluable to water demand forecasting and would greatly improve burst detection precision.

Up to now, data of exogenous processes for water demand forecasting was obtained from external sources (such as weather institutes or statistical agencies). However, the multitude of installed pressure and flow sensors in the network present a new, internal data source. These sensors can all be considered as real-time exogenous factors, as they reflect all of the exogenous processes, without having to identify what is the underlying cause of these processes. Hence, instead of using a short-term forecast of water demand based on forecasted exogenous processes, a water demand nowcast per sensor or DMA water balance based on exogenous flow and pressure sensors within the distribution network can be used. Where forecasting water consumption typically relies on the seasonal nature of collective human water consumption, water demand nowcasting not solely accounts for seasonal water consumption, but also various other water demand patterns caused by exogenous processes, such as weather, holidays, or valve position changes.

This observation becomes especially relevant regarding burst detection. When solely forecasting the seasonal water consumption, significant deviations between the forecasted and the measured water demand will contain many false positive burst alarm caused by non-seasonal water demand due to exogenous processes. Nowcasting water demand at a specific location and based on exogenous sensors will result in a significantly reduction of the false positive rate, as not only seasonal water demand, but also diverging water demand due to exogenous processes can be taken into account. Consequently, when the nowcast deviates from the measured water demand at a specific location, a burst alarm is triggered.

When investigating the measurements of a flow sensor situated close to a burst, the sensor will record a corresponding water demand pattern. However, since most exogenous flow and pressure sensors used in the nowcasting of this sensor will not detect this local burst, the nowcast will reflect the normal diurnal water demand pattern. The resulting difference between measured and nowcasted water demand will thus signal that a bursts has occurred. However, if a more widespread event, such as a holiday, causes a non-diurnal water demand pattern in the investigated sensor, most exogenous sensors will also show a similar pattern. Therefore the nowcasted and measured water demand will not deviate, and this event will thus not trigger a burst warning. The nowcasted water demand based on sensors in proximity as exogenous regressors will more accurately reflect actual water demand compared to water consumption forecasts based on exogenous methods, and thus allows for robust, high certainty and high precision burst detection, without needing vast historical data sets.

The objective of this study was to investigate and evaluate a water demand nowcasting method based on exogenous data from sensors in proximity to the nowcasted sensor or water balance. Our exogenous water demand nowcasting method is compared with a univariate water demand forecasting method that does not depend on data of exogenous processes and is based on RANdom SAmple Consensus (RANSAC) weighted linear regression using up to 20 weeks of past flow measurements (Fischler and Bolles 1981). The water demand nowcast is constructed from the signals from multiple flow and pressure sensors in the distribution network as exogenous factors in a RANSAC Bayesian ridge regression model (MacKay 1992). A 95% prediction uncertainty interval is determined for both methods, to evaluate the uncertainty of the forecasted and nowcasted water consumption. Where other methods reduce exogenous false positives by finding sensor signals with a relatively high distance compared to the signal of other exogenous sensors (Wu et al. 2018), exogenous nowcasting uses the nowcast's uncertainty interval to determine burst occurrence. Three data sets were subjected to the forecasting and nowcasting methods, after which the model efficiency scores were calculated in order to evaluate their performance.

3.2. Materials and Methods

The exogenous nowcast method and the univariate forecast method were applied to a DMA water balance (data set DMA1), a city-wide sub-DMA water balance (data set DMA1.1) and a single flow sensor (data set Q1) (Figure 3.1), sampled each five minutes from 01/06/2017 up to 01/11/2019, except for DMA1.1, which was sampled from 22/02/2018 up to 01/11/2019, since this DMA was not operational before this date. All data sets were provided by the Dutch drinking water company Vitens. DMA1 is situated in a mainly rural area with a population of more than 100,000 inhabitants spread over 800 km2 . DMA1.1 covers the largest city within DMA1 of more than 30,000 inhabitants and sensor Q1 is located near one of the water production facility within DMA1. For the exogenous nowcast method, data from up to 42 sensors from within DMA1 were used as the exogenous regressors (17 pressure sensors, 25 flow sensors of which 12 industrial water demand flow sensors). The data sets of these sensors were also sampled each five minutes from 01/06/2017 up to 01/11/2019.

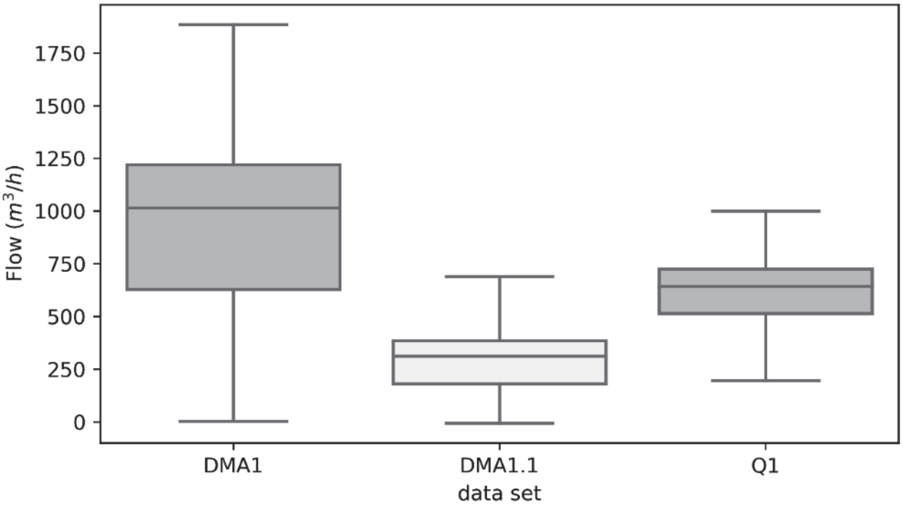


Figure 3.1 - Boxplots of the data sets of DMA1, DMA1.1 and Q1.

3.2.1. Univariate Water Demand Forecast

In order to forecast up to one week of water demand for a district metering area (DMA) or at a flow sensor in the net using a univariate water demand forecasting method, past measurements from that DMA or sensor are required. For each time t up to one week in the future, a forecast $\hat{y}t$ can be made based on past measurements. For that we took those measurements from the preceding 20 weeks that correspond with the same day in the week and the same time on that day. The corresponding linear regression problem is formulated as:

$$Y = X\beta + \epsilon \tag{3.1}$$

$$\hat{\beta} = ((WX)^T X)^{-1}(WX)^T Y \tag{3.2}$$

Here, $N = 20$ are the number of prior measurements considered, $Y = [y_1 y_2, \dots, y_N]^T$ is an N-dimensional vector with measured water demands for $1,2, \dots N$ weeks prior to time t, ϵ is the corresponding residual vector, $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & N \end{bmatrix}$ is the regressor matrix, and $\hat{\beta}$ is the parameter vector with weighted least squares estimates of the regression coefficients, which are the intercept and slope of the line fitting the 20 data points. Since more recent water demand has more predictive value compared less recent water demand, exponentially weighted least squares is applied to rely relatively more on the most recent measurements, instead of ordinary or generalized least squares. This weighting is achieved by using the diagonal weighting matrix $W = \begin{bmatrix} w_{1,1} & 0 & \cdots & 0 \\ 0 & w_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & w_{N,N} \end{bmatrix}$, where $w_{i,i} = (1 - p)^{N-i+1}$ with $i = 1,2, \dots, N$ and weighting factor $p = 0.2$.

To ensure robust regression, RANdom SAmple Consensus (RANSAC) is applied in order to eliminate the outliers from the training data measurements (Fischler and Bolles 1981). This is achieved by selecting all sets s consisting of every possible combination of N_s rows from X (keeping them in the original order), the corresponding N_s values from Y, and N_s rows and columns from W with $2 \le N_s \le N$. Each resulting combination $\tilde{X}$, $\tilde{Y}$, and $\widetilde{W}$, is used in Eq. (3.2), resulting in a corresponding estimate of the parameter vector $\hat{\beta}_s$. The "optimal" parameter vector $\hat{\beta}_{opt}$, that maximizes the RANSAC cost function, and accompanying inlier combinations of regressors X_{opt}, responses Y_{opt} and weights W_{opt} can be found from all sets s using the RANSAC cost function for each set s:

$$\hat{\beta}_{opt}(s) = \underset{\hat{\beta}_s}{\operatorname{argmax}} \sum_{i=1}^{N_s} \begin{cases} \widetilde{W}_{i,i} & if\ \left(\tilde{y}_{s\,i} - \tilde{X}_{i\,s}\hat{\beta}_s\right)^2 \le \delta_d, \\ 0 & otherwise. \end{cases} \tag{3.3}$$

Outliers are excluded from the regression, if the squared residuals $\left(\tilde{y}_{s\,i} - \tilde{X}_{i\,s}\hat{\beta}_s\right)^2$ are larger than a residual threshold δ_d chosen as the Median Absolute Deviation (MAD) of the responses $\delta_d = \mathrm{MAD}(Y) = \mathrm{median}(|Y - \mathrm{median}(Y)|)$. If RANSAC or missing data results in less than $N_{min} = 12$ inliers, a residual threshold $\delta_d = 2 * \mathrm{MAD}(Y)$ is used instead. If this more tolerant threshold still results in less than $N_{min} = 12$ inliers, RANSAC is not used, as RANSAC apparently does not help to improve the linear fit. In that case, all 20 measurements are used. Regarding the application of water demand forecasting, using multiples of $\mathrm{MAD}(Y)$ were chosen as the RANSAC residual threshold δ_d, since this is assumed to result in robust results when responses Y have low noise. Thus, the cost function in Eq. (3.3) makes use of the exponential weights $\widetilde{W}_{i,i}$ in order to penalize past measurements, since recent measurements are assumed to strongly resemble future water demand. The lower threshold of 12 inliers was used to ensure sufficient data is retained to fit the model. Combined with the exponential weights, this ensures sufficient measurements from the recent past are still taken into account.

Estimates of the predicted value $\hat{y}_t$ and variance $\Sigma_{\hat{y}_t}$ as well as the 95% prediction uncertainty interval $[\hat{y}_{t*}, \hat{y}_t^*]$ can be calculated based on the squared residuals Σ_ϵ, significance level 0.05, and $\widetilde{N}_{opt} - 2$ degrees of freedom (Chatfield 1993):

$$\Sigma_\epsilon = \left\| Y_{opt} - X_{opt}\hat{\beta}_{opt} \right\|_2^2 \tag{3.4}$$

$$\Sigma_{\hat{y}_t} = \left(1 + X_t\left(X_{opt}^T X_{opt}\right)^{-1} X_t^T\right)\Sigma_\epsilon \tag{3.5}$$

$$\hat{y}_t = X_t\hat{\beta}_{opt} \tag{3.6}$$

$$[\hat{y}_{t*}, \hat{y}_t^*] = \left[\hat{y}_t \pm t_{0.975,\widetilde{N}_{opt}-2}\sqrt{\Sigma_{\hat{y}_t}}\right] \tag{3.7}$$

3.2.2. Exogenous Water Demand Nowcast

The nowcasted water demand $\hat{y}_t$ was constructed using windows with $N = 2016$ measurements, corresponding to one week of sensor data sampled every five minutes, using P flow and pressure sensors in the same DMA. In the case of DMA-wide water demand nowcasting, data from inflow, outflow, and water production location sensors were excluded as regressors, as these are constituents of the DMA water mass balance (Hutton and Kapelan 2015a). For every window of 2016 measurements, sensor signals with a standard deviation smaller than $5\ kPa$ or $5\ m^3h^{-1}$ were excluded from the analysis, as signals with a small standard deviation contain no or hardly any information with added value to the nowcasting process. Consequently, these non-persistently exciting signals are omitted from the analysis as these hardly contain useful information and lead to increased multicollinearity.

The nowcasted water demand is constructed using a RANSAC Bayesian ridge regression model. For each week of N measurements, a real-time estimate $\hat{y}_t$ for a specific sensor or water balance can be calculated, based on past measurements with a sampling interval of five minutes. Thus, in this case: $Y = [y_1 y_2, \dots, y_N]^T$, and $X = \begin{bmatrix} 1 & x_{1,2} & \cdots & x_{1,P} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,2} & \cdots & x_{N,P} \end{bmatrix}$ for P exogenous regressors. The resulting Bayesian ridge prediction is formulated as:

$$Y = X\beta + \epsilon, \qquad \epsilon \sim \mathcal{N}\left(0, \hat{\Sigma}_\epsilon\right) \tag{3.8}$$

$$Y \sim \mathcal{N}\left(X\hat{\beta}, \hat{\alpha}\right) \tag{3.9}$$

$$\beta \sim \mathcal{N}\left(\hat{\beta}, \hat{\lambda}^{-1} I_P\right) \tag{3.10}$$

$$\hat{\beta} = \left(X^T X + \hat{\lambda} I_P\right)^{-1} X^T Y \tag{3.11}$$

Here, ϵ is the residual vector, α the variance of the noise, and λ the Tikhonov regularization penalty. Weakly informative Gaussian priors were used for the uncertainty in the regression coefficients, $\lambda \sim \Gamma(10^{-6}, 10^{-6})$ and noise variance $\alpha \sim \Gamma(10^{-6}, 10^{-6})$, with initial guesses $\hat{\lambda}_0 = 1$ and $\hat{\alpha}_0 = \frac{1}{\Sigma Y}$ in order to ensure fast and accurate optimization of the regression parameters, before solving for $\hat{\beta}$, $\hat{\lambda}$, and $\hat{\alpha}$ (MacKay 1992; Tipping 2001). By not setting the regularization penalty λ beforehand, but treating it as a random variable, it can be automatically tuned to the data, concurrently with α, λ, and β.

Ridge regression was chosen for its capacity to reduce multicollinearity caused by sensors displaying similar diurnal water demand patterns. Without regularization, thus with $\lambda = 0$, this would lead to a nearly singular matrix $X^T X$. Regularization improves efficiency of the nowcasting and reduces variance, by introducing a small amount of bias (Pacchin et al. 2019). For the practical application of water demand nowcasting, the small amount of bias introduced is deemed acceptable in the bias-variance tradeoff, as it prevents the prediction from being over-dependent on the signal of a single exogenous sensor. Bayesian LASSO (least absolute shrinkage and selection operator) regression was also considered, but was ultimately rejected, as prioritizing a low number of regressors resulted in overfitting on just a few exogenous sensors, which makes the prediction highly sensitive to local anomalies in a few exogenous sensors. By relying on multiple exogenous sensors under ridge regularization, a more robust prediction was created.

Similar to using the cost function (Eq. 3.3), RANSAC was used to remove outliers from the training data, retaining at least 90% inliers of all measurements ($N_{min} = 1814$). Every possible inlier combination s of N_s rows from X and the corresponding N_s values from Y with $2 \leq N_s \leq N$ is subjected to the Bayesian ridge regression model (Eqns 3.8-3.11) in order to find the corresponding regression coefficient estimates $\hat{\beta}_s$ and their precision $\hat{\lambda}_s$. The "optimal" parameter vector $\hat{\beta}_{opt}$, accompanying precision $\hat{\lambda}_{opt}$, and set s of inlier combination of regressors X_{opt} and responses Y_{opt} can be found using the RANSAC cost function (Eq. 3.3), disabling the weighting of more recent data by using $w_{i,i} = 1$ for $i = 1,2, \dots N_s$. A residual threshold $\delta_d = 0.2MAD(Y)$ was used, unless this results in less than 90% inliers ($N_{min} = 1814$), in which case $\delta_d = MAD(Y)$ was used. If neither resulted in more than 90% inliers or, due to missing data points, less than 90% of the total window size is available, RANSAC was not used. Removing a small fraction of data may indicate some outliers were present in the data. However, when RANSAC disregards a large fraction of the data (≥ 10%), this most likely indicates an anomalous signal that cannot be appropriately fitted by the chosen model. In this case, prioritizing fitting the model to the data instead of editing the data to fit the model will most likely explain more of the phenomena present in the data.

In order to construct a reliable nowcast of the water demand, the model should be fitted on the basis of representative data with minimal outliers. If anomalous events occur in the training data, masking these as outliers will benefit the model more than the commonly used weighting or replacement (Eliades and Polycarpou 2012; Ye and Fenner 2014). An additional advantage compared to similar methods is that allowing masking of a small percentage of data also ensures that the method does not struggle from a small percentage of missing data points, as these will be masked (Wu et al. 2018).

The resulting inlier regressor matrix X_{opt}, response vector Y_{opt}, and model parameter vector $\hat{\beta}_{opt}$ and $\hat{\lambda}_{opt}$ are used to calculate the sum of squared residuals Σ_{ϵ} and response estimate $\hat{y}_t$ (Eqns 3.4 and 3.6, respectively), as well as the response estimate's variance $\Sigma_{\hat{y}_t}$ and 95% prediction uncertainty interval $[\hat{y}_{t*}, \hat{y}_t^*]$:

$$\Sigma_{\hat{y}_t} = \left(1 + X_t\left(\tilde{X}_{opt}^T\tilde{X}_{opt} + \hat{\lambda}_{opt}I_P\right)^{-1}X_t^T\right)\sum\nolimits_{\epsilon} \tag{3.12}$$

$$[\hat{y}_{t*}, \hat{y}_t^*] = \left[\hat{y}_t \pm z_{0.975}\sqrt{\Sigma_{\hat{y}_t}}\right] \tag{3.13}$$

As the regression coefficients change slowly, it suffices to fit the water demand model only once per day. However, real-time predictions can still be constructed at any time t based on the last fitted regression coefficients $\hat{\beta}_{opt}$ and regularization penalty $\hat{\lambda}_{opt}$. To illustrate this approach, water demand nowcasting was performed every five minutes from the latest fitted model, which was updated every day at midnight.

3.3. Results and Discussion

The univariate forecasting and exogenous nowcasting method were applied to the three data sets, DMA1, DMA1.1 and Q1. For data set DMA1, the results of both methods were compared with a so called Dynamic Bandwidth Monitor (DBM), a univariate forecasting method developed and currently in use by drinking water company Vitens (Fitié 2014) (Figure 3.2). To evaluate and compare the model efficiencies between the methods applied to the same data set, the Normalized Root Mean Squared Error (NRMSE, Eq. 3.14, where values closer to 0% indicate better performance) was calculated; To compare the model efficiencies using different data sets, the Nash-Sutcliffe model efficiency (NS, Eq. 3.16, where a value closer to 100% indicates better performance) was calculated (Table 3.1).

$$NRMSE = \frac{1}{\mu_Y}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} * 100\% \qquad (3.14)$$

$$NS_p = \left(1 - \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|^p}{\sum_{i=1}^{N}|y_i - \mu_Y|^p}\right) * 100\% \qquad (3.15)$$

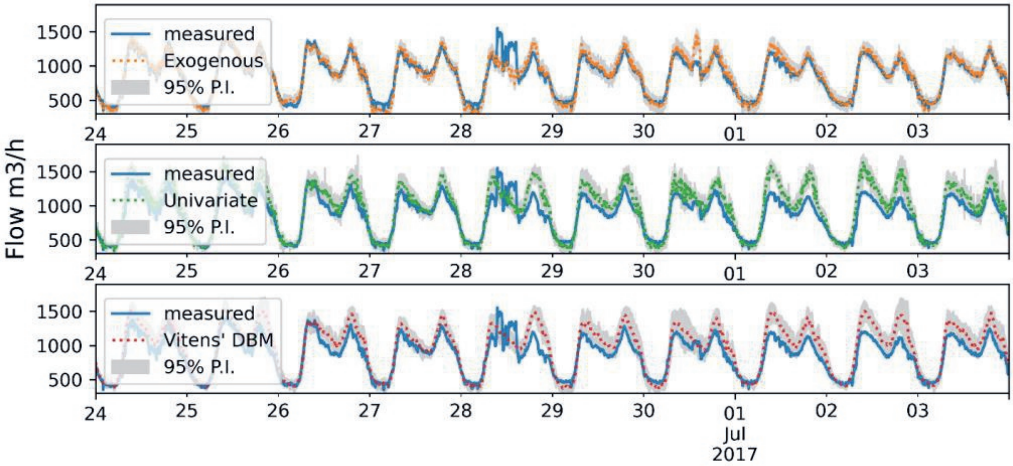


Figure 3.2 – Comparison between measured water demand and exogenous water demand nowcast (top), univariate water demand forecast (middle) and Vitens DBM forecast (bottom) for data set DMA1, along with their respective 95% prediction uncertainty intervals from 24th of June 2017 up to the 3rd of July 2017.

Table 3.1 - Model performance scores.

MODEL	DATA SET	NS$_1$ (%)	NRMSE (%)
EXOGENOUS	DMA1	90.6	6.5
UNIVARIATE	DMA1	75.1	11.4
VITENS' DBM	DMA1	77.7	10.0
EXOGENOUS	DMA1.1	82.7	10.4
UNIVARIATE	DMA1.1	74.6	14.1
EXOGENOUS	Q1	83.5	6.0
UNIVARIATE	Q1	57.9	11.0

Burst detection can be done by evaluating the actual flow measurements with respect to the prediction uncertainty intervals of the predicted water demand. Where some studies rely on manually selected or validation data based burst detection thresholds (Huang et al. 2018; Wu et al. 2018), exogenous water demand nowcasting relies on the calculated prediction uncertainty intervals.

The performance of the nowcasting method compared univariate methods is illustrated in Figure 3.2. Where the univariate forecasts deviates from the measured flow due to non-seasonal exogenous processes, and thus trigger significantly more false positive burst alarms, the exogenous nowcast is able to account for these phenomena and thus prevent these false alarms. Within DMA1, at 16:20 on the 28th of June a lengthwise tear burst occurred in a 630mm PVC pipe dating out of 1976 and at 14:20 on the 30th of June a ‘simulated burst’ occurred in a T-junction between 400mm PVS pipes dating from 1989 and 1994 when water was lost during placement of new pipes. Both bursts are detected correctly by the exogenous nowcasting method. Although the burst on the 28th is also detected by the univariate methods, the difference is less pronounced and the burst on the 30th is not detected by the univariate methods. The reduced burst alarm rate of the exogenous method and the combination of these water demand prediction methods are detailed in Appendix 3A.

The exogenous nowcasting method also outperforms the univariate methods regarding reduced alarm rate (Table 3.1). A possible reason for the very good performance of the exogenous method is its real-time nature in contrast with the one-week-ahead forecasts of the univariate methods. However, regarding real-time burst detection, the forecast window size is not relevant. The univariate method more often significantly deviates from the measured flow (Appendix 3A), as these deviations could be caused by any of many exogenous processes, ranging from holidays, festivities, extreme weather to unexpected peak water consumption (Figure 3.2).

The better performance of the nowcasting method is less pronounced when looking at DMA1.1. DMA1.1 reflects the water demand in a large city, while the majority of the exogenous sensors in DMA1 that are used to nowcast the DMA1.1 water balance are situated in more rural areas. The resulting difference in demographics influences water demand, making the rural exogenous sensors used suboptimal for predicting the DMA1.1 city water demand. For DMA1 and Q1, the exogenous sensors reflect water demand from both rural and city areas, which may explain the better performance.

The investigated DMA’s from Vitens contained enough sensors with signals that could serve as exogenous regressors. However, sensor density differs between DMA and water company. The sensitivity of the exogenous water demand nowcasting method with respect to the number of exogenous regressors considered was also investigated by applying our method to the Q1 data set for different number of exogenous regressors. Of the 40 exogenous sensors data sets available for sensor Q1, 30 were used in fitting the model, as the remaining 10 either did not significantly contribute to the Bayesian ridge analysis or had a too low variance to be included in the analysis. In order to determine how much each sensor contributed to the exogenous prediction, the mean of the absolute regression coefficient estimates ($\mu(\beta_i) = \underset{t}{\mathrm{mean}}\left(\left|\beta_{opt,i}(t)\right|\right)$ for $i = 1,2, \dots, P$) was determined for each of these sensor signals. Data set Q1 was again subjected to the exogenous method, where in each consecutive iteration the regressor with the smallest μ_i was removed. For less than three regressors, the method was not able to produce a prediction for at least 95% of all measurements, thus the resulting number of sensors investigated was ranging from 30 to 3. The NRMSE, Mean Absolute Percentage Error (MAPE, Eq. 3.16) and mean of the 95% prediction uncertainty interval bandwidth over the duration of the data set ($\mu(95\% P.I.)$) were calculated for each number of exogenous regressors used by comparing the respective predictions with the actual measurements (Figure 3.3).

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100\% \tag{3.16}$$

Using less regressors may result in a lower precision of burst detection, since there may be insufficient data on the local consumption pattern present in a limited number of regressors. In addition, using less regressors may result in a lower recall of burst detection, especially when a burst occurs that is reflected in the data of all regressors. Including more regressors reduces this possibility and increases recall. This result may also explain the few instances of increase in NRMSE and MAPE when including more sensors, instead of the expected decrease. Consequently, from the top panel in Figure 3.3, thus for the given period of data set Q1, approximately 13-20 sensors are needed to obtain appropriate water demand predictions. Consequently, this analysis facilitates the choice of sensor density for “optimal” detection. The other two data sets, DMA1 and DMA1.1, were subjected to the same approach and showed similar results (Figure 3.3), middle and bottom panel, respectively).

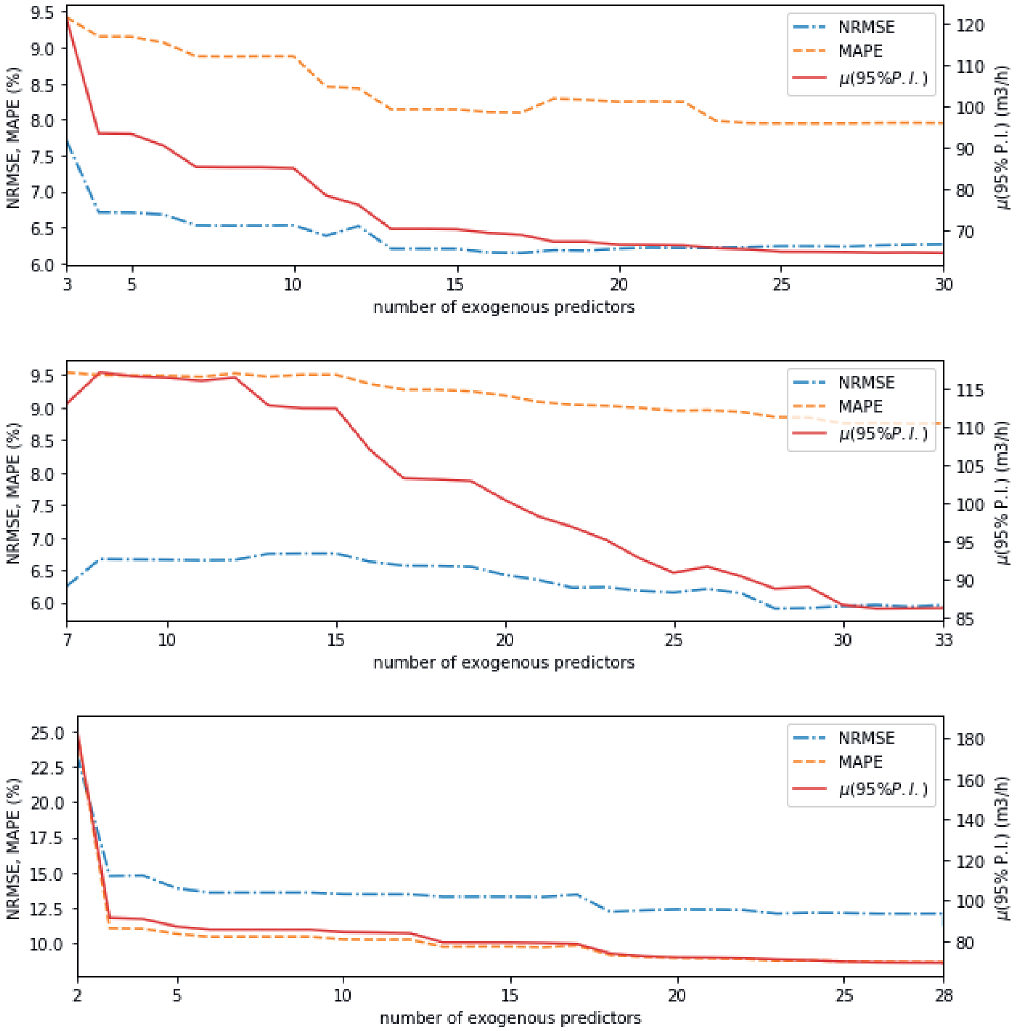


Figure 3.3 – Model performance (NRMSE, MAPE and average prediction uncertainty interval) of the exogenous method for data set Q1 (top), DMA1 (middle) and DMA1.1 (bottom) as a function of the number of exogenous sensors used as regressors.

3.4. Conclusions

Exogenous nowcasting is a more robust and accurate alternative to univariate water demand forecasting based on historical data. An advantage of nowcasting based on exogenous sensors in the distribution net is that no exogenous processes that influence the water demand have to be identified and no data from these processes need to be available. Our novel exogenous method performs significantly better than both currently used univariate methods for both data sets, regarding reduced false positive rate (Figure 3.2, Table 3.1).

By combining both the exogenous nowcasting and one of the univariate forecasting methods, a high precision tool is created that only detects a burst when the actual measurements lie outside both the exogenous and univariate 95% prediction uncertainty interval (Appendix 3A).

Notation

The following symbols are used in this paper:

N	scalar		number of measurements
P	scalar		number of exogenous regressors
Y	vector	$N{\times}1$	responses
X	matrix	$N{\times}(P+1)$	regressors
β	vector	$(P+1){\times}1$	regression coefficients
W	matrix	$N{\times}N$	exponential weights of responses
p	scalar		exponential weighting factor $0 < p < 1$
λ	scalar		regularization penalty
α	scalar		noise variance
t	scalar		time
ϵ	vector	$N{\times}1$	residuals
Σ_a	scalar		variance of random variable a
s	-		ordered set of inliers with N_s elements, where $2 \leq N_s \leq N$
$\tilde{a}$ or a_s	-		belonging to the inlier set s
a_{opt}	-		belonging to the “optimal” inlier set s as determined by RANSAC
$\hat{a}$	-		estimate of variable a
$t_{a,b}$	scalar		t-value with significance level a and b degrees of freedom
z_a	scalar		z-value with significance level a
$[\hat{y}_{*t}, \hat{y}_t^*]$	scalar, scalar		95% prediction uncertainty interval for predicted response $\hat{y}_t$
$\mu(95\%P.I.))$	scalar		mean of the prediction uncertainty interval over time
μ_Y	scalar		mean of vector Y
$RMSE$	scalar		Root Mean Squared Error
APE	scalar		Absolute Percentage Error
$MAPE$	scalar		Mean Absolute Percentage Error
NS	scalar		Nash-Sutcliffe score
$\mu(\beta_i)$	scalar		mean of absolute regression coefficient $\lvert\beta_i\rvert$ over time

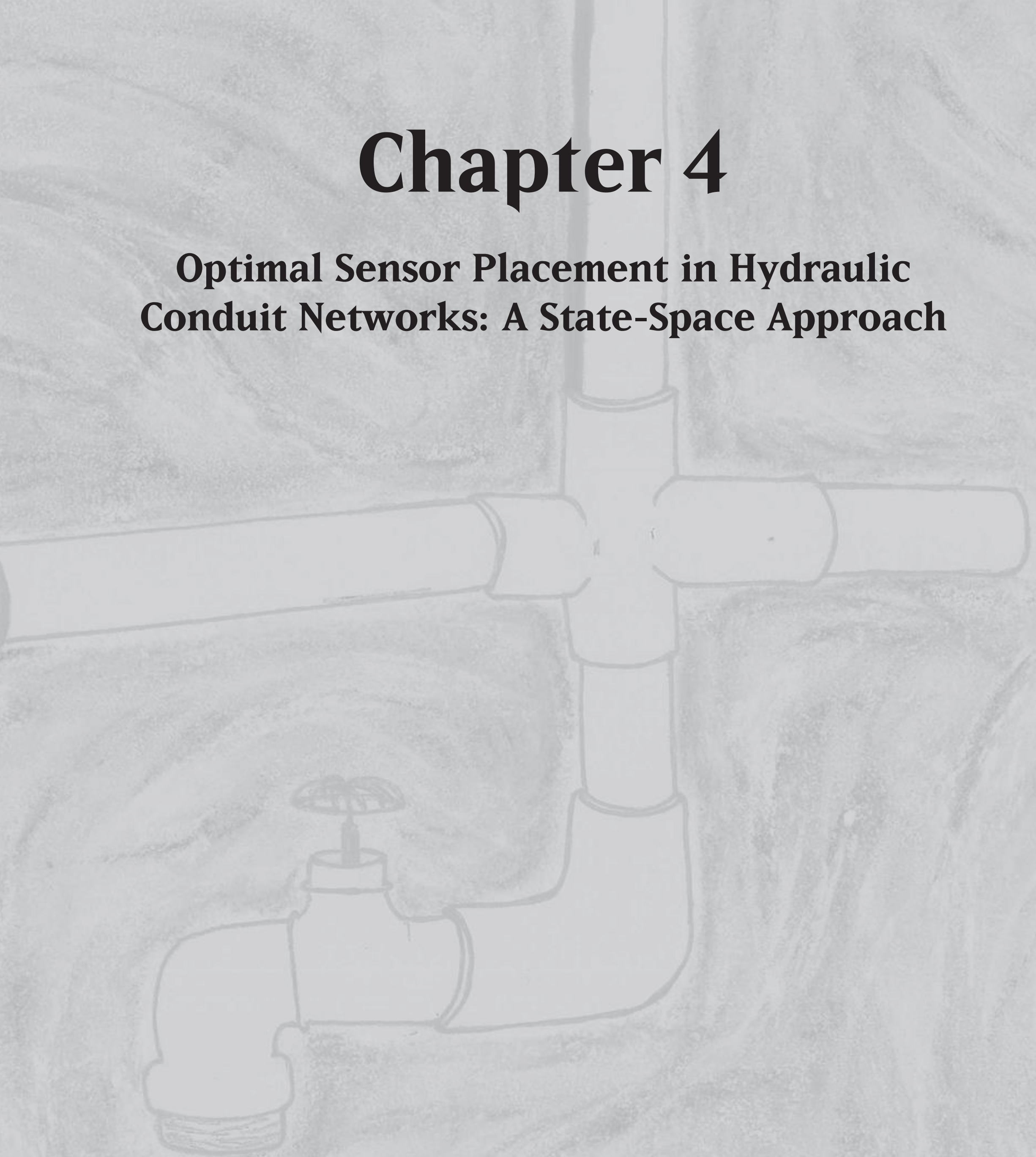
Appendix 3A

The combination of exogenous and univariate methods was also considered. Here we detail the percentage of identified anomalous measurements that significantly deviate from the nowcasted or forecasted water demand over the data set of DMA1. On the diagonal we present the results for the individual methods and in the non-diagonal entries the results for intersection of methods. the non-diagonal entries contain the results of a combination of the methods. By using both an exogenous and a univariate method, a high precision tool is created that only detects a burst when the actual measurements lie outside both the exogenous and univariate 95% prediction uncertainty interval.

NUMBER OF DETECTED ANOMALIES	**EXOGENOUS**	**UNIVARIATE**	**VITENS' DBM**
EXOGENOUS	2140 (2.0%)	824 (0.8%)	851 (0.8%)
UNIVARIATE	-	11215 (10.7%)	6603 (6.3%)
VITENS' DBM	-	-	10191 (9.7%)

Chapter 4

Optimal Sensor Placement in Hydraulic Conduit Networks: A State-Space Approach



A version of this chapter has been published as:
Geelen, C.V.C., Yntema, D.R., Molenaar, J., Keesman, K.J., 2021. Optimal Sensor Placement in Hydraulic Conduit Networks: A State-Space Approach. Water 13(21), 3105.
https://doi.org/10.3390/w13213105

4. Optimal Sensor Placement in Hydraulic Conduit Networks: A State-Space Approach

Abstract

Conduit bursts or leakages present an ongoing problem for hydraulic fluid transport grids, such as oil or water conduit networks. Better monitoring allows for easier identification of burst sites and faster response strategies, but heavily relies on sufficient insight in the network's dynamics obtained from real-time flow and pressure sensor data. This paper presents a linearized state-space model of hydraulic networks suited for optimal sensor placement. Observability Gramians are used to identify the optimal sensor configuration by maximizing the output energy of network states. This approach does not rely on model simulation of hydraulic burst scenarios or on burst sensitivity matrices, but instead determines optimal sensor placement solely from the model structure, taking into account the pressure dynamics and hydraulics of the network. For a good understanding of the method, it is illustrated by two small water distribution networks. The results show that the best sensor locations for these networks can be accurately determined and explained. A third example is added to demonstrate our method to a more realistic case.

4.1. Introduction

Hydraulic models are an essential tool for ensuring safe, reliable, affordable, and continuous delivery of fluids such as water or oil to end-users (Chaudhry 2014). Due to the physical size and complexity of most conduit networks, the actual operational conditions of these grids are hard to monitor (Díaz et al. 2016). Besides serving as a digital twin of the network, model simulations of the real system can be used to predict and forecast in real-time network flows and pressures under varying hydraulic scenarios, valve configurations, or conduit leakages (Sophocleous et al. 2017; Qi et al. 2018; Di Nardo et al. 2019; Conejos Fuertes et al. 2020). In addition, distribution system modelling can help optimize network design, sensor and actuator placement, or facilitate operation of the network through testing of control strategies of pumping and valve configurations (Farley et al. 2010; Sarrate et al. 2012; Bonada et al. 2014; Díaz et al. 2016; Steffelbauer and Fuchs-Hanusch 2016; Sophocleous et al. 2017; Boatwright et al. 2018; Qi et al. 2018; Di Nardo et al. 2019). Models therefore act as an active tool for network management and leakage control, instead of the classical passive approach of solely reacting when a defunct asset is detected.

In order to successfully deploy these models, accurate and up-to-date insight into the network is required in the form of real-time measurements from sensors placed throughout the network. Sensor placement, operation and maintenance is costly, meaning there is a tradeoff between network information gain and sensor costs. A wireless sensor network of as few as possible flow and pressure sensors at key positions within the network is of vital importance for optimal network management. Therefore, optimal sensor placement poses an ongoing challenge in hydraulic conduit networks.

Various studies have been conducted with the goal of maximizing the diagnostic performance of a system under budgetary constraints by means of applying optimal sensor placement. Current methods often consider optimal sensor placement with regards to leak detectability by simulating hydraulic scenarios with leakages (Nagar and Powell 2004; Sarrate et al. 2014; Díaz et al. 2016; Boatwright et al. 2018; Marchi et al. 2018). These studies are mostly based on simulating leakages at various locations within a hydraulic network. The change in each state (pressure/flow), as a consequence of these leakages, is captured in a binarized sensitivity matrix, represented by a Jacobian or forward finite differences matrix. The potential sensor locations are then ranked based on the number of burst locations for which significant sensitivities are found (Pudar and Liggett 1992; Quintiliani et al. 2020). This approach can be expanded in several directions, such as by using artificial bursts achieved by opening fire hydrants in various areas of the network instead of model simulated bursts (Farley et al. 2010), by considering demand uncertainty (Steffelbauer and Fuchs-Hanusch 2016), by not binarizing the sensitivity matrix (Bonada et al. 2014; Cugueró-Escofet et al. 2017), by using optimization techniques instead of iterative techniques to determine the optimal sensor location (Sarrate et al. 2014), or in combination with a leakage localization algorithm (Boatwright et al. 2018). In addition to these studies on hydraulic networks, also observability and sensor placement studies on electrical networks have been performed (Xu and Abur 2004; Qi et al. 2015; Johnson and Moger 2021).

Although practical and efficient, the performance of existing sensor placement techniques based on virtual leakage simulation is highly dependent on the accuracy of the hydraulic model. The estimated sensor placement is very sensitive to uncertainty in demand estimates, model parameters, measurement noise, and asset properties, and has to consider different leakage locations and fluid loss rates, in order to provide accurate sensor placement suggestions (Quiñones-Grueiro et al. 2018). Considering all these factors does not only result in a high dimensional problem and extensive simulations of hydraulic scenarios, but also in high cumulative uncertainties, which exponentially worsens when considering simultaneous placement of multiple sensors. Recent publications on the topic of tackling this high dimensional and highly uncertain problem all suggest to invest more research into these uncertain factors and focus on development of smart optimization strategies to reduce the computational load (Casillas et al. 2013; Steffelbauer and Fuchs-Hanusch 2016; Cugueró-Escofet et al. 2017; Boatwright et al. 2018; Qi et al. 2018). Additionally, an extra source of uncertainty in existing theories is that pressure changes are assumed to take place instantaneously, which especially for larger networks is too rough an assumption (Giustolisi et al. 2008; Boatwright et al. 2018)**.**

The more accurate a hydraulic model is, the better it can be used for e.g. burst localization or optimal sensor placement. However, in order to achieve more accurate hydraulic models, additional sensor placement is necessary. Besides burst detection as a criterion for optimal sensor placement, how well a hydraulic model can be built from a DMA based on sensor placement, could also function as a metric for optimal sensor placement. A possible metric to express how an additional sensor can best contribute to the accuracy of a hydraulic network is observability (Díaz et al. 2016).This concept from systems theory, including some background information is detailed in Appendix II – State-Space Representation and Observability.

The objective of this study is to investigate the observability of conduit networks and optimal sensor placement designs, only considering the structure of the hydraulic network model, and without a dependence on dynamic network simulations. By not relying on simulation of hydraulic scenarios, no computationally expensive high dimensional optimization is required. In this study, a linearized hydraulic network model is presented in state-space form, with as model states the pressures in all network junctions and the flows through all network pipes. Model outputs are the model states corresponding to network junctions and/or conduits where a pressure or flow sensor is installed. Based on a likely hydraulic scenario with corresponding stationary network flows calculated with an EPANET model, the original non-linear hydraulic network model was linearized. This linearization step enables conventional observability analysis (Kalman 1963; Kwakernaak and Sivan 1972) and optimal sensor placement based on maximizing the output energy of observability Gramians (Georges 1995). Current research often focuses on smart optimization to reduce the computational load associated with simultaneous placement of multiple sensors (Casillas et al. 2013; Steffelbauer and Fuchs-Hanusch 2016; Cugueró-Escofet et al. 2017; Boatwright et al. 2018; Qi et al. 2018). This study, however, aims to present an alternative sensor placement framework that starts with state-space modelling. The advantages of using a state-space model for sensor placement in hydraulic conduit networks are demonstrated by two illustrative examples and one more realistic example with corresponding EPANET models (Rossman 2000).

We also show how the suitability of each conduit and junction in the model for sensor placement can be mapped on a graph of the network. This visualization allows for easy identification of optimal regions in the network for sensor placement. This visual information can be used to combine observability function-based optimal sensor placement with other network-specific knowledge regarding sensor placement. The state-space methodology in this paper has been developed using open source software and is equipped with the capacity to transform EPANET model files into state-space models using a Python 3 algorithm.

4.2. Methods

In order to perform optimal sensor placement in an hydraulic network based on systems theory, in this study the network characteristics and dynamics are presented in state-space form. Given the network characteristics, such as conduit lengths, diameters and roughness, flows and pressures within the hydraulic system can be modelled using continuity and momentum equations for unsteady, nonuniform flow of a slightly compressible fluid in slightly elastic conduits. For each conduit, these assumptions lead to the following set of hyperbolic partial differential equations (Watters 1984; Chaudhry 2014):

$$\frac{\partial}{\partial t}\begin{pmatrix} p \\ V \end{pmatrix} + \begin{bmatrix} V & \rho c^2 \\ \rho^{-1} & V \end{bmatrix} \frac{\partial}{\partial x}\begin{pmatrix} p \\ V \end{pmatrix} = \begin{pmatrix} 0 \\ -g\ sin(\theta) - \dfrac{fV|V|}{2D} \end{pmatrix} \tag{4.1}$$

Here, p is the pressure in Pa, V is the flow velocity in $\frac{m}{s}$, ρ is the mass density of the transported fluid in $\frac{kg}{m^3}$, c is the elastic wave velocity in $\frac{m}{s}$, g is the acceleration due to gravity in $\frac{m}{s^2}$, θ is the angle the conduit makes with the horizontal, with the angle taken positive if the conduits slopes upwards in the flow direction, f is the Darcy-Weisbach friction factor (dimensionless), and D is the diameter of the inside of the conduit in m. The distinction between the magnitude of flow velocity $|V|$ and directional flow velocity V is made to allow for flow in both directions through a conduit. For a thorough observability analysis of systems described by hyperbolic partial differential equations, we refer to (Dager and Zuazua 2006).

The slope term $g\sin(\theta)$ is relatively small for most applications and may be neglected. Even if the slope is taken into account, the term will be interpreted as a disturbance and will therefore not influence optimal sensor placement, which solely relies on flow and pressure dynamics as well as sensor configurations.

Also, in many applications, the convective acceleration terms $V(\partial p/\partial x)$ and $V(\partial V/\partial x)$ are small compared to the other terms and may therefore be neglected (Chaudhry 2014).

However, in this study we assume slightly compressible fluid in slightly elastic conduits and thus changes in pressure and flowrate with distance $\frac{\partial}{\partial x}\binom{p}{V}$ are not zero. Furthermore, with $p = \rho g(H + z_0)$ and $V = \frac{Q}{A}$, Eq (1) can be expressed in terms of the piezometric head $H = \frac{p}{\rho g} - z_0$ above a specified level z_0, and volumetric flow rate Q with conduit's cross-sectional area $A = \frac{1}{4}\pi D^2$ (Izquierdo et al. 2004; Chaudhry 2014). In this state transformation both ρ and A are assumed to be constant. The variation of ρ and A is still indirectly taken into account by using a finite elastic wave velocity c. The elastic wave velocity c is a function of various properties of the transported fluid as well as the conduit, but is assumed constant within a pipe, since the changes within a single conduit are assumed small (Ramos et al. 2004). For water transport without air bubbles through PVC conduits, the wave velocity is estimated as $c = 1200\frac{m}{s}$ (Chaudhry 2014).

Equation (4.1) makes use of the empirical Darcy-Weisbach equation to describe friction losses as $\frac{f|V|}{2D}V$, where the Darcy-Weisbach friction factor f is a function of the Reynold's number. When solely considering water transport and assuming constant temperature and viscosity, such as is the case in water distribution networks, the friction losses are not dependent on the Reynold's number according to the empirical Hazen-Williams equation (Chaudhry 2014). Expressed in volumetric flow rate and piezometric head, the Darcy-Weisbach friction related flow loss $\frac{8}{\pi^2}\frac{f|Q|}{D^3}Q$ is replaced by the Hazen-Williams friction related flow loss $\frac{\pi}{4}\frac{10.67g|Q|^{0.852}}{C^{1.852}D^{2.8704}}\,Q$, where C is the conduit-specific dimensionless Hazen-Williams roughness coefficient. Implementing all these assumptions, we can rewrite Eq. (4.1) in the form (Chaudhry 2014):

$$\frac{\partial}{\partial t}\binom{H}{Q} + \begin{bmatrix} 0 & \frac{4}{\pi}\frac{c^2}{gD^2} \\ \frac{\pi}{4}gD^2 & 0 \end{bmatrix}\frac{\partial}{\partial x}\binom{H}{Q} = \begin{pmatrix} 0 \\ -\frac{\pi}{4}\frac{10.67g|Q|^{0.852}}{C^{1.852}D^{2.8704}}\,Q \end{pmatrix} \tag{4.2}$$

We will apply this system of hyperbolic partial differential equation to a conduit network with n_i junctions and n_{ij} conduits, connecting junctions i and j. Assuming $\frac{\partial}{\partial x}\binom{H_{ij}}{Q_{ij}}$ along conduit ij may be approximated by $\binom{\Delta H_{ij}/L}{\Delta Qij/L}$, where L is the length of the conduit in the flow direction, we arrive for the head in junction $i = 1,2,\dots,n_i$ at the following equation (Zhang 2020):

$$\frac{dH_i}{dt} = \sum_{j=1}^{\deg(i)}\left(\frac{4}{\pi}\frac{c^2}{gD_{ij}^2L_{ij}}\left(Q_{ij(i)} - Q_{ij(j)}\right)\right) \tag{4.3}$$

The difference between the flow in a conduit ij at conduit start i and end j is a consequence of the slight compressibility of the fluid and the slight elasticity of the conduit. For fluid transport, the difference between flow at beginning and end of a conduit is usually very small. Commonly, large pressure changes, as a result of high elastic wave velocity c, are assumed immediate. Hence, both the low flow variation within a conduit and the rapid pressure changes motivate the assumption that pressure changes are instantaneous, implying that the head in each junction is always in steady state, thus $\frac{dH_i}{dt} = 0$ and thus $Q_{ij(i)} = Q_{ij(j)} \equiv Q_{ij}$ (Chaudhry 2014). Consequently, under these assumptions the dynamics of the system would be solely governed by the momentum equation (4.2):

$$\frac{dQ_{ij}}{dt} = \frac{\pi}{4}\frac{gD_{ij}^2}{L_{ij}}\left(H_i - H_j\right) - \frac{\pi}{4}\frac{10.67g|Q|_{ij}^{0.852}}{C_{ij}^{1.852}D_{ij}^{2.8704}}Q_{ij} \tag{4.4}$$

Although this equation is very suitable for calculation of hydraulic scenarios and thus for performing optimal sensor placement through the use of burst simulations, significant and measurable pressure transients do occur (Ramos et al. 2004). Since modern sensors can operate under sampling frequencies higher than once per second, the damping oscillations as a consequence of pressure transients and friction in the conduits can be detected. Since large pressure transients, also referred to as water hammers, can cause conduit wear and bursts, it is important to be able to identify where, how often, and to what extent these transients occur, in order to identify their causes and adopt a mitigation strategy. For optimal sensor placement based on observability analysis, we take into account these transients by assuming $\frac{dH_i}{dt} \neq 0$, and assuming a linear relationship between change in flow rate and flow rate throughout each conduit in the flow direction $i \rightarrow j$:

$$Q_{ij(i)} - Q_{ij(j)} = \varepsilon Q_{ij} L \tag{4.5}$$

The relative flow gradient ε in m^{-1} is small, since the compressibility of fluids and elasticity of conduits are very small in most hydraulic conduit networks. As ε is unknown, in this study we assume it is unknown-but-bounded. Thus, ε is defined on an interval that represents the uncertainty in the values of the compressibility and elasticity. For details about the relative flow gradient, see Appendix 4A. This assumption (5) yields a system consisting of a linear continuity and a non-linear momentum equation:

$$\begin{gathered} \frac{dH_i}{dt} = \sum_{j=1}^{\deg(i)} \left(\frac{4}{\pi}\frac{c^2\varepsilon}{gD_{ij}^2}Q_{ij}\right) \\ \frac{dQ_{ij}}{dt} = \frac{\pi}{4}\frac{gD_{ij}^2}{L_{ij}}\left(H_i - H_j\right) - \frac{\pi}{4}\frac{10.67g|Q|_{ij}^{0.852}}{C_{ij}^{1.852}D_{ij}^{2.8704}}Q_{ij} \end{gathered} \tag{4.6}$$

Notice that Eq (4.6) is nonlinear with regards to volumetric flow Q_{ij}. However, for small perturbations from a specific hydraulic scenario, as a result of steady state computations in EPANET with steady state pressures $\bar{H}$ and flows $\bar{Q}$, the model can be linearized around that hydraulic scenario. Thus, we assume $\left|Q_{ij}\right|_{ij}^{0.852} Q_{ij} \approx \left|\bar{Q}_{ij}\right|_{ij}^{0.852} Q_{ij}$. Consequently, the dynamics of the system are then expressed in terms of the 'resistance' $\mathcal{X}_{ij} = \frac{4}{\pi}\frac{c^2\varepsilon}{gD_{ij}^2}$, 'conductance' $\mathcal{Y}_{ij} = \frac{\pi}{4}g\frac{D_{ij}^2}{L_{ij}}$, and 'friction' $\mathcal{Z}_{ij} = -\frac{\pi}{4}\frac{10.67g|\bar{Q}|_{ij}^{0.852}}{C_{ij}^{1.852}D_{ij}^{2.852}}$ constants:

$$\begin{aligned} \frac{dH_i}{dt} &= \sum_{j=1}^{\deg(i)} \left(\mathcal{X}_{ij}Q_{ij}\right) \\ \frac{dQ_{ij}}{dt} &= \mathcal{Y}_{ij}\left(H_i - H_j\right) + \mathcal{Z}_{ij}\, Q_{ij} \end{aligned} \tag{4.7}$$

Simulation of a steady-state hydraulic scenario is thus required in order to obtain estimates for the linearization points $\bar{Q}_{ij}$. Although EPANET hydraulic simulations do not include pressure dynamics, these dynamics are still taken into account in the state-space model Eq. (4.7).

Notice that Eq (3) and thus also (7a) describe the pressure change as a result of gradients in the flow rates. In steady state, for incompressible fluid and non-elastic networks, the continuity equation at each node is given by: $\frac{dH_i}{dt} = \sum_{j=1}^{\deg(i)}\left(Q_{ij}\right)/A_i = 0$. Small deviations in the flow rates through a node, as a result of changing boundary conditions, may lead to small increases or decreases of the pressure in the node, which are also covered by Eq (5). Consequently, as a result of our approximations, for large changes in the hydraulic scenario , new steady states need to be calculated, using e.g. the EPANET model (Rossman, 2000).

To put these equations in matrix-vector form, we introduce the state vector $x := \left[H_1, H_2, \dots, H_{n_i}, Q_1, Q_2, \dots, Q_{n_{ij}}\right]^T \in \mathbb{R}^n$ with $n = n_i + n_{ij}$, where the first n_i elements contain the heads H_i, and the remaining elements the flows Q_{ij}. We further introduce the output vector $y = \left[y_1, y_2, \dots, y_p\right]^T \in \mathbb{R}^p$ with $p = p_i + p_{ij}$, where the first p_i elements contain the heads H_i of those junctions equipped with a pressure sensor and the remaining elements contain the flows Q_{ij} of those conduits equipped with a flow sensor. In what follows, we assume a pressure sensor is always placed in a junction and a flow sensor halfway on a conduit. Eq (4.7) can thus be represented in the following form:

$$\begin{aligned} \frac{d}{dt}x(t) &= \boldsymbol{A}x(t) + \boldsymbol{B}u(t) \\ y(t) &= \boldsymbol{C}x(t) + \boldsymbol{D}u(t) \end{aligned} \tag{4.8}$$

where $u(t)$ contains the boundary conditions and in what follows $\boldsymbol{D} = \boldsymbol{0}$. The dynamics of the system are determined by the system specific parameters $\mathcal{X}_{ij}$, $\mathcal{Y}_{ij}$, and $\mathcal{Z}_{ij}$, which make up the elements of the $n \times n$ matrix $\boldsymbol{A}$. The positions of the sensors are specified via the $p \times n$ matrix $\boldsymbol{C}$. For applications of optimal sensor placement, only the system dynamics (matrix $\boldsymbol{A}$) and the sensor locations (matrix $\boldsymbol{C}$) are required. For model simulations, however, the $n \times m$ input matrix $\boldsymbol{B}$ would also be required and would contain inputs such as height differences between junctions, minor losses (valves, pumps), storage junctions (tanks), and the set values of flow or pressure at water sources and sinks (demand or reservoir junctions), and thus at the boundaries of the system. Thus, for the intended goal of optimal sensor placement based on state-space methodology, matrices $\boldsymbol{B}$ and $\boldsymbol{D}$ do not need to be specified and no temporal discretization of the system, Eqs. (4.7, 4.8), is required.

The output vector y is dependent on the p_i junctions with head sensors and the p_{ij} conduits with flow sensors, resulting in a binary pseudo-diagonal output matrix $\boldsymbol{C}$ with $p = p_i + p_{ij}$, where those elements of $\boldsymbol{C}$ are 1 if a sensor is present at that junction or in that conduit. For any chosen sensor configuration and accompanying output vector y and output matrix $\boldsymbol{C}$, the network is observable if the $pn \times n$ observability matrix $\mathcal{O}$ has full rank:

$$\mathcal{O} = \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{CA} \\ \boldsymbol{CA}^2 \\ \vdots \\ \boldsymbol{CA}^{n-1} \end{bmatrix} \tag{4.9}$$

A system that is observable allows a full reconstruction of the states over time from given input-output data. However, for large networks, the observability matrix $\mathcal{O}$ may be ill-conditioned, which would lead to an observability analysis that does not lead to accurate conclusions (Grubben and Keesman 2018). If the eigenvalues of matrix $\boldsymbol{A}$ all have negative real parts, the system is called (asymptotically) stable. In that case, for each sensor configuration, the $n \times n$ observability Gramian $W_{\mathcal{O}}$ of the network can be calculated, after solving the discrete Lyanupov equation $A^T W_{\mathcal{O}} + W_{\mathcal{O}} A = -C^T C$, and is given by:

$$W_{\mathcal{O}} = \int_0^\infty \left(e^{A^T \tau} C^T C e^{A\tau} \right) d\tau \tag{4.10}$$

If $W_{\mathcal{O}}$, a symmetric matrix and unique solution to the discrete Lyapunov equation, is positive definite, that is, has all eigenvalues larger than zero, then the system defined by $\boldsymbol{A}$ and $\boldsymbol{C}$ is observable.

For linear, time-invariant systems, such as given by Eqs. (4.7, 4,8), the sensitivity of the output y with respect to the initial state $x(0)$ is given by $\boldsymbol{C}e^{\boldsymbol{At}}$ (Grubben and Keesman 2018). Therefore, the observability Gramian $W_{\mathcal{O}}$ from Eq (4.10) can be interpreted as a Fisher Information Matrix, and can thus be understood as a measure of information content. Its inverse, apart from a scaling factor, represents the uncertainty in the estimates of the states (Keesman 2011). In the following, a norm of the observability Gramian $W_{\mathcal{O}}$ will be used as a measure for network observability. In this study, the smallest eigenvalue of $W_{\mathcal{O}}$ is chosen as norm, instead of a summarizing functional based on “optimality criteria” from optimal experiment design (Wald 1943; Pronzato and Pázman 2013). Georges showed that the eigenvalue-optimality criterion can be used to determine which system configuration, defined by y as a result of the choice of matrix $\boldsymbol{C}$, maximizes the observability (Georges 1995). This is achieved by quantifying the information content or “output energy” $\mathcal{E}(y)$

associated with each different sensor configuration, based on the real-valued non-negative eigenvalues λ_{W_O} of the corresponding observability Gramian W_O:

$$\mathcal{E}(y) = \min_{k=1,\ldots,n} \lambda_{W_{O,k}} \tag{4.11}$$

The smallest eigenvalue λ_{W_O} corresponds to a combination of states which are least observable. Choosing a sensor configuration that maximizes this minimum eigenvalue ensures maximum observability of this combination of least observable network states, thereby realizing the most meaningful increase in network observability. The sensor configuration that maximizes the output energy $\mathcal{E}(y)$ is the optimal sensor configuration y_{opt} that maximizes the network's observability:

$$y_{opt} = \underset{y}{\operatorname{argmax}}\big(\mathcal{E}(y)\big) \tag{4.12}$$

Although the exact magnitude of the observability index, in this case the smallest eigenvalue, of a specific sensor configuration y might not be preserved after our approximations, eigenvalue-optimality still allows for comparison of the observability index of different sensor configurations.

4.3. Case Study Results

In order to illustrate the power of the state-space representation of hydraulic conduit networks introduced in Section 2, we will apply the proposed method for optimal sensor placement to two hydraulic models of water distribution networks. Since the aim is to show the essence of the method, we restrict ourselves to small networks. In what follows, we assume a constant relative flow gradient $\varepsilon = 10^{-3}\ m^{-1}$. See Appendix 4BAppendix 4B for an analysis of the effect of the value of ε on output energy and optimal sensor placement.

4.3.1. Example 1: Triangular Network

The small triangular network we study here is sketched in Figure 4.1 (left) and its properties are specified in Table 4.1. The network consists of three junctions $i = 1,2,3$ that are connected in a loop via conduits $ij = 12,23,13$ (Table 4.2). An additional conduit $ij = 41$ connects a reservoir $i = 4$ with constant head $H_4^0 = 243.84m$ to node $j = 1$. The outgoing reservoir flow is assumed to be known, either inferred from measured reservoir volume or directly measured with a flow sensor on conduit $ij = 41$. The eigenvalue decomposition of the state matrix $\boldsymbol{A}$ of the triangular network is detailed in Appendix 4C, showing that the system is asymptotically stable. The question, however, is: where could one extra sensor be best positioned?

Table 4.1 - Network conduit properties of the triangular network

Conduit	Length [m]	Diameter [m]	Roughness [-]	Flow [m^3/s]	$\mathcal{X}_{ij}$ [1/m^2/s]	$\mathcal{Y}_{ij}$ [m^2/s^2]	$\mathcal{Z}_{ij}$ [s^{-1}]
12	1524	0.2032	120	2.50e-2	4.53e3	2.09e-4	-4.85e-2
13	914.4	0.1524	80	1.10e-2	8.05e3	1.96e-4	-1.10e-1
23	243.8	0.3048	200	-1.48e-3	2.01e3	2.93e-3	-5.29e-4
41	304.8	0.3048	100	4.86e-2	2.01e3	2.35e-3	-3.74e-2

Table 4.2 – State matrix A of the triangular network including network topology

	STATE	**CONDUIT**				**JUNCTION**		
STATE		**12**	**13**	**23**	**41**	**1**	**2**	**3**
CONDUIT	**12**	-4.85e-2	0	0	0	2.09e-4	-2.09e-4	0
	13	0	-1.00e-1	0	0	1.96e-4	0	-1.96e-4
	23	0	0	-5.29e-4	0	0	2.93e-3	-2.93e-3
	41	0	0	0	-3.74e-2	-2.35e-3	0	0
JUNCTION	**1**	-4.53e3	-8.05e3	0	2.01e3	0	0	0
	2	4.53e3	0	-2.01e3	0	0	0	0
	3	0	8.05e3	2.01e3	0	0	0	0

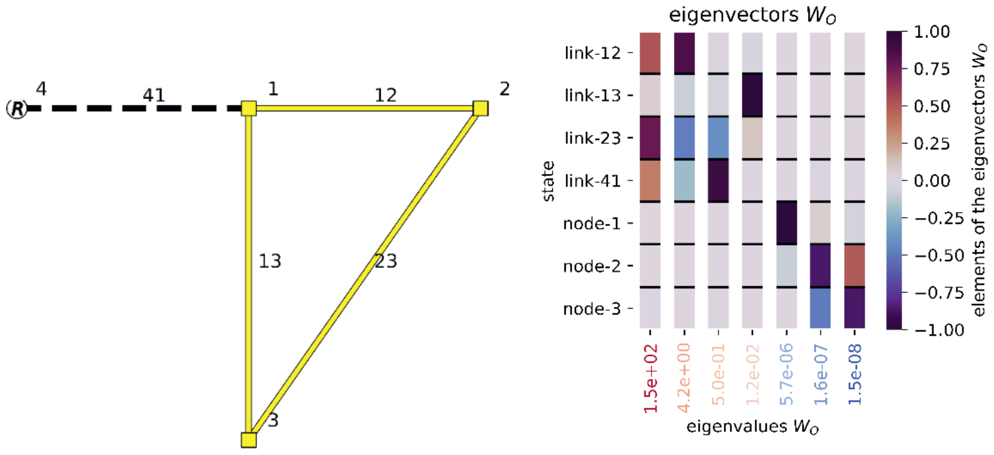


Figure 4.1 - Left: schematic overview of the triangular network, where R is the reservoir and conduit 41 has a flow sensor. Right: Eigenvalue decomposition of the corresponding observability Gramian W_O of the triangular network with only a flow sensor in conduit 41, where each column represents an eigenvector of W_O and each column label at the bottom the corresponding eigenvalue λ_{W_O}.

Since the triangular network is small, the corresponding observability matrix is well-conditioned. Eigenvalue decomposition of the observability Gramian W_O reveals that the three smallest eigenvalues are significantly smaller than the others (Figure 4.1, right). Especially regarding the two smallest eigenvalues, the corresponding weights of the states of node 2 and node 3 are significantly higher than the weights of the other states in the eigenvectors associated with these two smallest eigenvalues. This indicates that the heads in nodes 2 and 3 are significantly less observable compared to the other states and that placing a sensor in either of these nodes will greatly improve the observability of the least observable part of the network. Singular value decomposition of the observability matrix will give the same result, but W_O is less prone to ill-conditioning for large networks, and thus presents a more robust indicator for optimal sensor placement.

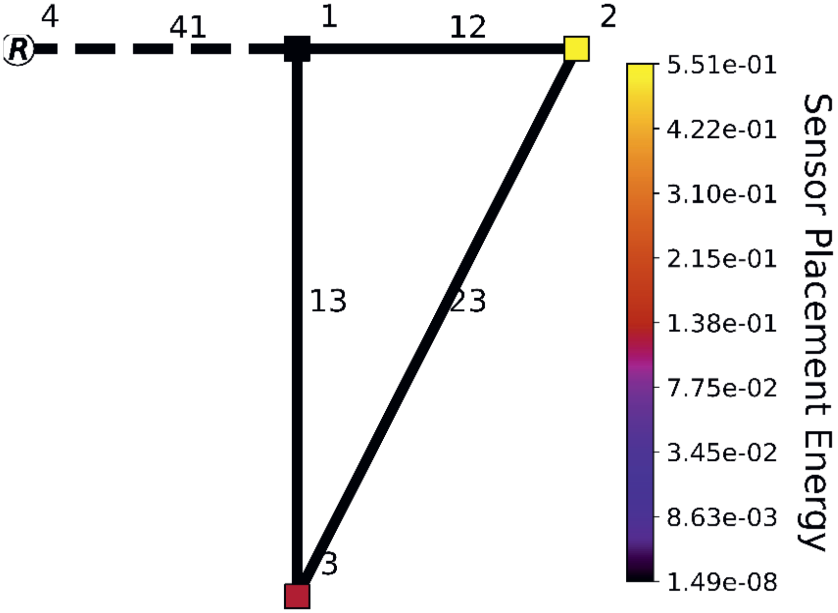


Figure 4.2 - Triangular network, where R represents a Reservoir, squares are junctions, and lines are conduits. Black dashed lines indicate a conduit with a flow sensor. Each additional possible sensor junction and conduit is colored based on the square root of the output energy (Eq. 4.11) corresponding with sensor placement in that specific junction or conduit.

For observability-based sensor placement, six options and thus six different $\boldsymbol{C}$ matrix were considered: a head sensor in one of the nodes or a flow sensor in one of the conduits other than conduit 41, since the flow in conduit 41 is already metered. Notice from Eq. (4.10) that the observability Gramian W_O is defined in terms of an inner product. In order to best visualize the output energy differences between various sensor placements, a square root color scale was used, so as to put emphasis on the comparison between the output energies (Eq. 4.11) of the different sensor placements (Figure 4.2). As can be seen from Figure 4.2, sensor placement in junction 2 maximizes the output energy (smallest eigenvalue), closely followed by junction 3. As shown in Figure 4.1 and discussed above, this is in line with expectations, since nodes 2 and 3 were the states responsible for the smallest eigenvalues of the observability Gramian.

4.3.2. Example2: Net1 case study

In order to perform observability-based sensor placement based on linear state-space models, estimates of the flow $\bar{Q}_{ij}$ through the network are required, using steady state hydraulic simulation of the system. However, these flows can differ significantly between scenarios. Therefore an additional investigation was performed to determine the effect of hydraulic scenarios on resulting optimal sensor location. Optimal placement of one additional sensor for the EPANET chlorine decay model named 'Net1' was also considered (Figure 4.3, left) (Rossman 2000). Net1 is a network with one reservoir, tank, and pump, where the flow from/to the reservoir and the tank are assumed measurable either directly or indirectly from monitoring the reservoir and tank volumes. Depending on the time of day and the tank water volume, reservoir 9 is decoupled from the network and tank 2 will act as a water source instead of a sink (Figure 4.3, right). Scenarios with and without reservoir will result in a different optimal sensor locations. Therefore, placement of one additional sensor in Net1 was investigated with and without reservoir, at 08:00 and 20:00, respectively. Analysis at intermediate time instants, and thus for different supplies and demands with corresponding steady state values of H_i and Q_{ij}, did show different values of the output energy. However, this did not lead to changes in the optimal sensor location.

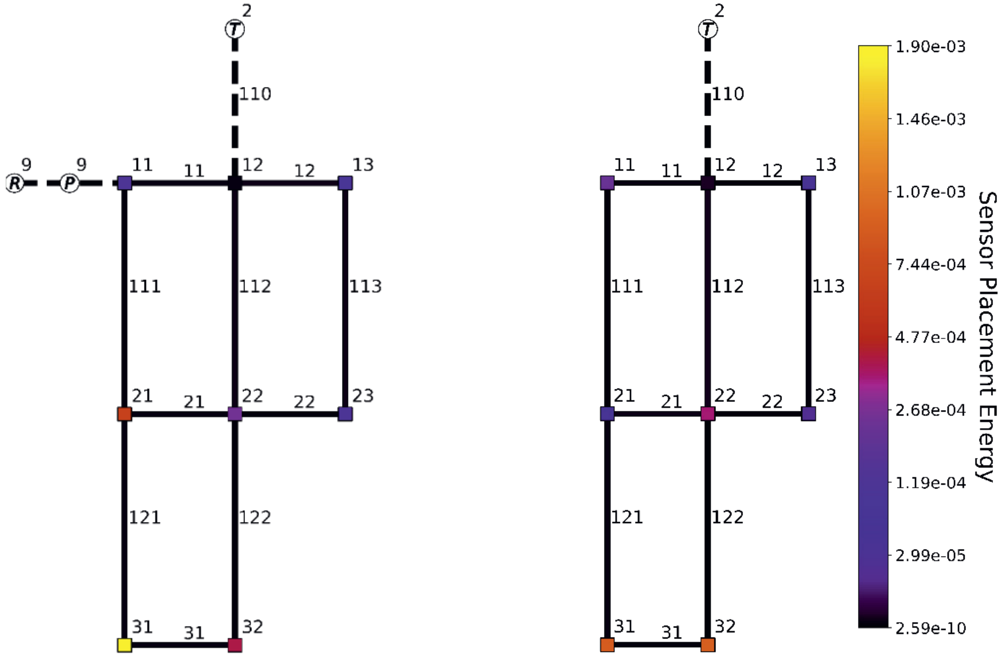


Figure 4.3 – Optimal sensor placement for net1 network with one tank (T), Reservoir (R) and pump (P) at 08:00 (left) and 20:00 (right). Black dashed lines indicate a conduit with a flow sensor. Each additional possible sensor junction and conduit is colored based on the square root of the energy corresponding with sensor placement in that specific state. At 20:00, conduit 9 is closed, thus decoupling reservoir 9 from the network.

Placement of one sensor was considered in addition to the existing flow sensors at conduit 110 and 9 (if the valve on conduit 9 is open). We found for the case at 08:00 that a head sensor in junction 31 is optimal regarding network observability, as seen from the maximum output energy of this sensor configuration compared to alternative sensor placements (Figure 4.3, left and right). Depending on the valve configuration in the network, junction 32 could also be considered for sensor placement (Figure 4.3, right). However, junction 31 is found to be optimal for both network configurations, whereas junction 32 is not significantly more suited for sensor placement compared to 31 and is significantly less suitable for the network configuration where the reservoir is connected to the network (Figure 4.3, left). In both cases the optimal sensor location is at the south end (bottom) of Net1. This is to be expected, since the original Net1 network only contains sensors in the north (top) of the network, so that an additional sensor in the south enables network-wide insight. The fact that our placement procedure leads to optimal positions that are very close to each other, although the flow conditions are rather different, indicates that the linearization step does not significantly impact sensor placement performance. Since the valve configuration of the network for other time instances is similar to the configurations at 08:00 or 20:00, with only slight differences in network pressures and flows, optimal sensor placement for other time instances yields the same optimal sensor placement results.

If pressure changes are assumed instantaneous, and thus $\frac{dH_i}{dt} = 0$, only Eq (4) would remain for analysis of optimal sensor placement. Consequently, only flow sensor placement will be regarded optimal using this approach. In this case, the best choice is to place a flow sensor in the conduit with the largest resistance. Since both pressure and flow dynamics of the hydraulic system are included in the state-space model (Eq 7), factors such as pressure wave velocity will effect sensor placement and thus result in more robust placement and investigation of pressure sensor placement in addition to flow sensor placement. In a practical sense, for full real-time reconstruction of all states, thus including the effect of water hammer, high speed (millisec. – sec.) sampling sensors are needed, which are not commonly used. However, high speed sampling (in the order of milliseconds to seconds) is not considered a challenge nowadays, and the results presented vote for this strategy.

4.3.3. Hanoi Network

The triangular network was used to illustrate the state-space methodology and the Net1 example network shows the robustness of state-space sensor placement with regards to changes in hydraulic scenario used for linearization. However, both networks are small theoretical networks. In order to investigate optimal sensor placement in a real network, the Hanoi network was used as a third case study. The Hanoi (Vietnam) network, is a drinking water distribution network with 34 conduits and 31 demand junctions, and is used as a benchmark for optimal network design application (Fujiwara and Khang 1990; Bi et al. 2015).

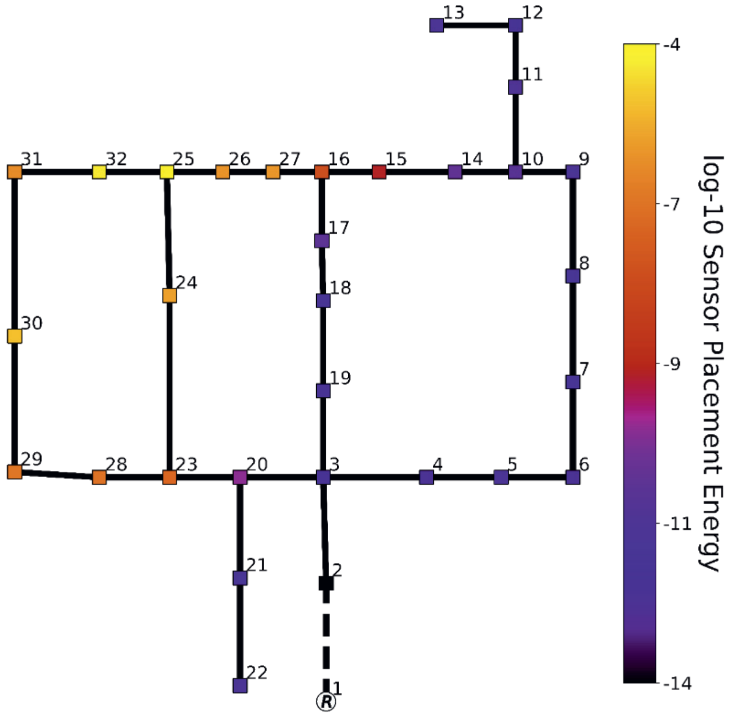


Figure 4.4 - Comparison of possible pressure sensor placements in the Hanoi network. Possible pressure sensor junctions (squares) are colored based on 10-log output energy associated with placement of a sensor in that junction. Those conduits (black lines) attached to a reservoir (R) are already metered (dashed black lines).

Placement of one additional pressure sensor in the Hanoi network was successfully performed using the state-space methodology (Figure 4.4). When assuming the reservoir outflow conduit already contains a flow sensor, placement of a single pressure sensor in junction 25 is deemed optimal for increasing the observability of the network's least observable regions. A sensor on the border of the first and second network loop is therefore deemed optimal. Since the boundary of second and third loop is already metered via the flow sensor at the network reservoir, this configuration allows for metering all three loops as thoroughly as possible when placing just a single pressure sensor. This in turn will result in greatly improving the observability of each region of the network. In practice this means that placement of one additional sensor will greatly benefit reconstruction of all network states (flows and pressures) and therefore will greatly supplement all network-wide methods and models, such as leakage detection algorithms or digital twins based on hydraulic models.

4.4. Conclusions

Optimal sensor placement is a well-studied topic within the literature on smart water grids. However, the focus of these studies often lies on reducing placement uncertainty as well as more computationally efficient optimalization of the involved calculations. The method presented in this study expands the often used burst detectability-centric sensor placement criterion to an observability-based criterion. Our approach ensures that placement of additional sensors will provide more information about the entire network and will help improve hydraulic models or digital twins of the water distribution process. Using a state-space approach that takes flow as well as pressure dynamics into account, and does not rely on dynamic simulations, optimal sensor placement can be performed with limited computational efforts. Results based on three case studies indicate a robust sensor placement performance solely based on network observability. Additionally, the effect of piece-wise linearization of the system, as a result of changing hydraulic scenarios, is shown to not significantly impact sensor placement.

Notation

The following symbols are used in this paper:

p	Pa	pressure
V	ms^{-1}	Flow velocity
ρ	kgm^{-3}	Mass density of transported fluid
c	ms^{-1}	Elastic wave velocity
g	ms^{-2}	Gravitational acceleration
θ	rad	Angle of conduit versus horizontal
f		Darcy-Weisbach friction factor
D	m	Inside diameter of conduit
z	m	Elevation
A	m^2	Cross-sectional area of conduit
H	m	Piezometric head
Q	m^3s^{-1}	Volumetric flowrate
C		Hazen-Williams roughness coefficient
ε	m^{-1}	Flow gradient
$\mathcal{O}$		Observability matrix $pn \times n$
$W_{\mathcal{O}}$		observability Gramian $n \times n$
$\mathcal{E}$		Eigenvalue optimality output energy
$\lambda_{W_{\mathcal{O}}}$		Eigenvalues of the observability Gramian $n \times 1$
x		State vector $n \times 1$
y		Output vector $p \times 1$
y_{opt}		Output vector associated with the optimal sensor configuration $p \times 1$
u		Input vector
$\boldsymbol{A}$		State matrix $n \times n$
$\boldsymbol{B}$		Input matrix
$\boldsymbol{C}$		Output matrix $p \times n$
$\boldsymbol{D}$		Feedthrough matrix
i		Network junction
ij		Network conduit connecting junction i and j
n		Total number of states
n_i		Number of junction head states
n_{ij}		Number of conduit flow states
p		Total number of states whose corresponding asset contains a sensor

Appendix 4A

Let us analyze the approximation of flow differences in a pipe, $\frac{Q_{ij(j)}-Q_{ij(i)}}{L} = \varepsilon Q_{ij(i)}$ with $\varepsilon \in \mathbb{R}$, in some more detail. This approximation is derived as follows. Assume the flow at the end points of a pipe is given by: $Q_{ij(j)} = (1+\varepsilon L)Q_{ij(i)}$. Then,

$$\frac{Q_{ij(j)}-Q_{ij(i)}}{L} = \frac{(1+\varepsilon L)Q_{ij(i)}-Q_{ij(i)}}{L} = \varepsilon Q_{ij(i)} \qquad (4A.1)$$

with ε the relative flow change per meter. Let the unsteady, nonuniform flow of a slightly compressible fluid in slightly elastic conduits, after linearization of the friction term around $\bar{Q}$, be described by the hyperbolic partial differential equation:

$$\frac{\partial}{\partial t}\begin{pmatrix} H \\ Q \end{pmatrix} + \begin{bmatrix} 0 & \bar{\mathcal{X}} \\ \bar{\mathcal{Y}} & 0 \end{bmatrix} \frac{\partial}{\partial x}\begin{pmatrix} H \\ Q \end{pmatrix} = \begin{pmatrix} 0 \\ -\bar{\mathcal{Z}}Q \end{pmatrix} \qquad (4A.2)$$

Where "resistance" $\bar{\mathcal{X}} = \frac{4}{\pi}\frac{c^2}{gD^2}$, "conductance" $\bar{\mathcal{Y}} = \frac{\pi}{4}gD^2$ and "friction loss" $\bar{\mathcal{Z}} = \frac{\pi}{4}\frac{10.67g|\bar{Q}|^{0.852}}{C^{1.852}D^{2.8704}}$ are constants. After spatial discretization and defining the boundary conditions: $H(0,t) \coloneqq H_0$ and $Q(L,t) \coloneqq Q_1$

$$\frac{d}{dt}\begin{pmatrix} H(L,t) \\ Q(0,t) \end{pmatrix} + \begin{bmatrix} 0 & \bar{\mathcal{X}} \\ \bar{\mathcal{Y}} & 0 \end{bmatrix} \begin{pmatrix} \dfrac{H(L,t)-H_0}{L} \\ \dfrac{Q_1-Q(0,t)}{L} \end{pmatrix} = \begin{pmatrix} 0 \\ -\bar{\mathcal{Z}}Q(0,t) \end{pmatrix} \qquad (4A.3)$$

For easy of notation, we define: $H(t) \coloneqq H(L,t)$ and $Q(t) \coloneqq Q(0,t)$. Then,

$$\frac{d}{dt}\begin{pmatrix} H(t) \\ Q(t) \end{pmatrix} = \begin{bmatrix} 0 & \dfrac{\bar{\mathcal{X}}}{L} \\ -\dfrac{\bar{\mathcal{Y}}}{L} & -\bar{\mathcal{Z}} \end{bmatrix} \begin{pmatrix} H \\ Q \end{pmatrix} + \begin{pmatrix} 0 & -\dfrac{\bar{\mathcal{X}}}{L} \\ \dfrac{\bar{\mathcal{Y}}}{L} & 0 \end{pmatrix} \begin{pmatrix} H_0 \\ Q_1 \end{pmatrix} \qquad (4A.4)$$

Using the approximation: $\frac{Q_1-Q(t)}{L} = \varepsilon Q(t)$, gives

$$\frac{d}{dt}\begin{pmatrix} H(t) \\ Q(t) \end{pmatrix} = \begin{bmatrix} 0 & \varepsilon\bar{\mathcal{X}} \\ -\dfrac{\bar{\mathcal{Y}}}{L} & -\bar{\mathcal{Z}} \end{bmatrix} \begin{pmatrix} H \\ Q \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ \dfrac{\bar{\mathcal{Y}}}{L} & 0 \end{pmatrix} \begin{pmatrix} H_0 \\ Q_1 \end{pmatrix} \qquad (4A.5)$$

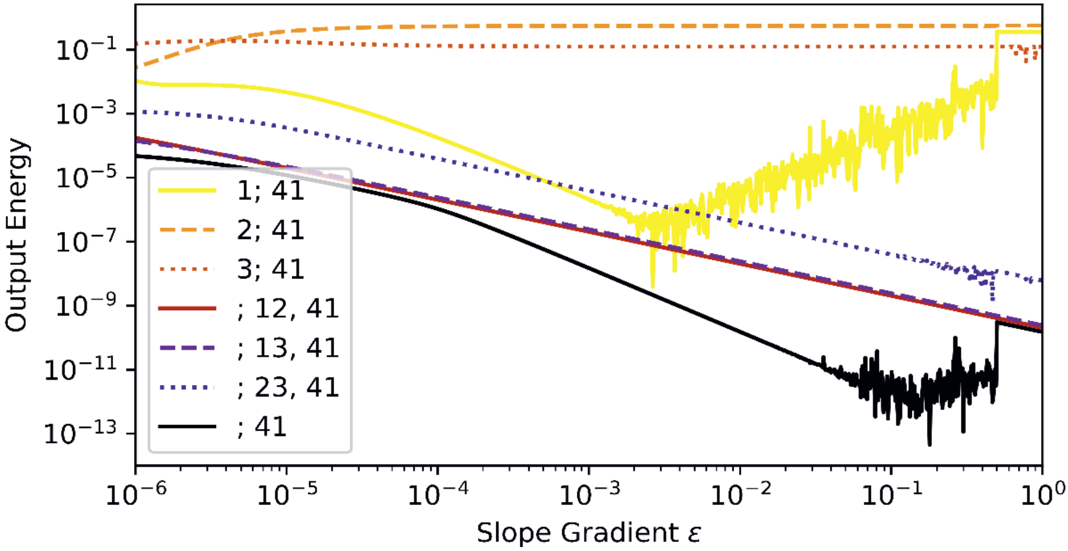
Both (A.4) and (A.5) are a two-dimensional linear time invariant system of the form $\frac{d}{dt}x(t) = \boldsymbol{A}x(t) + \boldsymbol{B}u(t)$. The eigenvalues $\lambda_{\boldsymbol{A}}$ of the system matrix $\boldsymbol{A}$ in (A.5) are given by,

$$\lambda_{\boldsymbol{A}} = -\frac{\bar{\mathcal{Z}}}{2} \pm \frac{1}{2}\sqrt{\frac{L\bar{\mathcal{Z}}^2 - 4\varepsilon\bar{\mathcal{X}}\bar{\mathcal{Y}}}{L}} \qquad (4A.6)$$

Hence, choosing $\varepsilon = 1/L$ will give the eigenvalues or poles of system (A.4). Consequently, for $\varepsilon \geq \frac{L\bar{Z}^2}{4\bar{\mathcal{X}}\bar{\mathcal{Y}}}$, $4\bar{\mathcal{X}}\bar{\mathcal{Y}} \geq \bar{Z}^2 L^2$ and all variables positive, both systems are asymptotically stable and have the same time constant $\frac{1}{|Re(\lambda_{\boldsymbol{A}})|}$. The approximate system becomes unstable for $\varepsilon < 0$.

Appendix 4B

As expected, placing a sensor in the triangular network, in addition to the sensor in conduit 41, will always result in a higher output energy, as an additional sensor will increase system observability, independent of assumed value of the flow gradient ε. For values of ε within the interval $[10^{-6}, 1]$, the sensor configuration with a pressure sensor at node 2 or 3 maximizes the output energy. Therefore, in the case studies we chose $\varepsilon = 10^{-3} m^{-1}$, a good estimate regarding the application of optimal sensor placement.



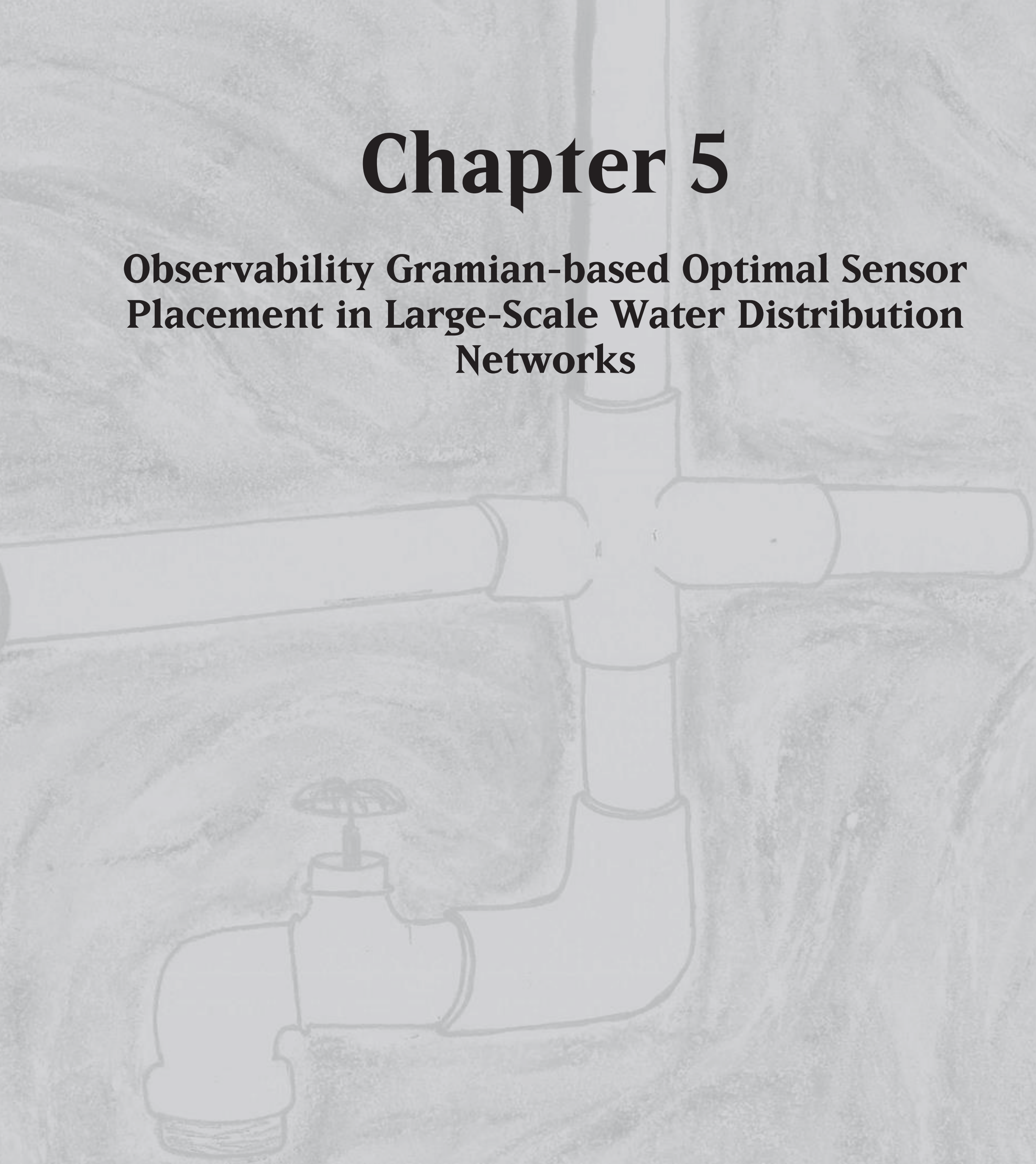
Appendix 4C

The eigenvalues $\lambda_{\boldsymbol{A}}$ and eigenvectors $v_{\boldsymbol{A}}$ (row-wise) of the state matrix $\boldsymbol{A}$ of the triangular network. Notice from the eigenvalues that the system is asymptotically stable and will show oscillatory behavior, as expected. The "fastest" characteristic mode is related to the heads in the three nodes (see 5th row of eigenmatrix).

	node 3	-0.702 - 0.032j	-0.702 + 0.032j	-0.191 - 0.029j	-0.191 + 0.029j	0.813	0.69 + 0.008j	0.69 - 0.008j
	node 2	0.709	0.709	-0.116 - 0.073j	-0.116 + 0.073j	0.23	0.716	0.716
	node 1	-0.018 + 0.046j	-0.018 - 0.046j	0.971	0.971	-0.534	0.102 + 0.012j	0.102 - 0.012j
	link 41	-0.001 - 0.001j	-0.001 + 0.001j	-0.004 + 0.028j	-0.004 - 0.028j	-0.023	-0.019 + 0.002j	-0.019 - 0.002j
	link 23	0 - 0.037j	0 + 0.037j	-0.002 - 0.002j	-0.002 + 0.002j	0.019	-0.003 + 0.001j	-0.003 - 0.001j
	link 13	0.001 - 0.001j	0.001 + 0.001j	0.001 - 0.001j	0.001 + 0.001j	-0.014	-0.001	-0.001
$v_{\boldsymbol{A}}$	**link 12**	-0 + 0.001j	-0 - 0.001j	0.001 - 0.003j	0.001 + 0.003j	0.004	-0.005 + 0.001j	-0.005 - 0.001j
	$\lambda_{\boldsymbol{A}}$	-0.003 + 0.112j	-0.003-0.112j	-0.025 + 0.08j	-0.025-0.08j	-0.091	-0.025 + 0.003j	-0.025-0.003j

Chapter 5

Observability Gramian-based Optimal Sensor Placement in Large-Scale Water Distribution Networks



A version of this chapter has been submitted for publication as:

Geelen, C.V.C., Yntema, D.R., Molenaar, J., Keesman, K.J.
Observability Gramian-based Optimal Sensor Placement in Large-Scale Water Distribution Networks

5. Observability Gramian-based Optimal Sensor Placement in Large-Scale Water Distribution Networks

Abstract

In order to reduce water loss from water distribution networks, measurements from flow and pressure sensors throughout the network are indispensable. In order to maximize the information gain from placement of additional sensors, smart sensor placement is essential. Current optimal sensor placement methodologies often ignore pressure dynamics and rely on simulating virtual bursts in a hydraulic model of the network, in order to determine pressure sensor placement by burst sensitivity maximization. In addition to improved burst detection, additional sensors also provide information for network reconstruction, hydraulic model improvement, and digital twin formation. Here we present a new sensor placement methodology for real-scale networks, based on a linearized state-space representation of the hydraulic network, pressure dynamics, and observability Gramians. The methodology is augmented with a network skeletonization approach, a robust optimality criterion, and a comparison between sequential and simultaneous sensor placement routines. The new methodology was applied to two real-scale networks and shows promising results.

5.1. Introduction

On average, drinking water distribution systems lose 30% of produced water (Liemberger and Wyatt 2019). One of the ways water companies aim to reduce water loss is by using reactive leakage control, which is based on timely detection, localization, and repair of burst assets. Besides relying on customer response to identify bursts, network flow and pressure sensor measurements present a valuable source of information for reactive leakage control. Sensor measurements can be used to detect medium to large burst events by monitoring water demand per District Metering Area (DMA) and tracking pressure transients, and detect smaller leaks based on changes in minimum night flow per DMA (Di Nardo and Di Natale 2012; Lee et al. 2016; Farah and Shahrour 2017; Wu and Liu 2017). These burst detection techniques are most effective on the DMA-level. Although network inflows and DMA borders are well-metered, less information is known about the flow and pressure inside each DMA (Scozzari et al. 2021).

In addition to burst detection, sensor data can be used to construct and validate a hydraulic model of the distribution network and help build a digital twin, which in turn allows for real-time support in design and operation of the water grid (Conejos Fuertes et al. 2020; Scozzari et al. 2021). Especially for operation of a digital twin, information about the flows and pressures inside each DMA is important in addition to already metered DMA boundaries (Wei et al. 2020). Since sensor placement and upkeep are costly, there is a tradeoff between the accuracy and resolution of leakage detection algorithms and hydraulic network models on one side and sensing costs on the other. Optimal sensor placement is essential for maximizing burst detection capacity as well as network insight and observability under budgetary constraints. However, current sensor placement methods focus on DMA formation or burst detectability (Giustolisi et al. 2008; Casillas et al. 2013; Steffelbauer and Fuchs-Hanusch 2016; Cugueró-Escofet et al. 2017; Boatwright et al. 2018; Quiñones-Grueiro et al. 2018).

Common methods for sensor placement often make use of virtual leakage simulations. In the existing approaches one usually calculates the difference between simulated burst and burst-free reference simulations for a number of locations in the network. Those locations showing high pressure sensitivity to the largest range of possible leakage locations are then considered optimal for sensor placement (Pudar and Liggett 1992). Recent work expands on these ideas by taking into account uncertainty in sensor data, model parameters, network properties, demand estimates, and/or leakage size, in addition to the use of smart optimization algorithms (Steffelbauer and Fuchs-Hanusch 2016; Cugueró-Escofet et al. 2017; Boatwright et al. 2018; Qi et al. 2018; Quiñones-Grueiro et al. 2018). However, tackling the combination of all these uncertainties remains a challenge (Casillas et al. 2013; Steffelbauer and Fuchs-Hanusch 2016; Cugueró-Escofet et al. 2017; Boatwright et al. 2018; Qi et al. 2018). Another problem stems from the fact that commonly pressure dynamics are ignored (Giustolisi et al. 2008; Boatwright et al. 2018). This may lead to less accurate optimal sensor placement, since measuring pressure transients has been shown to contribute to accurate and timely burst localization (Srirangarajan et al. 2012; Lee et al. 2016). Leak localization turns out to seriously depend on the accuracy of the hydraulic model, just as is the case for smart placement of sensors (Fuchs-Hanusch and Steffelbauer 2017).

An alternative and novel way of optimal sensor placement is based on the observability of the system (Díaz et al. 2016). Originally observability analysis investigates whether network flows and pressures can or cannot be inferred from measurements from a selected number of sensors (Kalman and Buey 1961; Kwakernaak and Sivan 1972). Nowadays, observability analysis also focusses on the numerical aspects, and therefore observability becomes a measure of how well states can be reconstructed from input-output measurements (Grubben and Keesman 2018). Traditional sensitivity-based optimal sensor placement solely focuses on maximizing burst detectability by placement of additional pressure sensors (Steffelbauer and Fuchs-Hanusch 2016; Cugueró-Escofet et al. 2017; Boatwright et al. 2018; Qi et al. 2018; Quiñones-Grueiro et al. 2018). However, many water companies focus on detecting bursts based on DMA water balances calculated from flow sensors on the boundaries of that DMA, and therefore do not yet make full use of additional pressure sensors within a DMA. In addition to these two, sensor placement based on network observability focuses on maximizing the information about the entire network, allowing for improvement of hydraulic models, burst localization, and digital twin formation, in addition to improved burst detectability (Scozzari et al. 2021).

In this study we propose and investigate an observability Gramian-based sensor placement methodology suited for large, real-scale water distribution networks. The present work starts from the system theoretical framework, with its technical details (Chapter 4), and explores it further in three distinct ways. First, instead of solely taking into account the observability of the least observable part of the network, we now use an optimality criterion that ensures network-wide observability is included. Second, the hydraulic network models of large-scale water distribution networks are reduced in size using network skeletonization. Lastly, sequential as well as simultaneous placement of multiple pressure sensors are investigated.

Two real-scale water distribution network models were used to determine the suitability of the methodology for large networks.

5.2. Methods

5.2.1. State-Space Representation

In order to determine optimal sensor placement, a hydraulic model of the water distribution network is required. The flows and pressures within the hydraulic system, using the continuity and momentum equation for unsteady, nonuniform flow of a slightly compressible fluid in slightly elastic conduits, are described by:

$$\frac{\partial}{\partial t}\begin{pmatrix} H \\ Q \end{pmatrix} + \begin{bmatrix} 0 & \frac{4}{\pi}\frac{c^2}{gD^2} \\ \frac{\pi}{4}gD^2 & 0 \end{bmatrix} \frac{\partial}{\partial x}\begin{pmatrix} H \\ Q \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{\pi}{4}\frac{10.67g|Q|^{0.852}}{C^{1.852}D^{2.8704}}\,Q \end{pmatrix} \tag{5.1}$$

a set of hyperbolic partial differential equations, derived from (Watters 1984, p. 46; Chaudhry 2014). In what follows, we will use a network model in state-space form, derived from Eq. (5.1) for optimal sensor placement (Chapter 4) and based on the following spatially discretized continuity and momentum equations:

$$\frac{dH_i}{dt} = \sum_{j=1}^{\deg(i)} \left(\frac{4}{\pi} \frac{c^2 \varepsilon}{g D_{ij}^2} Q_{ij} \right) \tag{5.2}$$

$$\frac{dQ_{ij}}{dt} = \frac{\pi}{4} \frac{g D_{ij}^2}{L_{ij}} \left(H_i - H_j \right) - \frac{\pi}{4} \frac{10.67 g \left| Q_{ij} \right|^{0.852}}{C_{ij}^{1.852} D_{ij}^{2.8704}} Q_{ij} \tag{5.3}$$

Here, H_i is the piezometric pressure head in m in junction i, Q_{ij} is the volumetric flowrate in $m^3 s^{-1}$ in the conduit connecting junctions i and j, c is the elastic wave velocity in ms^{-1}, g is the acceleration due to gravity 9.81 ms^{-2}, D_{ij} is the diameter of the inside of the conduit ij in m, L_{ij} is the length of conduit ij in m, C_{ij} is the Hazen-Williams roughness coefficient of conduit ij, and ε is the relative flow gradient in m^{-1} (Chapter 4) (Chaudhry 2014). The constants 10.67, 0.852, 1.852, and 2.8704 stem from the empirical Hazen-Williams equation (Chaudhry 2014). The distinction between the magnitude of volumetric flowrate $\left|Q_{ij}\right|$ and the directional volumetric flowrate Q_{ij} is made to allow for a flow in both directions through a conduit. This system of spatially discretized partial differential equations is linearized around a constant magnitude of volumetric flowrate $\left|Q_{ij}\right| = \left|\bar{Q}_{ij}\right|$. We combine the pressure heads of all n_i junctions, denoted as $H_1, H_2, \dots, H_{n_i}$ and the flows of all n_{ij} conduits, denoted as $Q_1, Q_2, \dots, Q_{n_{ij}}$, in the state vector $x := \left[H_1, H_2, \dots, H_{n_i}, Q_1, Q_2, \dots, Q_{n_{ij}}\right]^T \in \mathbb{R}^{n_i + n_{ij}}$. The p states that have a sensor placed in the corresponding asset are put in the output vector $y \in \mathbb{R}^p$. This leads to the following linearized system of spatially discretized partial differential equations in state-space form:

$$\frac{d}{dt} x(t) = \boldsymbol{A} x(t) + \boldsymbol{B} u(t) \tag{5.4}$$

$$y(t) = \boldsymbol{C} x(t) + \boldsymbol{D} u(t) \tag{5.5}$$

The matrix $\boldsymbol{A}$ contains the derivatives of the functions at the right-hand side of Eqns (5.2-5.3) with respect to each of the states, and is also known as the Jacobi matrix. Given a supply-demand pattern, only one steady-state simulation is needed to determine the flow rate $\bar{Q}_{ij}$ in each of the links and thus to define $\boldsymbol{A}$. The matrix $\boldsymbol{C}$, or vector in case of a single sensor placement, only contains zeros and ones. The ones indicate which states (pressure in nodes and/or flows in links) are directly measured by the sensors, and thus these entries are user-defined. In the following, optimal sensor placement is solely based on system observability and thus only on state matrix $\boldsymbol{A}$ and output matrix $\boldsymbol{C}$. Consequently, all this implies that the inputs $u(t)$, input matrix $\boldsymbol{B}$, and feedthrough matrix $\boldsymbol{D}$ are not relevant for sensor placement analysis.

Representing a hydraulic model in state-space form allows for sensor placement based on observability, without requiring dynamic simulations of virtual leaks. The resulting optimization is therefore of lower dimensionality and thus expected to be less computationally complex. Prior work detailed an observability-based method of optimal sensor placement based on a linearized state-space representation of a hydraulic network, using both flow as well as pressure dynamics (Chapter 4). The inclusion of pressure dynamics and independence of hydraulic model simulations reduces the amount of sources of uncertainty and ensures more elaborate dynamics are taken into account, while being computationally lighter due to the lower complexity of this dynamic simulation-free strategy. The ability to measure pressure transients traversing the network has been shown to be of great importance for detecting and preventing bursts, illustrating that the inclusion of pressure dynamics is an important criterion to take into account when performing optimal sensor placement (Kim et al. 2016; Lee et al. 2016).

5.2.2. Output Energy Optimality Criteria

We compared the observabilities of different patterns of sensor locations and accompanying output vectors y, using an observability-dependent output energy function $E(y)$. This output energy is calculated using the eigenvalues λ_{W_O} of the observability Gramian $W_O = \int_0^\infty \left(e^{A^T\tau} C^T C e^{A\tau}\right) d\tau$, with $\boldsymbol{A}$ (stable) and $\boldsymbol{C}$ given in Eqns (5.4-5.5) and $n = n_i + n_{ij}$, since these contain information about the observability of the investigated network. Grubben and Keesman (2018) showed that for linear, time-invariant systems, as in Eqns (5.4-5.5), the sensitivity of the output y with respect to the initial state $x(0)$ is given by Ce^{At}. Consequently, the observability Gramian (W_O) can be interpreted as a Fisher Information Matrix, and thus relates to the uncertainty in the estimates of the states (Keesman, 2011). In other words, The Gramian eigenvalues λ_{W_O} are a measure of how observable the network is given a certain configuration of sensors. A large eigenvalue usually corresponds to a network region that is well observable, whereas a small eigenvalue indicates the existence of a badly observable region with low information content. Different sensor configurations can thus be compared based on the accompanying Gramian eigenvalues λ_{W_O}, which can thus be used to determine the optimal sensor placement.

In Chapter 4 our sensor placement method was demonstrated using three examples of water distribution networks, maximizing the output energy function $E_E(y) = \min\limits_{k=1,\dots,n}\left(\lambda_{W_O,k}\right)$ using the eigenvalue optimality criterion, with $y = \boldsymbol{C}x$ containing the p selected states that are measured as defined by $\boldsymbol{C}_{p\times n}$. The eigenvalue optimality criterion can be used to score various sensor configurations using the Gramian eigenvalues, where each configuration is deemed as strong as the least observable network region. Using this criterion maximizes the observability of the worst observable region. In the present paper we scaled this sensor placement method up for real-scale networks and placement of multiple sensors. For relatively small networks, maximizing the eigenvalue optimality criterion $E_E(y)$ is a suitable criterion, since the network is often well observable in most states, with the exception of one badly observable region. The minimum eigenvalue will thus increase most by placing a sensor in the conduit or junction corresponding to this least observable network state. However, larger networks often have more than one badly observable regions. Then, placement of an additional sensor in one of those regions will not be sufficient to increase observability of the whole network. This suggests that a more general optimality criterion is required.

In experimental design studies a similar issue is often encountered, leading to so-called A-, D-, (modified) E, or T-optimal designs (Pronzato and Pázman 2013). In the following, we will explore the use of some of these criteria for our optimal sensor placement problem. A first possible choice could be the condition number optimality criterion $E_C(y) = \min_{k=1,\dots,n}(\lambda_{W_O,k}) / \max_{k=1,\dots,n}(\lambda_{W_O,k})$, also known as the modified E-criterion, which aims to minimize the difference in observability between most and least observable regions of the network. However, this suffers from the same problem, as this approach only considers most and least observable network regions, while ignoring the observability of all other network states.

Two criteria that do take the observability of all network states into account, are the trace optimality criterion $E_T(y) = \sum_{k=1}^{n}(\lambda_{W_O,k})$ and the determinant optimality criterion $E_D(y) = \prod_{k=1}^{n}(\lambda_{W_O,k})$. Due to the significant differences between smallest and largest eigenvalue of the Gramian observability, trace optimality is not significantly influenced by the smallest eigenvalue and does thus not significantly reflect the observability of the least observable states. Determinant optimality, on the other hand, indeed takes the observability of all states into account, since each state has a significant contribution to the determinant output energy $E_D(y)$. That is the reason that we preferred to use the latter criterion.

A technical detail in our analysis is that, in order to prevent numerical inaccuracy due to computational platform precision limits, the $\log_{10}$ of the output energies is taken. For the determinant-based output energy, this results in $E(y) = \log_{10}\left(\prod_{k=1}^{n}(\lambda_{W_O,k})\right) = \sum_{k=1}^{n}\left(\log_{10}(\lambda_{W_O,k})\right)$.

5.2.3. Skeletonization

Placement of multiple sensors in large networks is computationally intensive, due to the large number of combinations of possible sensor locations. Therefore, skeletonization was used, a novel method for model reduction. The approach consists of two steps. First, we apply skeletonization, and use optimal sensor placement in this reduced network. This step will roughly indicate which network region may mostly benefit from sensor placement. In the second step we apply optimal sensor location to that region of the original non-skeletonized network. Since the sole purpose of the present work is to perform optimal sensor placement, the effect of this skeletonization on simulation of hydraulic scenarios is ignored. Contrary to traditional skeletonization techniques, the present one allows for fusion of demand junctions, thus realizing larger reductions in model size.

Network skeletonization is based on two rules. First, the demand of non-reservoir dead-end junctions is added to the demand of the adjacent junction, after which the dead-end junction and its connected conduit are removed (Figure 5.1, left). Second, starting with the conduit with the smallest "resistance" L_{ij}/D_{ij}^2 (Eq. 5.2), if its two adjacent junctions both have a degree of two, the conduit is fused with the adjacent conduit ij with the smallest resistance, and the resulting disconnected junction is removed from the model (Figure 5.1, right).

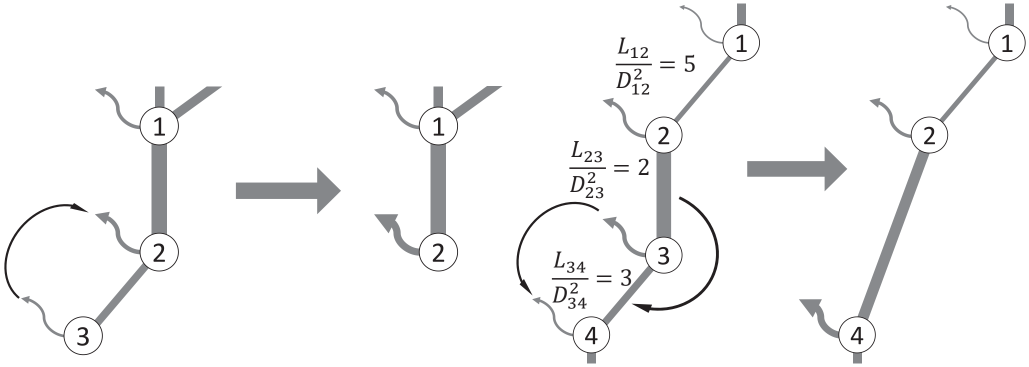

1
2
3
1
2
$\frac{L_{12}}{D_{12}^2} = 5$
$\frac{L_{23}}{D_{23}^2} = 2$
$\frac{L_{34}}{D_{34}^2} = 3$
1
2
3
4
1
2
4

Figure 5.1 – Network skeletonization schematic examples. Left) Dead end node demand (node 3, curved arrow) is fused with preceding node's demand (node 2, curved arrow), after which dead end node and conduit are removed. Right) Starting with the in series conduit with the lowest resistance (2→3), it is fused with the lowest resistance adjacent conduit (3→4). The same is done for fusing demands (curved arrows).

If this removed junction had a demand, it is added to the adjacent junction also connected to the fused conduit. This process is not performed if the junction to be removed is a reservoir. When fusing conduits in this fashion, the new length is the sum of the conduit lengths. The Hazen-Williams roughness coefficient and the diameter are taken as the mean of the constituent conduit properties weighted by their "resistance" L_{ij}/D_{ij}^2.

The first rule removes dead ends from the network, since placement of a flow or pressure sensor in a dead end junction is unlikely, due to a low information density in network extremities. The second rule is used to combine conduits in series. The rules for determining the properties of the combined conduits are chosen so as to be weighted in favor of those conduits with a high resistance L_{ij}/D_{ij}^2, in order to ensure the combined resistance is not underestimated. Skeletonization is started with small resistance conduits, in order to not underestimate the contribution of high resistance assets, which are fused last if at all. In practice, this means wider conduits are often fused with narrow conduits. The skeletonized network will then be subjected to optimal sensor placement, in both a sequential and simultaneous sensor placement approach. The effects of this skeletonization approach will be also discussed in the next section.

5.3. Results and Discussion

5.3.1. Hanoi Network

In order to validate the performance of the skeletonization, optimal sensor placement in the Hanoi network was compared with placement in the skeletonized Hanoi network, assuming a relative flow gradient $\varepsilon = 10^{-3} m^{-1}$ (Figure 5.2). See appendix A for an analysis of the effect of the choice of flow gradient ε on output energy and optimal sensor placement. The Hanoi (Vietnam) network, is a drinking water distribution network with 34 conduits and 31 demand junctions, and is used as a benchmark for optimal network design application (Fujiwara and Khang 1990; Bi et al. 2015). Since this network has a relatively low number of assets, optimal sensor placement can also be performed for the original network without significant computations demands (Figure 5.2).

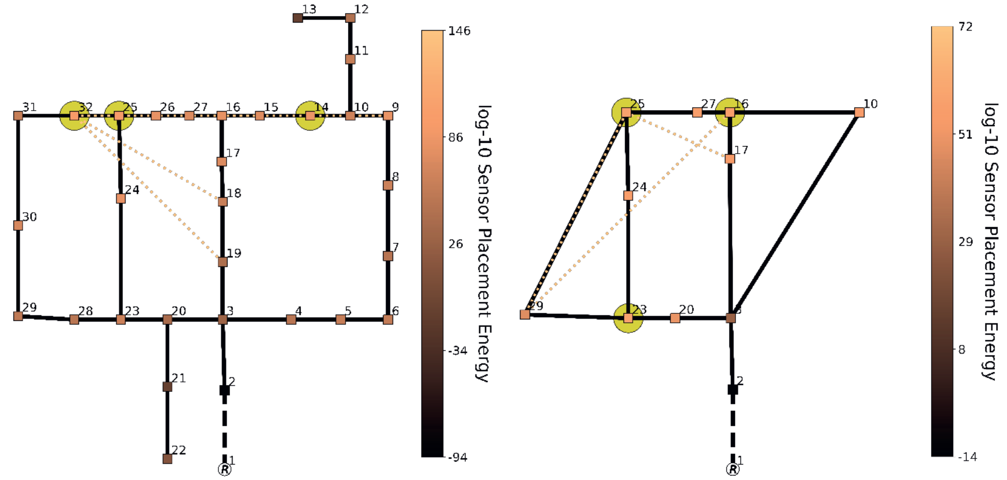


Figure 5.2 - Optimal sensor placement of a single sensor in the Hanoi network (left) and skeletonized Hanoi network (right). Junctions (squares) are colored based on the log_{10} determinant output energy of sensor placement at that asset, where the three assets with the highest output energies are highlighted. Black dashed line represents metered connection from reservoir (R) and colored dotted lines indicate the three sensor combinations for simultaneous placement of two sensors and are colored based on the log_{10} determinant output energy corresponding to these sensor combinations.

Looking at placement of a single sensor in the original Hanoi network, we expect that sensor placement in a network extremity, like junctions 13 and 22, will not greatly contribute to observability. This observation supports the skeletonization assumption of removing dead-end junctions and conduits (Figure 5.2). Additionally, placing sensors in close proximity to each other will hardly contribute to improvement of observability. This can be seen looking at the low output energy of sensor placement in junction 2 in addition to the already metered reservoir outflow conduit 1 (Figure 5.2). The skeletonized network confirms that sensor placement close to existing sensors is not advisable.

Regarding optimal sensor placement, results indicate that a sensor in the North-West and North-East of the network will be favorable (Figure 5.2). Since the reservoir outflow in the South of the network is already metered, sensor placement in the North of the network is in line with expectation. Looking at the skeletonized network, optimal placement in junctions 25 and 16 correspond with the North-West and North-East regions identified in the original network. Optimal placement in junction 23 in the skeletonized network at first sight does not seem to match a region identified in the original network. However, both optimal placements, 32 in original network and 23 in skeletonized network, belong to the same branch. For both the original network and the skeletonized network, determining optimal sensor placement was successful, leading to similar conclusions However, in general there is a tradeoff between the level of skeletonization and optimal sensor placement accuracy.

In order to further investigate effects of skeletonization as well as to determine optimal sensor positions in the skeletonized Hanoi network, simultaneous placement of two pressure sensors was also investigated (Figure 5.2). Pressure sensor placement in junctions 2 or 3 close to already existing sensors is hardly useful, due to the low added value of additional sensor close to existing sensors. Additionally, network extremities 22, 21, 13 and 12 are also poor choices for sensor placement. This effect is also seen in the skeletonized network, which thus manages to capture the same behavior, while greatly simplifying the model and computational demand.

Looking at simultaneous placement of two sensors in the Hanoi network, it is clear that placing sensors in the North-West of the network is preferable. The South-East section of the Hanoi network already contains a flow sensor, thus no additional sensors are required in this area. Our analysis indicates that sensors in both the leftmost and middle conduit loop of the network will have maximal effect.

Placing a first sensor in the North-West of the original Hanoi network at junction 32 and a second sensor in the East at junctions 9, 18, or 19, are the three best sensor combinations. Results of sensor placement in the skeletonized Hanoi network show a similar result, where the North-West placement corresponds to skeletonized junctions 25 and 29, and the East placement corresponds to skeletonized junctions 16 and 17. The indicated optimal combination of skeletonized junctions 25 and 29, both in the West of the network, does differ from the optimal placements indicated for the original Hanoi network. Due to the assumptions of skeletonization, minor discrepancies are expected, as there is a tradeoff between computational demand and sensor placement accuracy. Skeletonization can be used to indicate which placement combinations are optimal, after which a more detailed optimal sensor placement analysis can be ran based solely on those junctions in the original network corresponding to the indicated optimal skeletonized junctions.

We would like to emphasize that in practice there is often only a small difference between the sensor placement combinations that are close to optimal. The reason is that there is a region in which an extra sensor can be added at several nodes with nearly the same effect.

5.3.2. Balerma Network

In a second example, we studied optimal placement of one, two, or three sensors in a real-size and complex irrigation water distribution network based on the existing network in the Sol-Poniente irrigation district in Balerma in the province of Almeria (Spain), assuming a relative flow gradient $\varepsilon = 10^{-3}m^{-1}$ (Reca and Martinez 2006; Bi et al. 2015). The network consists of 43 water demand junctions (hydrants) fed from four different reservoirs via 454 conduits. The outflows of the four reservoirs are assumed to be already metered.

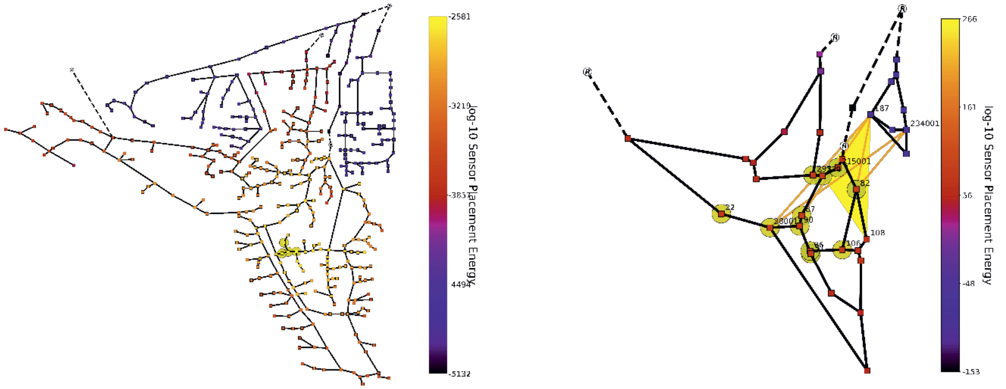


Figure 5.3 – Optimal sensor placement of a single sensor in the Balerma network (left) and skeletonized Balerma network (right). All junctions (squares), the five optimal sensor combinations (orange lines), and the optimal simultaneous placement of three sensors (yellow triangle), are colored based on the log_{10} Determinant output energy of sensor placement at that asset. The 5% best single sensor placements are encircled to highlight them. Black lines are conduits and black dashed lines are the reservoir (R) outflows assumed to be already metered.

The results of placing one additional pressure sensor in the original and skeletonized Balerma network are shown in Figure 5.3. Similar to the optimal sensor placement in the original and skeletonized Hanoi network, dead ends are not useful for sensor placement. Sensor placement in the network loops in order to accurately determine flow through each network loop is the preferred option. This also holds for the Balerma network, where optimal sensor placement is preferred in the central loops of the network instead of in the South of the network, which is furthest away from the reservoir outflow sensors in the North of the network. The North-East section of the Balerma network is connected to the rest of the network via a reservoir. Regarding optimal sensor placement, this essentially means that this North-Eastern section can be considered a separate network DMA. Due to its small size and its less mazed structure, sensor placement is preferred in the larger and more mazed central network. The skeletonized network confirms these results, although the 5% highest output energies are more spread out over a larger region of the network. The skeletonization method greatly reduces computational demand at the cost of placement accuracy.

From our calculations we found that simultaneous placement of two sensors led to placing one sensor in each of these two sub-DMA's of the skeletonized Balerma network. Optimal placement of the first sensor coincides with the optimal locations for single sensor placement in the South-Western DMA (junctions 229, 82, 28001, or 87, 91). The second sensor should then preferably be placed in the smaller North-Eastern DMA (junction 187 or 234001), to ensure that the observability of both DMA's improves.

When trying to optimally place three sensors in the skeletonized Balerma network, one has to cope with 6545 possible combinations. We found that the top 5% best combinations of sensor placement combinations included 193 possible options. The optimal combination is shown in Figure 5.3. Similar to placement of two sensors, there is a clear preference for a single sensor in the North-East section of the network, at junction 187 or 234001. Junctions 187 and 234001 lie at the intersection of two conduit loops, and are therefore preferred over other sensor locations closer to reservoir sensors or only part of a single loop. The other two sensors are optimally placed in the South-West section of the network, to ensure that both North-East and South-West sections of the network are metered. Due to the larger size of the South-West section, placement of two sensors in this network regions is expected to be advantageous. A second sensor is thus ideally placed in either junction 229 or 28001, which is in line with expectations, since both are located in the center of the South-West DMA. However, these optimal sensor combinations cover a wide range of possible third sensor locations. It turns out that one additional sensor in the South-West subnetwork already greatly increases observability, resulting in only a limited increase in observability from placement of a second sensor in the South-West network section. We found that a variety of possible sensor locations resulted in high output energies. This is confirmed by a limited increase in output energy compared to placement of two sensors. Placement of sensors in junctions 87, 229, and 108 is optimal and leads to a $\log_{10}$ determinant output energy of 266. This value is significantly higher than the values found for the nine next best sensor combinations, which score between 260.7 and 262.1.

5.3.3. Choice of Optimality Criteria

Placement of more than two sensors, either simultaneous or sequential, and the choice for the determinant optimality criterion were evaluated for the skeletonized Hanoi and Balerma networks (Figure 5.4).

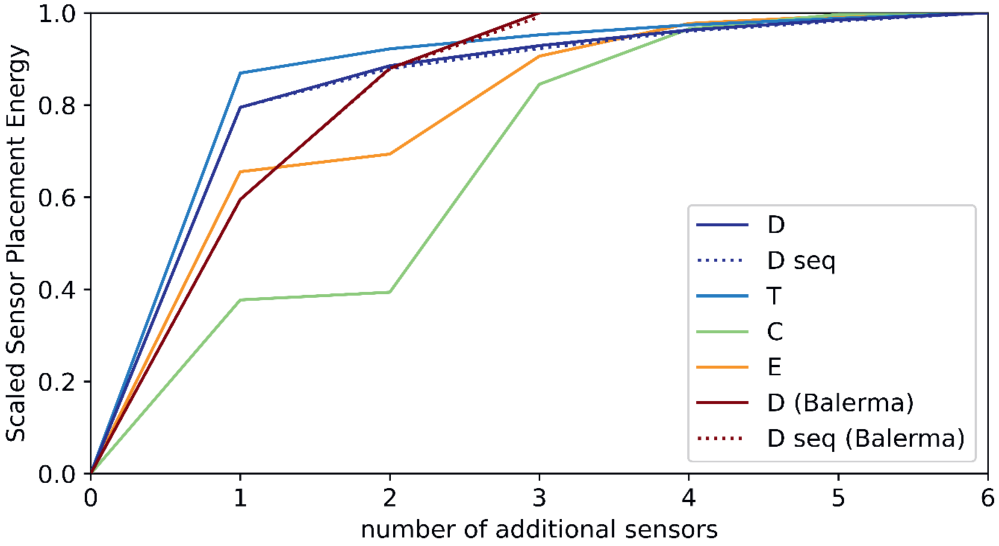


Figure 5.4 – Maximum scaled log_{10} output energy of four different optimality criteria, based on simultaneous and sequential (seq) placement of one to six additional sensors in the skeletonized Hanoi network and skeletonized Balerma network. Note that these sensors are added in addition to the sensors at reservoir outflows. The optimality criteria investigated are: Determinant (D), Trace (T), Condition number (C), and Eigenvalue (E). Each criterion's output energies are scaled by setting the output energy associated with the sensor configuration before placement of additional sensors at zero and setting the maximum output energy of simultaneous placement of six (Hanoi) or three (Balerma) sensors.

The maximum output energy under the eigenvalue optimality criterion ($E_E(y)$) significantly increases with placement of additional sensors, up to placement of three additional sensors. Since eigenvalue output energy reflects the observability in the least observable part of the network, all the badly observable regions of the skeletonized Hanoi network have been covered by sensors with optimal placement of three sensors in addition to the reservoir outflow sensor at conduit 1. Since the network boasts three loops, indeed three sensors are required to ensure that flows and pressures in each loop can be determined accurately. This explains the limited increase in eigenvalue and condition number output energy for placing four or more additional sensors. For smaller networks, where a sensor in each loop might be feasible, eigenvalue and condition number optimality might be most suitable. However, for mazed real-scale networks, these criteria focus only on the least observable regions, and likely neglect impact of sensor placement on the entire network.

The maximum output energy of the determinant and trace criteria show the most significant increase in maximum output energy for placement of one additional sensor. The observability of the best observable regions will be initially higher and will not significantly increase since they were already well observable. Because these criteria also take into account the best observable region(s) of the network, the maximum output energy of these criteria thus only significantly increase, until the whole network is well observable. Placement of one additional sensor thus no longer significantly increases network-wide observability.

The maximum energy using the trace criterion does not show significant relative increases after placement of more than one sensor. Using the determinant criterion, a more significant relative increase in maximum output energy is observed for placement of more than one additional sensor. Contrary to the trace criterion, the determinant criterion is more robust regarding inclusion of least observable network regions in the output energy calculation. This ensures a more robust sensor placement, and thus a relatively larger energy increase when placing more than one additional sensor. The condition number and eigenvalue criteria also show relatively strong increases in maximum output energy for placement of more sensors. However, as discussed in Section 5.2, a higher output energy of these criteria does not necessarily correspond with the most robust sensor placement regarding the goal of increasing network-wide observability, especially in more mazed and larger networks. Therefore we underline the choice for the determinant criterion for robust placement of multiple sensors in larger networks.

Also, simultaneous placement of multiple sensors was compared with sequential placement of multiple sensors under the determinant criterion (Figure 5.4). Sequential sensor placement results in a lower maximum output energy scores compared to simultaneous placement. However, for both the skeletonized Hanoi and Balerma networks, this effect is relatively small. For both networks, sequential placement of multiple sensors instead of simultaneous placement does not result in significantly worse output energies. For larger and more mazed networks, sequential sensor placement might result in significantly worse output energies compared to simultaneous placement. However, this can be easily tested for the case with two sensors, to have a first indication of the effects of simultaneous or sequential sensor placement.

5.4. Conclusions

A linearized state-space method for optimal sensor placement in hydraulic conduit networks was adapted to application of large networks and for simultaneous placement of multiple sensors, without requiring dynamic simulations of hydraulic scenarios. These adaptations consisted of a a) more robust optimality criteria based on Gramian observability of the network for sensor placement, b) skeletonization method in order to reduce network size and thus computational demand of optimal sensor placement, and c) an investigation of simultaneous and sequential placement of multiple sensors and visualization of these results. The skeletonization method proved to be effective in reducing the size of the network. Results indicate that one extra sensor, placed in a number of neighboring junctions, leads to very similar observability. Due to this freedom in sensor placement, further network skeletonization can be applied with acceptable loss in placement accuracy.

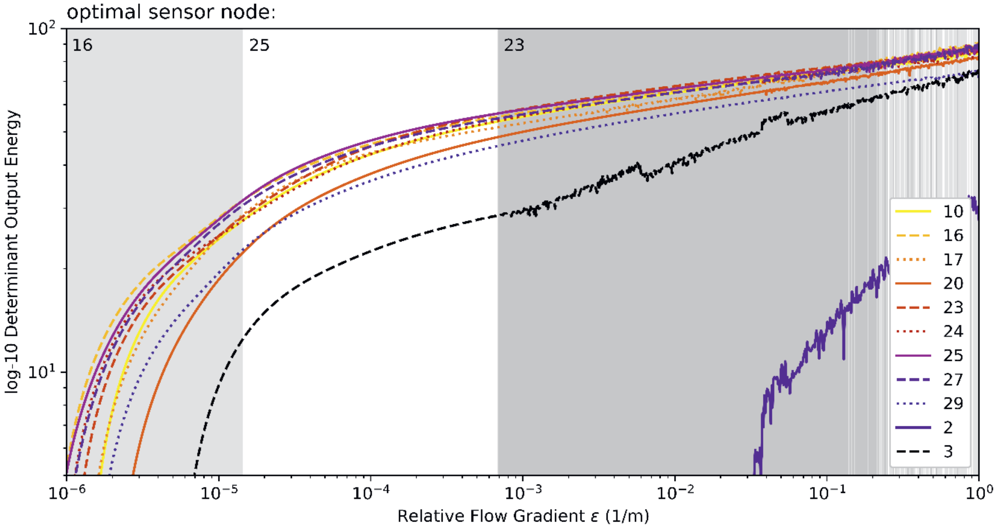
For smaller networks, where sensor placement in each least observable region is possible, the analysis will benefit from eigenvalue optimality, due to the emphasis put on increasing observability of these least observable regions. For larger, more mazed networks, sensor placement in all weakly observable regions is less feasible. Placement of a sensor in just a few of these regions will not necessarily increase network-wide insight, and is therefore hardly useful. The determinant criterion, which focuses on increasing overall network observability, is more robust and suitable for these networks.

Notation

The following symbols are used in this paper:

H	m	Piezometric head
Q	m^3s^{-1}	Volumetric flowrate
i		Network junction
ij		Network conduit connecting junction i and j
n		Total number of states
n_i		Number of junction head states
n_{ij}		Number of conduit flow states
p		Total number of states whose corresponding asset contains a sensor
c	ms^{-1}	Elastic wave velocity
g	ms^{-2}	Gravitational acceleration
D	m	Inside diameter of conduit
C		Hazen-Williams roughness coefficient
ε	m^{-1}	Relative flow gradient
$\mathcal{O}$		Observability matrix $pn \times n$
$W_{\mathcal{O}}$		observability Gramian $n \times n$
$E_k(y)$		Output energy function based on an optimality criterion k
$\lambda_{W_{\mathcal{O}}}$		Eigenvalues of the observability Gramian $n \times 1$
x		State vector $n \times 1$
y		Output vector $p \times 1$
y_{opt}		Output vector associated with the optimal sensor configuration $p \times 1$
u		Input vector
$\boldsymbol{A}$		State matrix $n \times n$
$\boldsymbol{B}$		Input matrix
$\boldsymbol{C}$		Output matrix $p \times n$
$\boldsymbol{D}$		Feedthrough matrix

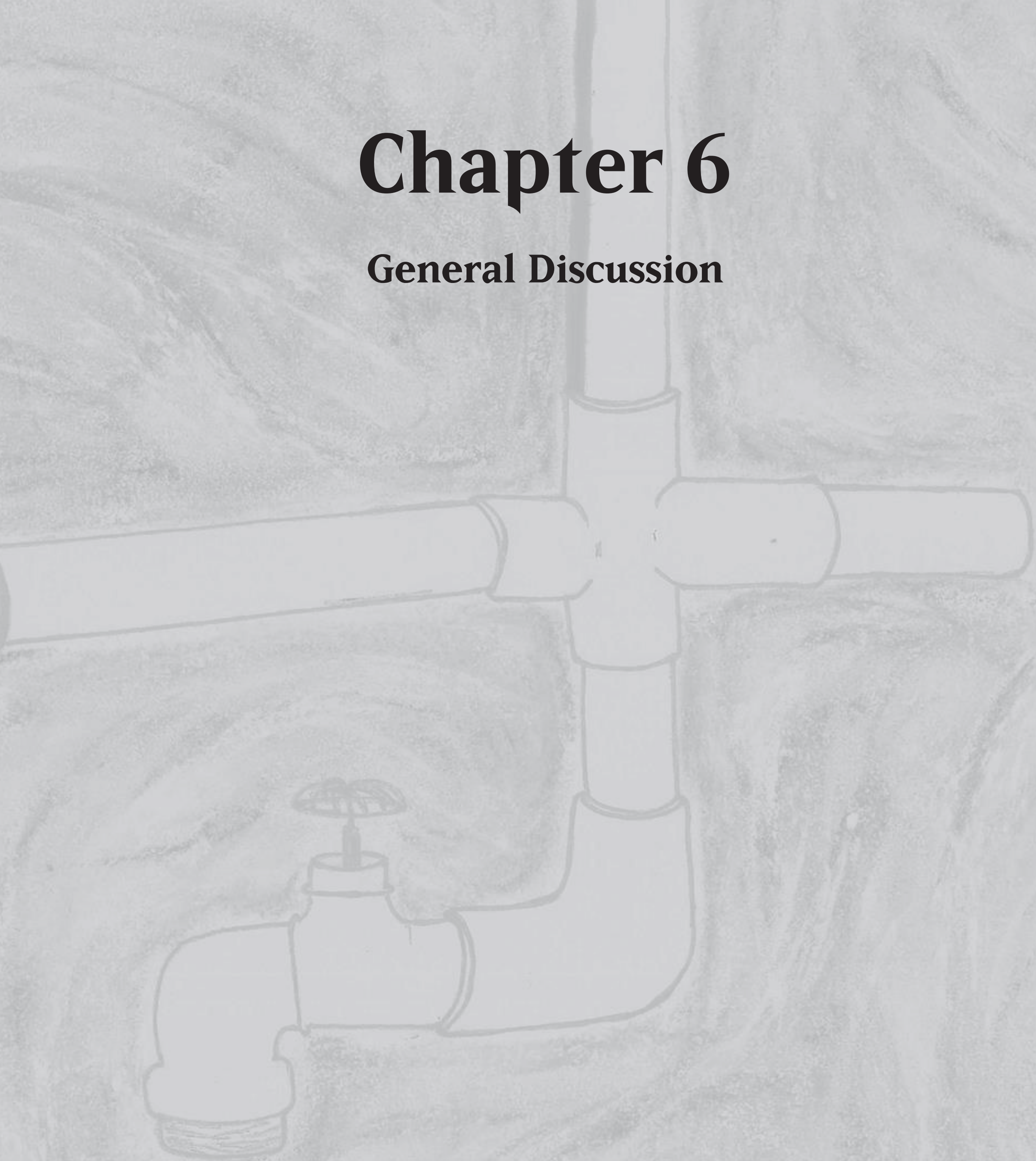
Appendix 5A



The effect of the choice of relative flow gradient ε on the output energy of pressure sensor placement in any of the nodes from the skeletonized Hanoi network was also evaluated, where background color indicates which node is the optimal sensor location. As can be seen in the graph and as is indicated by the background color, the sensor configurations with a pressure sensor at nodes 16, 23, or 25 maximizes the output energy, without significant differences between which of these nodes is chosen, independent of which value of the relative flow gradient ε is used from the interval $[10^{-6}, 1]$. Therefore, in the here discussed case study networks, we chose $\varepsilon = 10^{-3}m^{-1}$.

Chapter 6

General Discussion



6. General Discussion

6.1. Monitoring Support for Water Distribution Systems Based on Pressure Sensor Data

An important factor to consider when developing a burst detection algorithm, is the false positive rate (Wu and Liu 2017). Knowing that a significant fraction of burst alarms is false results in distrust of the detection algorithm, reduces response intensity, and leads to skepticism. Non-burst anomalies in real-time flow or pressure sensor signals should thus be suppressed and not trigger burst alarms, since no immediate emergency response is required. However, these anomalies could strain the network or indicate causes that are potentially harmful for the network. Each significant deviation from intended pressure and flow is an interesting anomaly and a potentially problematic one. Therefore the non-burst anomalies, and especially the recurring abnormalities, should be detected, tracked, and investigated. Traditional methods for burst detection often do not detect these anomalies, due to a focus on burst detection or insufficient labelled data to train supervised models (Li et al. 2019). Current research does investigate the capacity of anomaly detection in various forms of water distribution time series data (Li et al. 2019; Apostol et al. 2021; Deng and Hooi 2021). However, no further suggestions for modes of action are given to remedy the underlying cause of these potentially harmful anomalies. As a first step towards remedying the causes of these anomalies, identifying those anomalies that occur on a regular basis will point towards the most prevalent anomalous patterns, which in turn helps facilitate identification of the underlying causes. Timely response to recurring anomalous pressure or flow patterns will mitigate continuous network strain, identify faulty assets, detect cyber-physical attacks, and optimize and streamline network operation.

The method described in Chapter 2 has the added benefit of identifying and tracking these recurring anomalous patterns, and will thus serve as an early warning and decision support system for optimal network management, operation and proactive leakage control. Better monitoring of the water distribution process can thus help ensure safe and reliable drinking water, without relying on customers as surrogate sensors used for reactive leakage control. Additionally, stringent anomaly monitoring will help prevent bursts and thus lowers costs of water distribution.

6.2. Classification of Daily Patterns in Sensor Measurements for Improved Insight into the behavior of Water Distribution Systems

Low quality of training data in leakage registration does pose a problem for the implementation of smart and supervised burst detection methodologies (Castro-Gama and Agudelo-Vera 2019; Scozzari et al. 2021). Due to, amongst others, missing labels, mislabeling of bursts, and inductive bias, leakage databases suffer from a high fraction of incorrect burst labels. Due to the significantly lower fraction of burst events compared to non-burst events, training supervised models is highly sensitive to incorrect burst labels (Nyambura 2020).

Important information can be gained from looking at data anomalies, but also from the regularity/seasonality contained within the data. Although it is not always possible to directly point out a specific seasonal attribute for an observed phenomenon, the more time series are investigated, the more it becomes clear that there are various levels of regularity and seasonality at play, be it hourly, weekly, monthly, yearly, or on another time scale. Various visualization techniques are capable of bringing these patterns to the foreground (Billings and Jones 2008; Chena and Boccelli 2014; Arandia et al. 2016; Zubaidi et al. 2018). If one is able to determine whether time series data may be considered as regular, it is also possible to recognize non-regular data possibly stemming from bursts or other interesting disruptions of the water distribution process. Due to the power and accessibility of supervised machine learning algorithms, this regularity does not have to be explicitly formulated, as algorithms will be able to discern this from training data (Brentan et al. 2017; Navarrete-López et al. 2019). Due to interest from and easy cooperation with a large Dutch drinking water company in the central-East of the Netherlands, high resolution data from flow and pressure sensors, sampled every few seconds, as well as leakage databases were available for the present research project. Since supervised labels, in the form of a leakage database, were available, training of supervised methods was in principle possible. However, the attempt to develop such a supervised method was ultimately unsuccessful, as model validation was poor, for which two likely reasons were identified:

1. Time series data from pressure and flow sensors is “1D” in the sense that sensor positions are fixed and data is only available from those fixed sensor points, whereas a burst can be recorded in any section of the network and therefore can be seen as “2D” data. Therefore large bursts in sensor dense network regions are detectable via sensor time series data, but smaller bursts in sensor diffuse areas are not detectable by training a model solely based on sensor data. This means that there will be a certain bursts intensity threshold for detection. Identification of medium to large bursts is certainly possible. The resolution of the data thus determines the performance of the burst detection algorithm.

2. When investigating the time series data, some major anomalies in flow and pressure were not recorded as bursts in the database. Some of these anomalies could be explained by valve operations, repair or maintenance work. Limited information about these events was available, making it difficult to validate the origin of time series anomalies. Flow increase events deemed likely caused by bursts were also not recorded as such in the database.

In view of this experience with supervised burst detection, it was concluded that the leakage database was not suitable for the intended application Instead an unsupervised approach was chosen, in order to cluster days of regular water use, and thus identify the outlier days not added to a cluster as anomalies, such as conduit bursts.

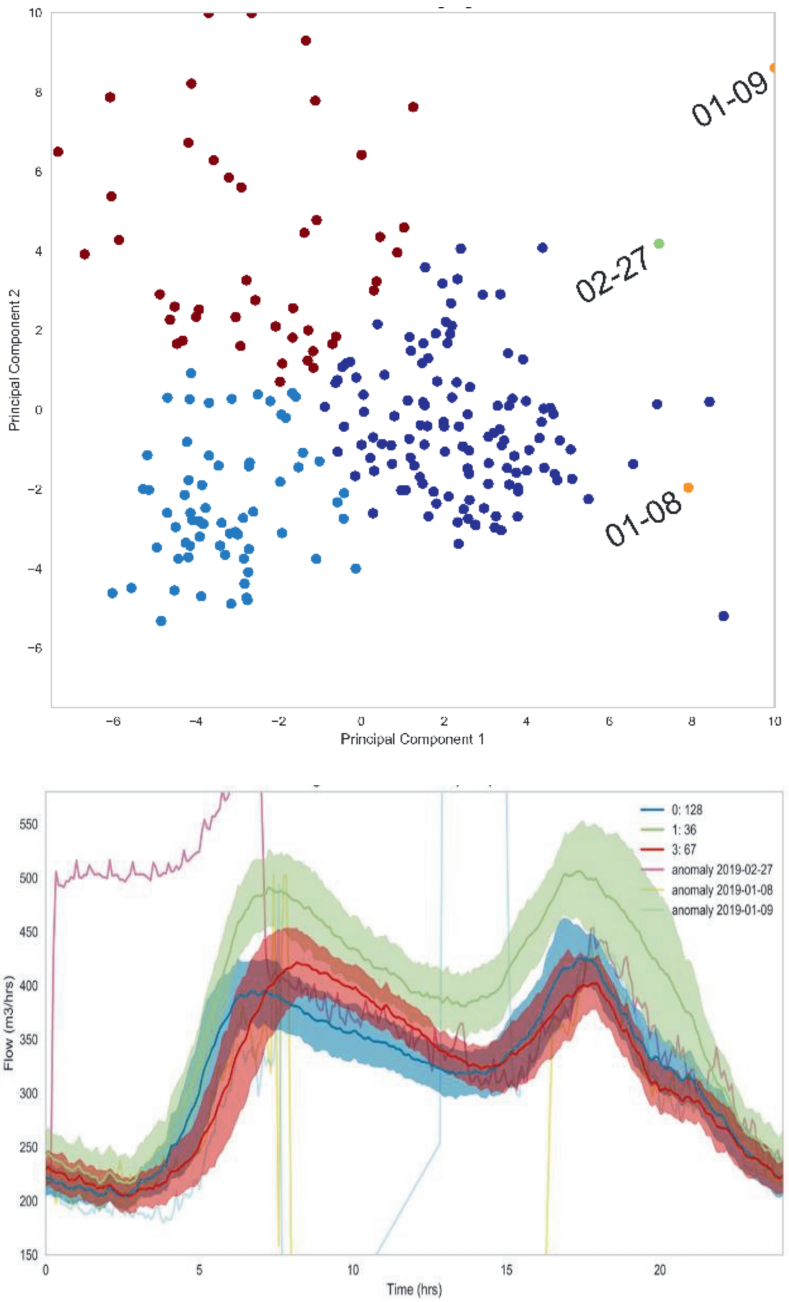


Figure 6.1 – Unsupervised learning for detection of days of anomalous water demand, based on data from January up to August. A) Clustering of days of flow sensor signals, based on the featurized flow time series measurements. Found clusters are 0 (East, navy blue), 3 (South-West, sky blue), 1 (North, brown). B) Mean flow per day per cluster, including standard deviation bandwidth. Source: (Nyambura 2020).

A clustering analysis based on 2019 flow data yielded three distinguishable clusters of regular water consumption. These clusters roughly corresponded to weekdays (Figure 6.1, cluster 0), weekends (Figure 6.1, cluster 3), and “holiday” water demand (Figure 6.1, cluster 1). People tend to wake up later in the weekend, and water demand on e.g. festivals or holidays is relatively higher compared to regular week and weekend days. Knowing which days of water demand are deemed regular helps specify which days are suitable to train water demand forecasting models. Directly or indirectly via forecasting models, identification of days of regular water demand will thus also facilitate identification of days with abnormal water demand. However, further investigation are needed to identify if these days contain leakages or other anomalous events.

6.3. Burst Detection by Water Demand Nowcasting Based on Exogenous Sensors

Using the insight obtained from the investigation into what defines a day of regular water demand described in Section 6.2, a burst detection algorithm was developed, with a focus on distinguishing bursts from other abnormal flow events. Due to the unsuccessful attempts to use the information present in burst databases, the traditional method of burst detection via DMA water balance forecasting was adapted since this approach is expected to be more robust and suffer less from known pitfalls. Four important concerns for current burst detection methods were identified, leading to the exogenous nowcasting method described in Chapter 3.

- The first concern is the effect of exogenous processes on water demand. Since water demand varies with various environmental and socio-demographic factors, such as weather, pandemic or holiday effects, predicting water demand solely based on historical data will not be able to account for these exogenous effects. Including data of these exogenous processes can greatly improve demand forecasts (Wu and Liu 2017).

- Second, in order to create a forecast based on exogenous predictors, forecasts of the exogenous predictors are required, which potentially adds high uncertainty. By nowcasting instead of forecasting water demand, no exogenous forecasts are required.

- Third, many water demand forecasting methods rely on DMA water balances in order to forecast water demand per DMA (Wu and Liu 2017). The nowcasting method can also be used to predict flow of a single sensor, provided that the sensor does measure a certain degree of seasonality under normal conditions.

- Fourth and most importantly, burst detection methods should be of high precision, as high false burst alarm frequencies undermine the method (Xu et al. 2020; Apostol et al. 2021). Due to the seasonality present in most sensor data, nowcasting allows harnessing of the regularity and seasonality as well as shared anomalies within sensor data, in order to suppress non-burst anomalies.

6.4. Case study: Exogenous Nowcasting in DMA of North-Western Dutch Drinking Water Company

After publication of the exogenous nowcasting methodology described in Chapter 3, a water utilities company in the North-West of the Netherlands reached out to express interest in investigating the added value of the nowcasting method for monitoring of their network. It was decided to investigate the added value of Exogenous Nowcasting via a case study of a single sensor-dense DMA and a data set spanning the first eight months of 2019 (Figure 6.2). Goal of this case study was to see if known burst could be detected using the exogenous method, and if it could indicate additional bursts or leakages not found by currently employed methods.

The water balance over the investigated DMA 1, was nowcasted based on available sensors within DMA 1, excluding sensors that were in very close proximity of each other, since these sensors showed near identical and highly correlated signals.. The water balance solely consisted of one metered inflow in the North of the DMA (Figure 6.2). Since all remaining exogenous sensors are positioned far downstream from the DMA 1 inflow, nowcasting results were weak. It was estimated that less than 8% of supplied water ends up in DMA 1a, less than 21% in DMA 1c, and thus at least 71% of supplied water is delivered to DMA 1b. Since no additional sensors are present in DMA 1b, no information was available on the water demand patterns in this section of the DMA. If DMA 1a and 1c were completed by placing sensors on the remaining unmetered connections between 1a and 1b as well as 1b and 1c, sub-DMA's 1a and 1c could have been investigated separately. However, this was not the case and thus in this study no attempt was made to nowcast water balances over 1a or 1c. As an alternative for nowcasting the DMA 1 water balance, the flow through those sensors with a high signal to noise ratio was nowcasted, using all other sensors in the DMA as regressors. Excluding aforementioned highly correlated sensor signals, two nowcasts could thus be made, for sensors α and β (Figure 6.2).

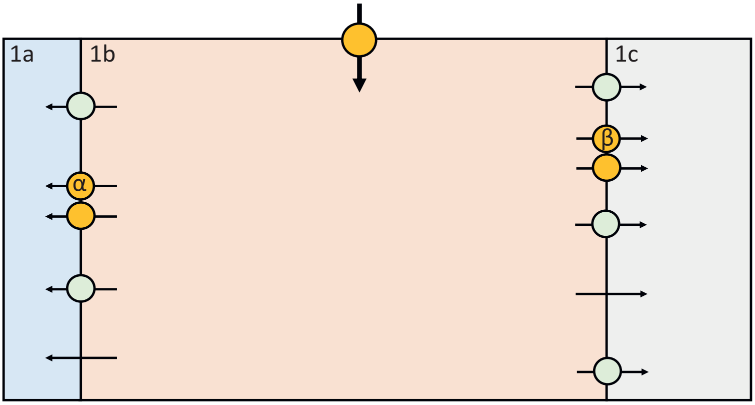


Figure 6.2 – Schematic overview of DMA 1, consisting of three sub-DMA's 1a, 1b, and 1c. Water is supplied from one pipe in the North of the DMA. Arrows indicate conduits connecting sub-DMA's, circles indicate flow and/or pressure sensors, where orange sensors showed a high signal to noise ratio, and green sensors showed a low signal to noise ratio. Nowcasts were made for flow sensors α and β.

Nowcasts of both sensors showed various suspected bursts and other anomalous demand patterns. A list of these anomalies, including graphs of the measurements and nowcasts, was made. The two most suspect cases were discussed with the water company (Figure 6.3). The first suspected burst was confirmed, but there was no knowledge about the second event. Although no further actions were taken regarding these cases, metering DMA's 1a and 1c has been finished since then, thus effectively dividing DMA 1 into three separate DMA's to facilitate better monitoring of these areas. Since the sensor signals as well as the water balance over DMA 1 were not actively analyzed, the water company also improved their dashboards for viewing these sensor signals, in order to facilitate and promote future analysis of collected sensor data.

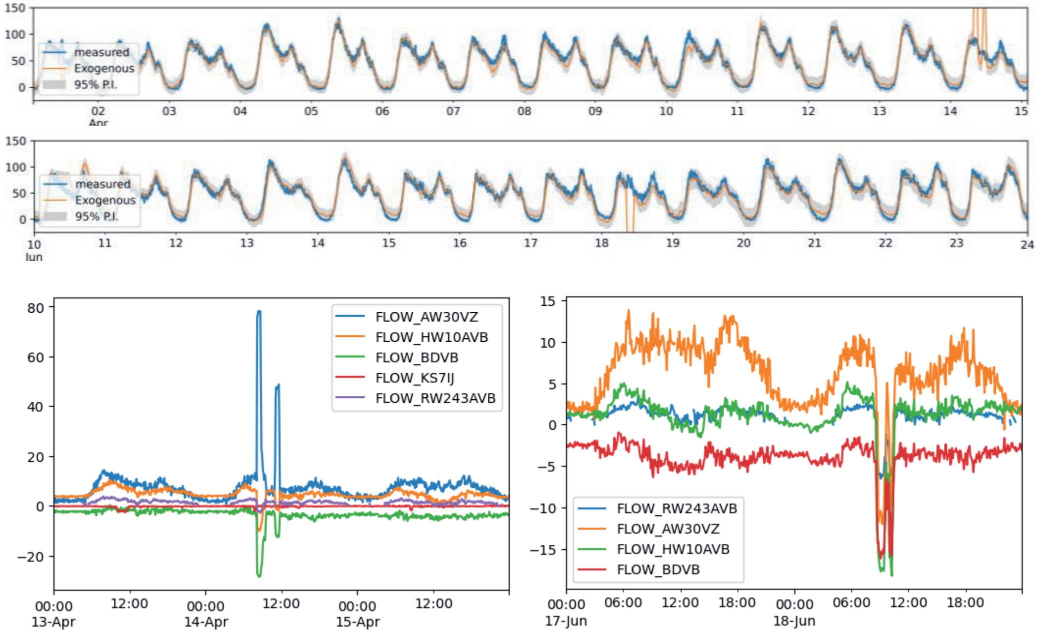


Figure 6.3 – Top: Sensor measurements of 2019 and exogenous nowcast, including 95% prediction confidence interval (P.I.), of flow measurements (m^3/h) from April (first graph) and June (second graph). Suspected bursts, where the nowcast significantly differs from the measurements, are seen on April 14th (first graph) and June 18th (second graph). Bottom: Flow measurements (m^3/h) of other sensors that were sensitive to these April and June events are shown on the bottom left and right, respectively.

Although being an intriguing case study, it failed in its premise of comparing performance of burst detection via either nowcasting or the current methodology. Results were interesting, but no hard conclusions could be drawn from the presented results, as the suspect events identified by nowcasting were not scored or validated. The water company concluded that the timespan of the large data set did not contain any large bursts that were picked up by their current detection methodology, and therefore no comparison of burst alarms could be made. The confirmed case of a small burst on June 18th was recorded in a burst database, but this burst was not detected by the current detection methodology. Since performance could not be scored, no comparison was possible. A list of possible past bursts is of no added value to a water company, if these past events did not lead to discernible problems. In order to score the performed case study, sufficient validation of the method is required. Since not all found events were associated with burst, maintenance, or valve operation databases, no validation was performed. Additionally, no follow-up investigation was performed for the found suspected burst that was not present in the records (Figure 6.3, April 14th).

However, the water company concluded that renewed interest in DMA's and real-time monitoring and displaying of sensor signals time series data on dashboards was warranted. Two other important conclusions were drawn. Firstly, when one gets the opportunity to prove the performance of developed algorithms, one should ensure that beforehand clear agreements are made as to how the results will be presented. This includes agreements on how to score and validate performance, and may also mean that adequate validation data, such as leakage databases, are made available. If the required validation data is confidential, agreements should be made about validation by employed experts. By knowing what to expect, results can become more clear and interpretable, and everyone involved will be better equipped to assess the added value of the presented methodology.

Secondly, in order to mine the information still locked within the collected sensor data, more emphasis should be put on curation of databases of burst, maintenance, and valve action records. In order to make the step to supervised burst detection, localization, and other smart applications, these databases are essential for training and validation of supervised algorithms. As this study shows, current databases are as of yet not always up to the standard required for successfully training various supervised algorithms. That said, the other side of this coin is that the best supervised algorithms are those that can largely overcome these disadvantages. Therefore, further research will help mine more information from the currently collected sensor data.

This insight is in line with the agenda of Dutch drinking water companies. Via participation in research and case studies like the present one, the status quo is continuously challenged in order to improve network monitoring and control. This is also exemplified by the creation of additional DMA’s and the fact that an improved sensor data viewing infrastructure was rolled out after the present case study.

6.5. Prediction of Pipe Degradation Based on Inline Ultrasonic Inspections

Besides up-to-date records, improving network monitoring depends on improving the amount of information that is being measured in the distribution network. Besides currently installed sensors, various other techniques can be used to obtain more information, such as various equipment-based measuring devices. Since cement pipes become brittle over time, inspection can help better assess speed of asset degradation, as well as help identify broken assets. Cement pipes become brittle due to calcium leaching from the material, thought to mainly be caused by conditions outside the pipe, such as groundwater level and acidity (Mainguy et al. 2000; Delgadillo et al. 2016).

Non-collinear wave mixing of ultrasonic signals is a non-destructive technique for measuring the leaching depth of cement. A major advantage of this inspection technique is that a pipe can be inspected without requiring access to both inner and outer wall. This means that it can be mounted on a Pipeline Inspection Gauge (PIG) capable of travelling through the pipelines, passively propelled by the water flow. With the help of an ultrasonic inspection company, a trajectory of 2 km of pipes from a central-Southern Dutch drinking water company was inspected, resulting in an ultrasonic response signal time series for every ~1.3 centimeter of pipe, for all eight sensors mounted on the PIG. This significant amount of noisy data then required further analysis, before providing insight in the condition of the inspected water mains (Kakes 2019; Delgadillo et al. 2020).

Each recorded ultrasonic response consists of multiple reflections of the transmitted signal, where the first two oscillations are the first generation responses of front and back wall reflections, and later reflections are the later generation responses (Figure 6.4). Using the maximum absolute derivative, a rough estimate of the time of occurrence of the first and most intense response of the front wall reflection was identified. After isolation of 10 μs windows, corresponding solely to this front wall reflection, a second, back wall reflection window was isolated in a similar fashion from the remainder of the signal. The difference in time of the maxima of both responses is a rough estimate of the cement condition. However, if the frequencies of both responses can be determined, a more accurate time difference can be identified based on the analytical envelope of both responses. Using a bandpass filter based on a Wiener filter and Hilbert transform, an analytical envelope of the front and back wall reflections was made (Delgadillo et al. 2016). The maxima of these analytical envelopes can be used to determine an exact time difference between both reflections, which in turn can be converted to a measure of leaching depth of the cement.

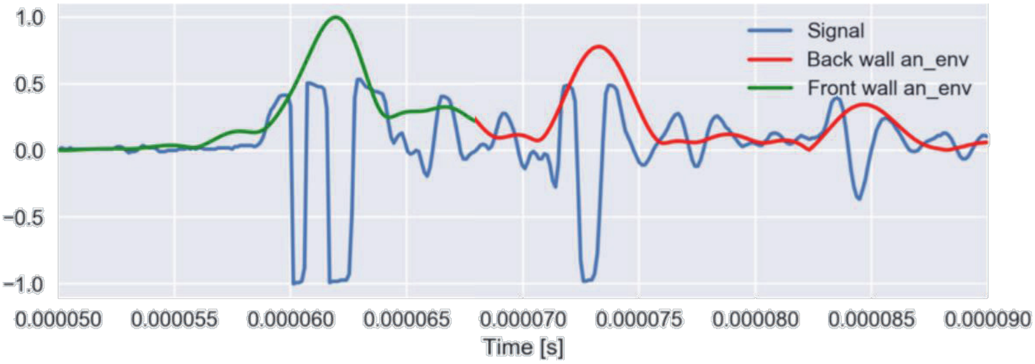


Figure 6.4 - Ultrasonic signal (mV), along with analytical envelopes for the front- and back wall reflection. The third local maximum in analytical envelope corresponds to the second front wall reflection. Adapted from (Kakes 2019).

Due to the high resolution of the thus obtained leaching depth and the noisy nature of the technique, a moving window mean in addition to outlier filters allow for a more interpretable measure of leaching depth (Figure 6.5) (Delgadillo et al. 2020).

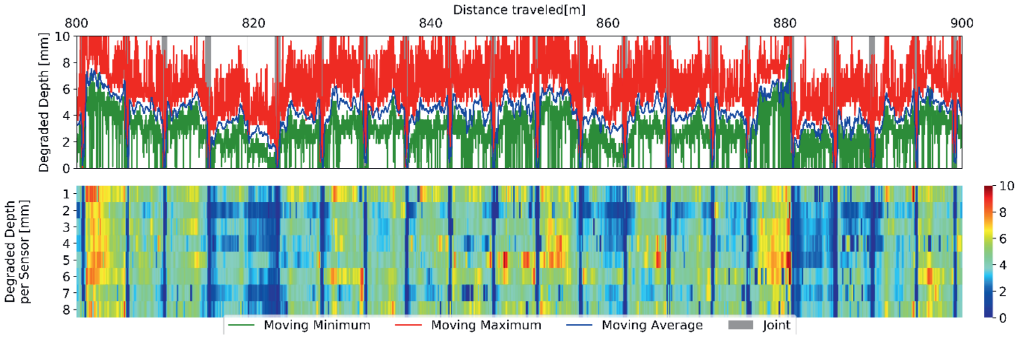


Figure 6.5 - Top: Moving window leaching depth, averaged over all sensors, along a part of the pipe trajectory. Joints between pipes are marked in grey. Bottom: Leaching depth for all eight sensors installed on the PIG (Kakes 2019).

Based on data on various features of environmental conditions, such as soil type or distance to nearest tree, and the leaching depth measurements, it was attempted to train a Random Forest model to predict leaching depth based on these environmental features. However, no workable model could be trained based on the trajectory data set. One inspected trajectory was insufficient to capture the underlying complex interactions and relations governing cement degradation.

Pipe bursts are often caused by a combination of various factors, amongst which inside and outside stress and pipe degradation, of which pipe degradation itself is also influenced by many factors (Barton et al. 2019). Although environmental parameters, such as soil type, pH, groundwater level, as well as distance of the pipe to trees, roads, and buildings, are assumed to play a major role, especially in the degradation of the outside of cement-based pipes, predicting remaining lifetime of pipes will remain difficult and require data on all of the involved processes. The current set of features did not contain enough information to capture the complexity of the underlying process. A qualitative distinction between critical leaching and acceptable leaching was made, based on a leaching depth threshold. All but two features spanned just two different values along the inspected trajectory, due to the low resolution of GIS features, such as soil type and soil subsidence. The only quantitative, non-binary features were distance to nearest tree and ground coverage, both in meters.

These two quantitative features were assigned a combined feature importance of 97% in the random forest model, meaning that the model almost solely fitted the qualitative leaching on these quantitative features. This enabled the model to reach a respectable validation precision of 97%, while consistently mislabeling certain data points, independent of which random and shuffled 80/20 data split was used for model training and validation. However, when training a model using the first 80% of the trajectory and validating with the last 20% of the data, precision dropped to 5%. The artificially high 97% precision was most likely caused by a combination of overfitting on the continuous features in addition to the spatial autocorrelation at smaller lags present in the degradation profile. Data shuffling enabled the model to fit on unique combinations of tree distance and ground coverage all over the trajectory, which due to the data shuffling was enough to interpolate these overfitting results to the validation data set. When there is only training data from the first 80% of the trajectory, this autocorrelation effect is no longer seen, resulting to far worse but more realistic precision. Future attempts to use supervised learning in predicting pipe degradation will benefit from a preemptive investigation to spatial autocorrelation present in the data, which can be analyzed by creating a semi-variogram.

Although the degradation measurements were of high spatial resolution (~cm-scale), there was a lack of features and within-feature variation (~km scale, qualitative features). Additionally, the difference between many of these environmental conditions over the length of the pipes was neglectable, meaning that significantly more inspections of various environmental conditions are required before inspection data can be used to train a supervised model. This does mean variations in pipe age, material and inside flow and pressure regime have to be taken into account in the model, before more inspection data can positively contribute to pipe condition assessment.

6.6. Sensor Placement

Robust records of past bursts, maintenance and valve operations are essential with regards to improving leakage control and network management, as these databases are required in order to train the supervised methods of the future. However, labels alone are not sufficient in order to train e.g. leakage detection algorithms, as these labels need to be coupled to training data. If insufficient high-frequency sensors are installed in the network, monitoring DMA- and network-wide water quantity and quality will remain difficult (van Summeren et al. 2016). In order to obtain high resolution insight into the water distribution process in each section of the drinking water networks, smart placement of flow and pressure sensors is essential.

Different water companies may handle different criteria regarding where to best place additional sensors. Besides flow sensor placement for DMA formation, placement of additional sensors within each DMA is essential in order to facilitate robust, accurate, and high resolution burst detection and localization. Where currently burst detection mainly hinges on prolonged and unexpected increases in DMA water demand detected by flow sensors, pressure sensors have been shown to be effective in detecting (burst-induced) pressure transients at the exact moment of pipe burst (Srirangarajan et al. 2012; Lee et al. 2016). If an accurate hydraulic model is available, accurate burst localization will also be possible (Fuchs-Hanusch and Steffelbauer 2017). However, based on smart installation of pressure sensors and/or the correlations between the signals of these sensors, rough localization of burst can be performed, further emphasizing the importance of optimal pressure sensor placement (Quevedo, Casín et al. 2011; Ponce et al. 2014). Besides pipe bursts causing pressure transients, pressure sensor signals contain many pressure anomalies, which are as of yet not actively detected and analyzed. However, as has been shown in Chapter 2, even non-burst anomalies are valuable indications of network performance, and can be leveraged as early warning and decision support system to improve the water distribution process. For detection and localization of both burst and non-burst anomalies, taking into account pressure dynamics in optimal sensor placement is therefore critical.

The static and definite placement of flow or pressure sensors will benefit from taking into account future uses for the additional sensor data thus made available. One of the most crucial factors to take into account regarding future data goals, is a more proactive management of the water grid, without relying on a reactive approach based on customer complaints. Data sources are currently not sufficiently mined and combined, meaning relying on customers as surrogate sensors remains a necessity to facilitate the current reactive strategy (Scozzari et al. 2021). Placement of additional sensors can tackle some of these issues. By placing additional sensors based on network observability, the thus obtained sensor data will help improve hydraulic models as well as boost burst detectability.

6.7. Concluding Remarks

As mentioned before, in the preface of the book “ICT for Smart Water Systems”, editor S. Mounce refers to the water industry as generally “data-rich but information poor” (Scozzari et al. 2021). They argue that collected data is currently underused and undervalued, where only the tip of the information iceberg in the data sea is currently visible.

Collected data might be underused, but is not undervalued, as water companies are acutely aware that there is a wealth of research available regarding data and model driven approaches to optimize drinking water distribution. Since company resources are limited, not every opportunity can be taken, and water companies seem to be unable to see the forest through the trees sometimes, regarding which novel techniques to pursue and what research to participate in. Especially for smaller water companies, this means potentially missing relevant innovations which can result in lagging behind the innovations in the field of water distribution. Where digital twins, supervised learning, and AI are hot topics in the water sector, many companies are still far removed from this stage and are still more concerned about starting and/or focusing on DMA formation. Since novel research builds upon a foundation of prior research, in order to participate in novel research, water companies first have to ensure that all prerequisite prior innovations have been realized.

However, for most current challenges faced by water companies, supervised machine learning may not be the most pressing. A recurring problem when facing challenges regarding water distribution, is the absence of labelled data. Water companies are very interested in data driven strategies to improve insight in the water distribution and have a strong preference for machine learning applications, as the potential power of these algorithms are widely known and recognized. However, for many applications of supervised learning, no adequate training data labels are available, or label curation has only recently been started, meaning no extensive records are available as of yet.

Let us for example consider burst detection. Although large amounts of e.g. flow and pressure sensor data are collected in real-time, often no class labels suitable for training supervised algorithms are recorded. Due to the regular nature of flow and pressure sensor time series data, past data can be used to facilitate anomaly detection. However, as mentioned in Chapter 2, not all anomalies are indeed bursts. In order to employ robust and high precision burst detection and localization, extensive records of past bursts are required to validate the anomaly detection methods. However, these databases should take into account various considerations:

1) Although these data sets are often referred to as “Big Data”, the number of data points and data features are not always sufficient to enable machine learning algorithms to learn the complex underlying processes. Where the Googles and Facebooks of the world can safely claim that they work with big data, for water companies operating in a single Dutch city or province, this is not the case. Although more than 10 years of burst records are available, companies themselves often point out that the registration stringency only increased some years ago, essentially marking results from before that time substandard. Add to that the (fortunately) low burst frequency of a specific asset or region of assets, and the number of labels per said asset or region will be hard to classify as big data.

2) A big challenge in water distribution is the detection of so-called silent leaks: bursts that do not cause massive flooding of streets and do not deprive homes of water, but bursts that go unnoticed, while still losing significant amounts of water over time. These undetected and thus silent bursts are often visible in the time series data, but are not detected and will therefore not appear in a leakage database. Moreover, if these bursts are found and repaired at a later date, without knowing the exact date of when the burst originated, no correct burst record can be added to the database. Instead, the burst might be mislabeled as originating from the discovery date. Many bursts cause significant water hammers when they originate. However, these water hammers can thus not be validated or used for model training, if the burst records are filed at other dates.

3) It is difficult to accurately assess water loss flowrate, and therefore it is hard to distinguish which burst record in the database should have been detectable based on sensor data and which burst most likely will require additional data sources besides sensor data to identify. There exists a grey area of detectability of bursts that in size fall between undetectable drip-leaks and meters-long gushing cracks in main transport pipes. More pressure transient-based burst detection will mitigate some of these difficulties, although that would require accurate hydraulic models and high frequency sensor data.

4) The effect of missing labels or mislabeled burst records is worsened by the imbalance in the frequency and distribution of burst and non-burst scenarios. Since there are many more days on which a given pipe does not burst, compared to days on which it does, the fraction of bursts is small.

Although creative solutions and supervised model variants have been developed to deal with most if not all of these problems, it still will always involve a lot of database curation and validation and manual labelling by drinking water distribution experts, which are often not the people developing these algorithms. It is therefore important to limit labelling requirements, as already illustrated by recent research using active learning to achieve laudable accuracy under minimal labelling burden (Russo et al. 2020; Tomar and Burton 2021).

A priority is often to maintain the status quo regarding burst frequency and repair response, and new technologies are sometimes perceived as potential risks to current admirable network operation and management, and are thus sometimes regarded with suspicion or scepsis. Although Dutch drinking water distribution boasts a low non-revenue water fraction, the number of bursts in ageing pipes are increasing and costs associated with network maintenance present the largest and fastest growing investment category (Vewin 2017)(Figure 6.6). This should indicate that the status quo alone may not be sufficient for continuous and safe drinking water distribution, and a more proactive leakage control strategy is required. Although on the long term supervised models are more suitable than their unsupervised counterparts, for direct application as well as some specific applications, unsupervised models provide a suitable alternative.

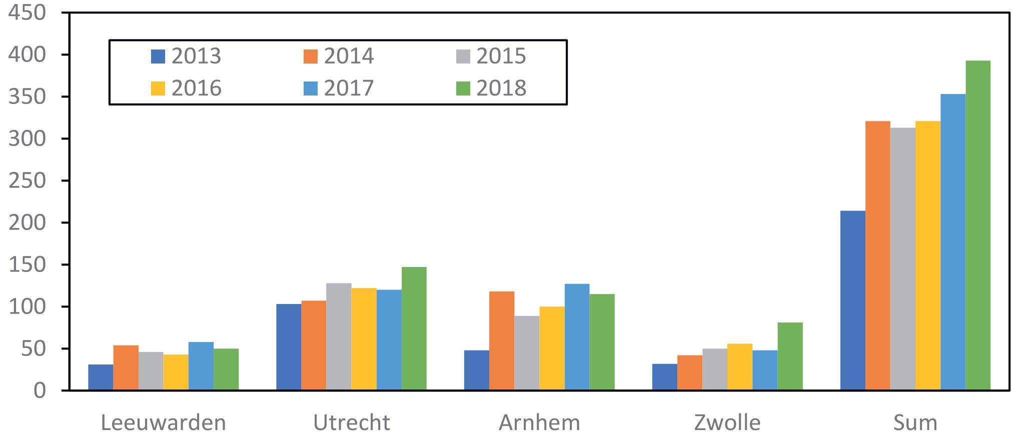


Figure 6.6 - Number of (reported) bursts per year, for various Dutch cities supplied by a central-Eastern Dutch drinking water company (www.waterstoring.nl)

There is a large societal relevance to providing safe and reliable drinking water. It is not only a basic need, but also the sixth of the seventeen sustainable development goals ("clean water and sanitation"). Dutch drinking water grids belong to the best of the world, even for Western standards (Vewin 2017). Although not as growth-dependent as large parts of the USA's water distribution, which has been compared to a Ponzi scheme, the high and increasing investments required to manage the distribution nets do warrant attention (Marohn 2011, 2013). The ageing infrastructure is at higher risk of leakages, which besides water loss potentially cause drinking water contamination and damage to other infrastructure (van Summeren et al. 2016).

Climate and drought concerns have sparked interest in a separate distribution system for potable and non-potable water (Killion 2011). Due to the high costs associated with centralized water distribution, such as current drinking water infrastructure, decentralized retaining and redistribution of non-potable water could potentially greatly alleviate demand for potable water (Vewin 2017). Due to the inability to retain water between seasons, groundwater wells are often threatened by low groundwater levels during summer (Van Huijgevoort et al. 2020). Retaining and reusing more precipitation and surface water during winter, in addition to separate and decentralized distribution of non-potable water, offers interesting possibilities for alleviating pressure on existing groundwater wells, especially during dry summer periods.

6.8. Outlook

Monitoring support and burst detection: Monitoring support is not dependent on labelled training data, but makes use of real-time unsupervised learning via expanding and moving windows. Future research could experiment with consensus clustering in order to better combine different time windows of clustered recurring patterns. If monitoring support is ran in parallel per pressure or flow sensor, a database of thus encountered recurring patterns can be created. All recurring patterns encountered in the time series data of a single flow or pressure sensor could then be classified using the database, or added to the database if not encountered before. Classification would allow for fast responses as soon as recurrences are detected, since the database will be able to point to possible causes and mitigation strategies. If certain causes are overrepresented in the detected recurring anomalous patterns among all sensor data, this could point to a weak spot in network operation, and be used to steer investments and innovation in the direction most beneficial for remedying recurring abnormalities.

Although trying to explicitly capture the “rules” that define a certain day of water demand as regular is hard, it would certainly help to obtain insight in consumer demand and water distribution. However, the underlying “rules” of water demand are highly complex and everchanging, and can therefore not always be explicitly articulated (Li et al. 2019). These spatial-temporal correlations and rules can be taught to supervised machine learning algorithms, even if these rules cannot be formulated in advance.

Having classified what normal water demand on week or weekends looks like, more accurate anomaly detection can be performed (Wu and Liu 2020). With increasing awareness for database curation and validation, as well as placement of additional sensors, future research may again focus on supervised modelling of network bursts (Castro-Gama and Agudelo-Vera 2019).

Moving window correlations between sensor signals can provide interesting insight in the water distribution process, since changes in these correlations are caused by changes in the routes most of the water in the network takes. These correlation changes thus indirectly point towards changes in water demand, presence of bursts, or changes in valve states. The exogenous sensor signals that are used as regressors in the nowcasting method from Chapter 3 show a similar capacity. Since the nowcast is fitted every day at 00:00, these regressors change over time. If a certain sensor signal suddenly starts contributing significantly to the nowcast of a neighboring sensor or water balance, this could indicate a change in the routes the water takes within the DMA. Although all these effects are summarized and converted into interpretable information in the nowcast, part of the information iceberg remains locked in these regressors. The regressor changes can therefore potentially provide more indirect information about the water distribution process and are more readily available than high accuracy digital twins of each DMA. Research focused on extracting concrete information from these regressor changes or changes in inter-sensor signal correlations therefore are promising as a tool for studying the water distribution process. Additionally, these correlations and regressors show potential as features for future supervised learning methods with a focus on network monitoring and early warning.

By expanding the training data with more inspected trajectories in future studies, more variation in soil and environmental parameters are to be expected. More variation in features and additional features in the form of pipe properties and internal parameters will thus enable the model to be trained on the actual relationship between environmental parameters and cement condition. Although it was thus not yet possible to predict cement leaching based on surrounding conditions, the developed processing pipeline to convert inspection signals to a measure of leaching depth does facilitate fast and easy interpretation of inspection results. A check of spatial autocorrelation is also advisable for degradation inspections of high resolution.

Sensor placement: Since placement of sensors, especially pressure sensors, is rapidly becoming cheaper, metering the inside of each DMA in addition to the boundaries is rapidly becoming more viable. However, where to place these new sensors can be difficult to determine, especially for DMA's or large sizes. In order to maximize the information that becomes available of placement of a limited number of additional sensors, optimal sensor placement is of utmost importance. Similar to burst localization, virtual burst simulation can be used to determine the sensor placement that optimizes burst detectability. Optimal sensor locations are those junctions in the network that are most sensitive to as wide an area of possible burst sites as possible.

Applications such as burst or contamination localization as well as digital twin construction, require an accurate hydraulic model of the network (van Thienen and Morley 2018). Robust and accurate hydraulic models can be used to combine multiple data sources, such as water quantity and quality data, household and industrial water demand measurements, and valve and pump status data. Using digital twin technology, real-time simulation of the network can be performed, and bursts, contaminations, or other deviations can be detected in real-time. Moreover, a digital twin allows for a more proactive strategy, by allowing for simulation of e.g. contamination events or specific hydraulic scenarios (van Summeren et al. 2016; Conejos Fuertes et al. 2020; Scozzari et al. 2021).

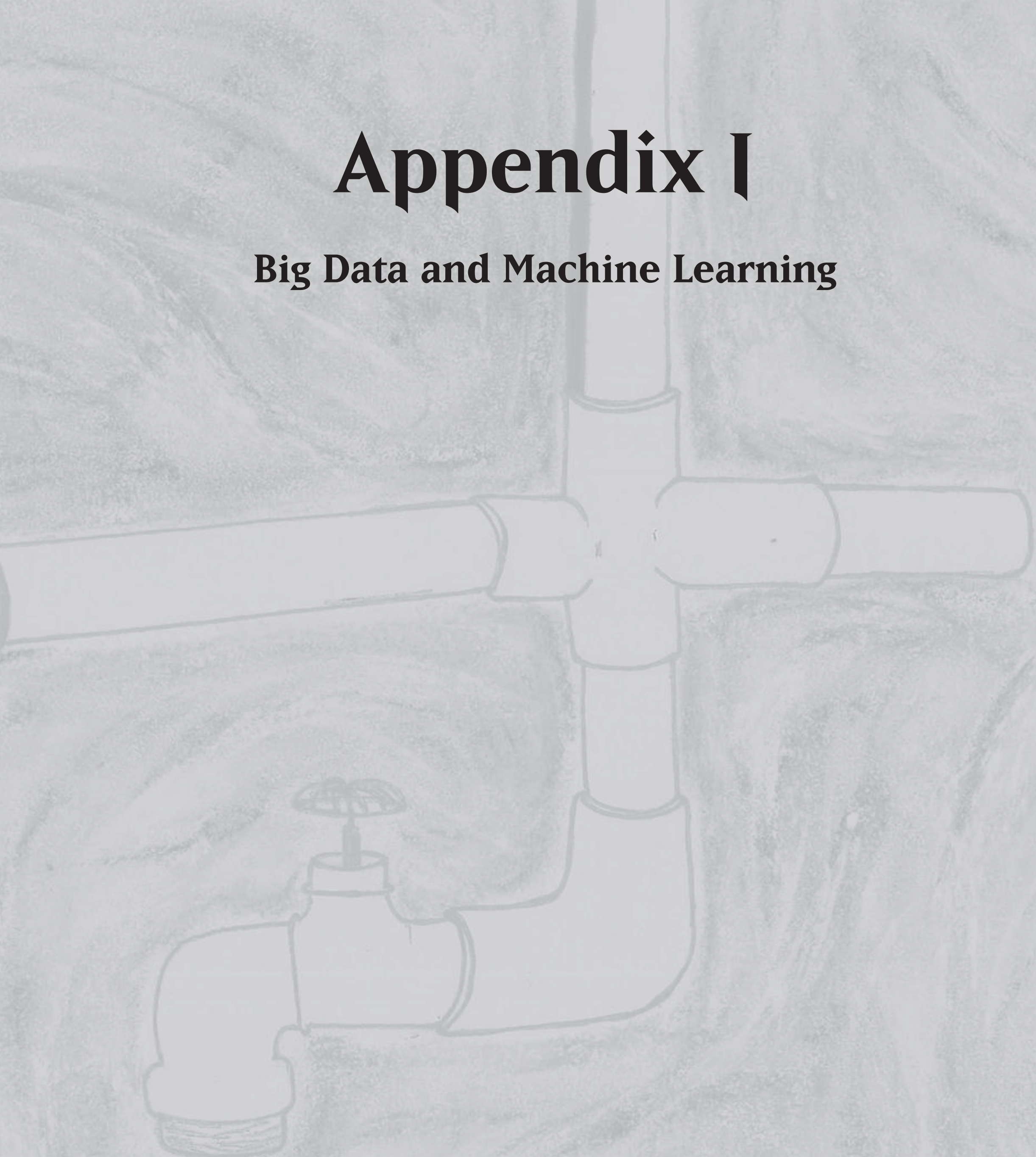
Current hydraulic models often have to estimate water demand per household, since limited real-time data about water use per water connection is available, if any (Nagar and Powell 2004; Meseguer et al. 2014; Cugueró-Escofet et al. 2017; Verde and Torres 2017). Smart meters, sensors per household that measure water demand in real-time, and potentially also offer possibilities for measuring network pressures at the connection point, offer the potential to change this in the near future (Scozzari et al. 2021). Future work may focus on determining observability-based optimal placement of water quality sensors or smart meters. Optimal placement of flow, pressure, quality, and demand sensors increases certainty of network properties and demands, and thus allows for more accurate hydraulic models, which expedite digital twin establishment.

However, since not all households have sensors as of yet, assumptions have to be made. These demand estimates, as well as assumptions about the resistance of the network pipes, also add uncertainty to the hydraulic models. When running network operation simulations, such as the virtual burst simulations required for sensitivity-based optimal sensor placement, the consequences of these compounding uncertainties need to be considered (Steffelbauer and Fuchs-Hanusch 2016; Cugueró-Escofet et al. 2017; Boatwright et al. 2018; Qi et al. 2018; Quiñones-Grueiro et al. 2018). Although the potential of running various network operation scenarios is undeniable, it remains important to validate and evaluate the assumptions made. Before placement of smart meters on a large scale and sufficiently validated hydraulic models are commonplace, optimal sensor placement not relying on simulation offers a more robust route to achieve optimal sensor placement.

In order to ensure sufficient water flow for each customer, sufficient network pressure is required. Since current control over network pressure is limited, overpressure is used to meet this criterion. Due to high energy costs associated with pressurizing the network, water companies have expressed interest in a more optimal pressure regime not relying on overpressure (van Summeren et al. 2016). Besides increasing monitoring capabilities of the network, the same state-space representation methodology in Chapters 4 and 5 can be used to determine optimal placement of actuators, such as pumps or valves. This would increase the control that can be exerted over the network. In a similar fashion to observability-based placement, future research may focus on using controllability to determine multi-objective optimal actuator placement.

Appendix I

Big Data and Machine Learning



Appendix I – Big Data and Machine Learning

1. Eigenvectors and Eigenvalues

Let us consider a system of n differential equations, defined as $\frac{d}{dt}x(t) = Ax(t)$. If A is not a diagonal matrix, this coupled system cannot be solved separately for each variable x. Ideally, we would rather solve a decoupled system $\frac{d}{dt}z(t) = \Lambda z(t)$ based on a diagonal matrix Λ, with elements λ_1, λ_2, ..., λ_n, since this would greatly simplify the involved matrix calculus and would provide the simple solution $z_i(t) = z_i(0)e^{\lambda_i t}$ for $i = 1, \dots, n$ (Miller et al. 2013). Using $Av = \lambda v$ (eigenvalue decomposition) it is possible for diagonalizable system matrices A to find the so-called eigenvectors v and accompanying eigenvalues λ required to transform the original system into a decoupled system. Each eigenvectors v of a square matrix A is a nonzero vector whose direction remains unchanged when applying the linear transformation $f(x) = Ax$, but whose magnitude will be scaled by a factor λ, called the eigenvalue. Amongst others, eigenvalue decomposition of a matrix A allows for easier matrix calculation and provides intuitive insight in the behavior of a system.

2. Principal Component Analysis

Due to the increased computational capacity, it is difficult to easily interpret and summarize the multitude of big data collected. Fortunately, there are techniques available that help summarize these large data sets into more compact and workable sizes. A widely used and well known method to achieve this so-called dimensionality reduction is Principal Component Analysis (PCA). PCA can achieve this dimensionality reduction in an unsupervised way by projecting a data set into a subspace, while preserving the original structure, relationships, and essence of the data set. The resulting reduced or “summarized” projection can then still be used for further machine learning predictions, without a significant loss in accuracy.

PCA is defined by a linear transformation or “projection” of a high dimensional vector space X into a lower dimensional space v. The axes of this new space are called the principal components of the original vector space, where each principal component is a vector that maximizes the variance of the data while being orthogonal to each preceding principal component. We can find the principal components v of the $N \times p$ data points, represented by the data matrix X, as well as accompanying explained variance λ captured by these components, from the $p \times p$ covariance matrix $\mathrm{C} = \frac{1}{\mathrm{N}}X^T X$ via $Cv = \lambda v$. Therefore, the principal components are found by performing eigenvalue decomposition of the covariance matrix C of the data set, where the resulting eigenvectors v are the principal components and accompanying eigenvalues λ are a measure for the variance explained by the corresponding eigenvectors.

For example, let’s consider a two dimensional data set containing three clusters of data, where each data point consist of a variable x_1 and x_2 (Figure I.1, left, black circles) and calculate the principal components of this data sets (Figure I.1, left, red and blue line). If we would solely consider the thus created first principal component axis (figure I.1, left, red line) and project each data point on this one dimensional subspace (red line), we can still clearly distinguish the three clusters of data points based on their position on this first component (Figure I.1, right, red dots), without the need for the original two dimensional data (x_1 and x_2).

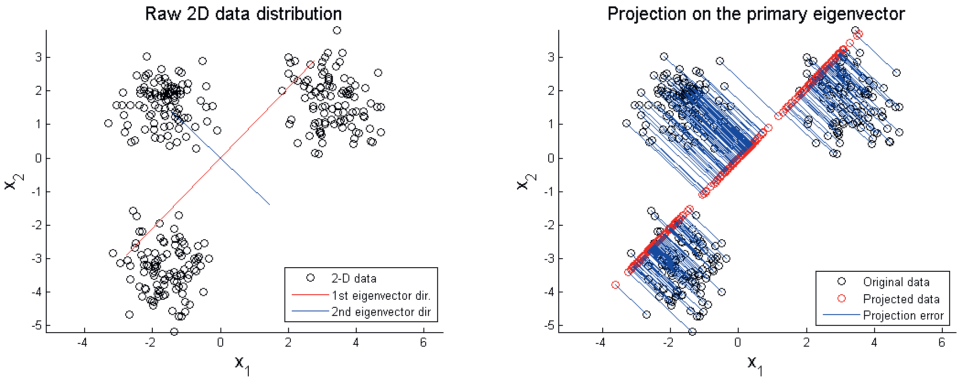
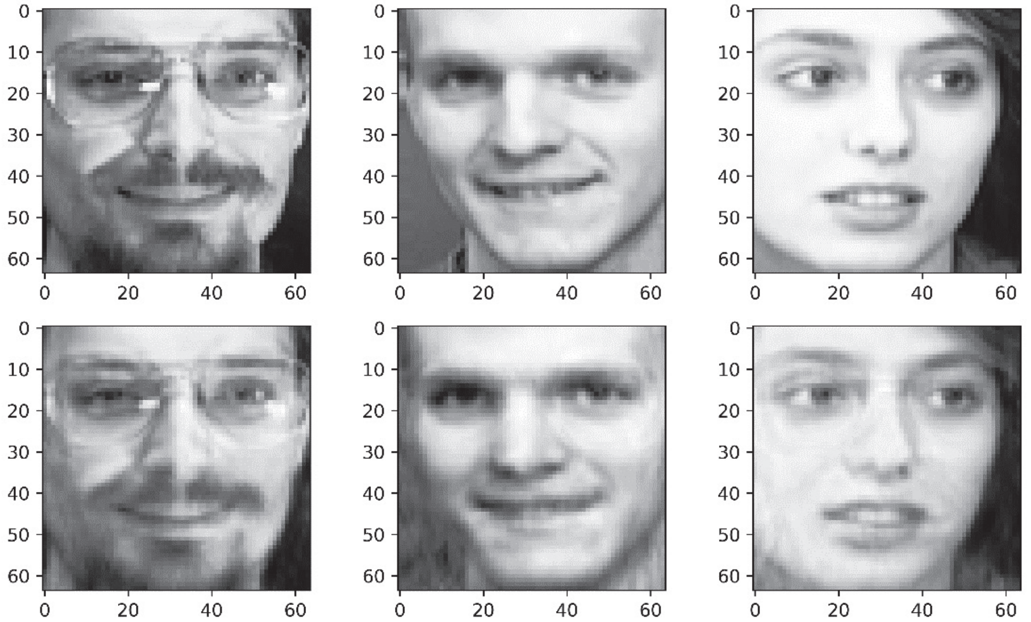


Figure I.1 - Example of PCA projection. Left: data set where the two principal components of the data are drawn. Right: data set is projected onto first principal component (Sujin 2014).

A more practical example would be to consider a data set of 64x64 pixels black and white pictures of human faces, from the Olivetti database at AT&T Laboratories, Cambridge (Figure I.2, top row). Each pixel can be considered a variable of this data set with a value between 0 (white) and 1 (black), meaning that for each picture 64x64=4096 variables are recorded. If we apply PCA and want to retain at least 95% of the variance present in the original data set of 400 64x64 pictures, we only need 123 components instead of the original 4096 variables. This means we need only require 3% of the size of the original data set, without a significant loss of information. When translating these 123 components back to 64x64 pictures, we can even verify by eye that these reduced pictures (Figure I.2, bottom row) are still quite similar to the original pictures (Figure I.2, top row).



```
from sklearn import datasets # load faces data set             Python3 code example
from sklearn.decomposition import PCA # load PCA method
from matplotlib import pyplot as plt # load figure plotting method
faces = datasets.fetch_olivetti_faces() # data set of 400 64x64 pictures of faces
pca = PCA(n_components=0.95, whiten=True).fit(faces.data) # 95% variance components
fig, axes = plt.subplots(2, 3, figsize=(10,6)) # create a new 2x3 plot
for i_plot, i_face in enumerate([160, 45, 79]): # three 4096-variable pictures
    projected_face = pca.transform(faces.data[[i_face]]) # 123-variable projection
    reconstructed_face = pca.inverse_transform(projected_face) # reconstructed face
    axes[0, i_plot].imshow(faces.images[i_face], cmap="Greys_r") # plot original
    axes[1, i_plot].imshow(reconstructed_face.reshape(64, 64), cmap="Greys_r")
plt.show() # show final result, both original and 123 feature reconstruction
```

Figure I.2 - Top, First Row: Three faces from the Olivetti Database. Top, Second Row: Reconstructed faces based on 123 principal component projection. Bottom: Python code used to construct the top images.

3. Featurizing

As shown in the PCA example above, summarizing high dimensional data sets can be realized based on techniques such as PCA. However, if a clear goal has been formulated, often no high dimensional data is required. For example, when trying to distinguish cows from horses based on 4096 pixel pictures, we could apply PCA to pictures of various cows and horses and compare 100+ components. Alternatively, we could record each animal's weight and height differentiate the two species based solely on these two "features". The intrinsic differences between the two animal species is already contained in the two features of animal height and animal weight. In other words, knowing what we expect to analyze from a given data set, smartly choosing features will summarize the intrinsic properties of the data set without requiring all original data values. This process is called feature extraction. Regarding water distribution, when investigating the daily pattern of water use, it is often more interesting to investigate e.g. mean, peak, and standard deviation of daily water consumption, instead of analyzing the thousands of flow measurements collected each day. A robust set of chosen features often captures the intrinsic information contained in a large set of data and can therefore be a useful first data processing step, before considering additional dimensionality reduction techniques, such as PCA.

4. Machine Learning

Machine learning refers to computer algorithms that improve automatically through experience and by the use of data. Generally two types of machine learning are distinguished:

- Supervised learning: algorithms that learn to map an input (data) to an output (label) based on example input-output pairs (labelled training data).
- Unsupervised learning: algorithms that learn to distinguish patterns from unlabeled data.

4.1. Classification

Supervised machine learning, or classification, refers to algorithms that use training data and corresponding labels in order to train a model that can then automatically label new as of yet unseen unlabeled data. Supervised learning is a form of “learning with teacher”, where first the algorithm is taught what the correct answer (label or class) is for a set of training data, after which the algorithm is supposed to be able to correctly classify unlabeled data it has never seen before (Figure I.3). A common process is to train a supervised algorithm based on (multiple sets of) 80% of the training data, so that the remaining 20% can be used to validate the performance of the trained algorithm. The availability of labels thus also allows for a robust and clear method for scoring the performance of supervised methods.

If the assigned classes are qualitative, we speak of classification. If instead we try to automatically label data points with one or more quantitative values, we speak of regression. The easiest example of regression is fitting a line through a data set containing two variables, linear or nonlinearly.

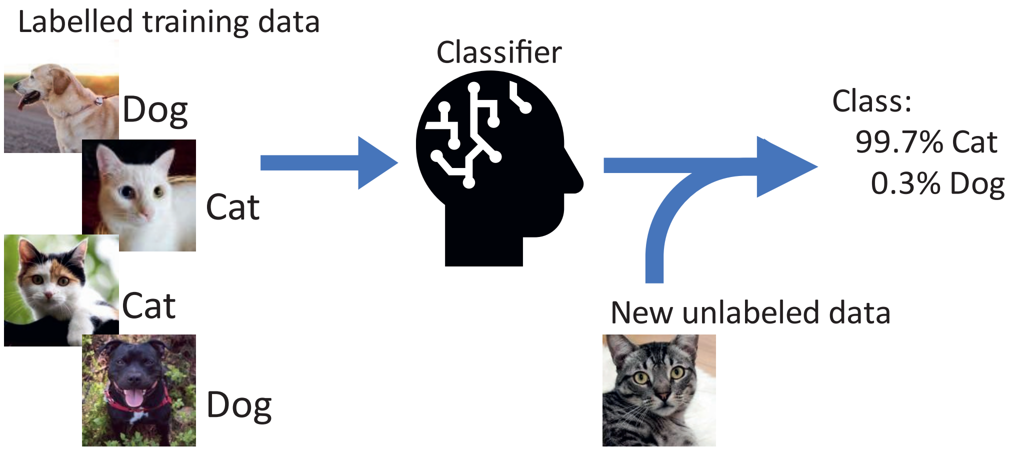


Figure I.3 – Example of a supervised machine learning algorithm.

4.2. Clustering

Unsupervised learning, or clustering, is a method to distinguish patterns in data by dividing a data set in groups of similar samples, without being told beforehand what patterns or groups to expect. Unsupervised algorithms try to identify structure in the data by "learning without teacher". The most common form of unsupervised learning is clustering, where sample vectors from a data set are grouped into clusters based on the similarities in their measurement vectors.

One possible method for clustering discussed in this thesis is density-based spatial clustering of applications with noise (DBSCAN) (Ester et al. 1996). Density-based clustering appoints data to clusters if part of a contiguous region of relative and significant higher point density in the data space compared to other contiguous regions, where data points not in a cluster are deemed outliers (Sander 2010). For example, let us consider a data set where we retain two principal components after reducing data set dimensions using PCA. DBSCAN can be illustrated by visualizing our reduced data, where we plot each data point with the first principal component on the horizontal axis and the second on the vertical axis (Figure I.4). DBSCAN divides all data points into three categories based on two model parameters: a neighborhood radius ϵ and a minimum points per neighborhood density $MinPts$. First, core points are those data points with at least $MinPts$ other points within a radius ϵ (Figure I.4, black circles). Second, each point within the radius of a core point that is itself not a core point, is a boundary point (Figure I.4, green circles). Lastly, all other points are deemed outliers (Figure I.4, orange circles). As can be seen from Figure I.4, clustering helps distinguish groups of data with similar data.

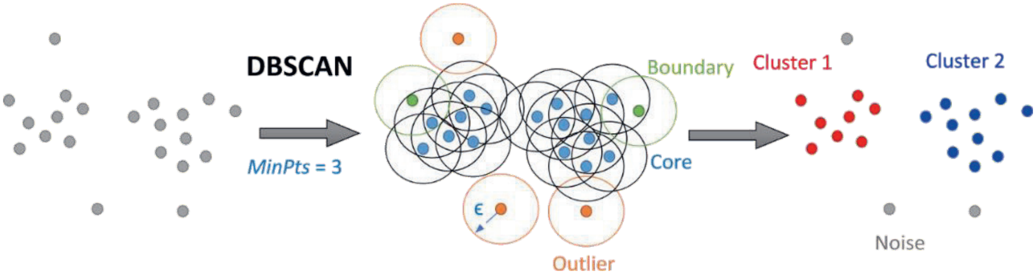


Figure I.4 - Example of clustering algorithm DBSCAN. Adapted from (Khater et al. 2020).

Since no labels are available and thus often no expected and correct expected clustering is known, the performance of clustering algorithms is less robust compared to supervised algorithms. Scoring metrics for clustering performance not based on expected clustering outcome do exist in the form of “internal” evaluation metrics, where the clustering results are scored based on how well clusters are defined and separate from each other in the data space. An example of such a metric is the Calinski-Harabasz index, which is the ratio of the sum of between-clusters dispersion and of within-cluster dispersion for all clusters (Caliñski and Harabasz 1974). If the resulting clusters are well apart, the index will be higher compared to when clusters overlap in the measurement space. Similarly, dense clusters result in a higher index compared to more sparse clusters. The index therefore scores the clustering results based on how distinct and well defined each cluster is within the data space.

4.3. Machine Learning Caveats

Although the potential of machine learning and especially supervised is undeniable, these methods do suffer from some caveats that should be carefully considered at method deployment.

4.3.1. Generalization Errors

Generalization errors occur when the model is not capable of correctly capturing the general relation between input data and output labels. There are two types of generalization errors, caused by either high bias due to underfitting or high variance due to overfitting.

An algorithm is biased when consistently misclassifying a particular input data point, even if trained using other but equally good training data sets. This generalization error is caused by underfitting, where a model is too simple/inflexible or informed by too few features to correctly capture the entire and more complex general relation between input data and output labels.

An algorithm has high variance for a particular input data point if it predicts different outputs when trained on different, but equally good training data sets. This generalization error is caused by overfitting, where the model is too complex/flexible or informed by too many features. The model thus wrongly (also) assumed the noise in the training data to be (part of) the true relation between input and output, making it perform poorly on data other than the training data set by learning more than the simple relation between input data and output labels.

There is a tradeoff between bias and variance, and the chosen supervised algorithm must be able to correctly weight the effect of both bias and variance according to indented outcome.

4.3.2. Process Complexity and Interpretability

Highly complex algorithms are capable of capturing even very complex processes. However, since these models are high variance, attention should be paid not to overfit a model. If the true relation between input data and output classes is highly complex and involves complex interactions, a large amount of training data is required for an algorithm to learn this relationship entirely. Additionally, mislabeled or noisy input data will worsen this effect, since many algorithms will only perform well under a low fraction of incorrect training data. Lastly, highly complex models are not or hardly interpretable. These models may be capable of correctly predicting complex relations, but do not provide insight in what exactly these relations are or what underlying interactions govern them.

4.3.3. Imbalanced Classification

Another bias common in classification problems is an imbalance in class distributions, where one class is significantly more prevalent than another. An easy example would be the burst detection problem: there are significantly more days on which a specific pipe does not burst, compared to days on which it does. This imbalance in the training data is especially relevant when, like in this example, we are more interested in the underrepresented class. To overcome this imbalance, more training data is required, in order to ensure the underrepresented class is also sufficiently present in the training data. Second, low mislabeling of the underrepresented class is permitted. And third, care should be taken when selecting the appropriate supervised method, as some assume equal class distributions.

4.3.4. Inductive Bias

Even if sufficient and correctly labelled training data is available, not every possible class may be represented in that data set. For example, training a supervised algorithm to distinguish photos of different animals is possible. However, if you only have photos of cats and dogs available when training the algorithm, it will always mislabel shark photos as showing either a cat or a dog. Therefore emphasis should be put on identifying and ensuring each possible class is sufficiently present in the training data.

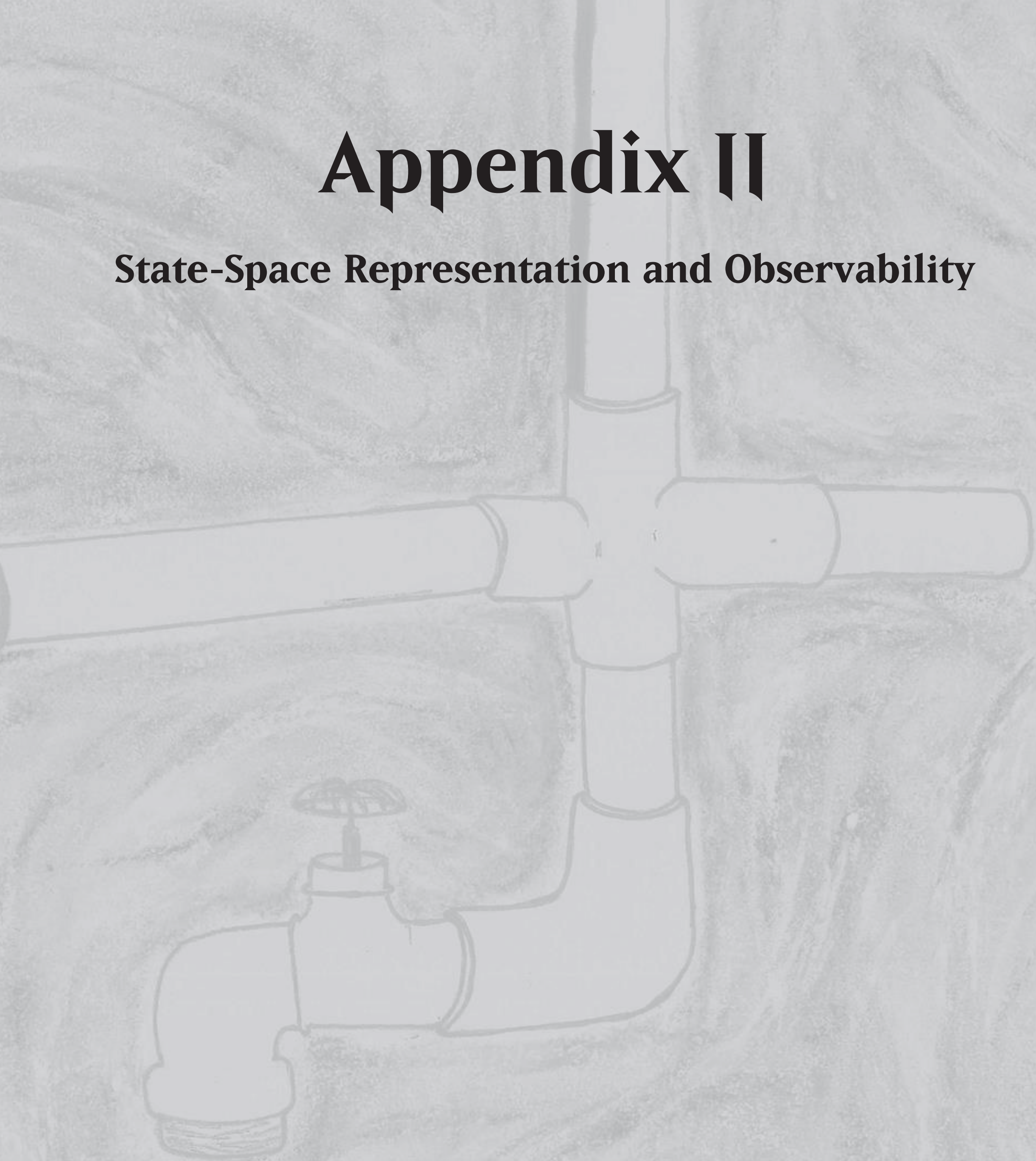
4.3.5. Unsupervised-Specific Caveats

Unsupervised learning often focuses solely on clustering one given data set and can be used when no labels are available to group each data point into clusters. Since no labelled data is considered, mislabeled or insufficient training data are not a problem, and since no novel unseen data needs to be processed, generalization errors are of a lesser concern. Another advantage of unsupervised methods is their low inductive bias. If a certain class is not present in the training data of a supervised algorithm, this class will thus never be correctly labelled. However, when clustering data points belonging to a new, never before seen label, these data point can become a novel cluster and will thus be successfully distinguished from the prior encountered clusters.

However, since no labels are available, validation of the unsupervised method is difficult. In other words, unsupervised learning will always yield results, although effort has to be taken in order to ensure these results are interpretable and offer practical relevance. Unsupervised learning can only group data, and will never be able to determine which group belongs to what class, if any. Besides scoring using internal metrics, such as the abovementioned Calinski-Harabasz index, part of the clustered data can be manually labelled by field experts. Based on these labels, validation can be performed. Manual labelling thus allows for proper validation of unsupervised algorithms. However, manual labelling often is time and expertise intensive, suffers from human bias, and may not be feasible for some data sets.

Appendix II

State-Space Representation and Observability



Appendix II – State-Space Representation and Observability

1. State-Space Representation

A state-space model representation of a physical dynamic system contains input $u(t)$, output $y(t)$ and state $x(t)$ variables, where the inputs influence the system from the outside, states are latent variables of the system, and outputs are those (combinations of) states that are measurable. Linear, time-invariant and finite-dimensional systems can be written in the following matrix notation:

$$\begin{gathered}\frac{d}{dt}x(t) = \boldsymbol{A}x(t) + \boldsymbol{B}u(t) \\ y(t) = \boldsymbol{C}x(t) + \boldsymbol{D}u(t)\end{gathered} \qquad (II - a)$$

In this equation, state matrix $\boldsymbol{A}$ describes the dynamics of the system, input matrix $\boldsymbol{B}$ shows how the inputs affect the system, output matrix $\boldsymbol{C}$ shows which (combinations of) states are measurable, and feedthrough matrix $\boldsymbol{D}$ presents the direct influence of inputs on the output (as opposed to indirect influence, since inputs $u(t)$ influence states $x(t)$ via the input matrix $\boldsymbol{B}$, which influence the outputs $y(t)$ via the output matrix $\boldsymbol{C}$). This state-space notation, in short system (***A, B, C, D***), helps to "algebraically summarize" an entire system of equations, and can be used to quickly analyze system properties and behavior, based solely on the matrices and vectors of the state-space representation. For the theory of infinite-dimensional systems, such as the hyperbolic partial differential model in Chapters 4 and 5, we refer to (Curtain and Zwart 1995).

2. Observability and Observability Gramian

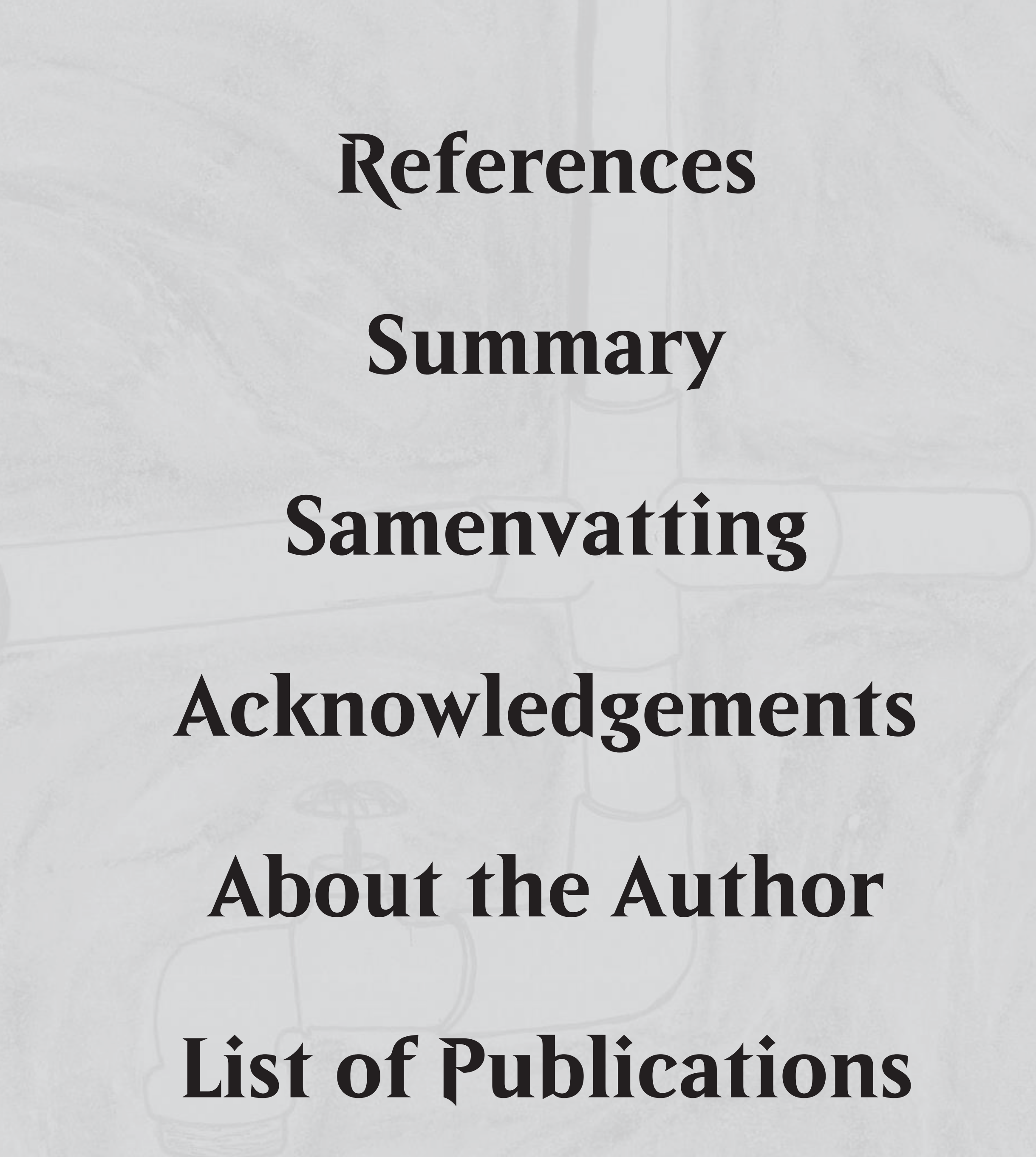
State-space model representation can be used to quicky determine system observability, which investigates the question: How well can I reconstruct each state of the system based solely on those output states I can measure? In other words, observability is a measure for how well you can know what happens in the entire system, while only being able to measure a select few parts of the system. The so-called system's observability matrix $\mathcal{O}$ can be calculated from the $n \times n$ state matrix $\boldsymbol{A}$ and the output matrix $\boldsymbol{C}$:

$$\mathcal{O} = \begin{bmatrix} \boldsymbol{C} \\ \boldsymbol{CA} \\ \boldsymbol{CA}^2 \\ \vdots \\ \boldsymbol{CA}^{n-1} \end{bmatrix} \qquad (II - b)$$

A system is fully observable if $\mathrm{rank}(\mathcal{O}) = n$, meaning that each column of $\mathcal{O}$ is linearly independent and thus each state $x(t)$ can be inferred from the measured outputs $y(t)$. For larger systems, matrix $\mathcal{O}$ can become very large, leading to numerical problems. Alternatively, for stable matrix $\boldsymbol{A}$, the smaller $n \times n$ observability Gramian $W_{\mathcal{O}}$ is given by (Chen 1999):

$$W_{\mathcal{O}} = \int_0^\infty \left(e^{A^T\tau} C^T C e^{A\tau}\right) d\tau \qquad (II - c)$$

Not only is the observability Gramian more compact, the same information can be obtained from this matrix. Thus, this Gramian matrix contains information about how well states can be reconstructed from input-output data over time. Applying eigenvalue decomposition to the observability Gramian $W_{\mathcal{O}}$ can be used to provide more information about the observability of combinations of states.



References

Summary

Samenvatting

Acknowledgements

About the Author

List of Publications

References

Adamowski J, Fung Chan H, Prasher SO, et al (2012) Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resour Res 48:1–14. https://doi.org/10.1029/2010WR009945

Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T (2015) Time-series clustering - A decade review. Inf Syst 53:16–38. https://doi.org/10.1016/j.is.2015.04.007

Al Qahtani T, Yaakob MS, Yidris N, et al (2020) A Review on water leakage detection method in the water distribution network. J Adv Res Fluid Mech Therm Sci 68:152–163. https://doi.org/10.37934/ARFMTS.68.2.152163

Amran TST, Ismail MP, Ahmad MR, et al (2018) Monitoring underground water leakage pattern by ground penetrating radar (GPR) using 800 MHz antenna frequency. IOP Conf Ser Mater Sci Eng 298:. https://doi.org/10.1088/1757-899X/298/1/012002

Anele AO, Hamam Y, Abu-Mahfouz AM, Todini E (2017) Overview, comparative assessment and recommendations of forecasting models for short-term water demand prediction. Water (Switzerland) 9:. https://doi.org/10.3390/w9110887

Apostol E-S, Truică C-O, Pop F, Esposito C (2021) Change Point Enhanced Anomaly Detection for IoT Time Series Data. Water 13:1633. https://doi.org/10.3390/w13121633

Arandia E, Ba A, Eck B, McKenna S (2016) Tailoring seasonal time series models to forecast short-term water demand. J Water Resour Plan Manag 142:1–10. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000591

Babel MS, Shinde VR (2011) Identifying Prominent Explanatory Variables for Water Demand Prediction Using Artificial Neural Networks: A Case Study of Bangkok. Water Resour Manag 25:1653–1676. https://doi.org/10.1007/s11269-010-9766-x

Bai Y, Wang P, Li C, et al (2014) A multi-scale relevance vector regression approach for daily urban water demand forecasting. J Hydrol 517:236–245. https://doi.org/10.1016/j.jhydrol.2014.05.033

Bakker M, Vreeburg JHG, Van De Roer M, Rietveld LC (2014) Heuristic burst detection method using flow and pressure measurements. J Hydroinformatics 16:1194–1209. https://doi.org/10.2166/hydro.2014.120

Barton NA, Farewell TS, Hallett SH, Acland TF (2019) Improving pipe failure predictions: Factors effecting pipe failure in drinking water networks. Water Res 164:. https://doi.org/10.1016/j.watres.2019.114926

Benítez R, Ortiz-Caraballo C, Preciado JC, et al (2019) A short-term data based water consumption prediction approach. Energies 12:1–24. https://doi.org/10.3390/en12122359

Bi W, Dandy GC, Maier HR (2015) Improved genetic algorithm optimization of water distribution system design by incorporating domain knowledge. Environ Model Softw 69:. https://doi.org/10.1016/j.envsoft.2014.09.010

Billings RB, Jones CVTA-TT- (2008) Forecasting Urban Water Demand

Boatwright S, Romano M, Mounce S, et al (2018) Optimal Sensor Placement and Leak/Burst Localisation in a Water Distribution System Using Spatially-Constrained Inverse-Distance Weighted Interpolation. 3:282–273. https://doi.org/10.29007/37cp

Bonada E, Jordi Meseguer, Tur JMM (2014) Practical-Oriented Pressure Sensor Placement for Model-Based Leakage Location in Water Distribution Networks. Informatics Environ Data Model Integr a Heterog Hydro World

Branisavljević N, Kapelan Z, Prodanović D (2011) Improved real-time data anomaly detection using context classification. J Hydroinformatics 13:307–323. https://doi.org/10.2166/hydro.2011.042

Breen J (2006) Levensduurverwachting van bestaande PVC Leidingen

Brentan BM, Luvizotto E, Herrera M, et al (2017) Hybrid regression model for near real-time urban water demand forecasting. J Comput Appl Math 309:532–541. https://doi.org/10.1016/j.cam.2016.02.009

Brynych A (2018) Off-line Internal Inspection of Pipelines – An Important Tool for Investment Decisions. Pipeline Technol. J.

Bui XK, Marlim MS, Kang D (2020) Water network partitioning into district metered areas: A state-of-the-art review. Water (Switzerland) 12:. https://doi.org/10.3390/W12041002

Caliñski T, Harabasz J (1974) A Dendrite Method Foe Cluster Analysis. Commun Stat 3:1–27. https://doi.org/10.1080/03610927408827101

Candelieri A (2017) Clustering and support vector regression for water demand forecasting and anomaly detection. Water (Switzerland) 9:. https://doi.org/10.3390/w9030224

Casillas M V., Puig V, Garza-Castañón LE, Rosich A (2013) Optimal sensor placement for leak location in water distribution networks using genetic algorithms. Sensors (Switzerland) 13:14984–15005. https://doi.org/10.3390/s131114984

Castro-Gama M, Agudelo-Vera C (2019) BTO report: Data Quality Control. Niewegein

CBS Hoeveel wegen zijn er in Nederland? https://www.cbs.nl/nl-nl/visualisaties/verkeer-en-vervoer/vervoermiddelen-en-infrastructuur/wegen. Accessed 8 Jun 2021

Chatfield C (1993) Calculating interval forecasts. J Bus Econ Stat 11:121–135

Chaudhry MH (2014) Applied Hydraulic Transients, 3rd Ed. Springer New York

Chen CT (1999) Linear System Theory and Design, 3rd edn. Oxford University Press, New

York

Chena J, Boccelli DL (2014) Demand forecasting for water distribution systems. Procedia Eng 70:339–342. https://doi.org/10.1016/j.proeng.2014.02.038

Christ M, Braun N, Neuffer J, Kempa-Liehr AW (2018) Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing 307:72–77. https://doi.org/10.1016/j.neucom.2018.03.067

COB (2018) Kabels en Leidingen Overzicht.pdf. https://www.cob.nl/magazines-brochures-en-nieuws/verdieping/september-2018/kabels-en-leidingen/. Accessed 8 Jun 2021

Conejos Fuertes P, Martínez Alzamora F, Hervás Carot M, Alonso Campos JC (2020) Building and exploiting a Digital Twin for the management of drinking water distribution networks. Urban Water J 17:704–713. https://doi.org/10.1080/1573062X.2020.1771382

Cugueró-Escofet M, Puig V, Quevedo J (2017) Optimal pressure sensor placement and assessment for leak location using a relaxed isolation index: Application to the Barcelona water network. Control Eng Pract 63:1–12. https://doi.org/10.1016/j.conengprac.2017.03.003

Curtain RF, Zwart HJ (1995) An Introduction to Infinite Dimensional Linear Systems Theory. Springer-Verlag, New York

Dager R, Zuazua E (2006) Wave propagation, observation and control in 1-d flexible multi-structures. Mathématiques Appl 50:227

De Coster A, Pérez Medina JL, Nottebaere M, et al (2019) Towards an improvement of GPR-based detection of pipes and leaks in water distribution networks. J Appl Geophys 162:138–151. https://doi.org/10.1016/j.jappgeo.2019.02.001

Delgadillo HH, Geelen CVC, Kakes R, et al (2020) Ultrasonic inline inspection of a cement-based drinking water pipeline. Eng Struct 210:110413. https://doi.org/10.1016/j.engstruct.2020.110413

Delgadillo HH, Loendersloot R, Akkerman R, Yntema D (2016) Development of an inline water mains inspection technology. IEEE Int Ultrason Symp IUS 2016-Novem:7728471. https://doi.org/10.1109/ULTSYM.2016.7728471

Deng A, Hooi B (2021) Graph Neural Network-Based Anomaly Detection in Multivariate Time Series

Di Nardo A, Di Natale M (2012) A design support methodology for district metering of water supply networks. Water Distrib Syst Anal 2010 - Proc 12th Int Conf WDSA 2010 870–887. https://doi.org/10.1061/41203(425)80

Di Nardo A, Di Natale M, Di Mauro A, et al (2019) Calibration of a water distribution network with limited field measures: The case study of Castellammare di Stabia (Naples, Italy). Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect

Notes Bioinformatics) 11353 LNCS:433–436. https://doi.org/10.1007/978-3-030-05348-2_36

Díaz S, González J, Mínguez R (2016) Observability Analysis in Water Transport Networks: Algebraic Approach. J Water Resour Plan Manag 142:04015071. https://doi.org/10.1061/(asce)wr.1943-5452.0000621

Eliades DG, Polycarpou MM (2012) Leakage fault detection in district metered areas of water distribution systems. J Hydroinformatics 14:992–1005. https://doi.org/10.2166/hydro.2012.109

Ester M, Kriegel H-P, Sander J, Xu X (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise

Farah E, Shahrour I (2017) Leakage Detection Using Smart Water System: Combination of Water Balance and Automated Minimum Night Flow. Water Resour Manag 31:4821–4833. https://doi.org/10.1007/s11269-017-1780-9

Farley B, Mounce SR, Boxall JB (2010) Field testing of an optimal sensor placement methodology for event detection in an urban water distribution network. Urban Water J 7:345–356. https://doi.org/10.1080/1573062X.2010.526230

Farley B, Mounce SR, Boxall JB (2012) Development and field validation of a burst localisation. J Water Resour Plan Manag 145:

Fischler MA, Bolles RC (1981) Random sample consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Commun ACM 24:381–395. https://doi.org/10.1145/358669.358692

Fitié JH (2014) Dynamic Bandwidth Monitor. In: Vitens. https://github.com/Vitens/DBM

Folkman S (2018) Water Main Break Rates In the USA and Canada: A Comprehensive Study

Froelich W (2016) Daily urban water demand forecasting - Comparative study. Commun Comput Inf Sci 613:. https://doi.org/10.1007/978-3-319-34099-9_49

Fuchs-Hanusch D, Steffelbauer D (2017) Real-world Comparison of Sensor Placement Algorithms for Leakage Localization. Procedia Eng 186:499–505. https://doi.org/10.1016/j.proeng.2017.03.262

Fujiwara O, Khang DB (1990) A two-phase decomposition method for optimal design of looped water distribution networks. Water Resour Res 26:539–549. https://doi.org/10.1029/WR026i004p00539

Fulcher BD, Jones NS (2014) Highly comparative feature-based time-series classification. IEEE Trans Knowl Data Eng 26:3026–3037. https://doi.org/10.1109/TKDE.2014.2316504

Gelazanskas L, Gamage KAA (2014) Demand side management in smart grid: A review and proposals for future direction. Sustain Cities Soc 11:22–30. https://doi.org/10.1016/j.scs.2013.11.001

Georges D (1995) Use of observability and controllability gramians or functions for optimal sensor and actuator location in finite-dimensional systems. In: Proceedings of the IEEE Conference on Decision and Control

Giustolisi O, Savic D, Kapelan Z (2008) Pressure-Driven Demand and Leakage Simulation for Water Distribution Networks. J Hydraul Eng 134:. https://doi.org/10.1061/(asce)0733-9429(2008)134:5(626)

Grubben NLM, Keesman KJ (2018) Controllability and observability of 2D thermal flow in bulk storage facilities using sensitivity fields. Int J Control 91:1554–1566. https://doi.org/10.1080/00207179.2017.1321782

Hagos M, Lansey KE, Jung D (2016) Optimal meter placement for pipe burst detection in water distribution systems. J Hydroinformatics 18:741–756. https://doi.org/10.2166/hydro.2016.170

Huang P, Zhu N, Hou D, et al (2018) Real-time burst detection in District Metering Areas in water distribution system based on patterns of water demand with supervised learning. Water (Switzerland) 10:1–16. https://doi.org/10.3390/w10121765

Hutton C, Kapelan Z (2015a) Real-time burst detection in Water Distribution Systems using a Bayesian demand forecasting methodology. Procedia Eng 119:13–18. https://doi.org/10.1016/j.proeng.2015.08.847

Hutton CJ, Kapelan Z (2015b) A probabilistic methodology for quantifying, diagnosing and reducing model structural and predictive errors in short term water demand forecasting. Environ Model Softw 66:87–97. https://doi.org/10.1016/j.envsoft.2014.12.021

Izquierdo J, Pérez R, Iglesias PL (2004) Mathematical models and methods in the water industry. Math Comput Model 39:1353–1374. https://doi.org/10.1016/j.mcm.2004.06.012

Johnson T, Moger T (2021) A critical review of methods for optimal placement of phasor measurement units. Int Trans Electr Energy Syst 31:1–25. https://doi.org/10.1002/2050-7038.12698

Kabir G, Tesfamariam S, Sadiq R (2015) Predicting water main failures using Bayesian model averaging and survival modelling approach. Reliab Eng Syst Saf 142:498–514. https://doi.org/10.1016/j.ress.2015.06.011

Kakes R (2019) Prediction of Pipe Degradation based on Inline Ultrasonic Inspections. MSc Thesis supervised by: Geelen, CVC; Delgadillo, HH; Yntema DR; Kiewidt, LW. Wageningen

Kakoudakis K, Behzadian K, Farmani R (2017) Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K - means clustering. Urban Water J 9006:1–6. https://doi.org/10.1080/1573062X.2016.1253755

Kalman RE (1963) Mathematical Description of Linear Dynamical Systems. J Soc Ind Appl Math Ser A Control 1:. https://doi.org/10.1137/0301010

Kalman RE, Buey R (1961) A new approach to linear filtering and prediction theory. Trans ASME, J Basic Eng 83:95–108

Keesman KJ (2011) Sytem Identification, An Introduction, Advanced Textbooks in Control and Signal Processing

Khater IM, Nabi IR, Hamarneh G (2020) A Review of Super-Resolution Single-Molecule Localization Microscopy Cluster Analysis and Quantification Methods. Patterns 1:100038. https://doi.org/10.1016/j.patter.2020.100038

Killion SM (2011) Design and Modeling of Infrastructure for Residential and Community Water Reuse. 1–192

Kim Y, Lee SJ, Park T, et al (2016) Robust leak detection and its localization using interval estimation for water distribution network. Comput Chem Eng 92:1–17. https://doi.org/10.1016/j.compchemeng.2016.04.027

KNMI (2020) Achtergrondinformatie Neerslag. https://www.knmi.nl/kennis-en-datacentrum/achtergrond/achtergrondinformatie-neerslagtekort. Accessed 8 Jun 2021

Kotsiantis SB, Pintelas PE (2004) Recent Advances in Clustering : A Brief Survey. Methods 1:73–81

Kozłowski E, Kowalska B, Kowalski D, Mazurkiewicz D (2018) Water demand forecasting by trend and harmonic analysis. Arch Civ Mech Eng 18:140–148. https://doi.org/10.1016/j.acme.2017.05.006

Kwakernaak H, Sivan R (1972) Linear Optimal Control Systems, Vol. 1. Wiley-interscience, New York

Laucelli DB, Simone A, Berardi L, Giustolisi O (2017) Optimal Design of District Metering Areas for the Reduction of Leakages. J Water Resour Plan Manag 143:04017017. https://doi.org/10.1061/(asce)wr.1943-5452.0000768

Lee SJ, Lee G, Suh JC, Lee JM (2016) Online Burst Detection and Location of Water Distribution Systems and Its Practical Applications. J Water Resour Plan Manag 142:04015033. https://doi.org/10.1061/(asce)wr.1943-5452.0000545

Leu S Sen, Bui QN (2016) Leak Prediction Model for Water Distribution Networks Created Using a Bayesian Network Learning Approach. Water Resour Manag 30:2719–2733. https://doi.org/10.1007/s11269-016-1316-8

Li D, Chen D, Jin B, et al (2019) MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 11730 LNCS:703–716. https://doi.org/10.1007/978-3-030-30490-4_56

Liemberger R, Wyatt A (2019) Quantifying the global non-revenue water problem. Water Sci Technol Water Supply 19:831–837. https://doi.org/10.2166/ws.2018.129

Liu Z, Kleiner Y (2013) State of the art review of inspection technologies for condition assessment of water pipes. Meas J Int Meas Confed 46:1–15. https://doi.org/10.1016/j.measurement.2012.05.032

MacKay DJC (1992) Bayesian Interpolation. Neural Comput 4:415–447. https://doi.org/10.1162/neco.1992.4.3.415

Mainguy M, Tognazzi C, Torrenti JM, Adenot F (2000) Modelling of leaching in pure cement paste and mortar. Cem Concr Res 30:83–90. https://doi.org/10.1016/S0008-8846(99)00208-2

Malm A, Ljunggren O, Bergstedt O, et al (2012) Replacement predictions for drinking water networks through historical data. Water Res 46:2149–2158. https://doi.org/10.1016/j.watres.2012.01.036

Marchi A, Dandy GC, Boccelli DL, Masud Rana SM (2018) Assessing the Observability of Demand Pattern Multipliers in Water Distribution Systems Using Algebraic and Numerical Derivatives. J Water Resour Plan Manag 144:04018014. https://doi.org/10.1061/(asce)wr.1943-5452.0000909

Marohn C (2011) The Growth Ponzi Scheme, Part 4. In: Strong Towns. https://www.strongtowns.org/journal/2011/6/16/the-growth-ponzi-scheme-part-4.html. Accessed 16 Jul 2021

Marohn CL (2013) Suburban ponzi scheme. Leadersh Manag Eng 13:181–189. https://doi.org/10.1061/(ASCE)LM.1943-5630.0000234

Martínez-Codina Á, Cueto-Felgueroso L, Castillo M, Garrote L (2015) Use of pressure management to reduce the probability of pipe breaks: A bayesian approach. J Water Resour Plan Manag 141:. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000519

McInnes L, Healy J, Astels S (2017) hdbscan: Hierarchical density based clustering. J Open Source Softw 2:205. https://doi.org/10.21105/joss.00205

Meseguer J, Mirats-Tur JM, Cembrano G, et al (2014) A decision support system for on-line leakage localization. Environ Model Softw 60:331–345. https://doi.org/10.1016/j.envsoft.2014.06.025

Millán-Roures L, Epifanio I, Martínez V (2018) Detection of anomalies in water networks by functional data analysis. Math Probl Eng 2018:. https://doi.org/10.1155/2018/5129735

Miller H, Jerison D, French J (2013) M.I.T. 18.03 : Extra Notes and Exercises. M.I.T. 18.03 Ordinary Differ. Equations

Mounce SR, Khan A, Wood AS, et al (2003) Sensor-fusion of hydraulic data for burst detection and location in a treated water distribution system. Inf Fusion 4:217–229.

https://doi.org/10.1016/S1566-2535(03)00034-4

Mounce SR, Mounce RB, Jackson T, et al (2014) Pattern matching and associative artificial neural networks for water distribution system time series data analysis. J Hydroinformatics 16:617–632. https://doi.org/10.2166/hydro.2013.057

Nagar AK, Powell RS (2004) Observability analysis of water distribution systems under parametric and measurement uncertainty. Jt Conf Water Resour Eng Water Resour Plan Manag 2000 Build Partnerships 104:55. https://doi.org/10.1061/40517(2000)213

National Research Council (2006) Drinking water distribution systems: Assessing and reducing risks. National Academies Press

Navarrete-López C, Herrera M, Brentan BM, et al (2019) Enhanced water demand analysis via symbolic approximation within an epidemiology-based forecasting framework. Water (Switzerland) 11:. https://doi.org/10.3390/w11020246

Nyambura H (2020) Classification of Daily Patterns in Sensor Measurements for Improved Insight in Water Distribution System Behavior. MSc Thesis supervised by: Geelen, CVC; Yntema DR; Keesman, KJ. Leeuwarden

Okeya I, Hutton C, Kapelan Z (2015) Locating pipe bursts in a District Metered Area via online hydraulic modelling. Procedia Eng 119:101–110. https://doi.org/10.1016/j.proeng.2015.08.859

Pacchin E, Gagliardi F, Alvisi S, Franchini M (2019) A Comparison of Short-Term Water Demand Forecasting Models. Water Resour Manag 33:1481–1497. https://doi.org/10.1007/s11269-019-02213-y

Papageorgiou EI, Poczeta K, Laspidou C (2015) Application of Fuzzy Cognitive Maps to water demand prediction. IEEE Int Conf Fuzzy Syst 2015-Novem: https://doi.org/10.1109/FUZZ-IEEE.2015.7337973

Ponce MVC, Castañón LEG, Cayuela VP (2014) Model-based leak detection and location in water distribution networks considering an extended-horizon analysis of pressure sensitivities. J Hydroinformatics 16:649–670. https://doi.org/10.2166/hydro.2013.019

Pronzato L, Pázman A (2013) Design of Experiments in Nonlinear Models: Asymptotic Normality, Optimality Criteria and Small-Sample Properties. Linear Notes Stat 212:404

Pudar RS, Liggett JA (1992) Leaks in Pipe Networks. J Hydraul Eng 118:. https://doi.org/10.1061/(asce)0733-9429(1992)118:7(1031)

Puust R, Kapelan Z, Savic DA, Koppel T (2010) A review of methods for leakage management in pipe networks. Urban Water J 7:25–45. https://doi.org/10.1080/15730621003610878

Qi J, Sun K, Kang W (2015) Optimal PMU Placement for Power System Dynamic State Estimation by Using Empirical Observability Gramian. IEEE Trans Power Syst 30:2041–2054. https://doi.org/10.1109/TPWRS.2014.2356797

Qi Z, Zheng F, Guo D, et al (2018) Better Understanding of the Capacity of Pressure Sensor Systems to Detect Pipe Burst within Water Distribution Networks. J Water Resour Plan Manag 144:04018035. https://doi.org/10.1061/(asce)wr.1943-5452.0000957

Quevedo, Casín JJ, Cugueró, Escofet MÀ, Pérez, Magrané R, et al (2011) Leakage location in water distribution networks based on correlation measurement of pressure sensors. In: 8th IWA Symposium on Systems Analysis and Integrated Assessment. IWA, San Sebastian, pp 290–297

Quiñones-Grueiro M, Bernal-de Lázaro JM, Verde C, et al (2018) Comparison of Classifiers for Leak Location in Water Distribution Networks ∗. IFAC-PapersOnLine 51:407–413. https://doi.org/10.1016/j.ifacol.2018.09.609

Quintiliani C, Vertommen I, Van Laarhoven K, Van Der Vliet J (2020) Optimal Pressure Sensor Locations for Leak Detection in a Dutch Water Distribution Network †. 1–9. https://doi.org/10.3390/environsciproc2020002040

Ramos H, Covas D, Borga A, Loureiro D (2004) Surge damping analysis in pipe systems: modelling and experiments. J Hydraul Res 42:413–425. https://doi.org/10.1080/00221686.2004.9641209

Ray C, Benammar ASO (2002) Mean shift: A robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24:603–619

Reca J, Martinez J (2006) Genetic algorithms for the design of looped irrigation water distribution networks. Water Resour Res 42:1–9. https://doi.org/10.1029/2005WR004383

Rosario-Ortiz F, Rose J, Speight V, et al (2016) How do you like your tap water? Science (80-) 351:912–914. https://doi.org/10.1126/science.aaf0953

Rossman LA (2000) EPANET 2

Russo S, Lürig M, Hao W, et al (2020) Active learning for anomaly detection in environmental data. Environ Model Softw 134:104869. https://doi.org/10.1016/j.envsoft.2020.104869

Sander J (2010) Density-Based Clustering. In: Sammut C, Webb GI (eds) Encyclopedia of Machine Learning. Springer US, Boston, MA, pp 270–273

Sarrate R, Blesa J, Nejjari F, Quevedo J (2014) Sensor placement for leak detection and location in water distribution networks. Water Sci Technol Water Supply 14:795–803. https://doi.org/10.2166/ws.2014.037

Sarrate R, Nejjari F, Rosich A (2012) Sensor placement for fault diagnosis performance maximization in Distribution Networks. 2012 20th Mediterr Conf Control Autom

MED 2012 - Conf Proc 110–115. https://doi.org/10.1109/MED.2012.6265623

Schreiber T, Schmitz A (1997) Discrimination power of measures for nonlinearity in a time series. Phys Rev E - Stat Physics, Plasmas, Fluids, Relat Interdiscip Top 55:5443–5447. https://doi.org/10.1103/PhysRevE.55.5443

Scozzari A, Brozzo G (2017) Making use of continuous measurements for change detection purposes: An application to water distribution networks. I2MTC 2017 - 2017 IEEE Int Instrum Meas Technol Conf Proc. https://doi.org/10.1109/I2MTC.2017.7969738

Scozzari A, Mounce S, Han D, Soldovieri F (2021) ICT for Smart Water Systems: Measurements and Data Science

Slaats N (2015) Voorspellingsmodel conditiebepaling AC- leidingen

Snider B, McBean EA (2020) Improving Urban Water Security through Pipe-Break Prediction Models: Machine Learning or Survival Analysis. J Environ Eng 146:04019129. https://doi.org/10.1061/(asce)ee.1943-7870.0001657

Sophocleous S, Savić D, Kapelan Z (2019) Leak Localization in a Real Water Distribution Network Based on Search-Space Reduction. J Water Resour Plan Manag 145:04019024. https://doi.org/10.1061/(asce)wr.1943-5452.0001079

Sophocleous S, Savić DA, Kapelan Z, Giustolisi O (2017) A Two-stage Calibration for Detection of Leakage Hotspots in a Real Water Distribution Network. Procedia Eng 186:168–176. https://doi.org/10.1016/j.proeng.2017.03.223

Srirangarajan S, Allen M, Preis A, et al (2012) Wavelet-based Burst Event Detection and Localization in Water Distribution. J Signal Process Syst 72:1–16. https://doi.org/10.1016/j.cell.2014.06.027.H3K4me3

Steffelbauer DB, Fuchs-Hanusch D (2016) Efficient Sensor Placement for Leak Localization Considering Uncertainties. Water Resour Manag 30:5517–5533. https://doi.org/10.1007/s11269-016-1504-6

Sujin J (2014) Basics and Examples of Principal Component Analysis (PCA)

Tipping ME (2001) Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 1:211–244

Tomar A, Burton H V. (2021) Active learning method for risk assessment of distributed infrastructure systems. Comput Civ Infrastruct Eng 36:438–452. https://doi.org/10.1111/mice.12665

Tran VQC, Le D V., Yntema DR, Havinga PJM (2021) A Review of Inspection Methods for Continuously Monitoring PVC Drinking Water Mains. IEEE Internet Things J 1–1. https://doi.org/10.1109/jiot.2021.3077246

Van Huijgevoort MHJ, Voortman BR, Rijpkema S, et al (2020) Influence of climate and land use change on the groundwater system of the veluwe, the netherlands: A historical and future perspective. Water (Switzerland) 12:1–16.

https://doi.org/10.3390/w12102866

van Rijsbergen CJ (1979) {I}nformation {R}etrieval, 2nd edn. Butterworths, London

van Summeren J, van Leerdam T, Blokker M (2016) Investeringen en prestaties van sensornetwerken in het drinkwaterdistributienet. Nieuwegein

van Thienen P, Morley M (2018) CALLISTO: Comparison and combination of leakage-Detection and Localization Methods. In: KWR. https://www.kwrwater.nl/en/projecten/callisto/. Accessed 16 Jul 2021

Verde C, Torres L (2017) Modeling and Monitoring of Pipelines and Networks

Vewin (2017) Drinkwaterstatistieken

Vewin (2016) Dutch drinking water is of top quality as shown by international research. https://www.vewin.nl/english/News/Paginas/Dutch_drinking_water_is_of_top_quality_as_shown_by_international_research_20.aspx#:~:text=Infrastructure and drinking water prices&text=In the United States of America%2C 22%25 of the total,75 to 80 years old. Accessed 8 Jun 2021

Vitens Groeiende Watervraag. https://www.vitens.nl/over-water/groeiende-watervraag. Accessed 8 Jun 2021

Wald A (1943) On the Efficient Design of Statistical Investigations. Ann Math Stat 14:134–140. https://doi.org/10.1214/aoms/1177731454

Wang L, Zhang H, Niu Z (2012) Leakage prediction model based on RBF neural network. Adv Intell Soft Comput 114:451–458. https://doi.org/10.1007/978-3-642-03718-4_56

Watters GZ (1984) Analysis and control of unsteady flow in pipelines, 2nd edn. USA: Butterworths

Wei Z, Pagani A, Fu G, et al (2020) Optimal Sampling of Water Distribution Network Dynamics Using Graph Fourier Transform. IEEE Trans Netw Sci Eng 7:1570–1582. https://doi.org/10.1109/TNSE.2019.2941834

Wu Y, Liu S (2017) A review of data-driven approaches for burst detection in water distribution systems. Urban Water J 14:972–983. https://doi.org/10.1080/1573062X.2017.1279191

Wu Y, Liu S (2020) Burst Detection by Analyzing Shape Similarity of Time Series Subsequences in District Metering Areas. J Water Resour Plan Manag 146:1–12. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001141

Wu Y, Liu S, Smith K, Wang X (2018) Using Correlation between Data from Multiple Monitoring Sensors to Detect Bursts in Water Distribution Systems. J Water Resour Plan Manag 144:04017084. https://doi.org/10.1061/(asce)wr.1943-5452.0000870

Wu Y, Liu S, Wu X, et al (2016) Burst detection in district metering areas using a data driven clustering algorithm. Water Res 100:28–37. https://doi.org/10.1016/j.watres.2016.05.016

Xu B, Abur A (2004) Observability analysis and measurement placement for systems with PMUs. 2004 IEEE PES Power Syst Conf Expo 2:943–946. https://doi.org/10.1109/psce.2004.1397683

Xu Q, Chen Q, Ma J, Blanckaert K (2013) Optimal pipe replacement strategy based on break rate prediction through genetic programming for water distribution network. J Hydro-Environment Res 7:134–140. https://doi.org/10.1016/j.jher.2013.03.003

Xu W, Zhou X, Xin K, et al (2020) Disturbance Extraction for Burst Detection in Water Distribution Networks Using Pressure Measurements. Water Resour Res 56:1–17. https://doi.org/10.1029/2019WR025526

Xu Y, Zhang J, Long Z, Chen Y (2018) A novel dual-scale deep belief network method for daily urban water demand forecasting. Energies 11:. https://doi.org/10.3390/en11051068

Ye G, Fenner RA (2014) Weighted Least Squares with Expectation-Maximization Algorithm for Burst Detection in U.K. Water Distribution Systems. J Water Resour Plan Manag 140:417–424. https://doi.org/10.1061/(asce)wr.1943-5452.0000344

Zhang Z (2020) Hydraulic Transients and Computations

Zubaidi SL, Gharghan SK, Dooley J, et al (2018) Short-Term Urban Water Demand Prediction Considering Weather Factors. Water Resour Manag 32:4527–4542. https://doi.org/10.1007/s11269-018-2061-y

Summary

The Netherlands is a country with a relatively high availability of fresh water sources. However, due to the drier summers and limited water storage capacity, drinking water winning from groundwater increasingly exceeds permitted levels in recent summer periods. Therefore, even for a country with an abundance of fresh water, drinking water production and distribution became a top priority. A redeeming feature of the Dutch drinking water is the excellent and well maintained conduit networks. Although the water losses of Dutch drinking water infrastructure are relatively low, maintaining this ageing infrastructure is costly. Reducing water loss by optimizing management of the distribution network therefore presents one of the most important responsibilities of the water companies.

Prevention of water loss is not only important in reducing the loss of a valuable resource, bursts in drinking water conduits can also lead to contamination of the drinking water and damage to surrounding infrastructure. Conduit bursts are often caused by a combination of various factors, of which not all are easily identified. Although conduit bursts are complex processes, action can be taken to limit water losses. Traditionally, customers are used as surrogate sensors in order to facilitate reactive leakage control. Recent methods focus on more proactive leakage control strategies, concerned with preventing water loss by optimizing grid operation and management. These methods make use of techniques, such as timely replacement of ageing and damaged pipes, identification of weak assets, optimizing network pressures and valve operations.

The information contained within the currently collected big data sets can be compared with an iceberg, where we only currently consider the tip visible above water, while not yet aware of the vast mass of information still hiding underwater. The value of this submerged information has become apparent to water distribution companies. A major challenge faced by drinking water companies is, therefore, how to expose the submerged iceberg of information in the data. Both machine learning algorithms and hydraulic models offer never seen before opportunities to extract concrete information from the real-time data.

Chapter 2 describes one such proactive leakage control algorithm, that can be used for timely detection and tracking of recurring patterns of pressure/flow anomalies. Recurring anomalous events of a similar nature can be indicative of misuse or malfunctioning of the water grid. Early warnings at first notice of recurring problems can ensure response strategies are initialized, thus facilitating timely remedying of the underlying causes, and thus preventing potential pipe bursts or asset malfunctions. Based on feature- or instance-based clustering of variable-size windows of prespecified anomalous sensor measurements, executed once per moving window, recurring patterns can be identified and visualized in a stacked area graph called a “fingerprint graph”. The method is completely unsupervised and can be run in real-time. Results based on two pressure sensor data sets indicate that high accuracy detection and tracking of recurring anomalous patterns is possible, achieving F1-scores of 92-94%.

Where detection of recurring anomalies is essential for proactive leakage control, timely detection of pipe bursts is essential for reactive leakage control. **Chapter 3** describes a water demand ‘nowcasting’ (predicting current demand) method capable of identifying expected water demand in real-time. A comparison between actual water demand and expected water demand will facilitate detection of water loss events. Due to the addition of exogenous signals from other sensors within the same DMA, non-burst deviations (Christmas, festivals, etc.) in water demand will be suppressed, thus increasing burst detection precision. The nowcasting method uses Bayesian ridge regression and Random Sample Consensus to provide demand estimate. Results based on three case studies indicate that exogenous nowcasting is indeed more accurate and robust, achieving Nash-Sutcliffe model efficiencies of 82.7%-90.6% compared to 57.9%-77.7% for traditional, historic measurement-based water demand forecasting methods.

Besides the robustness of burst detection methods, the accuracy of burst and anomaly detection is needed, as well. For high accuracy methods a high sensor density in the water grid is required. **Chapter 4** presents a linearized state-space model for optimal sensor placement in hydraulic fluid transport networks, based on network properties such as conduit resistances. The state-space formulation covers both pressure and flow dynamics. Network observability was used as a criterion for determining optimal sensor positions, without requiring dynamic simulation of hydraulic scenarios. The method is illustrated by two small case-study networks. Results indicate the robust sensor placement capacity of the methodology with regard to network observability.

In order to adapt the robust sensor placement methodology for real-scale networks, it was adapted to large networks and to simultaneous placement of multiple sensors. These adaptations, detailed in **chapter 5**, consisted of a) a more robust optimality criteria based on Gramian observability of the network for sensor placement, b) a skeletonization method in order to reduce network size and thus computational demand of optimal sensor placement, and c) an investigation of simultaneous and sequential placement of multiple sensors as well as visualization for these results was used. The adapted method was illustrated on two real-scale network models of the Hanoi, Vietnam and Balerma, Spain water networks. Results indicate skeletonization is sufficiently robust and thus warranted for faster use of the sensor placement method. Comparison of optimality criteria illustrates the robust nature of optimal sensor placement based on the determinant of the network’s observability Gramian for larger networks.

Samenvatting

Nederland staat bekend als land met veel beschikbaar zoet water. Toch worden de zomers er droger (uitzondering 2021), met steeds verdere dalingen van het grondwaterpeil tot gevolg. Waterland of niet, drinkwaterwinning en distributie blijft ook voor Nederland van groot belang. Gelukkig is de Nederlandse drinkwaterinfrastructuur goed onderhouden en van hoge kwaliteit, waardoor het waterverlies laag wordt gehouden. Echter zijn veel verouderde leidingen binnenkort aan vervanging toe, waardoor de toch al hoge kosten van het netwerkonderhoud blijven toenemen. Hierdoor blijven optimaal netwerkgebruik en waterverliesminimalisatie hoog op de agenda staan van de Nederlandse drinkwaterbedrijven.

Gesprongen of lekke leidingen zorgen niet alleen voor waterverlies, maar kunnen ook vervuiling van het drinkwater of schade aan omringende infrastructuur veroorzaken. De oorzaken van deze lekken zijn vaak veelvoudig en complex, wat detectie of preventie van waterlekken een lastig proces maakt. Traditioneel leunt een drinkwaterbedrijf op klanten om lekken te melden. Ook wordt er steeds meer ingezet of preventieve technieken met een focus op tijdig vervangen van verzwakte leidingen en een slimmer gestuurd netwerk qua e.g. waterdruk en sluitklep standen.

De informatie die gevangen zit in de big data aan huidige (sensor)meetwaarden, kan vergeleken worden met een drijvende ijsschots, waarvan slecht het topje boven het water uitsteekt, terwijl de grote meerderheid zich voor ons onderwater verborgen houdt. Dat deze verborgen kolos aan informatie van waarde is, wordt door waterbedrijven niet meer weersproken. Echter blijft het een uitdaging de volledige ijsschots boven water te krijgen. Gelukkig bieden artificiële intelligentie en hydraulische modellen veelbelovende mogelijkheden om zo veel mogelijk informatie te onttrekken aan de huidige en groeiende collectie verzamelde data.

Hoofdstuk 2 beschrijft een van deze veelbelovende algoritmes, dat gebruikt kan worden voor tijdige detectie en het volgen van terugkerende afwijkende patronen in huidige druk of debiet metingen. Herhaaldelijk optreden van eenzelfde afwijking kan duiden op foutief afgestelde of haperende onderdelen. Tijdige waarschuwingen voor deze terugkerende afwijkingen faciliteert tijdig ingrijpen, om lekken of andere consequenties te voorkomen. Het clusteren van lopende vensters van afwijkende druk of debiet tijdreeksen, gebaseerd op sensormetingen of afgeleide eigenschappen (“features”) hiervan, maakt het identificeren van deze terugkerende afwijkende patronen mogelijk. Deze terugkerende patronen worden zo gedetecteerd en gevisualiseerd in een gestapelde “vingerafdruk” grafiek. Deze methode kan live worden toegepast op binnenkomende tijdreeksen van druk of debiet metingen en is zelflerend zonder supervisie te vereisen. Twee casussen met druk datasets bevestigen de hoge accuraatheid van de methode, met F1-scores van 92-94%.

Naast het detecteren van terugkerende afwijkingen in druk of debiet data, is het tijdig detecteren van lekken met behulp van deze data ook zeer relevant. **Hoofdstuk 3** beschrijft een methode voor het real-time ‘nowcasten’ (voorspellen van huidige waarde) van de watervraag. Door gemeten en genowcaste watervraag te vergelijken, kunnen lekken (momenten met hoger dan verwachte watervraag) worden geïdentificeerd. Door ook de signalen van nabijgelegen sensoren mee te nemen in deze analyse, worden vals-positieve lekmeldingen onderdrukt. Evenementen zoals Kerst of festivals kunnen namelijk ook een verklaarbare watervraag toename teweegbrengen, zonder dat dit duidt op de aanwezigheid van een lek. De nowcasting methode maakt gebruikt van Bayesiaanse regressie met Tikhonov regularisatie alsook Random Sample Consensus om zo tot een robuuste nowcast te komen. Resultaten van drie casussen tonen de robuustheid van deze methode en bereiken Nash-Sutcliffe efficiëntie scores van 82.7%-90.6% vergeleken met 57.9%-77.7% voor traditionele voorspellende algoritmes slechts op basis van historische watervraag meetwaarden.

Herhaaldelijke patronen en lekken detecteren kent een hoge prioriteit, maar beide kunnen alleen met hoge accuraatheid worden toegepast as er voldoende (druk en debiet) sensoren geplaatst worden in het leidingnetwerk. **Hoofdstuk 4** introduceert een gelineariseerd toestandsruimte model van de dynamica van vloeistoftransport onder druk door een leidingnetwerk, bedoeld voor het bepalen van de optimale locaties voor sensorplaatsing in het netwerk. Deze modelrepresentatie maakt het mogelijk rekening te houden met de dynamica van drukverplaatsing en observeerbaarheid van het netwerk, zonder hydraulische scenario's dynamisch te simuleren. Deze methode wordt geïllustreerd aan de hand van twee kleine casusnetwerken, om zo sensorplaatsing inzichtelijk te maken.

Om sensorplaatsing via toestandsruimtemodellen te kunnen toepassen op echte leidingnetwerken, is deze methode geschikt gemaakt voor toepassing op grote netwerken en voor het simultaan plaatsen van meerdere sensoren. Deze aanpassingen, beschreven in **hoofdstuk 5**, bestaan uit a) een robuuster optimalisatiecriterium, b) een methode om netwerken te reduceren in grootte om de vereiste rekenkracht behoefte te verkleinen en c) een onderzoek naar de mogelijkheden van zowel sequentiële als simultane plaatsing van meerdere sensoren in een netwerk. Deze aangepaste methode is geïllustreerd aan de hand van twee grote werkelijke water transportnetwerken van Hanoi (Vietnam) en Balerma (Spanje). De resultaten laten zien dan de netwerk reductie methode robuust is en dus toegepast mag worden voor snelle sensorplaatsingsresultaten. Vergelijking van de optimalisatiecriteria bevestigt dat sensor plaatsing aan de hand van de determinant van de waarneembaarheids-Gramiaan het meest geschikt is voor grotere netwerken.

Acknowledgements

After four years of work, my PhD research is done, with this thesis as a result. Although quite an endeavor, there were a lot of people who made it an enjoyable four years. Here I hope to thank all these people.

Karel, thank you for being a super involved promotor, who always had time for a meeting and who I spoke to at least as much as my daily supervisor. Your positive attitude, patience and great ideas made my PhD way easier. I hope we can keep working together in my future endeavors. Doekle, thank you for the interesting discussions, as often about the research as about current affairs. With such a solution-oriented attitude and down-to-earth mindset, I think you are an excellent supervisor, independent of whichever field of study. Thanks for the excellent supervision, good conversations, and sailing trips. Jaap, thank you for your critical feedback and sharp helicopter view regarding my research. Although we unfortunately had to meet via Teams during the second half of the PhD, every meeting was still valuable and much appreciated. Hans, whose short contributing time to my project nevertheless had a great impact on my research. Thank you for the programming advice and discussions in the sun on the roof atrium of the Radix building.

Thank you to my Paranymphs, Hector and Sam, for the help facilitating the defense. Thanks Sam for the very Dutch coffee breaks used to discuss current events as “realists”, and thanks Hector for our talks about jobs and water, as well as your contagious laughter.

Rutger Kakes and Hilda Nyambura, thank you for your excellent work as MSc thesis students. It is remarkable how fast you managed to learn Python programming and machine learning within the limited timespan of a MSc thesis project.

Gijs, Natascha, Raquel, Sandra, Yang, Hakan, Fabian, Qingdian, Sebastian, Sara, Assala, Evelyn, Ha, Sam, thanks for the great office environment we managed to create together. I will miss the flying nerf gun bolts, the cornucopian amounts of candy, as well as the advice we gave each other during challenging parts of our PhD’s.

Hector, Nandini, Cao Vinh, it was great to work together in the Smart Water Grids theme. The sailing trips with Doekle were really great and I hope we get the opportunity to undertake something similar again in the future.

Thank you Peter and colleagues of PWN. The easy communication with Peter and speed with which research could be applied in practice was impressive. The actions taken by PWN to complete current envisioned DMA’s and to create dashboards for easy access and analysis of collected data will definitely contribute to fast implementation of future research. The active participation in research and critical view of their own infrastructure makes PWN a very robust and future-proof organization. I also what to thank Dave, Tim, and Martin for their help and contributions to the PWN case study.

I would like to thank Vitens for access to their data infrastructure and helpful feedback and discussion sessions. Thank you Eelco, Mario, Sjoerd, Mario, Yvonne, and Johan for the

insightful discussions and interesting conversations about the newest technical innovations in the drinking water sector. The focus and clarity employed by Vitens regarding the possibilities of machine learning and big data, make them one of the fastest innovators in the Dutch drinking water sector.

Thank you Roel and colleagues of Brabant Water, for the opportunities to share my research at your internal congress and for the change to apply my software solutions to an interesting case study.

Also thanks to Marcel from Wavin, for the smart and supportive advice during Smart Water Grids theme meetings, Thomas from Evides, for the change to help evaluate Evides' AI progress, and Rudy and Frank of Acquaint, for the interesting discussions about the future of AI in water distribution.

Then I'd like to thanks the excellent support staff at Wetsus, who made working there easier, but more importantly, more gezellig too. Thank you canteenies Gerbren, Riet, Catherina, IT help John, Rienk, Fabian, and swimming buddy Philip. Thanks also to all other wetsus colleagues who made my stay there a success, sorry I can't name you all personally in this brief acknowledgements section.

Slopera, Molfetta, Martijn en VS110, Dank voor een goede balans tussen werk en vrije tijd.

Mam, pap, Stef en oma, bedankt voor jullie steun tijdens dit PhD avontuur. Jullie diverse en unieke blikken op het leven zijn een inspiratie voor me en hebben ongetwijfeld bijgedragen aan creativiteit in mijn wetenschappelijk denken.

Tot slot, Manouk, dank voor je liefde en begrip. Tijdens drukke werkweken keek ik altijd uit naar ons weekend samen, en het is mede dankzij jou dat deze onderneming tot een succes is gemaakt.

About the Author

Caspar Vincentius Carolus Geelen was born in the Dutch city Weert on the 19th of December 1992. After high school Het College in Weert, Caspar choose to study in Wageningen from 2011-2017. He obtained a Bachelor's and Master's degree in Biotechnology with a specialization in Process Engineering, as well as a Master's degree in Bioinformatics from the Wageningen University. It was during his Master Thesis of Biotechnology that the possibilities of a PhD were brought to his attention by Prof. Dr. Karel Keesman. Having peaked his interest, Caspar decided in 2017 to continue with the process engineering and data science skills developed in Wageningen in the form of a PhD at the Biobased Chemicals and Technology chair group of the Wageningen University, performed at Wetsus, Leeuwarden, under the guidance of promotor Karel Keesman. His research would focus on data and model solutions for better drinking water transport.

List of Publications

Geelen, C.V.C., Yntema, D.R., Molenaar, J., Keesman, K.J., 2019. Monitoring Support for Water Distribution Systems based on Pressure Sensor Data. Water Resour. Manag. 33, 3339–3353. https://doi.org/10.1007/s11269-019-02245-4 (**Chapter 2**)

Geelen, C.V.C., Yntema, D.R., Molenaar, J., Keesman, K.J., 2021. Burst Detection by Water Demand Nowcasting based on Exogenous Sensors. Water Resour. Manag. 35, 1183–1196. https://doi.org/https://doi.org/10.1007/s11269-021-02768-9 (**Chapter 3**)

Geelen, C.V.C., Yntema, D.R., Molenaar, J., Keesman, K.J., 2021. Optimal Sensor Placement in Hydraulic Conduit Networks: A State-Space Approach. Water 13(21), 3105. https://doi.org/10.3390/w13213105 (**Chapter 4**)

Reyes Lastiri, D., **Geelen, C.V.C.**, Cappon, H.J., Rijnaarts, H.H.M., Baganz, D., Kloas, W., Karimanzira, D., Keesman, K.J., 2018. Model-based management strategy for resource efficient design and operation of an aquaponic system. Aquac. Eng. 83, 27–39. https://doi.org/10.1016/j.aquaeng.2018.07.001

Delgadillo, H.H., **Geelen, C.V.C.**, Kakes, R., Loendersloot, R., Yntema, D., Tinga, T., Akkerman, R., 2020. Ultrasonic inline inspection of a cement-based drinking water pipeline. Eng. Struct. 210, 110413. https://doi.org/10.1016/j.engstruct.2020.110413

SENSE

Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment

DIPLOMA

for specialised PhD training

The Netherlands research school for the
Socio-Economic and Natural Sciences of the Environment
(SENSE) declares that

Caspar Vincentius Carolus Geelen

born on 19 December 1992 in Weert, Netherlands

has successfully fulfilled all requirements of the
educational PhD programme of SENSE.

Wageningen, 25 january 2022

Chair of the SENSE board	The SENSE Director
Prof. dr. Martin Wassen	Prof. Philipp Pattberg

The SENSE Research School has been accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW)

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

The SENSE Research School declares that **Caspar Vincentius Carolus Geelen** has successfully fulfilled all requirements of the educational PhD programme of SENSE with a work load of 36.5 EC, including the following activities:

SENSE PhD Courses

- Environmental research in context (2017)
- Research in context activity: 'Organising a 6-day Python course' (2018)

Other PhD and Advanced MSc Courses

- A Systems and Control Perspective on Privacy, Safety, and Security in large-scale Cyber-Physical Systems, Dutch Institute of Systems and Control (2017)
- Design of Experiments, Wetsus (2017)
- Introductory Course, Wetsus (2017)
- Communication Styles, Wetsus (2018)
- Presentation Course, Wetsus (2018)
- Supervision Course, Wetsus (2018)
- Talent Course, Wetsus (2019)
- Career Perspectives, Wetsus (2021)

Management and Didactic Skills Training

- Organising a two day course on 'Regression in Python' (2019)
- Teaching assistant in the MSc course 'Parameter Estimation and Model Structure Identification' (2018-2020)
- Teaching assistant in the MSc course 'Systems and Control Theory' (2018-2020)
- Supervising two MSc students with theses entitled 'Predicting pipe degradation based on environmental variables' (2019) and 'Classification of daily patterns in sensor measurements for improved insight in water distribution system behavior' (2020)

Selection of Oral Presentations

- *Water Distribution Management: Big & Fast Data*. Amsterdam International Water Week Conference, 31 October1 November 2017, Amsterdam, The Netherlands
- *SMART detection and real-time learning in water distribution: Pressure Sensor based Monitoring Support*. Benelux Conference DISC, 27-29 March 2018, Soesterberg, The Netherlands
- *Predicting Pipe Bursts using Data Science*. Brabant Water Congress, 3 April 2018, Den Bosch, The Netherlands
- *Processing of Big & Fast Sensor Data for Failure Prediction*. Wetsus Members Only Congress: Wetsus annual workshop day, 19 April 2018, Leeuwarden, The Netherlands

SENSE coordinator PhD education

Dr. ir. Peter Vermeulen

It is the moment of discovery, the triumph of the mind, and the end of this thesis.

This work was performed in the cooperation framework of Wetsus, European Centre of Excellence for Sustainable Water Technology (www.wetsus.eu). Wetsus is co-funded by the Dutch Ministry of Economic Affairs and Ministry of Infrastructure and Environment, the European Union Regional Development Fund, the Province of Fryslân and the Northern Netherlands Provinces. The authors thank the participants of the research theme 'Resource Recovery' for the fruitful discussions and their financial support.

Financial support from both Wageningen University and Wetsus for printing this thesis is gratefully acknowledged.

Cover design by Caspar Geelen
Cover art by Manouk Senhorst
Printed by: ProefschriftMaken || www.proefschriftmaken.nl