UNIVERSITY OF ZULULAND

MASTERS DISSERTATION

Active Leakage Management With Bayesian Networks

Author:
Felix SILWIMBA
201833166

Supervisors:
Dr. Surafel. L. TILAHUN
Prof. Sibusiso. S. XULU

A thesis submitted in fulfillment of the requirements for the degree of Master of Applied Mathematics

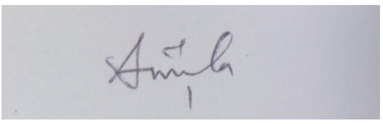
in the

Department of Mathematical Sciences

Declaration of Authorship

I, Felix SILWIMBA 201833166, declare that this thesis titled, " Active Leakage Management With Bayesian Networks", and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Submission Date: 10/12/2020

UNIVERSITY OF ZULULAND

Abstract

Faculty of Science and Agriculture
Department of Mathematical Sciences

Master of Applied Mathematics

Active Leakage Management With Bayesian Networks

by Felix SILWIMBA 201833166

Despite the existence of different active leakage management methods and models, their implementation in most developing countries' Water distribution systems (WDS's) remain limited. This is attributed to some limiting challenges to leakage management faced by these WDS. Some of these include; financial constraints, poor record-keeping systems, and inadequate technical skills and technology. This leads these distribution systems to adopt passive leakage management approaches that increase water losses and risk the destruction of neighboring infrastructure and water contamination. This study therefore presents a pipe leak monitoring and optimal maintenance sequence determining modelling framework that is applicable even in distribution systems with limited data.

The framework comprises of three models: a data adaptive Bayesian network (BN) model for predicting pipe leak probabilities used for leakage monitoring, a water loss estimation model for estimating pipe leak water losses and a linear programming model in which water loss estimates and pipe leak predictions are used to determine the optimal pipe leak maintenance sequence.

From the assessment and implementation example of the modeling framework which used simulated data, results indicate that the modeling framework is suitable for WDS with limited data.

Acknowledgements

I wish to express my sincere appreciation to my supervisors, Dr. Surafel. L Tilahun and Prof Sibusiso. S. Xulu for their guidance and patience through out this journey. To Prof. Bruce Bassett and the AIMS cosmology research group I am highly indented.

The financial support from the African Institute for Mathematical Sciences is truly appreciated. Without it, this project could not have reached its goal. Thank you to Umhlathuze municipality for the provision of sample data. Your kind gesture is highly appreciated.

Further more I would like to thank my elder brother Kalle Musalilwa Silwimba and my friend Blessing Nkatha for being emotionally supportive and helping my focus to maintain a bigger perspective. To the living waters church of the Nazarine in Manzini, I thank you for the online preaching services and bible studies. Your preaching and discussions nourished my spiritual life and kept me mentally and emotionally aligned through this journey. To my sait clair housemates, Precious Blessing Khumalo, Karabo Makola, Tshenolo Daumas, Evander Nyoni, Thabani Ncobo and Kiprono Koech. I thank you for being such great souls that my everyday was filled with motivation, laughter, exercise and learning. Thank you for creating an enabling environment.

Contents

List of Figures

List of Tables

List of Abbreviations

WDS	**W**ater **D**istribution **S**ystems
DMA	**D**istrict **M**etered **A**rea
NRW	**N**on **R**evenue **W**ater
BN	**B**ayesian **N**etworks

Chapter 1

Introduction

1.1 Background of Study

The problem of leakages is one that is experienced by every water distribution system (WDS). However, this is a challenge that should be closely monitored and lowered as much as possible because of its associated risks and costs. These include the loss of pumped water, revenue, and energy, the risk of water contamination, and damage the nearby infrastructure. Evidence suggests that 40% of water losses in distribution systems are attributed to leaks (Perdikou et al., 2014). This figure differs from one distribution system to another based mostly on their approach to leakage management.

Distribution systems adopt either the passive leakage control or active leakage control approach to leakage management. In the former leaks are only dealt with if they cause a dramatic loss to the supply or in response to public complaints. Whereas, in the latter, the WDS provides resources for leak detection and repair while closely monitoring the system for both seen and unseen leakages (WRDMAP, 2010). To keep leakages at their minimum an active leakage control approach is necessary.

One approach to active leakage management is through the three-phase technique – identify, locate and pinpoint (ILP) - described by (El-Zahab and, Zayed, 2019). In the first phase, one aims at simply identifying the existence of a leak. The minimum night flow is one of the most commonly used approaches in this phase. This phase is then followed by the **locate** phase which involves the identification of segments where the leak is. The suitable methods for this phase include leak noise correlation (LNC), step testing procedure and acoustic logging(AL). In the final phase of leak pinpointing, one seeks to determine the exact location of a leak with a certain level of accuracy. Some tools developed for this phase include: trace gas injection (TGT), pig mounted sensors (PMA), radar, and listening devices. These will be discussed in detail in chapter two.

Despite the necessity and advantages that come with active leakage control, not all distribution systems have adopted this leakage management policy. According to Water et al. (2012) WDS's in the developing world lack a systematic approach to leakage management, as such, they depend on passive leakage control policies. Frauendorfer and, Liemberger (2010) have also posited that passive leakage control is the approach practiced by most distribution systems in Asia. This is reflected in some of the high non-revenue water (NRW) losses from WDS's in the developing world. See Table 1.1 below for examples.

TABLE 1.1: NRW as percentage of total input volume for sample developing countries. The recorded high values are can be attributed to the passive approach towards water losses in most of these countries distribution systems.

Country/ Region	NRW(%)	Source
Central Africa Southern Africa	31 53	(Water-Operations-Partnerships, 2011)
Kosovo	58	(WorldBank, 2018)
Kenya India South Africa Malaysia	40 38 35 33	(Warming, 2019)
Zambia Tanzania	44 46	(Water-Operations-Partnerships, 2011)

This state of affairs can be attributed to some of the challenges faced by most WDS's in the developing world. Some of these include; the lack of a systematic way to check and identify leakages (Water et al., 2012; Farley et al., 2010), low funding, unskilled staff, ineffective record keeping, and lack of technology upgrades (Farley et al., 2010). Some of these challenges prevent them from obtaining the necessary requirements to implement the desired detection tools. Since, sensors, pressure loggers, or a sufficient skilled workforce to operate and maintain equipment are limited.

As such, this study presents an active leakage management tool that is applicable to distribution systems with limited resources. Which includes a data-adaptive Bayesian network model, a pipe leak water loss estimation function, and an optimization model.

1.2 Problem Statement

Despite the negative implications of passive leakage control, most WDS's in the developing world tend to adopt it. For example, Frauendorfer and, Liemberger (2010) indicated that "many utilities in Asia practiced passive leakage control". In addition, Water et al. (2012) indicated that the lack of systematic approaches to leakage management in the developing world leads to their dependency on passive observations for leakage awareness.

This state of affairs can be attributed to some of the challenges faced by most WDS's in the developing world. Some of these include; the lack of a systematic way to check and identify leakages (Water et al., 2012; Farley et al., 2010), low funding, unskilled staff, ineffective record keeping, and lack of technology upgrades (Farley et al., 2010).

This passive approach to leakage management exposes distribution systems to increased risks of water contamination, financial losses, reputational damage and threats to infrastructure in close proximity to the pipe. All these are not desirable outcomes.

For these reasons, this study presents an approach for systematic monitoring of pipe leaks and for suggesting their order of maintenance to minimize water losses. This approach also puts emphasis on flexibility of implementation on low-resourced distribution systems.

1.2.1 Aim of the Study

This study presents a modeling framework that promotes active leakage management in WDS with limited data.

1.2.2 Research Objectives

1. To assess the ability of the pipe leak prediction model to correctly learn model parameters through iterative data updates.
2. To Evaluate the pipe leak prediction model.
3. To provide an implementation example of the pipe leak monitoring and maintenance sequence determining modeling framework.

1.2.3 Significance of the Study

The use of the resulting modeling framework will enhance and promote active leakage management among WDS, thereby reducing the amount of fresh water lost to leaks. The greater demand for efficiency in WDS amidst global warming challenges and growing water demands justifies the need for a leakage monitoring tool that is also suitable for distribution systems in the developing world. Results based on the modeling framework will help the leakage management team in making maintenance decisions. This includes decisions like, the pipes to prioritize for inspection and maintenance, and replacement materials to prepare in readiness for possible pipe leaks. This will in-turn increase leakage awareness and response in a distribution system. This study also provides an opportunity to test the applicability of Bayesian network models to the management of pipe leaks.

1.3 Scope and Limitation of the Study

The main focus of this work was to develop a pipe leak management tool to enhance active leakage management in low resourced WDS. The tool is based on three models, a Bayesian network model for pipe leak monitoring, a water loss estimation model and an optimization model used together with the Bayesian network model to determine the optimal pipe leak maintenance sequence. The proposed tool is limited to establishing model variable relationships based on literature and not statistical exploration.

One of the limitations faced in this study is the availability of complete data on all variables of interest, in which case literature-based analysis will be used.

1.4 Research Outline

This dissertation contains six chapters which include the introductory chapter, literature review, theoretical concepts, development and implementation of the modelling framework, results and discussion and conclusion and recommendations. The outline is presented thematically below.

- Chapter One: This chapter introduces the problem of leaks and the associated risks and costs to water distribution systems. It then discusses passive and active leakage management policies, emphasizing the process they need for active leakage management to keep leakage levels at a minimum. It further alludes to the lack of implementation of active leakage policies in most developing countries and suggests a modeling framework for active leakage management even on distributions with limited data. It also highlights the research objectives, problem statement, scope of the study and the significance of the study

- Chapter Two: This chapter reviews extant literature that has bearing on leakage maintenance. It also covers relevant information on factors affecting leaks and their relationships to leak occurrences.

- Chapter Three: This chapter presents the theoretical concepts for the development of the pipe leak prediction model and defining the optimization problem for optimal maintenance sequence determining.

- Chapter Four:

  The chapter presents the development and implementation of the modeling framework for pipe leak prediction and determining the optimal maintenance sequence to minimize water losses. It provides an example development of the modeling framework on a real-world distribution system with limited data. Based on simulated data, discussions on assessing the modeling framework on parameter updating, pipe leak predictions, and an example implementation of trained models for decision making is provided.

- Chapter Five:

  Results and discussions of the assessment and implementation examples discussed in chapter four are presented in this chapter. It provides as part of the example implementation the key output results and their interpretations for pipe leak monitoring and optimal maintenance sequence determining.

- Chapter Six: This chapter summarises the study, highlights its findings and contributions, points out its recommendations based on the interpretation of findings, and suggests directions for future studies.

Chapter 2

Literature Review

2.1 Leakages in Water Distribution Systems

In water distribution systems, leakages refer to the physical losses of water from the system through pipes, joints, fittings, and overflowing service reservoirs. However, most significant losses occur from leaking pipes (pipe-bursts and joint raptures) (WHO, 2001).

Scholars have attributed the occurrence of pipe leaks in WDSs to a number of factors. Tables 2.1 and 2.2 below show some of these factors and their effects on leaks.

The occurrence of leakages has impacts on infrastructure, WDSs finances, customers and their health (WHO, 2001). Underground voids caused by water leaks lead to the collapse of infrastructures such as roads, buildings, and pipe bedding. Whereas, financial cost implications of leakages on WDS's include the costs of repairing and compensating for damaged infrastructure, increased pumping, treatment and electricity costs. With regards to WDS's customers, leakages cause inconveniences such as failed supply or reduced pressure on taps or other water-dependent appliances. More so, in cases of intermittent supply, air introduced in the system causes damage to meters and leads to the over-measurement of true consumption. Also, low-pressure systems and intermittent supply can lead to the infiltration of contaminants in the system, thus leading to the deterioration of water quality and health risks to consumers. Leakages also lead to increased administrative costs to deal with increasing consumer complaints and reputational damage.

2.2 Leakage Management

Generally, they are two different policies for pipe-leak management in water distribution systems namely: passive and active leakage management policies. In passive leakage management, WDS's have low-level commitment and activity towards the management of leaks. Whereas, in active-leakage management, WDS's are continuously monitoring the distribution system for signs of leaks (WRDMAP, 2010). To ensure the sustainable use of water resources and minimize the effects of leaks on the distribution system, its customers, and the neighboring infrastructure, adopting an active leakage management policy is preferred.

These policies will be discussed in detail in the section below and some example approaches and methods that support their implementation will also be discussed.

2.2.1 Passive Leakage Management

This is a leakage management policy in which only reported visual leaks or burst pipes that lead to dramatic pressure drops in the network are responded to for leakage maintenance (WRDMAP, 2010). According to Frauendorfer and, Liemberger (2010), about 90% of leaks in a distribution network are not visible on the surface as such most leakages take far too long to be noticed and repaired. This leads to increased exposure of the distribution system to the effects of leakages.

Despite the many shortfalls, passive leakage management minimizes the daily costs of leakage detection and can be applied in situations where the costs of leakage detection are high (Sharma, 2008). This makes it an attractive management approach, especially in distribution systems with low budgets.

2.2.2 Active Leakage Management

Active leakage management refers to the continuous monitoring of a distribution system for signs of leaks and to ensure continued detection and repair as leaks occur or as they are reported (WRDMAP, 2010). El-Zahab and, Zayed (2019) described a three-phase approach for leakage management involving the **identify** phase, **locate** phase then the **pinpoint** phase. Some of the common tools developed to help achieve these phases below. In addition, an additional leakage prediction class of tools that seeks to give future expectations about the leak status in a distribution system will be presented.

Identify Phase

This phase is aimed at discovering particular areas in the distribution system that have leaks without any information about their precise location. This is normally done using hydraulic models by monitoring and checking for anomalies in the flow, consumption, or pressure readings of a distribution system. One popular method used in this phase is the minimum night flow (MNF) described as a method for identifying leaks through measuring leakage flow rate by subtracting the legitimate night flow from the current flow rate into a district metered area (DMA) between 2 AM and 4 AM (Liemberger and, Farley, 2004).

However, Liemberger and, Farley (2004) also states that, identifying leakages with MNF is faced with challenges since the legitimate night flow is not always lowest in the early morning hours especially in distribution systems with intermittent supply where customers leave taps open to fill up storage containers.

Locate Phase

This phase involves the use of tools for isolating areas with leaks in the distribution system to narrow the sections searched for leaks. There is a variety of approaches and tools useful for this phase including, the step testing procedure, described by Boulos and, Aboujaoude (2011) and Puust et al. (2010), as a procedure that is conducted by systematically closing isolation valves during minimum night flow to isolate sections of a WDS within a DMA while continuously monitoring the water flow into that DMA. This effectively narrows down the leak manageable sections of a DMA and the number of pipes to be inspected for leaks. Other tools for locating leakages include, acoustic logging(AL) and leak noise correlation (LNC) that both use data from vibration sensors and hydrophones that are placed in accessible positions of a WDS's piping. The implementation of AL involves comparing sensor data obtained during the night to simulated data from a hydraulic model built with a no-leak assumption while adjusting to the distribution systems' conditions. LNC on the other hand, detects a leak by comparing the signals obtained from two sensors on either side of the leak.

Other model-based tools for locating leakages include, the Bayesian reasoning based model presented by Soldevila et al. (2016), a leakage detection tool based on a Bayesian probabilistic framework by Poulakis, Valougeorgis and, Papadimitriou (2003) and a Bayesian classifier based on a leak detection method presented by Soldevila et al. (2017). The model by Soldevila et al. (2016) detects leaks by analyzing pressure residuals based on a single fault assumption and Bayesian reasoning to explain the most likely cause of the observed residuals. The residuals are defined as the difference between pressure estimates obtained from a model simulated under free leak conditions and pressure records obtained from pressure sensors in a WDS. While the model by Poulakis, Valougeorgis and, Papadimitriou (2003) made use of pipe flow rate data obtained from sensors in a WDS. The model was aimed at obtaining optimal pipe flow and error models by maximizing the updated probability density function,

$$P(\theta,\sigma|\tilde{x}) = c_1 P(\tilde{x}|\theta,\sigma)\pi(\theta,\sigma), \tag{2.1}$$

obtained via Bayes theorem, where θ, σ are the parameters that specify the pipe flow and error models respectively, $\tilde{x}$ signifies the data on pressure heads and estimated flow rates, and c_1 is the normalizing constant. The authors applied the model to simulated data to detect both the location and severity of the leakage. This was done for both single and multiple leakage scenarios. With model errors below a certain threshold, the model was said to be capable of correctly identifying the location of the leakage and computing its severity. However with model errors above the given threshold, the diagnosis of leakages in the WDS was said to be impossible. In addition, by using a hydraulic simulator to calibrate the probability density functions for pressure residuals by considering demand uncertainty, leak size uncertainty, and sensor noise, Soldevila et al., 2017 estimated the location of leaks by applying Bayes theorem given for every time sample k, the posterior probability $P(l_i|r(k))$ that the leak l_i caused the observed residual vector $r(k) = (r_1(k) \cdots r_j(k))^T$ is given by equation 2.2.

$$P(l_i|r(k)) = \frac{P(r(k)|l_i)P(l_i)}{P(r(k))}, \qquad r = 1 \cdots N. \tag{2.2}$$

The residual vector $r(k)$ is defined as the difference between measurements provided by pressure sensors installed inside the DMA and estimations provided by a hydraulic model simulated under leak-free conditions, $P(r(k)|l_i)$ was considered the likelihood of the residuals $r(k)$ assuming that a leak l_i is active, $P(l_i)$ was considered the prior probability for the leak l_i, and $P(r(k))$ is a normalizing factor given by the total probability law.

The sensor dependency of most of the tools in this phase limits their implementation in low skilled and financed distribution systems. Also for a method like step-testing, there is a need for skilled personnel to open and close isolation valves.

Pinpointing Phase

This final phase to leak detection involves determining the exact location of leaks in the distribution system. It includes a variety of tools, such as, TGT and PMA by (Hunaidi et al., 2004), and GPR by (Eyuboglu et al., 2003). TGT makes use of a gas analyzer to detect leaks by analyzing the gas on the surface of the piping in search of a traceable, non-toxic, and lighter than air gas that is injected in the piping under pressure. The traceable gas on the surface is the gas that has escaped through the leaks in the piping and risen to the surface, hence evidence of leaks at those points in the WDS. Conversely, PMA uses microphones which are inserted into the piping system under pressure. These microphones detect leaks by analyzing acoustic signals as they are propagated through the piping by water flow. The trade-off for this precision in both methods is their high costs and requirements for highly skilled personnel (Puust et al., 2010). PMA also requires clean pipe interiors and this makes its application on old WDS a challenge. Ground-penetrating radar (GPR) is a leakage management approach implemented by analyzing measurements of the reflected electromagnetic waves that are directly sent into the ground. Here a skilled worker has to drive the GPR equipment on the ground above a pipe and constantly interpret the displayed results. Even though it is a good alternative for monitoring non-metallic and large diameter pipes, its performance can be affected by anomalies in the ground such as metal objects, cold temperatures, and in the event of leak detection in pipes buried near 2 meters deep (Puust et al., 2010).

The costs and skills required by these models make their implementation in WDS running on low budgets with limited skilled personnel difficult.

2.2.3 Leakage Prediction Models

Beyond locating pipe leaks that have already occurred, there is need to have a system of predicting future or expected leaks and their likely locations. There are leakage prediction tools which provide useful information for maintenance planning and help towards preemptive leakage management. Some of these tools are the Artificial Neural networks (ANN) model presented by Asnaashari et al. (2013), Weibull proportional hazard (WPH) model by Le Gat and, Eisenbeis (2000), incidence prediction model by A. Ogutu, Kogeda and, Lall (2016), and pipe and isolation valve failure prediction model by G. A. Demissie (2017). Firstly, the ANN was developed to predict failing water mains for purposes of assigning priorities for maintenance and repair. The model outperformed a multiple linear regression model with $R^2 = 0.94$ compared to $R^2 = 0.75$ based on data from a Canadian WDS. The authors emphasized that for ANN to create sufficiently diverse connections and improve its ability to create predictions, a large data set with sufficient variability is required. Secondly, the WPH model was used for forecasting failures in distribution systems based on maintenance records. The model linked the distribution of time to failure with the factors known to influence leaks. This led to the formulation of the survival function that was then solved using Monte Carlo simulations. Thirdly, the pipe leak incidence prediction model used Bayes theory to estimate the unknown parameter θ (the leakage incidence). The authors considered θ and other unidentified values as random variables with their probability distributions being derived from prior information and newly obtained data. This was done through the pipe deterioration equation given by.

$$G = \sum_{i=1}^{I} \lambda_i x_i, \tag{2.3}$$

where G is the pipe deterioration rate, i is the respective variable and I is the total number of variables all of which influence the pipe deterioration rate, and x is the factor position. Lastly, the pipe and isolation valve failure prediction model was based on three Bayesian network models, including a Bayesian belief network (BBN) to obtain soil corrosivity index while accounting for interdependencies among soil parameters, a dynamic Bayesian network (DBN) to predict annual and monthly pipe failure rates, and a Bayesian model averaging (BMA), to integrate climate projection data for future pipe failure forecasting. The variables in the BBN included soil resistivity, redox potential, soil sulfides content, and soil PH. The DBN model utilized pipe attributes data such as pipe material, pipe diameter, year of installation, pipe length, and operational data such as the number of residential and service connections. Whereas the environmental factors considered were temperature and precipitation.

Good quality data requirements by some predictive models like ANN makes their adoptions for use challenging especially in distributions with data collection and quality issues. Despite not being admirable, leakage data in some WDS's, especially those in developing countries, data on the distribution system is known to be unreliable, incomplete, and even unavailable (Dighade, Kadu and, Pande, 2014; G. A. Ogutu, Okuthe and, Lall, 2017). In such cases, models based on a Bayesian framework that allows for the incorporation of expert knowledge make them preferable (A. Ogutu, Kogeda and, Lall, 2016).

TABLE 2.1: Common causes of leakages in WDS's. Explored for the identification of possible pipe leak modeling variables.

Cause	Qualitative relation to pipe-leaks	Citations
Pipe material	Asbestos pipes have the least ruins with the most ruins occurring in iron pipes	(WHO, 2001; A. G. Ogutu, Kogeda and, Lall, 2018; G. A. Demissie, 2017; Saghi and, Aval, 2015)
Pipe age	leakages increase with increase pipe age	(WHO, 2001; Kabir, G. Demissie, et al., 2015; Farmani et al., 2017; G. A. Demissie, 2017; Saghi and, Aval, 2015; Parvizsedghy et al., 2017)
Pipe diameter	The smaller the pipe diameter the higher the risk of leakage incidences	(A. G. Ogutu, Kogeda and, Lall, 2018; Saghi and, Aval, 2015; Kabir, G. Demissie, et al., 2015; Farmani et al., 2017; Parvizsedghy et al., 2017)
Pressure	High pressure causes increases in leaks	(WHO, 2001; Saghi and, Aval, 2015)
Soil movement around pipe	Sudden soil movements or significant soil movements over time will lead to cracks and loose joints in the piping thus leading to leaks	(WHO, 2001; Saghi and, Aval, 2015)
Corrosion	The presence of corrosive elements such as salts in both the environment and water leads to the deterioration of system components such as the piping, valves and meters which in turn leads to leakages	(Saghi and, Aval, 2015; A. G. Ogutu, Kogeda and, Lall, 2018; Kabir, G. Demissie, et al., 2015)
Hit	In cases of digging by other organizations such as road maintenance, gas and electricity, these may directly damage piping systems, thus leading to leakages	(Saghi and, Aval, 2015)

TABLE 2.2: Common causes of leakages in WDS's. Explored for the identification of possible pipe leak modeling variables.

Cause	Qualitative relation to pipe-leaks	Citations
Water Hammer Hit	Direct pumping of water into the distribution system causes water hammer hit which may lead to breaks in the piping and joint, with eroded components being more susceptible.	(Saghi and, Aval, 2015)
Clogging on system components	The entry of solid material into the system leads to a build up in pressure which causes breaks, especially on the weak components	(Saghi and, Aval, 2015)
Temperature	High temperatures reduce the resistance of polythene based pipes. Also very low temperatures lead to tension in the piping, all of which makes the piping susceptible to breaks	(G. A. Demissie, 2017; Farmani et al., 2017; Saghi and, Aval, 2015)
Traffic loading	Vibrations and loading from high traffic on buried piping is thought to cause breaks and failures in the piping.	(A. G. Ogutu, Kogeda and, Lall, 2018; Saghi and, Aval, 2015)
Leakage control methods	Whether passive or active leak control	WHO, 2001
Incorrect installation	Unprofessional or lack of use of standard installation techniques leads to leaks from the onset which may increase over time	Parvizsedghy et al., 2017; Saghi and, Aval, 2015

Chapter 3

Theoretical Bases

This chapter highlights the theories and key concepts used in this study. The theories used to understand leak monitoring and maintenance are the Bayesian network theory and discrete linear programming.

3.1 Bayesian Networks

Bayesian networks are a modeling framework in which relationships among domain variables are modeled as probabilistic dependencies among random variables. These models are represented as directed acyclic graphs whose nodes are the domain variables of interest and the edges represent conditional dependencies among these variables. These dependencies are encoded as conditional probability distributions and can be discrete, continuous, categorical, and mixed.

The flexibility to span across different types of random variables, while encoding dependencies in a graphical structure gives BN's the ability to combine varying knowledge sources in the modeling process. It also makes them suitable for both classification and regression problems. In the same vein, their ability to represent variable dependencies in an interpretable manner accounts for why they are used different domains. For example, BNs have found use in disease diagnosis, prognosis, and treatment selection in medical studies, spam filtering, document classification, and semantic search in natural language processing, anomaly detection problems, etc. Other advantages of Bayesian Networks include their explicit treatment of uncertainty and support for decision analysis, suitability for use on small and incomplete data sets, the possibility to learn the network structure from data, and quick responses to queries given a defined model (Uusitalo, 2007).

The rest of this chapter provides more details on the concepts in Bayesian networks. After this, the definition and key concepts in modeling with Bayesian networks are presented.

3.1.1 Definitions and Key Concepts

Before moving to the definition of Bayesian networks, it will be useful to look over some key concepts.

Key terms

- Parent-Child: direct edge from parent to the child node. Eg, With $X \to Y$, then X is a parent of Y and similarly Y a child of X.
- Markov Blanket: set of node's encompassing the nodes' parents, its children, and its children's parents.
- Root nodes: nodes without parents.
- Leaf nodes: nodes without children.
- Intermediate nodes: non-leaf and non-root nodes in a Bayesian network. Configuration: a possible joint combination of the values of a node's parents that represent a distribution over the child node for non-root nodes. Otherwise, it refers to the possible values of a root node.
- Markov property: encoded in model structure so that, a node is only directly dependent on its parents and no other nodes in the graph.
- Explaining away: for a node with more than one parent, observing a subset of these parents reduces the probability that other parents caused the effect in the child node.

3.1.2 Conditional Independence in Bayesian Networks

The list below outlines a summary of three situations in which conditional independencies arise based on the topology of Bayesian networks.

a. Casual chains; $X \to Y \to Z$, that is, Y depends on X and Z depends on Y. In this case X is conditionally independent of Z given Y written as;

$$(X \perp\!\!\!\perp Z|Y)$$

b. Common Cause;

$$X \leftarrow Y \to Z,$$

that is both X and Z are dependent on Y and also implies that X is conditionally independent of Z given Y as given in a.

c. Common effect;

$$X \rightarrow Y \leftarrow Z$$

, Y depends on both X and Z. In which case X is conditionally dependent on Z given Y and can be summarized as,

$$(X \not\perp\!\!\!\perp Z|Y).$$

3.1.3 Definition

A Bayesian network is defined as a pair $\mathbf{B} = \langle G, \Theta \rangle$ where G is a directed acyclic graph structure whose nodes (vertices) $\{X_1, X_2, \cdots, X_n\}$ represent random variables and the edges between then represent conditional dependencies among these random variables (**Ben-Gal2007**). Each Bayesian network **B** represents a unique joint probability distribution over its' set of random variables.

The other component $\Theta \supset \theta_{x_i|\pi_i} = P_{\mathbf{B}}(x_i|\pi_i)$, for each realization x_i of X_i conditioned on π_i the set of parents of X_i, is the full set of parameters for the joint probability distribution induced by the network structure G. With the Markov property represented in the network, this joint distribution can be factorized into a product of conditional distributions of child nodes given their parents (Ben-Gal, 2008). See 3.1.

$$\begin{aligned} P_B(X_1, X_2, \cdots, X_n) &= \prod_{i=1}^{n} P_B(X_i|Pa(X_i)) \\ &= \prod_{i=1}^{n} \Theta_{X_i|Pa(X_i)} \end{aligned} \tag{3.1}$$

This allows for the building of a full joint probability model by only specifying the conditional probability distributions of every node in **B**.

3.2 Modelling With Bayesian Networks

The process of modeling with Bayesian networks involves the satisfaction of the following steps presented by Korb and, Nicholson (2010).

1. Identifying the variables of interest: this will define the nodes of the network and their representative values.

2. Structure: in this step, the qualitative relationships among model variables are defined. This can be achieved by leveraging expert knowledge or by learning the structure from sufficient data.

3. Quantifying relationships of connected nodes; this can be achieved by annotation of experts or by parameter estimation or both.

After accomplishing the three steps, a BN model for the phenomena of interest will have been defined. It is then used to answer queries about nodes in the network based on some other observed nodes (inference/prediction). Further insight is given on how one can get through each one of the three stages below.

3.2.1 Variable Identification and Value Defining

After establishing the phenomenon of interest, the first step to developing an appropriate Bayesian network is the identification of suitable representative variables and their respective value ranges.

Typically, this step would involve seeking answers to such questions as: what variables best represent the phenomena of interest? What data type are these variables (continuous, categorical or discrete counts)? What is the range for continuous variables, and what are the possible outcomes of the categorical variables?

3.2.2 Developing Network Structure

After identifying the variables representing the phenomena of interest, the next step is defining the relationships among these variables by defining the network structure for the Bayesian network.

One way to do this is by making use of known variable relationships from experts/domain knowledge. This can be achieved by reviewing various research work and reports involving the domain and variables of interest while paying attention in the presented relationships among these variables. The other is by learning the network structure from data through the use of score-based approaches. This means that the network structure is learned by using a defined structure scoring function that evaluates the probability of a network structure given the data while simultaneously penalizing structures with many parameters. The two commonly used scoring functions are the Bayesian information criterion (BIC) and the Bayesian score. The former is defined in 3.2 below while the latter is defined in 3.3 below. For different possible network structures, the scores are calculated and the network whose structure maximizes the scoring function is adopted as the model structure.

For a given structure G, under complete observed data D, the BIC score is as given by Beretta et al. (2018) as;

$$score_{BIC}(G|D) = l(\hat{\theta}_G|D) - \frac{logM}{2}Dim(G), \tag{3.2}$$

where, $l(\hat{\theta}_G|D) = M\sum_{i=1}^{n} I_{\hat{p}}(X_i, Pa^G(X_i)) - M\sum_i H_{\hat{p}}(X_i)$ with M being the number of training instances, $I_{\hat{p}}(X_i, Pa^G(X_i))$ being the mutual information of node X_i and the joint configurations of its parents, $H_{\hat{p}}$ being the mutual information about a single node's various possible outcomes and $Dim(G)$ being the number of free parameters in the network

The first term of 3.2 computes the log-likelihood of a graph G given the data D which, if maximized separately, tends to lead to over-fitting since the log-likelihood is maximal in a fully connected network. The second term in 3.2 counters this over-fitting behavior by penalizing highly parameterized models. Notice here that the BIC scores are computed by first estimating the maximum likelihood parameter estimates $\hat{\theta}$ for nodes according to the structure G. Based on the conditions on G and D, Cao, 2014 presents the Bayesian score as given by equation 3.3.

$$\begin{aligned} score_{BDE}(G|D) &= \frac{P(D|G)P(G)}{P(D)} \\ &\propto \int P(D|\Theta_G, G)P(\Theta_G|G)d\Theta_G \end{aligned} \tag{3.3}$$

Where, $P(D|\Theta_G, G) = l(\hat{\theta}_G|D)$ and $P(\Theta_G|G)$ is the prior distribution for the network parameters

3.3 Parameter Estimation

From the joint decomposition of Bayesian networks stated in 3.1, the parameters of interest are the $\Theta_{X_i|Pa(X_i)}$. For purposes of estimating these parameters from data, either a maximum likelihood or Bayesian parameter estimation approach can be adopted. The objective of the maximum likelihood approach is to choose parameter values such that the likelihood of the data is maximized. Whereas, the Bayesian parameter estimation approach seeks parameter values that maximize the posterior distribution of the parameters given the data.

Despite both approaches giving similar parameter estimates, Bayesian parameter estimates provide extra benefits by allowing the initiation of prior distributions in which expert knowledge can be incorporated. It also provides a natural framework for updating parameter estimates based on new observations. There will be more discussion on this in subsequent parts of this chapter

To exemplify these two parameter estimation approaches, consider a discrete Bayesian network with conditional probability tables (CPT). With consideration of M complete observations, the parameters in the CPT's can be estimated separately (Cao, 2014). In which case, the log-likelihood function for each CPT parameter $\theta_{x|\pi}$ is given by 3.8.

$$l(x_i|\pi_i;\theta_{x|\pi}) = \sum_{m=1}^{M} logP(x_i^m|\pi_i^m;\theta), \tag{3.4}$$

where, x_i^m is the X_i value of the m^{th} observation and π_i^m is the configuration of the parents of X_i on the m^{th} observation in the observed data. And can be simplified as;

$$\begin{aligned} \sum_{m=1}^{M} logP(x_i^m|\pi_i^m;\theta) &= \sum_{m=1}^{M} log\theta_{x_i^m|\pi_i^m} \\ &= \sum_{x_i^m,\pi_i^m} C(x_i^m,\pi_i^m) log\theta_{x_i^m|\pi_i^m} \end{aligned} \tag{3.5}$$

Maximum Likelihood Estimation

The goal in this approach is maximizing equation for θ, after which the maximum likelihood estimate $\hat{\theta}$ is given by,

$$\hat{\theta}_{x|\pi} = \frac{C(x_i^m,\pi_i^m)}{\sum_{x_i'\in X_i} C(x_i',\pi_i^m)}, \tag{3.6}$$

where, $C(x_i^m,\pi_i^m)$ is the number of times the value x_i co-occurs with the parents' configuration π_i and $\sum_{x_i'\in X_i} C(x_i^m,\pi_i^m)$ is the sum over joint co-occurrences of all possible states of X_i for each possible parent configuration π_i^m.

Bayesian Parameter Estimation

For this approach, the goal is to compute the posterior distribution $P(\theta_{x|\pi}|D)$ over the parameters $\theta_{x|\pi}$ given a data set D with M observations. This is achieved through Bayes theory by taking the product of the likelihood function (first quantity on the right-hand side of 3.7) and the prior distribution (last quantity of 3.7) given as;

$$P(\theta_{x|\pi}|D) = P(D|\theta_{x|\pi}) \cdot P(\theta_{x|\pi}) \tag{3.7}$$

Considering again, a discrete Bayesian network with its node CPT's, we obtain the likelihood function over all $x_i \in X_i$ for a particular parent configuration $\pi \in \boldsymbol{\pi}$ as;

$$P(D|\theta_{x|\pi}) = \prod_{j=1}^{k} \theta_{x_j|\pi}^{C(x_j,\pi)} \tag{3.8}$$

For a fixed π from possible parent configurations. This is noted as a multinomial likelihood whose conjugate prior is the Dirichlet distribution. Thus leading to our choice of the Dirichlet distribution as the prior distribution over $\theta_{x_i|\pi}, \quad x_i \in X_i$; given as;

$$P(\theta_{x_j|\pi}) \propto \prod_{j=1}^{k} \theta_{x|\pi}^{\alpha_j - 1} \tag{3.9}$$

with 3.8 and 3.9 we have in possession the quantities for deriving the posterior distribution. That is;

$$\begin{aligned} P(\theta_{x_j|\pi}|D) &\propto \prod_{j=1}^{k} \theta_{x_j|\pi}^{C(x_j,\pi)} \prod_{j=1}^{k} \theta_{x|\pi}^{\alpha_j - 1} \\ &\propto \prod_{j=1}^{k} \theta_{x|\pi}^{(C(x_j,\pi)+\alpha_j)-1} \end{aligned} \tag{3.10}$$

Notice that 3.10 is the density function for a Dirichlet distribution with parameters $(\alpha_1 + C(x_1, \pi), \cdots, \alpha_k + C(x_k, \pi))$. As such we have that our posterior distribution over $\theta_{x_j|\pi} x_j \in X_i$ is $Dir(\alpha_1 + C(x_1, \pi), \cdots, \alpha_k + C(x_k, \pi))$. To obtain parameter estimates for the network we use the expected value of the posterior distribution. That is;

$$\hat{\theta}_{x|\pi} = E(P(\theta_{x_j|\pi}|D)) \tag{3.11}$$

This framework provides a natural way of updating model parameters based on observed data. Take for instance, starting with a prior Dirichlet $Dir(\alpha_1, \cdots, \alpha_k$, which is then updated to get a $Dir(\alpha_1 + C[x_1, \pi], \cdots, \alpha_k + C[x_k, \pi])$ posterior based on observation counts $C[x_1, \pi]$ of $x_1 \in X_i$ to $C[x_k, \pi]$ of $x_k \in X_i$.

After this initial update, the posterior parameters can be further updated given a new set of observations. This is done by considering the previously computed posterior distribution parameters as the prior and updating them as was done after the first set of observations. In which case, the prior will be $Dir(\alpha_1 + C[x_1, \pi], \cdots, \alpha_k + C[x_k, \pi])$ and after observing an additional L observations with $L[x_1, \pi]$ for $x_1 \in X_i$ to $L[x_k, \pi]$ for $x_k \in X_i$. The new posterior distribution will be $Dir(\alpha_1 + C[x_1, \pi] + L[x_1, \pi], \cdots, \alpha_k + C[x_k, \pi] + L[x_k, \pi])$. In this fashion, we can keep updating model parameters as more data becomes available.

3.4 Query Based Inference

After the complete estimation of the model CPT parameters, the Bayesian network can now be used for prediction and diagnostic tasks accordingly. The accomplishment of these tasks is based on the computation of conditional probabilities for variables of interest (query variables), based on some subset of observed (evidence) variables from the model. More formally, given a BN model over a set of discrete random variables $\boldsymbol{X} = (X_1, \cdots, X_n)$ partitioned into query (X_q), evidence (X_e) and unobserved (X_u) variables, the computation of the conditional probability is presented as;

$$P(X_q|X_e = e) \propto \sum_{X_e=e,X_u} P(X_q, X_e = e, X_u)$$

which decomposes according to the BN structure as;

$$\sum_{X_e=e,X_u} P(X_q, X_e = e, X_u) = \sum_{X_e=e,X_u} P(X_q|Pa(X_q))P(X_e = e|Pa(X_e))P(X_u|Pa(X_u)). \tag{3.12}$$

One way to compute 3.12 is by using the variable elimination algorithm. This is implemented by first representing CPT's in the network as factors. What remains is exploiting the product and summing out operations on factors to systematically eliminate non-query variables from the factorized joint distribution over the network variables.

3.5 Optimization Model

An optimization model consist of three major components namely; decision variables which represent physical quantities controlled by decision-makers, the objective function that defines the evaluation criteria of the quantity being optimized, and constraints that enforce restrictions on the decision variables (Ferguson, 2000). In cases where the decision variables are restricted to discrete values, the problem is referred to as a discrete optimization problem. Its the type of problem that was of interest in obtaining the optimal sequence for pipe maintenance to minimize water losses.

After defining the three components, the optimization model can be presented in matrix form as,

$$
\begin{aligned}
\text{Optimize} \qquad & z = \mathbf{c}^T\mathbf{x} \\
\text{Subject to} \qquad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\
& \text{or} \\
& \mathbf{C}\mathbf{x} = \mathbf{d} \\
& \text{or} \\
& \mathbf{x} \geq \mathbf{0}
\end{aligned}
$$

where z is the quantity to be optimized; $\mathbf{x}$ is a vector of decision variables, $\mathbf{c}$ is the vector of constants associated with the decision variables in the objective function, $\mathbf{A}$ and $\mathbf{C}$ are constant matrices associated with the decision variable and, $\mathbf{b}$ and $\mathbf{d}$ are vectors containing the constraining values for each corresponding constraint.

Given this presentation, standard solvers for optimization problems can be used to obtain values for the decision variables that optimize the objective function. As such, the main task in optimizing the order of pipe leak maintenance for minimum water losses was the definition of decision variables, the objective function, and the constraints. See section 4.2.3 for an implementation example in which we present and discuss the expected pipe leak water loss minimization problem

Chapter 4

Modelling Approach

4.1 Introduction

This chapter provides descriptions for the development, implementation, and assessment of the pipe leak monitoring and maintenance sequence suggestion approach. The first section describes the development of model components under this approach, while concurrently providing their example development for a real-world WDS with limited data. The second section provides descriptions for the implementation of the modeling approach after developing its modeling components. Lastly, based on simulated data, the third section provides an assessment of the pipe leak prediction model and an example implementation of the modeling approach.

4.2 Model Development

This section provides an overview of the development processes for the models in the pipe leak monitoring and maintenance sequence suggestion approach. It provides details around the main stages and requirements for their development, and software choices. In addition to the development processes, an example development of the models based on a real-world distribution system with limited data was provided. However, most of the concepts on which the presented models are based are presented in the previous chapter - chapter three.

4.2.1 Pipe Leak Monitoring Model

For purposes of monitoring pipe leaks in WDS, the use of Bayesian networks with Bayesian parameter estimation was proposed. This is because they provide a framework with which model parameters for a defined network structure are continually updated as data becomes available. These features help to account for distribution systems with limited data, wherein, a model can be developed based on pipe leak theory and then have its parameters get continually updated as more data becomes available.

Figure 4.1 below presents the main stages in the development and continued update of model parameters for the pipe leak prediction model.

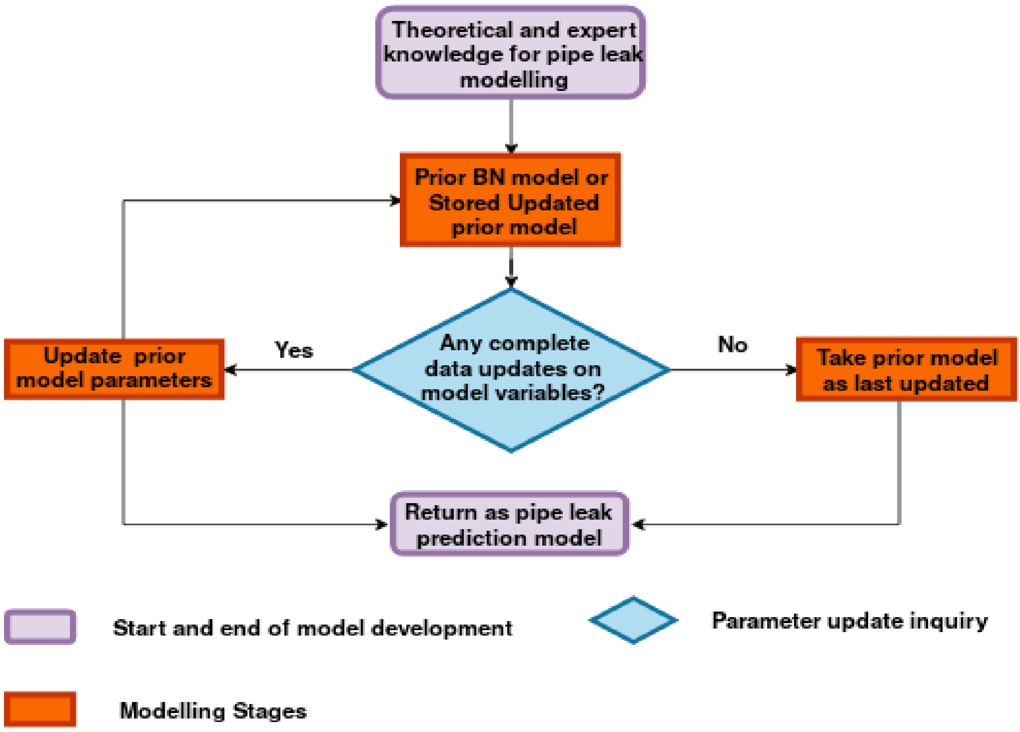


FIGURE 4.1: Main stages in developing and continually updating a pipe leak prediction model for pipe leak monitoring.

The first stage involves the review of theoretical and expert knowledge on pipe leak modeling to aid in defining the prior Bayesian network model. This includes reviewing literature and expert knowledge on factors affecting pipe leaks, their relationships to pipe leaks , and choosing the factors that apply to the distribution system of interest. This affords the opportunity to gather information to define the prior BN model for use in pipe leak prediction in that distribution system.

The second stage is dedicated to the development of the prior Bayesian network model based on the assessments in the first stage. This involves the development of the network structure, the choosing of node states, and prior parameter values for each node CPT.

In the third stage, data on the model, variables are used to update the network parameters as discussed in 5.2.1. In this stage, if data on model variables are updated, the network parameters of the prior model are updated and the posterior model is returned for use in pipe leak prediction. On the other hand, if no data updates are available the prior model as of the last data updates is returned for use in pipe leak prediction.

4.2.2 Water Loss Estimation

For purposes of estimating the amount of water lost from pipe leaks, the background leakage model presented by Germanopoulos (1985) and Giustolisi, Savic and, Kapelan (2008) was used. It is formulated as,

$$q_k = \begin{cases} \beta_k L_k (P_k)^{\delta_k} & \text{if} \quad P_k > 0 \\ 0 & \text{if} \quad P_k \leq 0 \end{cases} \tag{4.1}$$

where:

q_k = leakage flow rate from pipe ($\frac{m^3}{hr}$)
L_k = length of the k^{th} pipe (m)
P_k = average pressure through pipe k (m)
β_k = pipe degradation parameter
δ_k = constant dependent on pipe material

To calibrate (4.1) for use in pipe leak water loss estimation of a particular distribution system, the parameters δ and β need to be calibrated accordingly to suit the network. For this δ is calibrated using the information on pipe materials in the distribution system. According to Giustolisi, Savic and, Kapelan (2008), the range of values for δ_k is $0.5 - 2.5$ with plastic pipes exhibiting higher δ values due to their susceptibility to longitudinal splits. With metallic based pipe exhibiting lower values of δ due to the rigid nature of the material.

For the calibration of β, aggregated quantities of the terms in equation (4.1) are used. These include estimated water losses from the distribution system, the total length of its pipes, its average operational pressure, and a determined δ value based on the pipe materials in the system. To proceed with calibrating β, water losses from the distribution system can be estimated by either considering losses as a percentage of the total input volume as approached by Giustolisi, Savic and, Kapelan (2008), or by using recommended estimates such as those provided for WDS in South Africa by McKenzie (1999) while adjusting for pressure differences. The latter approach is adopted in this study since there was access to the recommended loss estimates. According to McKenzie (1999), distribution systems operating at 50m pressure head lose 60, 40, and 20 liters/kilometer/hour if they are respectively old, middle-aged, and young. In the event of a distribution system having an average pressure head reading different from 50m, adjustments for the pressure differences are made using

$$Flow_{p2} = Flow_{p1} \times PCF, \tag{4.2}$$

where:

$p1$ = pressure1 (m)
$p2$ = pressure2 (m)
$Flow_{p1}$ = flow at p1 ($\frac{m^3}{hr}$)
$Flow_{p2}$ = flow at p2 ($\frac{m^3}{hr}$)
PCF = pressure correction factor

and,

$$PCF = \left(\frac{p1}{p2}\right)^{\gamma} \tag{4.3}$$

where, γ is the power exponent and typically ranges from $0.5 - 2.5$ with suggestions of 1.5 for distribution systems with a mix of iron and plastic-based pipes, higher values for systems with mostly plastic pipes, and lower values for mostly cast iron-based piping systems (McKenzie, 1999).

Using the measured average pressure from the distribution over some time of interest and using equation (4.2) and the age of the system, the adjusted water loss estimate for the distribution system can be obtained.

The estimated water loss is then used together with, the total length of the piping, average pressure in the distribution system, and a determined δ value to estimate β from equation 4.1. These calibrated δ and β values are then used to define equation 4.1 for pipe leak water loss estimation in the distribution system of interest.

4.2.3 Maintenance Sequence Suggestion Model

For purposes of determining the pipe leak maintenance sequence that minimizes water losses, results from the pipe leak prediction and water loss estimation model are used to define the objective function of the optimization model. The other components of the model include; defining the decision variable and outlining corresponding constraints. These components as they relate to pipe leak maintenance sequence suggestions are discussed in detail in subsequent parts of this chapter.

Model Assumptions

For purposes of simplifying the optimization model the following modelling assumptions were made.

1. The time needed to attend to a leaking pipe is fixed, say T unit of time.
2. Due to limited manpower we cannot attend to more than one pipe at a time
3. The travel time between pipes is negligible and is not taken into consideration
4. The distance between pipes is not considered

Decision Variable

Consider an equally spaced discrete time interval $0, T, 2T, 3T, \cdots$, representing time intervals in which pipe fixing will occur. Define the decision variable x_k as the time interval in which pipe k is scheduled to be fixed. For example, $x_1 = 4$ means that pipe 1 is scheduled to be fixed in w^{th} time interval.

Objective Function

Results from both the pipe leak prediction and water loss estimation models are used to define the objective function. That is, let q_k be the estimated amount of water lost for some time T from pipe k if leakage happens and it is not fixed. In this case, a pipe k with $x_k = w$ will lose water for a time period 0 to $(w-1)T$. As such, the estimated water loss for this pipe will be $(w-1)q_k$. However, it is not definite that the pipes are leaking and thus the need to account for their probabilities of leaking through the pipe leak prediction model. In which case, the aim becomes defining and finding the pipe leak maintenance sequence that minimizes the total expected water losses. This is given for each pipe as ;

$$l_k = p_k q_k (x_k - 1) \tag{4.4}$$

where:

l_k = expected water loss from pipe k

from which the objective function (L) is composed as the sum of all the individual pipes' expected water loss defined as 4.5

$$\begin{aligned} L &= \sum_{k=1}^{M} l_k \\ &= \sum_{k=1}^{M} p_k q_k (x_k - 1) \end{aligned} \tag{4.5}$$

Inevitably, the desired result is the decision vector of pipe leak maintenance time intervals that minimize 4.5 and satisfy the constraints on the decision variables.

Constraints

The constraints related to the presented optimization problem are 4.6 and 4.7.

$$x_k = 1, 2 \cdots M \tag{4.6}$$

$$x_k \neq x_l \quad \forall \quad k \neq l \tag{4.7}$$

To enforce (4.7), the condition that no two pipes are fixed during the same time interval, we rewrite the constraint while incorporating the index of the pipes as;

$$|x_i - x_j| \geq \frac{|i - j|}{M} \tag{4.8}$$

Defined Optimization Model

Considering a system with M pipes, we define the optimization model as;

$$\begin{aligned} \text{Minimize} \qquad & L = \sum_{k=1}^{M} p_k q_k (x_k - 1) \\ \text{Subject to} \qquad & |x_k - x_l| \geq \frac{|k - l|}{M} \\ \text{and} \qquad & x_k, x_l \in \{1, 2, \cdots, M\} \end{aligned}$$

In this case the number of constraints in the objective function will be dependent on the number of decision variables (equivalent to the number of pipes in the WDS).

As a result of solving the optimization problem, we obtain the sequence of fixing pipes that minimizes pipe leak water losses. Since the distance between the pipes has been neglected, the present model definition is no different from simply fixing pipes in descending order of their expected water losses. However, the modeling framework has been chosen to facilitate future considerations for incorporating distance between pipes.

4.2.4 Model Defining on a Real-World WDS

For an example development of the leakage management tool on a real-world distribution system with limited data, Eskhawini's section H reticulation was used. This reticulation is located in Eskhawini township under uMhlathuze Municipality in KwaZulu-Natal Province of South Africa. The township was established in 1976 for middle-income residents (Ngubane, 2009). The province experiences warm weather patterns with minimum and maximum annual temperatures of 9.1^0C and 25.6^0C respectively (South African Weather Service, 2020). Based on the data obtained from Umthlathuze municipality, the recorded number of pipes in the reticulation was 931 with varying diameter sizes, pipe lengths, and pipe material types. See figures 4.4a, 4.4c and 4.4e for pipe counts with respect to pipe lengths, diameters and pipe material types.

Bayesian Network Model Development

While accounting for the reticulations' context, the construction of the Bayesian network pipe leak model involved four main stages, including the identification of model variables, defining variable states, identifying variable relationships, and defining prior network parameters to be updated later with observations. The preceding discussions present these model development stages as applied to the development of a pipe leak prediction model for Eskhawini reticulation.

Identifying Model Variables
The identification of variables for the Bayesian network model was done by reviewing the literature on factors affecting pipe leaks while considering the weather context and features of the reticulation. As a result, pipe diameter, pressure, pipe location, pipe length, strain, and previous breaks were identified as factors for inclusion in the model from the works of A. G. Ogutu, Kogeda and, Lall (2018), G. A. Demissie (2017), Kabir,

Tesfamariam, et al. (2015), and Saghi and, Aval (2015). Despite being a pipe leak affecting factor in freezing places, temperature was omitted due to its lack of significance in affecting pipe leaks in warmer climates. Soil type was also omitted under the assumption that all reticulation pipes of interest are buried under the same soil type.

Defining Variable States

After identifying suitable model variables, the next step was defining node variables. This involved, defining categories for numerical values and was done in two parts. The first part considered such variables as lengths, diameters and materials as data got from the Municipality. In this case, the values of pipe material were defined in three main categories of asbestos, plastic and steel. For both diameter and length, three state node values (high, medium, and low) were defined. The second part considered defining values for variables on which data could not be sourced from the Municipality.This included data on all other variables in the model but those identified. This also involved the review and assessment of how other scholars defined similar variables in modeling pipe leaks. See Table 4.1 for sample variable values as defined by other scholars.

Corresponding values for the identified model variables were defined by adapting some of the value definitions used by other scholars in Table 4.1, while adjusting their threshold values to represent the observed variable values or expected variable value ranges from obtained data. See Table 4.2 for the variable values as defined for the presented data-adaptive Bayesian network pipe leak prediction model.

Investigating Variable Dependencies

The third stage in developing the prior Bayesian network model involved investigating relationships among identified variables through a literature search. This was done by first exploring the relationships of pipe leaks with each pipe leak affecting factor. After this, the relationships among these factors were explored and exploited for use in defining the network structure of the Bayesian network model. This stage serves the role of providing the necessary information for defining the Bayesian network model apriori.

The relationships between leakages and the identified factors have been discussed and used in the works of other scholars on pipe leak prediction modeling. These include the positive relations of pipe leak occurrence with pipe age, pipe length, pressure, and previous breaks (Farmani et al., 2017;

TABLE 4.1: Variable values as defined by other scholars

Variable	Values	Thresholds	Source
Pipe diameter	small, medium, large		A. G. Ogutu, Kogeda and, Lall, 2018
Pipe diameter	100, 150, $\cdots$ 1650, 1950	As individual pipe diameters in mm	G. A. Demissie, 2017
Pipe Age	0-9,9-18,$\cdots$,88-96,96-105	As presented in years	G. A. Demissie, 2017
Pipe material	ACD, HDPE, Steel, upvc	-	A. G. Ogutu, Kogeda and, Lall, 2018
Pipe material	CI,DI,ST	-	G. A. Demissie, 2017
Location	Erf, exit	-	A. G. Ogutu, Kogeda and, Lall, 2018
Pipe length	small, medium, large	-	Kabir, Tesfamariam, et al., 2015
Pipe length	0-35, 35-55, $\cdots$, 240-310, 310-1190	As presented by intervals in meters	G. A. Demissie, 2017
Previous Breaks	0,1, $\cdots$,12,13	As actual count of previous breaks	G. A. Demissie, 2017
Pressure	low, medium, high		Kabir, Tesfamariam, et al., 2015
Pipe corrosion status	Yes, No	-	A. G. Ogutu, Kogeda and, Lall, 2018
Strain	low, medium, high	-	A. G. Ogutu, Kogeda and, Lall, 2018
Damage status	Yes, No	-	A. G. Ogutu, Kogeda and, Lall, 2018

Kleiner and, Rajani, 2002; Jafar, Shahrour and, Juran, 2010). With respect to pipe diameter, Kabir, G. Demissie, et al. (2015), A. G. Ogutu, Kogeda and, Lall (2018), and Farmani et al. (2017) reported a negative relationship between pipe diameter.

In addition, other inter-variable dependencies among the identified factors

TABLE 4.2: Variable values as defined for the pipe leak predictive model

Variable (symbol)	Values
Pipe diameter (D)	small, medium, large
Pipe Age (A)	0-15, 16-30, 31-45
Pipe material (M)	AC, Plastic, Steel
Location (Lo)	Erf, exit
Pipe length (L)	small, medium, large
Previous Breaks (Pb)	$0-7, 8-15, >15$
Pressure (P)	low, medium, high
Pipe corrosion status (C)	Yes, No
Internal strain (I)	low, medium, high
External strain (E)	low, medium, high
Pipe damage status (Dg)	yes, no

affecting leakages have been advanced by other scholars in their development of pipe leak prediction models. These include the effect of pipe material on the relationship of pipe age and failures. Jafar, Shahrour and, Juran, 2010 reported this relationship as being positive for pipe failures and age in iron pipes but with no effect in asbestos pipes. In their work, Adedeji et al., 2017 presented the effects of pipe diameter on the relationship between pressure and pipe leaks. In which the effects of pressure on leakages was highest in low diameter pipes and lowest in high diameter pipes.

Another chain of inter variable dependencies include those presented in the Bayesian network model by Kabir, Tesfamariam, et al. (2015) where pipe leak, pipe age, and pipe diameter first affect the physical features node, which then affects the failure risk index through structural integrity. Water pressure also affects the failure risk index through hydraulic integrity.

In the same vein, the modeled relationships in the pipe leak BN model by A. G. Ogutu, Kogeda and, Lall (2018) in which pipe material affects the pipes' corrosion status and partly affects the leakage probability of a pipe. Here, since plastic pipes do not experience corrosion, there is no chance of observing a corrosion-related leak event in plastic pipes. However, both iron and asbestos-based pipes have some chance of experiencing corrosion-related leak events. On the other hand, location affects the strain experienced by a pipe. This, together with the pipe's diameter affect the damage status of the pipe, which then affects the pipe leak probabilities. Here, small diameter pipes with high strain have the highest chance of becoming damaged. On the other hand, large diameter pipes with low

FIGURE 4.2: Caption

strain have the least chance of becoming damaged. Finally, corrosion and damage statuses affect the probability of leakage on a pipe. Leaks have the highest chance of occurring when both corrosion and damage status are highest and vice versa.
From the identified variables, their node values, their inter variable relationships, and their relationships to leaks a prior Bayesian network model were defined for Eshkhawini section H's water reticulation system. The model has prior Dirichlet parameters relating to some known relationships among variable values and uniform priors on CPT with limited information. See figure 4.3 for the presented prior Bayesian network model to be later updated with progressively obtained data on model variables.

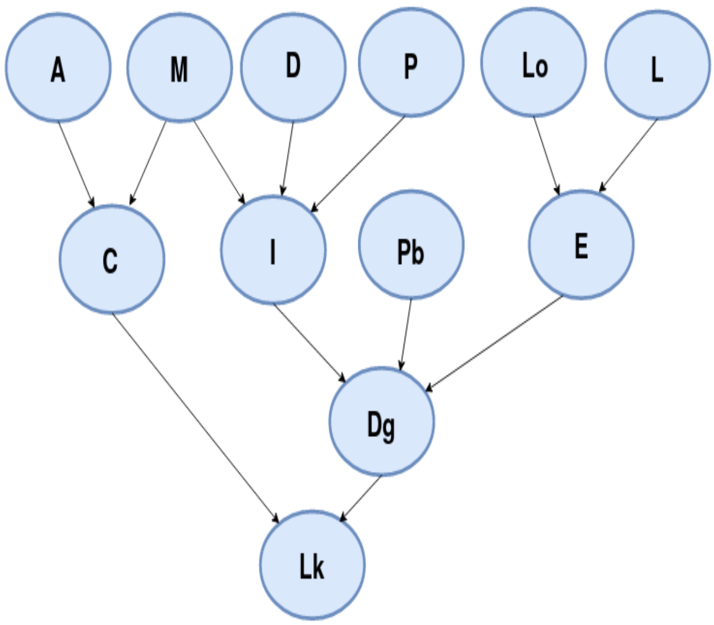


FIGURE 4.3: BN model structure for pipe leak predictions presenting variable relationships among factors affecting leaks with node names A, M, C, D, P, I, Lo, L, E, Pb, Dg, Lk representing pipe age, pipe material, corrosion status of a pipe, pipe diameter, pressure in the piping, the internal strain on pipe, pipe location, pipe length, the external strain on pipe, previous breaks and leak status of a pipe.

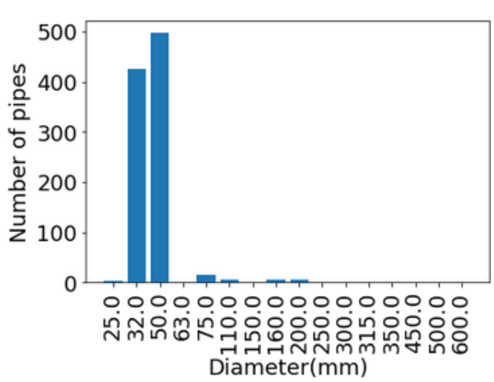
Figure 4.3 above presents the structure of the Bayesian network model for pipe leak predictions in Section H reticulation. The qualitative relationships encoded by the network structure include, the indicated dependency of pipes corrosion status on its age and its material type. The nodes from M, D, and P to I indicate the dependency of the internal stress experienced by a pipe on its diameter, material type, and the pressure in the system. Similarly, the nodes from Lo and L to E indicate that the external strain experienced by a pipe depends on its location and length. Nodes from I, E, and Pb to Dg indicate the dependency of the pipe damage status on the strains experienced by a pipe and its previous break history. Finally, the nodes from C and Dg to Lk indicate the dependency of pipe leak probability on the corrosion and damage status of a pipe.

Prior Model Parameters

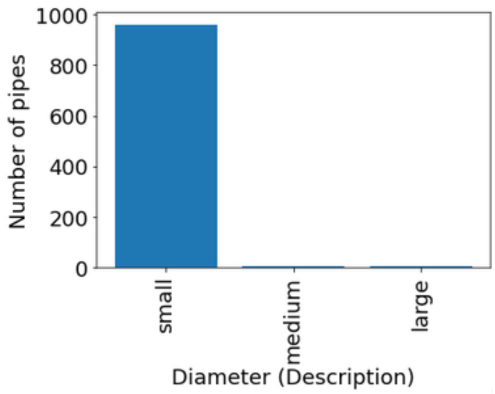
The final stage in defining the prior Bayesian network model for Eskhawini involved the definition/choice of prior model parameters. Having noticed that the model variables and their values represent multinomial distributions, their conjugate distribution (the Dirichlet) was chosen as the prior distribution for each node's multinomial distribution parameters. The distribution parameters were accordingly adjusted to represent observations and some theoretical and expert knowledge. This was done by first considering the nodes for which data on the reticulation was available. Based on the observed values of these variables, the prior Dirichlet parameters were defined. Whereas, for nodes with known relationships among their node values, their prior parameters were adjusted to reflect the known relationships. Lastly, for parameters with completely no information, their prior parameters were set as uniform Dirichlet priors. For the definition of the full network parameters, an effective sample size of 254, equal to the number of parameters was considered.

For prior Dirichlet parameter choices based on the transformed data variables presented in figures 4.4b, 4.4d and 4.4f, see tables A.2, A.6, A.3. The prior parameter choices for the other model variables are represented in tables A.7, A.8, A.19, A.4, A.9, A.11 and A.12 where they are defined by leveraging expert knowledge and in the cases were there was lack of information proper uniform priors were defined for corresponding conditional distributions. All parameter values scaled were accordingly to ensure the desired equivalent sample size.

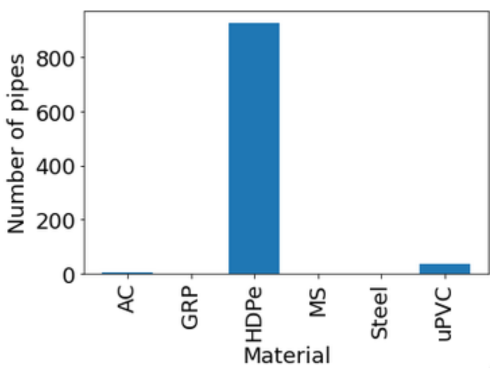
After fully defining the Bayesian network model, it can now be updated for use to make predictions on pipe leak behavior in the reticulation.



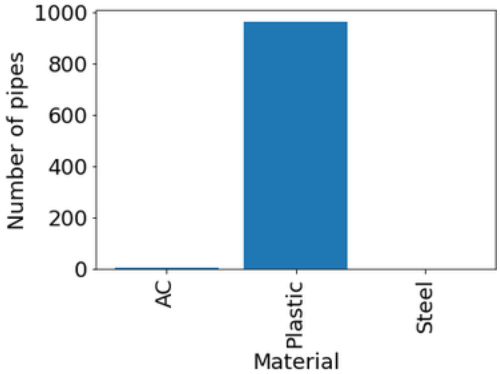
(A) Distribution of pipe counts on the raw pipe diameter values



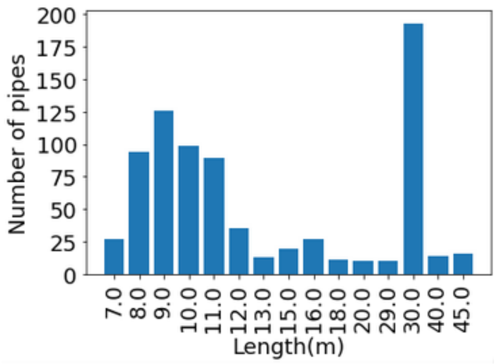
(B) Distribution of Pipe counts by pipe diameter descriptions.



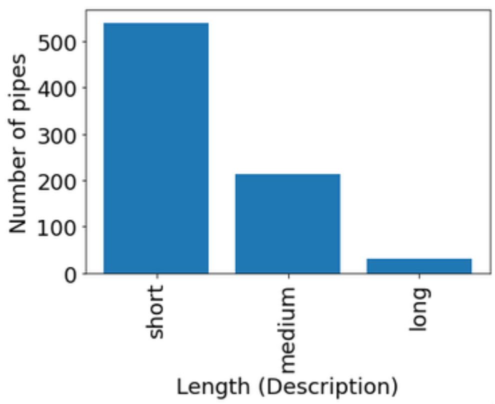
(C) Pipe count distributions on raw pipe material data



(D) Pipe counts distribution for pipe materials



(E) pipe counts distributions for raw pipe length data



(F) Pipe counts distribution based on descriptive categorised pipe lengths.

FIGURE 4.4: Pipe count distributions for the obtained modeling variables. With 4.4a,4.4c,4.4e indicating the distributions of the raw data 4.4b,4.4d,4.4f in the second column indicate the transformed data that will be used in developing the model.

Reticulations' Water Leak Losses

To calibrate the parameters δ_k and β_k in 4.1 for estimating water leak losses in Eskhawini section H, the high proportion of plastic-based pipes lead to the choice of 2 for δ. Whereas β_k was calibrated using the total length of the reticulation, average daily water pressure and the distributions water loss estimates computed using equations 4.1 and 4.2 as described in 4.2.2. From the data obtained for the distribution system, the calculated total pipe length for Eskhawini section H reticulation was $82606.97m$ and an average pressure head of $37.62m$ calculated from average hourly pressure readings over the period May 2018 to June 2019 was used as the pressure estimate for the distribution system. Using these figures the average total losses from the distribution system was estimated using 4.2.2, which was then used for the calibration of β. This way, the calibrated β value based on pressure, was 1×10^{-6}, leading to defining 4.1 for estimating pipe leak losses for Eskhawini section H's reticulation presented as equation 4.9 where δ is allowed to vary based on each pipe's material type.

$$q_k = (1 \times 10^{-6})LP^{\delta} \tag{4.9}$$

4.3 Model Implementation

After defining the pipe leak prediction and water loss estimation models, data from the reticulation can then be used to monitor pipe leaks through pipe leak predictions and estimate water losses. Further more, pipe leak predictions and water loss estimates are combined in an optimization model that is used to determine the optimal sequence in which to fix pipes to minimize pipe leak water losses that result from pipe leaks. See figure 4.5 for the implementation stages in using the developed models for pipe leak monitoring and optimal pipe leak maintenance sequence.

These stages involve the input of data on pipe characteristics and system pressure readings for distribution pipes of interest. Pipe characteristics refer to factors affecting pipe leaks in the developed Bayesian network model and those contributing to pipe leak water losses. For the pipe leak prediction model, there need not be complete observations on model variables to obtain pipe leak predictions.

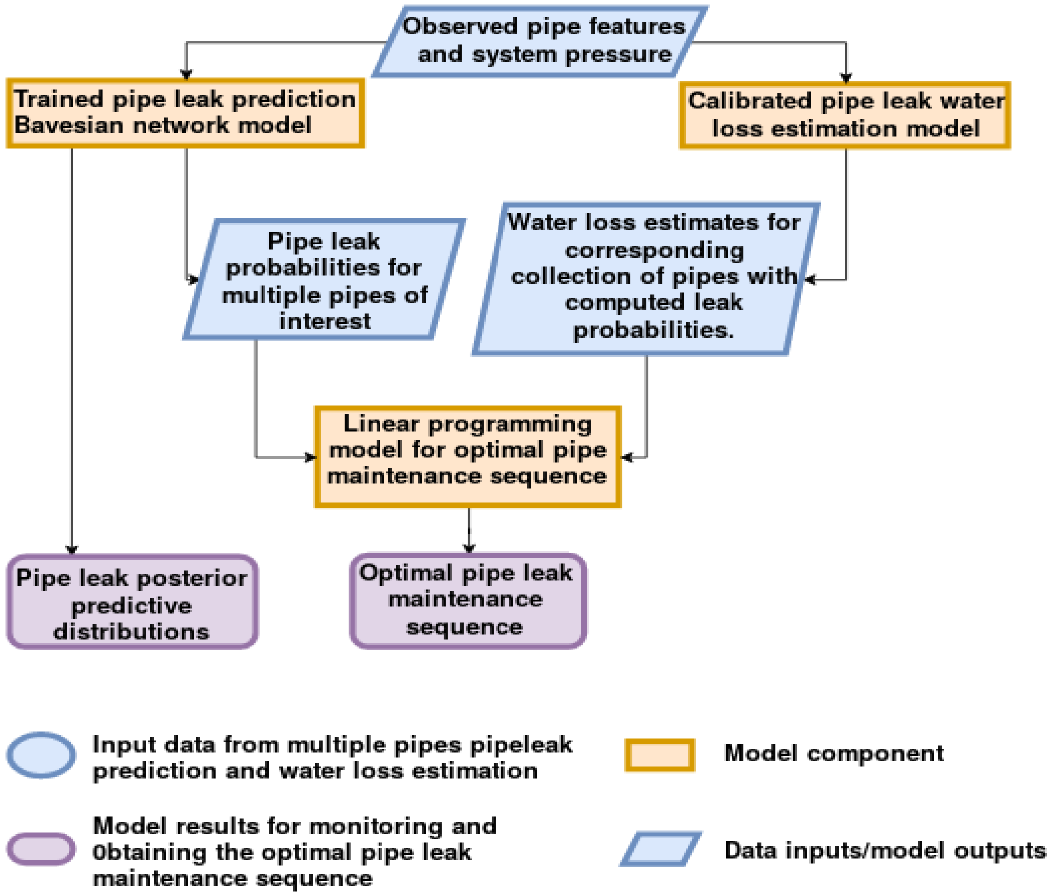


FIGURE 4.5: Flow chart of the developed pipe leak prediction and water loss estimation models. Indicating how the developed models are used to obtain pipe leak predictions used for pipe leak monitoring and to determine an optimal maintenance sequence based on observations on pipe characteristics and pressure measurements through the piping

In line with the objectives of monitoring pipe leaks and obtaining an optimal maintenance sequence, input data is used to obtain pipe leak posterior predictive distributions for monitoring pipe leaks. In addition, pipe leak probabilities are used in combination with estimated pipe leak water losses to determine the optimal pipe leak maintenance sequence.

4.4 Assessing the Modeling Framework

In this section, an assessment of the modeling framework is presented with a focus on the pipe leak prediction model. The water loss estimation model was also defined to further exemplify the implementation of the modeling framework and provide example model outputs for decision making.

With regards to assessing the pipe leak prediction model, two parts of the modeling process were of interest and they are parameter updating under the model development stage and model performance under the implementation stage. Before providing further details on the assessment of the pipe leak prediction model, a presentation of the data simulation process and the resulting example data was provided. Following this, a discussion on the assessment of parameter updating is presented. Thereafter, a discussion of model evaluation and the metric used for it are discussed. Lastly, for purposes of providing an example implementation of the modeling frame for monitoring pipe leaks and suggesting their optimal maintenance sequence, the calibration of a water loss estimation model is discussed.

4.4.1 Data Simulation

To assess the pipe leak prediction model and provide an example implementation of the modeling framework, a fully defined pipe leak Bayesian network model by (A. G. Ogutu, Kogeda and, Lall, 2018) was adopted. This allowed for the comparison of the updated parameters to the known parameter values. A. G. Ogutu, Kogeda and, Lall (2018) presented a Bayesian network model to assess leakage behaviour with the hypothesis that more leak events occur on road intersections and exits. In their work they model pipe leaks with as would be modeled for pipe leak predictions. The authors also published within their work all the model parameter values. For this reason their model was used to simulate data for use in assessing the proposed pipe leak prediction and maintenance sequence determining model.

See figure 4.6 for the model structure of the Bayesian network model upon on which data was simulated. See A.2 for the networks corresponding known model parameters (node conditional probability tables).

The nine variables represented in 4.6 include; soil type (S), material type (M), location (Lo), diameter (D), Strain(Sn), corrosion status (C),defective (Df), damaged status (Dg), and leak status(Lk). Each with their respective possible values are shown in table 4.3

The variable dependencies presented in figure 4.6 represent the joint effect of soil type and material type on the corrosion status of a pipe. The network also represents the effect of location on the strain experienced by a pipe. All these combine with the pipe diameter to determine the pipe's damage status. Finally, pipe's corrosive, defective, and damage statuses jointly determine the leakage status.

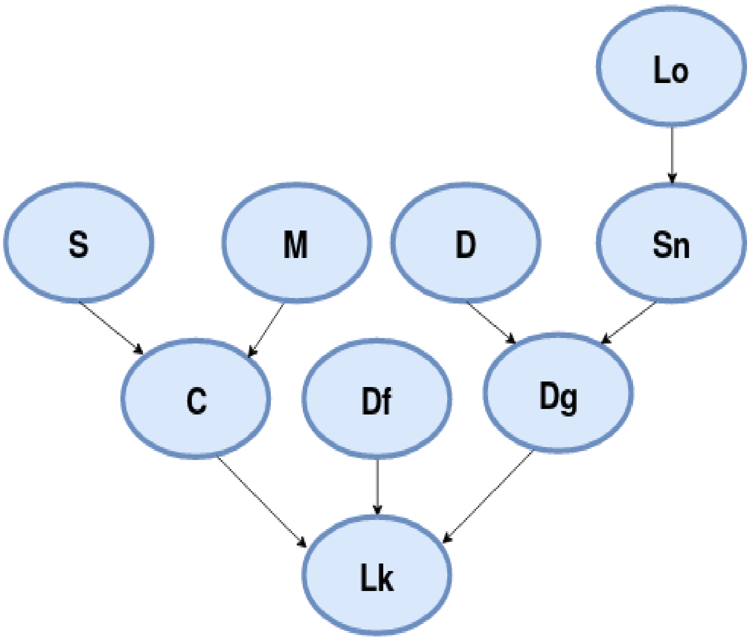


FIGURE 4.6: Bayesian network structure for data simulation from (A. G. Ogutu, Kogeda and, Lall, 2018)

Simulation Algorithm

Given a fully defined Bayesian network model from which random samples of its variables are to be obtained, the forward algorithm for sampling Bayesian networks was employed. The idea behind the algorithm is to sample each model variable given its parents' values according to the topological order of the network structure.

To achieve this we sample from the multinomial distributions represented by $P(X_i|\pi)$, for a child node X_i given a fixed configuration of its parents' nodes $\pi \in \boldsymbol{\pi}$. That is;

$$P(X_i|\pi) \sim P(X_i = x^1, \cdots, X_i = x^k|\theta_i, \cdots, \theta_k) \tag{4.10}$$

Notice that for root nodes $P(X_i|\pi)$ reduces to $P(X_i)$. With this in mind the forward algorithm proceeds as follows:

TABLE 4.3: Model variables, their possible values and descriptions

Variable (Symbol)	Values	Variable description
Soil type (S)	Dolomite = 0, Regular = 1	The type of soil under which the pipe is buried
Material type (M)	Pvc, ACD, Metallic	The kind of material from which the pipes are made
Diameter (D)	Low, medium, high	The diameter of a pipe
Location (Lo)	Erf, Exit	Intersection or non intersection road sections
Strain (Sn)	No strain, low, high	The amount of external strain experience by a pipe
Corrosion (C)	yes, no	The state of corrosion on a pipe
Defective (Df)	yes, no	whether or not a pipe had a defective at point of installation
Damaged (Dg)	yes, no	whether or not a pipe has suffered a form of physical damage
Leak (Lk)	positive, negative	Refers to the leakage status of a pipe

Algorithm 1: Forward algorithm

Result: Samples
initialization (defining the variable ordering according to Bayesian network structure);
samples $\leftarrow$ array(M, N);
for *1 to M* **do**
  sample $\leftarrow$ array$(1, N)$;
  for *1 to N* **do**
    sample $\hat{x}_i$ from $P(X_i|\pi)$ based on variable ordering;
    add $\hat{x}_i$ to sample;
  end
  add the sample to samples;
end

To simulate data from the Bayesian network represented in figure4.6 with its corresponding CPT's, forward sampling under the RBN function of the BNlearn library in R was used. See table4.4 for the first five rows of the simulated samples.

TABLE 4.4: Sample simulated data generated for assessing the modeling frame work

St	Mt	Di	Lo	Cr	Sn	Df	Dg	Lk
regular	HDPE	Medium	Erf	No	Low	False	No	Negative
dolomite	Upvc	Small	Exit	No	Low	False	No	Negative
regular	Upvc	Medium	Erf	No	Low	False	No	Negative
regular	HDPE	Small	Exit	No	Low	False	Yes	Negative
regular	HDPE	Small	Exit	No	High	False	Yes	Negative

A noteworthy characteristic of the data is its class imbalance for the target variable (pipe leaks). See figure 4.7 for the frequency comparison of negative against positive classes.

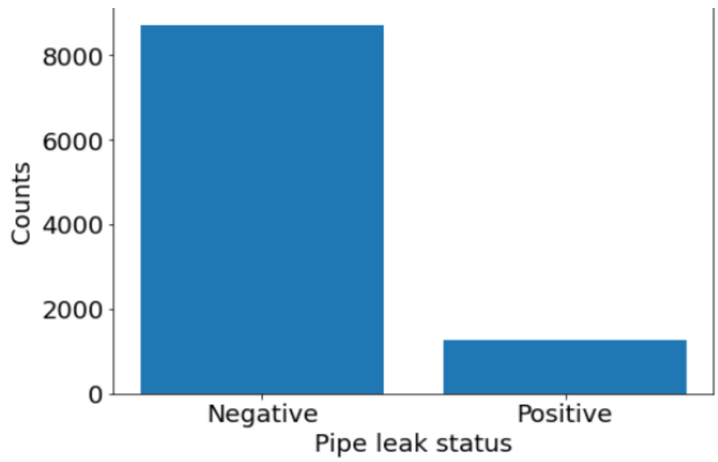


FIGURE 4.7: Pipe leak frequencies based on simulated training data. It indicates a high class imbalance against positive pipe leak status, thus influencing modeling decisions like the choice of the evaluation metric

4.4.2 Assessment and Implementation Descriptions

This subsection provides, descriptions of how the assessment of parameter updating and model evaluations were performed. Firstly, it was done by describing the assessment of parameter updating, with mention of the training sample size and how it was used to iteratively update model parameters. Secondly, it was done by describing model evaluation and the metric used for it. Thirdly, for the provision of an implementation example, illustrations for the calibration of the water loss estimation model are given. Lastly, descriptions for an example implementation of the modeling approach for pipe leak monitoring and maintenance suggestion were given.

The actual computations for the assessment of the pipe leak prediction model and implementation example leading to the results in Chapter 5 was done using python version 3.6. Specifically, the Pgmpy module with additional defined functions was used to assess parameter updating and evaluate model performance. For the implementation example, pipe leak predictions are obtained from the trained model with Pgmpy, the water loss estimation function was defined as expressed in section 4.2.2 and the optimization problem was solved using the Gekko API in python. The code is open for review on the authors GitHub account [1].

Assessment on Parameter Updating

For purposes of assessing parameter updating and model performance of the data-adaptive Bayesian network, 10000 data points were generated from the simulator. To simulate the updating of data in batches, the data is divided into 200 batches of equal size. Starting with proper uniform priors, the model parameters were then iteratively updated with each batch and their corresponding posterior distributions compared to the true model parameters from the data simulation model. Parameter updating was then assessed under the hypothesis that the 95% credible intervals of the posterior distributions will contain the true parameter values and narrow towards them as more data updates are performed.

Evaluation

To iteratively evaluate every model update, a test sample of 200 data points was generated and used to evaluate each updated model using the Matthews correlation coefficient (MCC). This metric is used for evaluating binary classification models trained on data set with imbalanced target labels (Boughorbel, Jarray and, El-Anbari, 2017). For instance, pipe leak data has very few leak events in comparison to non-leak events. MCC is computed based on quantities from the confusion matrix. See equation 4.11 for its formulae.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}. \tag{4.11}$$

[1] https://github.com/sillsphelyx/MscCode

where, $-1 \leq MCC \leq 1$ and the quantities, TP, TN, FP, FN are respectively the true positives, true negatives, false positives, and false negatives. Here, $MCC = 1$ represents perfect predictions, $MCC = 0$ represents random predictions and $MCC = -1$ represents completely wrong predictions.

The assessment of model parameter estimates as more updates are done, helps to establish the ability of the modeling framework to improve its parameter estimates as more data becomes available. Similarly assessing the evolution of the MCC metric for updated posterior models provides a way to monitor model performance as more data becomes available. Through this approach, every updated posterior model can be monitored for parameter estimation and evaluated for the prediction task. This is an important piece of information capable of guiding decision-makers on how much to trust the resulting model predictions.

Water Loss Estimation Model

To develop a water loss estimation model for an example implementation of the modeling framework on simulated data, additional data on pipe length and system pressure are required. For this, consider a hypothetical reticulation containing 100 pipes, with characteristics similar to those used in the pipe leak prediction model. The pressure values were adopted from (Strijdom, Speight and, Jacobs, 2017) and length values were kept below 300 meters for each pipe. See table A.22 for sample hypothetical input data containing pipe characteristics and hypothetical average pressure values for implementing the modeling approach.

Given the data in table A.22, the parameters of the water loss estimation model were calibrated as discussed in subsection 4.2.2. This was done using the pipe material types, lengths and pipeline pressure values.

Example Implementation on Simulated Data

To implement the modeling framework based on its assessed and developed pipe leak prediction model, and its calibrated water loss estimation model, data on each distribution system's pipe characteristics and average pressure head readings are required. For purposes of this implementation, consider the simulated and hypothetical pipe and pressure data in table A.22. With this data, implementation examples of the modeling framework on the two tasks of concern were given. Firstly, the examples were given on the pipe leak prediction model for purposes of pipe leak monitoring. Secondly, they were given on the water loss estimation model for results to use in conjunction with pipe leak predictions for determining the optimal pipe leak maintenance sequence.

Monitoring pipe leaks using the proposed modeling framework relies on the pipe leak prediction model. This is achieved by implementing the most updated Bayesian network model on pipe data to obtain each pipes corresponding pipe leak posterior predictive distributions. The posterior predictive distributions provide pipe leak probabilities while incorporating uncertainty from the Dirichlet distribution over the model parameters. In this way, each pipe's posterior predictive distribution can be viewed and used as a means to monitor the health of a pipe concerning leaks. As some of the model variables change, the posterior predictive distributions will be computed and adjusted accordingly.

After estimating both pipe leak probabilities and water losses for individual pipes, expected water loss estimates are computed for each pipe and used to define the objective function and constraints for the linear programming model presented in 4.2.3. Its optimal input values give the optimal sequence for pipe leak maintenance

4.5 Summary

This chapter discusses the pipe leak monitoring and maintenance suggestion approach that has been presented to enhance active leakage management in WDS's. The development of its model components is discussed with an example development for a real-world distribution system with limited data. Using simulated data discussions on how the pipe leak prediction model was assessed on parameter updating and pipe leak predictions have been discussed. Lastly, an example implementation of the modeling framework on simulated data was presented.

Chapter 5

Results and Discussion

5.1 Introduction

This chapter presents the key results from assessing the pipe leak prediction model and implementing the modeling framework on simulated data. It Starts with the presentation of assessment results from the development of the pipe leak prediction model, and results from the calibration of the water loss model. After this, results based on the implementation of the modeling framework and on the most recent parameter updates and observations on a collection of distribution pipes are presented. These indicate also the key model results for pipe leak monitoring and optimal maintenance sequence determination.

5.2 Assessment and Calibration

In this section, results from assessing the pipe leak prediction model on parameter updating and prediction performance are presented and discussed. The resulting calibrated water loss estimation model is presented and discussed. The first part of the section presents and discusses the assessment results while, the second presents the water loss estimation model.

5.2.1 Assessment Results

The two main tasks of concern in the pipe leak model are its ability to improve parameter estimates and prediction performance as more data is acquired. The first part of this section presents and discusses the results from the assessment of the modeling approach's ability to improve parameter estimates with progressive data updates. Secondly, results from assessing the updated posterior models on pipe leak predictions are presented.

Parameter Updating Results

To assess the modeling approach's ability to improve parameter estimates with progressive data updates, the updated posterior distributions for each CPT estimate are compared to their corresponding parameter CPT entries from the data simulation model.

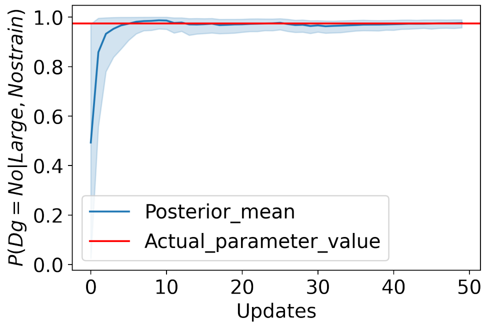
Generally, the 95% credible intervals for all the iteratively computed posterior distributions for each CPT parameter were observed to contain and narrow towards their corresponding CPT parameter values. Despite this, eight of the sixty nine iteratively updated model posterior estimates where slow to converge, relative to the other sixty one. See A.12h, A.13a, A.13b, 5.1d, 5.2a, 5.2b, 5.2c, 5.2d for 95% credible intervals of the posterior distributions for some CPT parameters over multiple data updates. In comparison to the posterior distributions in figures A.12h, A.13a, A.13b, 5.1d, those in figures 5.2a, 5.2b, 5.2c, 5.2d where slow to converge. For more plots on the updated posterior distributions for model parameter estimates see Appendix A.

The number of observations for each possible parent configuration affected the convergence of the posterior distributions related to that configuration. See table 5.1 for the counts corresponding to parent configurations of the posterior distributions given in A.12h to 5.2d as data updates were conducted. Notice that the counts corresponding to the parent configurations of the slow converging posterior distributions ie, the first four rows of table5.1, are smaller compared to the total number of observations and numbers of parent configuration observations with quicker converging parameters. This is consistent with the described effect of poor parameter estimation in parent configurations encountering data fragmentation (Karkera, 2014). In this case, eight out of the sixty-nine posterior distributions encountered this challenge.

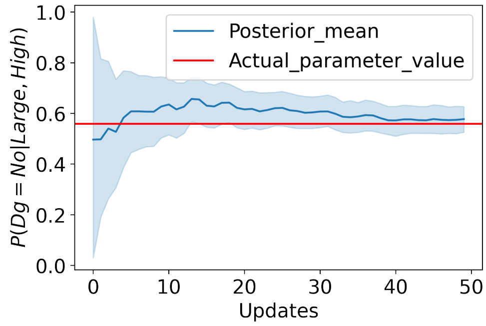
With regards to a data-adaptive parameter learning approach, these plots indicated that the Bayesian parameter updating approach is appropriate for iteratively updating Bayesian network parameters for pipe leak modeling. However, one needs to pay attention to parameters based on low counts since they are highly variable and can lead to biased predictions.

Posterior Model Evaluations

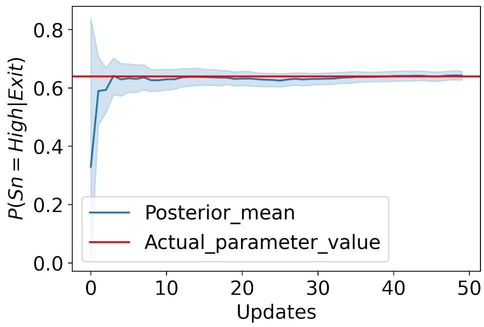
The results of the MCC scores for the iteratively updated posterior Bayesian network models indicated improvement in the models' prediction ability as



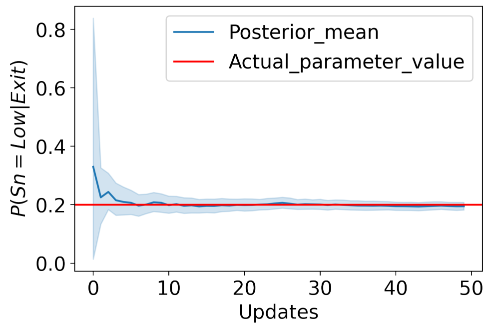
(A)



(B)



(C)



(D)

FIGURE 5.1: Example posterior means with corresponding 95% credible intervals for updating model parameters in comparison to the actual parameter values from the simulation model. With more data updates the posterior means approach the actual parameters as the credible interval narrows towards the mean. This is indicative of the usefulness of Bayesian parameter updating to continually update and improve the parameters of the pipe leak prediction model

its parameters got updated with more data batches. The *MCC* score improved from 0 for the uniform prior model to 0.571 after the fourth model parameter update. At this point no further improvements were observed in model performance as a result of parameter updating. See figure 5.3 for the *MCC* scores for the first twenty iteratively updated posterior models.

These results show the corresponding improvements in model performance as parameters get updated. However, the performance only improves to as much as the model represents pipe leak behavior. In which case,further improvement in model performance on parameter updating needs to be combined with adjustments to model variables in a way that represents the observed data points more accurately. As such, it is crucial for a system implementing this modeling framework to monitor the *MCC* scores as the model parameters get updated. This also provides decision-makers with useful information to judge the extent to which decisions based on the model results

Parent configurations	Updates					
	0	50	100	150	200	250
Cr_ACD_dolomite	2.0	48.0	88.0	130.0	184.0	221.0
Lk_yes_true_yes	2.0	3.0	5.0	8.0	10.0	14.0
Lk_yes_true_no	2.0	10.0	18.0	26.0	33.0	41.0
Lk_no_true_yes	2.0	125.0	272.0	394.0	539.0	696.0
Lk_no_false_no	2.0	6921.0	13765.0	20558.0	27485.0	34349.0
Cr_HDPE_regular	2.0	1389.0	2809.0	4273.0	5737.0	7199.0
Dg_small_high	2.0	1699.0	3385.0	5205.0	6843.0	8548.0
Sn_low_exit	3.0	1311.0	2647.0	3966.0	5226.0	6569.0
n_update	2.0	10002.0	20002.0	30002.0	40002.0	50002.0

TABLE 5.1: Sample counts on model parameters over a series of updates showcasing the effect of sample counts on parameter learning. In conjunction with figures A.12h, A.13a, A.13b, 5.1d, 5.2a, 5.2b, 5.2c, 5.2d, the negative effects of low sample counts on a parent configurations over the model updates are seen. Low counts are leading to slow and highly uncertain posterior parameter estimates.

can be trusted.

5.2.2 Water Loss Estimation

After calibrating δ and β based on example data and, equations 4.2 and 4.1, the resulting pipe leak water loss estimation function is given in equation 5.1.

$$q_k = (1.6 \times 10^{-8}) L P^{\delta} \tag{5.1}$$

The estimate $\beta = 1.6 \times 10^{-8}$ is within the same order of magnitude as the β estimates reported by Giustolisi, Savic and, Kapelan (2008). The defined model was then used to estimate background leak losses for each pipe. see table 5.2 for pipe leak water loss estimates based on example pipe data.

5.3 Key Model Outputs

After training the Bayesian network model and calibrating the pipe leak water loss estimation model, the two models are then used to monitor and suggest the pipe leak maintenance sequence using two key output results. These include the posterior predictive distribution and the optimal pipe leak maintenance sequence.

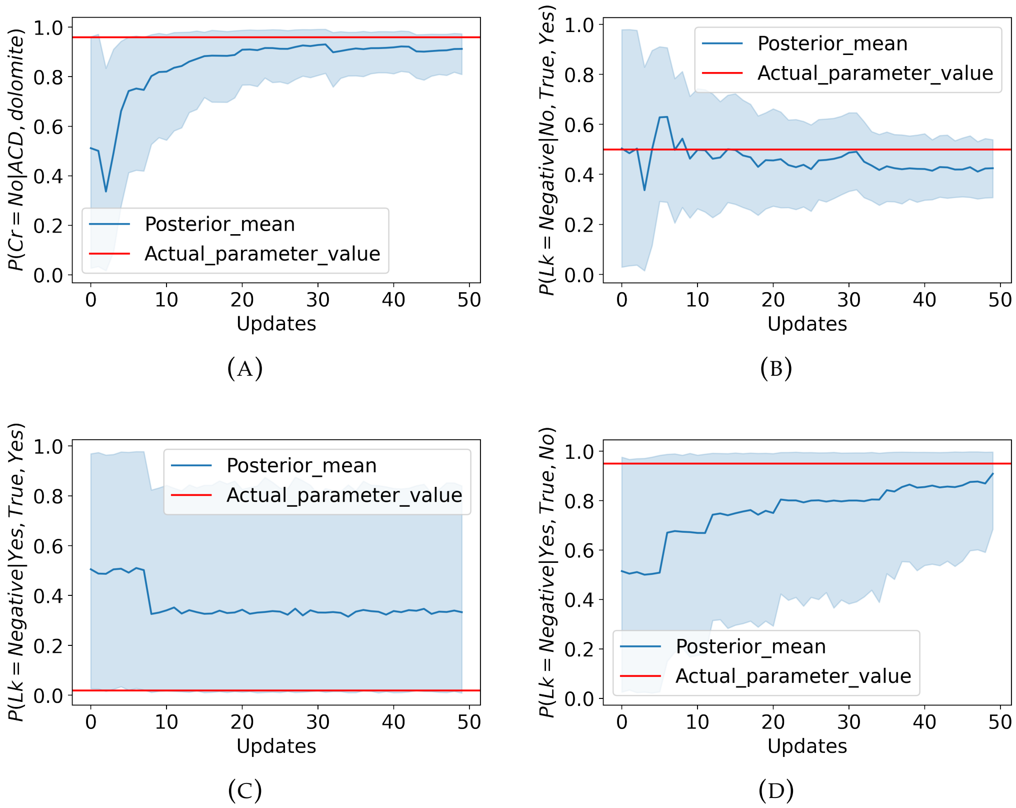


FIGURE 5.2: Example model posterior means and 95% credible intervals for model parameters with low update counts, compared to the actual parameter value from the simulation model. In comparison to the posterior distributions in 5.1, these distributions approach the actual parameters much slower and with wider credible intervals. This indicates the negative effects of small sample parent configuration observations on model parameter estimation. This is especially to be watched out for among nodes with many parents.

This section presents results based on the implementation of the previously trained BN model and calibrated water loss estimation model on ten hypothetical pipes.The results provide example model outputs and interpretations for monitoring pipe leaks and also provide an optimal pipe leak fixing sequence. The proceeding subsections include results on the pipes' posterior predictive distributions and the optimal pipe leak maintenance sequence.

5.3.1 Posterior Predictive Distributions

For purposes of monitoring pipe leaks, each distribution pipes' posterior predictive distribution is plotted as shown in figures 5.4a, 5.4b, 5.4c, and 5.4d. These predictions depend on observations made on model variables for each

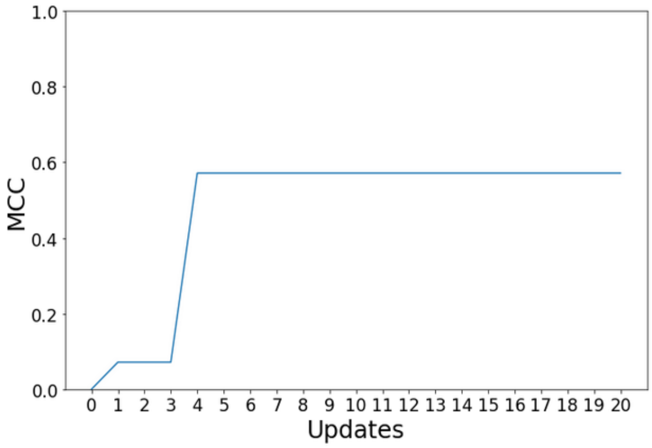


FIGURE 5.3: Mathews Correlation Coefficient over multiple model posterior updates. Indicating the improvement of model prediction performance on the first and fourth updates. The performance stays moderate at $MCC = 0.57$ and does not improve despite more data updates. This suggests the need for further model assessment and the addition of more pipe leak affecting explanatory variables to improve model performance on pipe leak predictions.

pipe of interest. As such, as some of the pipes model variables change their leakage probabilities change accordingly. By implication, leakages on pipes can be monitored based on the observed model variable values on each pipe.

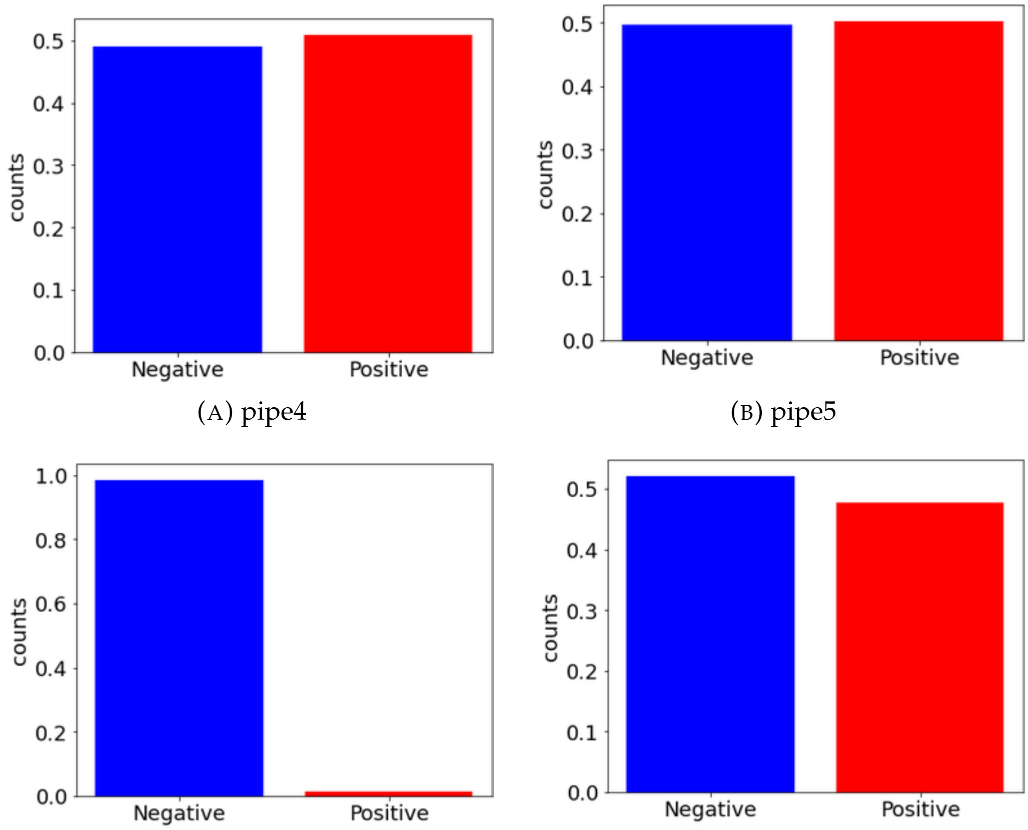
5.3.2 Optimal Maintenance Sequence

Using the estimated pipe leak water losses and individual pipe leak predictions, estimations of expected water losses for each pipe were calculated, see table 5.2.

These expected pipe leak water losses were then used to define the objective function and the corresponding decision variables as discussed in 3.5. After solving this, the sequence in Table 5.3 was obtained as the optimal maintenance sequence to minimize total water losses. Notice that despite starting with ten pipes the optimal sequence only enlists four pipes. This is so because all pipes with a leak probability of zero are omitted in uncovering the optimal sequence.

5.4 Concluding Remarks

This chapter presents results from the assessment, calibration, and implementation of the presented modeling framework based on simulated data.



(A) pipe4 (B) pipe5

(C) pipe9 (D) pipe10

FIGURE 5.4: Example pipe leak posterior predictive distributions for monitoring pipe leaks in WDS. In addition to predicting pipe leak status these distributions provide useful information on evidence based pipe leak probability distributions, while accounting for uncertainties in posterior parameters. For instance, with a decision threshold of 0.5, both pipe9 and pipe10 would predict pipe leak status as negative. However, for leakage monitoring with posterior distributions pipe10 would require attention since its positive leak probability is fairly high.

From the assessment of the BN model, the results indicate that the parameter updating approach is capable of iteratively improving the prior model parameters over multiple data updates. This is useful in a situation where pipe leak models are to be defined apriori. However, the highest MCC score for the posterior updated models was achieved and maintained at a value of 0.57. This Indicates only moderate performance on pipe leak predictions after parameter updating. A value that is likely to improve if more pipe leak affecting variables such as pressure is included in the model.

From the implementation of the trained and calibrated models on simulated data, the results indicate that the resulting model outputs can be used for monitoring pipe leak probabilities and uncovering the optimal sequence

	leak losses(l/hr)	leak probability	Expected leak losses(l/hr)
pipe1	1.8	0.000	0.000
pipe2	1.1	0.000	0.000
pipe3	4.2	0.000	0.000
pipe4	2.2	0.510	1.122
pipe5	2.4	0.503	1.207
pipe6	0.9	0.000	0.000
pipe7	1.4	0.000	0.000
pipe8	1.1	0.000	0.000
pipe9	0.7	0.013	0.009
pipe10	1.6	0.478	0.765

TABLE 5.2: Important quantities for optimal pipe leak maintenance sequence determination. This is obtained from the trained pipe leak prediction model and the calibrated water loss estimation model. Notice that even though pipe 3 has the highest water loss its 0 probability of leaking will rule it out from being prioritized for maintenance.

Optimal maintenance sequence
pipe5, pipe4, pipe10, pipe9

TABLE 5.3: Pipe leak optimal maintenance sequence table for guiding leakage maintenance team to minimize water losses

in which to maintain pipe leaks for minimized water losses.

Chapter 6

Conclusions and Future Works

6.1 Conclusions

This research aimed at developing a pipe leak monitoring and maintenance sequence suggestion modeling framework for leakage management in water distribution systems. From the assessment of the pipe leak prediction model and an example implementation of the modeling framework based on simulated data, it can be concluded that the framework is suitable for use in pipe leak monitoring and optimal maintenance sequence determination in WDS's.

Assessment results indicate the ability of the pipe leak model to update model parameter estimates to their respective true values based on iterative data updates. This is an important feature especially in systems with limited data. The literature-based prior model parameters must be adapted to a distribution system of interest by iteratively updating them based on iterative data updates on model variables. Based on the evaluation results of the pipe leak prediction model, a moderate performance was reported for the best performing updated models. Implementation results demonstrated the use of a developed pipe leak prediction model for monitoring pipe leaks by developing and interpreting their respective posterior predictive distributions. The results also demonstrated the calibration of a pipe leak water loss estimation model. The water loss estimates from this model are then, together with pipe leak probabilities, used in the optimization model for determining the optimal pipe leak maintenance sequence.

As a tool, the implementation of the modeling framework serves as an active leakage management tool. Therefore, considering its development flexibility, the framework will serve as an active leakage management tool even in distribution systems with data limitations which are mostly those in developing countries. Essentially, more distribution systems can adopt an active leakage management thereby reducing the amount of water lost from distribution systems. This also creates a data need for model improvement and is likely to improve data collections practices. Despite some technological, skills, and financial challenges faced by some distribution systems, active leakage management is still a possibility, provided that related data on pipe leak is collected and properly stored.

From the foregoing, water loss managers must consider implementing this modelling tool for pipe leak surveillance.

6.2 Recommendations for future studies

Th following are possible studies that may be investigated in the future:

1. Exploration of different model structures and variable combinations to optimize model predictions. This may be achieved through learning the optimal network structures for different combinations of pipe leak affecting variables and evaluating the performance of each resulting model. The model with the best variable combination and a theoretically abiding structure is then considered for use in prediction tasks.

2. Assessment of possible general rules for calibrating Bayesian network models for different reticulation's in a water distribution system. This may be done by first building and optimizing multiple pipe leak predicting Bayesian network models for various reticulation's. After which, the resulting models may then be assessed for similarities in model structures and variable combinations.

3. Extension of the optimization model to determine optimal pipe leak maintenance sequence by considering the distances between pipes in the distribution system.

Appendix A

Appendix A

A.1 Section H's Prior Parameters

This section presents the Dirichlet prior parameters for the conditional distribution tables for the nodes in the Bayesian network presented for pipe leak prediction in Eskhawini section H reticulation. These prior parameters are based on theoretical relationships between the model variables. For cases were relationships are not clear, uninformative proper priors where used.

In the case of the corrosion status (Cr) of a pipe, the lack of plastic pipes suffering any form of corrosion leads to every parent combination including $Mt = Plastic$ to have a value of 0 for $Cr = yes$. Also at older ages, both AC and steel pipes are more likely to be corroded than not. This has been encoded by slight deviations away from the uniform parameter setting to reflex the Age effect on corrosion in AC and steel pipes. Considering the factors affecting internal strain (iSn) experienced by a pipe, high pressure increases the amount of iSn experienced by a pipe, while low pressure reduces it. Accounting for diameter size, the strain is higher for small diameter pipes and lower in larger ones. TableA.8 shows the defined prior parameters indicating these relationships. To define the prior conditional probability table for external strain (eSn) we note the findings by A. G. Ogutu, Kogeda and, Lall, 2018 indicating that pipes on-road exits experience more external strain than those on Erf. In consideration of pipe length, the strain is even higher in longer pipes than it is in shorter ones. TableA.19 indicates the prior parameter definitions for eSn's CPT probabilities which reflect the described

Pipe Age (Ag)		
$\alpha(Ag = 0 - 15)$	$\alpha(Ag = 16 - 30)$	$\alpha(Ag = 31 - 45)$
1	1	1

TABLE A.1: Initial Dirichlet parameters for the multinomial CPT's

Pipe Material (Mt)		
$\alpha(Mt = AC)$	$\alpha(Mt = plastic)$	$\alpha(Mt = steel)$
0.012	2.978	0.009

TABLE A.2: Initial Dirichlet parameters for the multinomial CPT's

Pipe Diameter (Di)		
$\alpha(Di = small)$	$\alpha(Di = medium)$	$\alpha(Di = large)$
1.33	1.59	0.08

TABLE A.3: Initial Dirichlet parameters for the multinomial CPT's

Pressure		
$\alpha(Pr = low)$	$\alpha(Pr = medium)$	$\alpha(Pr = high)$
1	1	1

TABLE A.4: Initial Dirichlet parameters for the multinomial CPT's

Location	
$\alpha(Lo = Erf)$	$\alpha(Lo = exit)$
1	1

TABLE A.5: Initial Dirichlet parameters for the multinomial CPT's

Pipe Length (Le)		
$\alpha(Le = short)$	$\alpha(Le = medium)$	$\alpha(Le = long)$
1.98	0.89	0.08

TABLE A.6: Initial Dirichlet parameters for the multinomial CPT's

Corrosion status (Cr)			
Age	Material	$\alpha(Cr = Yes)$	$\alpha(Cr = No)$
0-15	AC	0.8	1.2
0-15	Plastic	0	2
0-15	Steel	0.8	1.2
16-30	AC	0.9	1.1
16-30	Plastic	0	2
16-30	Steel	0.9	1.1
31-45	AC	1.4	0.6
31-45	Plastic	0	2
31-45	Steel	1.4	0.6

TABLE A.7: Initial Dirichlet parameters for the multinomial CPT's

Internal strain (iSn)				
Diameter	Material	Pressure	$\alpha(iSn = Low)$	$\alpha(iSn = High)$
small	AC	low	1	1
small	AC	medium	1	1
small	AC	high	0.7	1.3
small	Plastic	low	1	1
small	Plastic	medium	1	1
small	Plastic	high	0.7	1.3
small	Steel	low	1	1
small	Steel	medium	1	1
small	Steel	high	0.7	1.3
medium	AC	low	1	1
medium	AC	medium	1	1
medium	AC	high	0.9	1.1
medium	Plastic	low	1	1
medium	Plastic	medium	1	1
medium	Plastic	high	0.9	1.1
medium	Steel	low	1	1
medium	Steel	medium	1	1
medium	Steel	high	0.9	1.1
large	AC	low	1	1
large	AC	medium	1	1
large	AC	high	0.9	1.1
large	Plastic	low	1	1
large	Plastic	medium	1	1
large	Plastic	high	0.9	1.1
large	Steel	low	1	1
large	Steel	medium	1	1
large	Steel	high	0.9	1.1

TABLE A.8: Initial Dirichlet parameters for the multinomial CPT's

Previous Breaks (PrB)		
$\alpha(PrB = 0 - 7)$	$\alpha(PrB = 8 - 15)$	$\alpha(PrB => 15)$
1	1	1

TABLE A.9: Initial Dirichlet parameters for the multinomial CPT's

Parents		external strain (eSn)	
Le	Lo	$\alpha(eSn = Low)$	$\alpha(eSn = High)$
short	Erf	1	1
short	Exit	1	1
medium	Erf	1	1
medium	Exit	1	1
long	Erf	1	1
long	Exit	1.5	0.5

TABLE A.10: Initial Dirichlet parameters for the multinomial CPT's

Damage status (Dg)				
eSn	iSn	PrB	$\alpha(Dg = Yes)$	$\alpha(Dg = No)$
Low	Low	0-7	1	1
Low	Low	8-15	1	1
Low	Low	>15	1	1
Low	High	0-7	1	1
Low	High	8-15	1	1
Low	High	>15	1	1
High	Low	0-7	1	1
High	Low	8-15	1	1
High	Low	>15	1	1
High	High	0-7	1	1
High	High	8-15	1.4	0.6
High	High	>15	1.6	0.4

TABLE A.11: Initial Dirichlet parameters for the multinomial CPT's

qualitative relationships.

The Damage status Dg of a pipe dependent on eSn, iSn, and PrB will firstly be the more likely true for pipes higher iSn and eSn regardless. While accounting for pipes' previous breaks, the probability is even larger in pipes with a higher number of previous breaks. See tableA.11 for the representative prior parameter values.

Finally, considering pipe leak status, the prior is set similar to the reflect the Corrosion, Damage influence in leak status from (A. G. Ogutu, Kogeda and, Lall, 2018). That is if both Cr and Dg are Yes the probability of leaking is at its highest. And if non is Yes then there is no chance of observing a leak in a pipe. See table A.12 for the complete initialized prior CPT parameters for Lk, reflecting the encoded parent-child relationships.

Leak status (Lk)			
Cr	Dg	$\alpha(Lk = Positive)$	$\alpha(Lk = Negative)$
Yes	Yes	1.7	0.3
Yes	No	1	1
No	Yes	1	1
No	No	0	2

TABLE A.12: Initial Dirichlet parameters for the multinomial CPT's

Soil type	
$P(S = Regular)$	$P(S = Dolomite)$
0.922	0.078

TABLE A.13: Initial Dirichlet parameters for the multinomial CPT's

A.2 Simulation Model Parameters

A.3 Sample Simulated Data

A.3.1 Sample Test Data

A.4 Posterior Parameter Updates

Figures A.1, A.2, A.3, A.4, A.5, A.6, A.7, A.9, A.10, A.11, A.12, A.13, present assessment results for Bayesian parameter updating on each node CPT. Each plot represents an individual parameter in the CPT.

Location	
$P(Lo = Erf)$	$P(Lo = Exit)$
0.633	0.367

TABLE A.14: Simulation model parameters for the location (Lo) node. Used for the simulation of training data and in assessing the iteratively updating posterior parameter distributions based on data updates.

Defective	
$P(Df = True)$	$P(Df = False)$
0.02	0.98

TABLE A.15: Simulation model parameters for the defective (Df) node. Used for the simulation of training data and in assessing the iteratively updating posterior parameter distributions based on data updates.

Material type			
$P(M = ACD)$	$P(M = HDPE)$	$P(M = Steel)$	$P(M = UPVC)$
0.057	0.340	0.157	0.446

TABLE A.16: Simulation model parameters for the material type (M) node. Used for the simulation of training data and in assessing the iteratively updating posterior parameter distributions based on data updates.

Diameter		
$P(D = Small)$	$P(D = Medium)$	$P(D = Large)$
0.546	0.350	0.104

TABLE A.17: ISimulation model parameters for the diameter (D) node. Used for the simulation of training data and in assessing the iteratively updating posterior parameter distributions based on data updates.

Parent	Strain		
Lo	$P(Sn = NoStrain)$	$P(Sn = Low)$	$P(Sn = High)$
Erf	0.36	0.52	0.12
Exit	0.16	0.20	0.64

TABLE A.18: Simulation model parameters for the Strain (Sn) node. Used for the simulation of training data and in assessing the iteratively updating posterior parameter distributions based on data updates.

Parents		Damage status	
D	Sn	$P(Dg = No)$	$P(eSn = Yes)$
Small	NoStrain	0.120	0.880
Small	Low	0.160	0.840
Small	High	0.640	0.360
Medium	NoStrain	0.027	0.973
Medium	Low	0.120	0.880
Medium	High	0.520	0.480
Large	NoStrain	0.025	0.975
Large	Low	0.080	0.920
Large	High	0.440	0.560

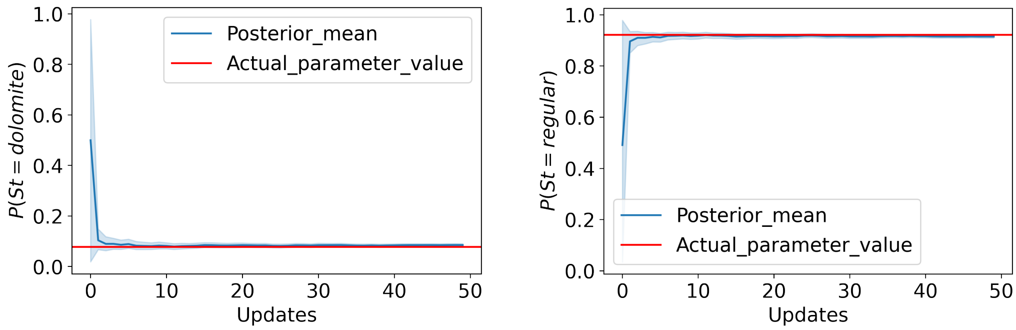
TABLE A.19: Simulation model parameters for the damage status (Dg) node. Used for the simulation of training data and in assessing the iteratively updating posterior parameter distributions based on data updates.

Parents		Corrosion	
Mt	St	$P(C = Yes)$	$P(C = No)$
ACD	Dolomite	0.040	0.960
ACD	Regular	0.290	0.710
HDPE	Dolomite	0.000	1.000
HDPE	Regular	0.000	1.000
Steel	Dolomite	0.040	0.960
Steel	Regular	0.290	0.710
Upvc	Dolomite	0.000	1.000
Upvc	Regular	0.000	1.000

TABLE A.20: SSimulation model parameters for the corrosion status (C) node. Used for the simulation of training data and in assessing the iteratively updating posterior parameter distributions based on data updates.

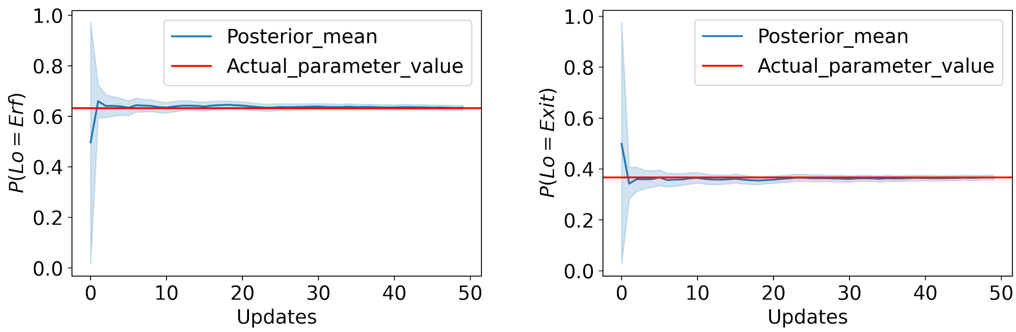
Parents			Leak	
Cr	Df	Dg	$P(Lk = Positive)$	$P(Lk = Negative)$
Yes	True	Yes	0.980	0.020
Yes	True	No	0.050	0.950
Yes	False	Yes	0.512	0.488
Yes	False	No	0.012	0.988
No	True	Yes	0.500	0.500
No	True	No	0.020	0.980
No	False	Yes	0.500	0.500
No	False	No	0.000	1.000

TABLE A.21: Simulation model parameters for the leak status (Lk) node. Used for the simulation of training data and in assessing the iteratively updating posterior parameter distributions based on data updates.



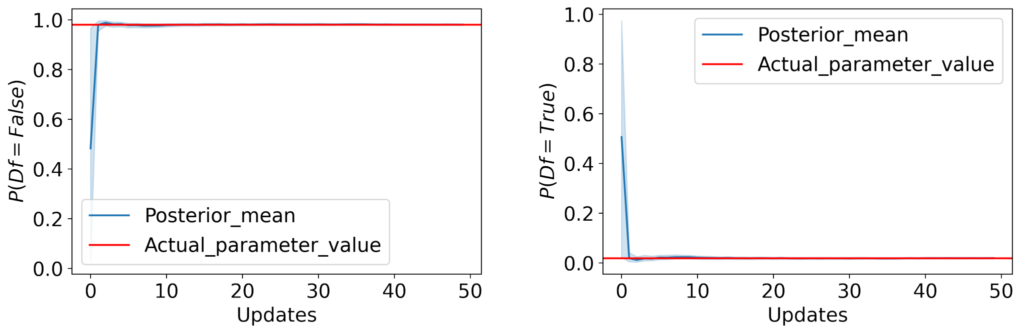
(A) (B)

FIGURE A.1: Mean and 95% credible intervals for Soil type CPT parameter posterior distributions based on 250 data updates



(A) (B)

FIGURE A.2: Mean and 95% credible intervals for location CPT parameter posterior distributions based on 250 data updates

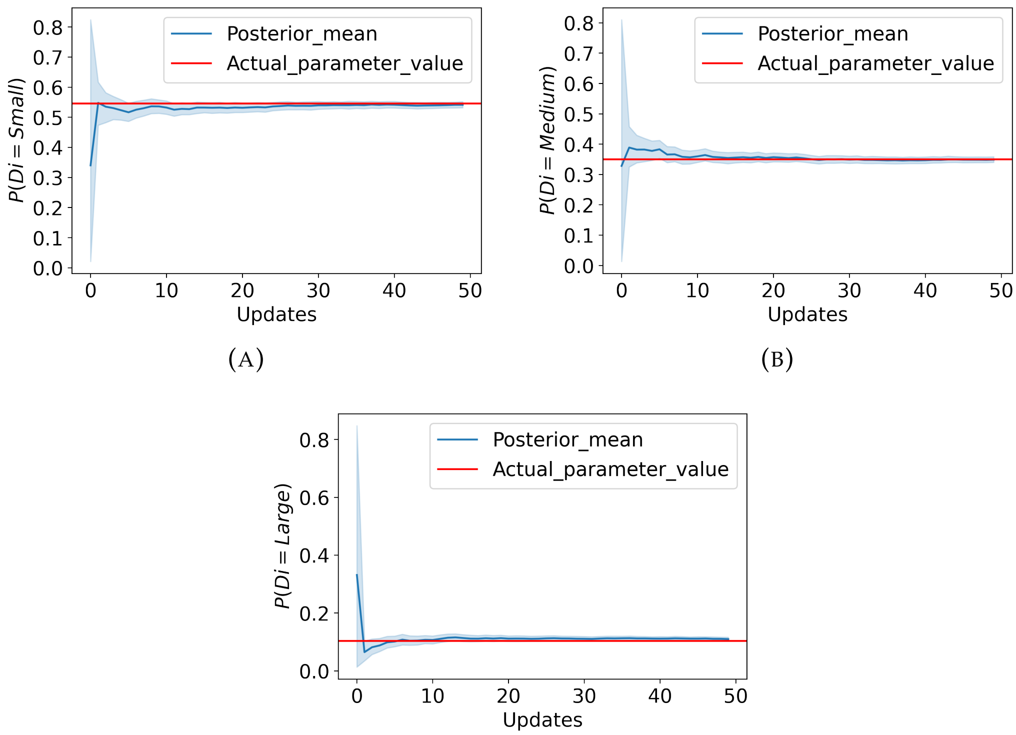


(A) (B)

FIGURE A.3: Mean and 95% credible intervals for defective CPT parameter posterior distributions based on 250 data updates

	St	Mt	Di	Lo	Cr	Sn	Df	Dg	Lk	Pr	Lt
0	regular	HDPE	Medium	Erf	No	Low	False	No	Negative	13.39	53.67
1	dolomite	Upvc	Small	Exit	No	Low	False	No	Negative	13.09	32.94
2	regular	Upvc	Medium	Erf	No	Low	False	No	Negative	14.05	133.99
3	regular	HDPE	Small	Exit	No	Low	False	Yes	Negative	20.98	20.23
4	regular	HDPE	Small	Exit	No	High	False	Yes	Negative	26.31	22.62

TABLE A.22: Sample input data for monitoring and maintenance suggestion approach. An example data set representing a section of pipes in a distribution system used for displaying an implementation example of the presented modelling framework.



(A)

(B)

(C)

FIGURE A.4: Mean and 95% credible intervals for Diameter CPT parameter posterior distributions based on 250 data updates

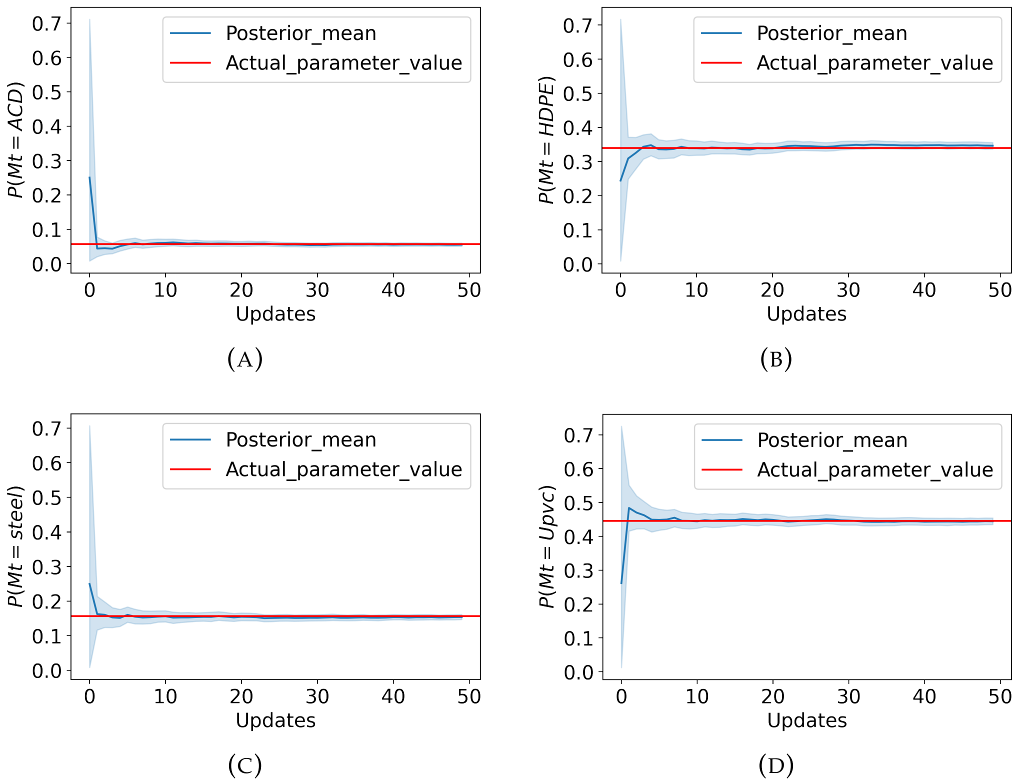
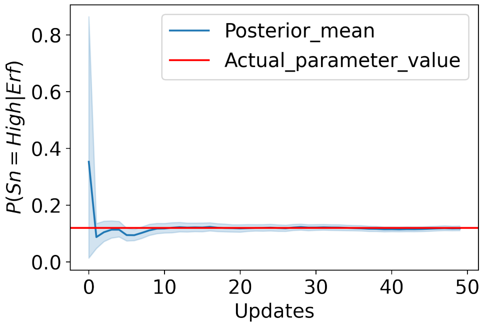
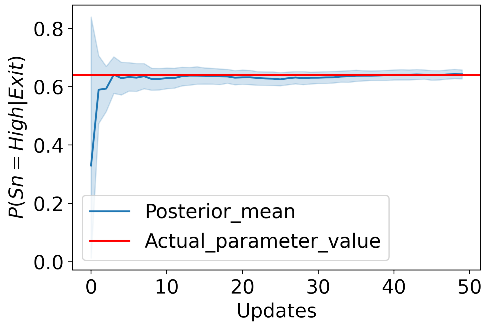


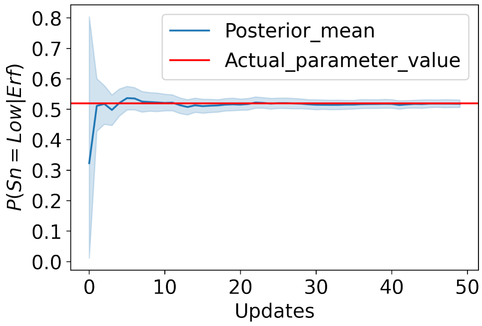
FIGURE A.5: Mean and 95% credible intervals for material type CPT parameter posterior distributions based on 250 data updates



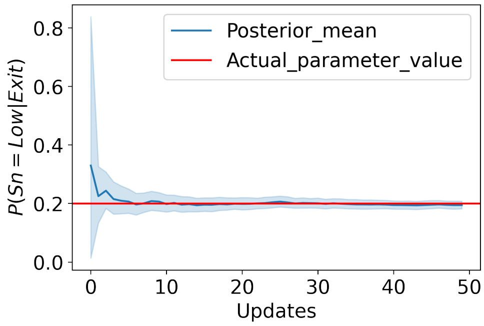
(A)



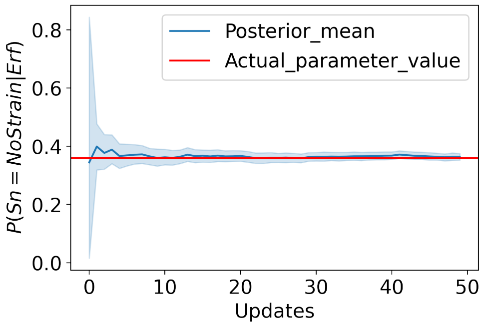
(B)



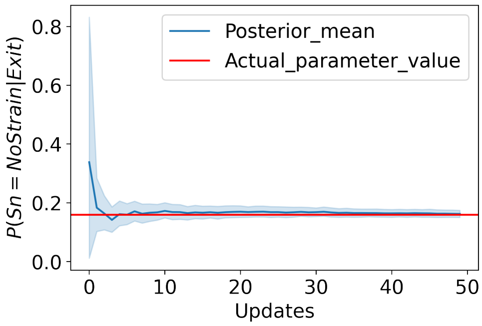
(C)



(D)

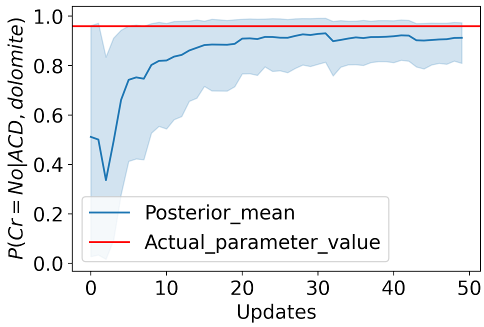


(E)

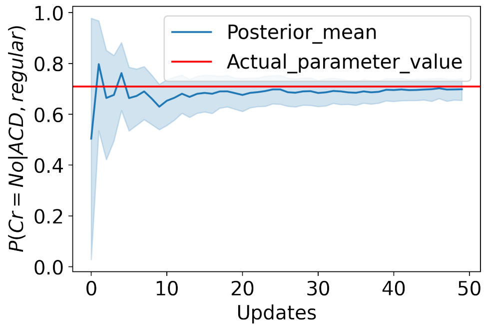


(F)

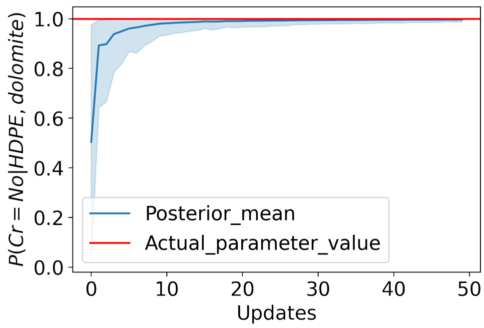
FIGURE A.6: Mean and 95% credible intervals for Strain CPT parameter posterior distributions based on 250 data updates



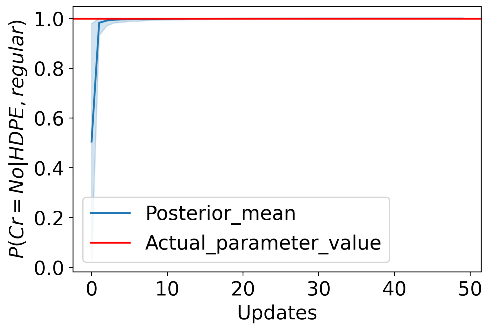
(A)



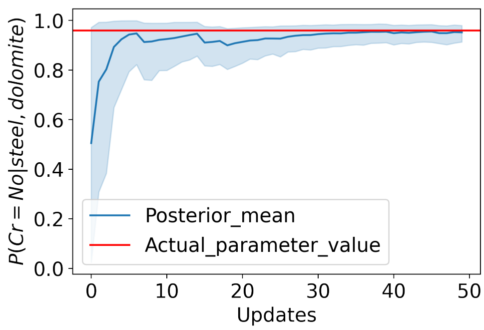
(B)



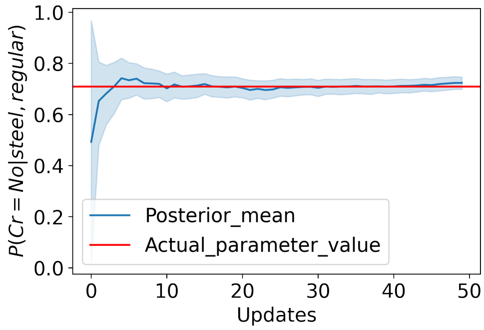
(C)



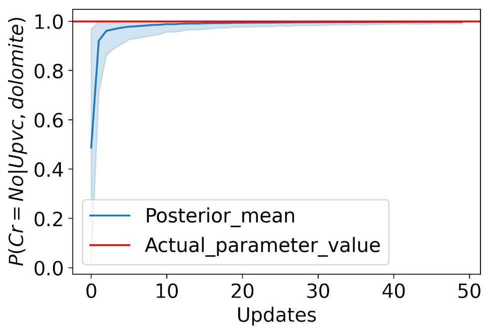
(D)



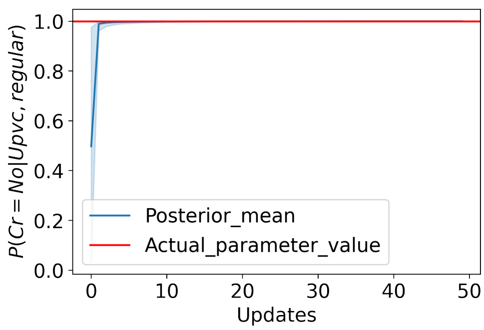
(E)



(F)



(G)



(H)

FIGURE A.7: Mean and 95% credible intervals for Corrosion CPT parameter posterior distributions based on 250 data updates

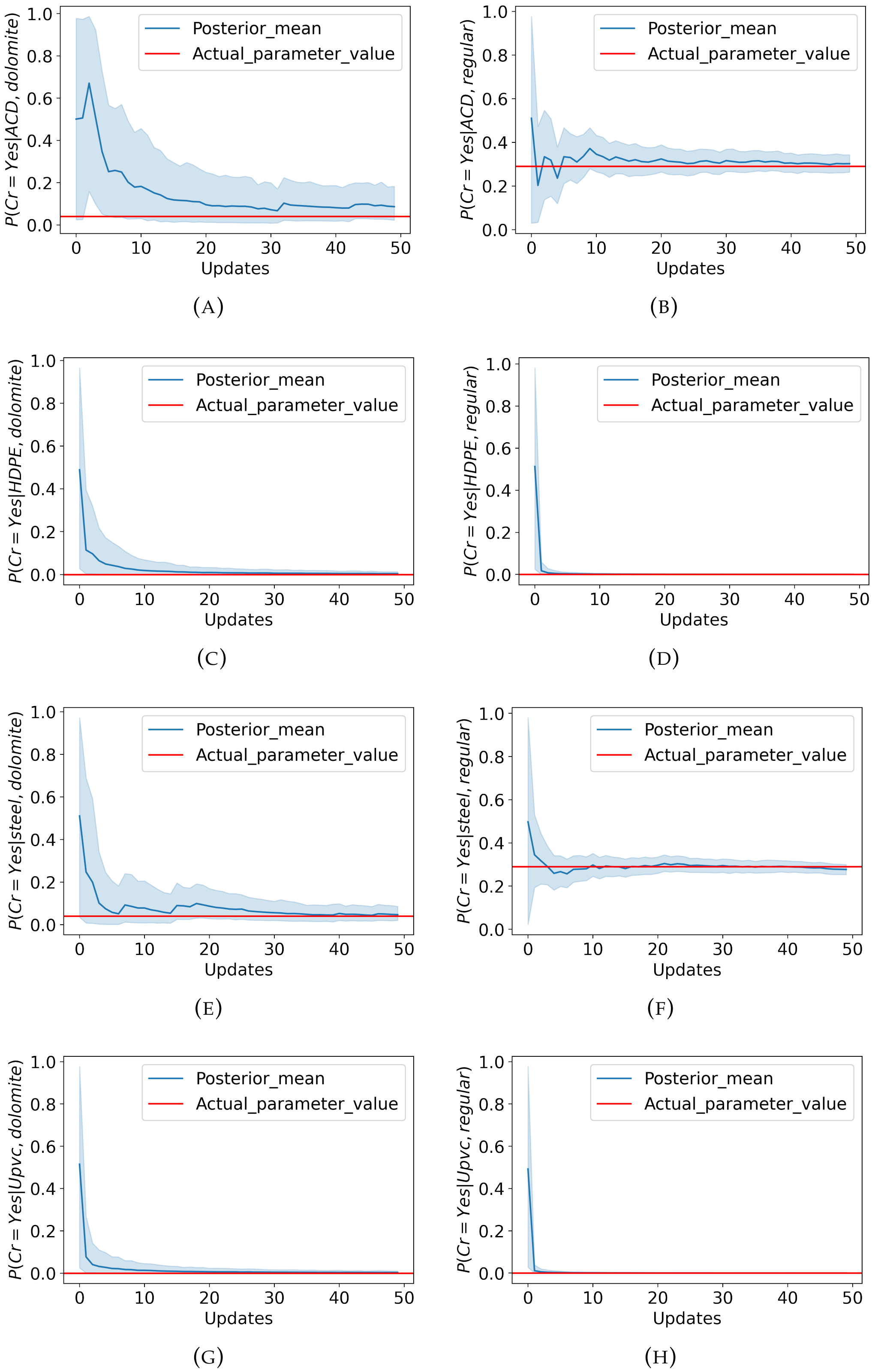


FIGURE A.8: Mean and 95% credible intervals for corrosion CPT parameter posterior distributions based on 250 data updates. Plot continuation for corrosion CPT parameters

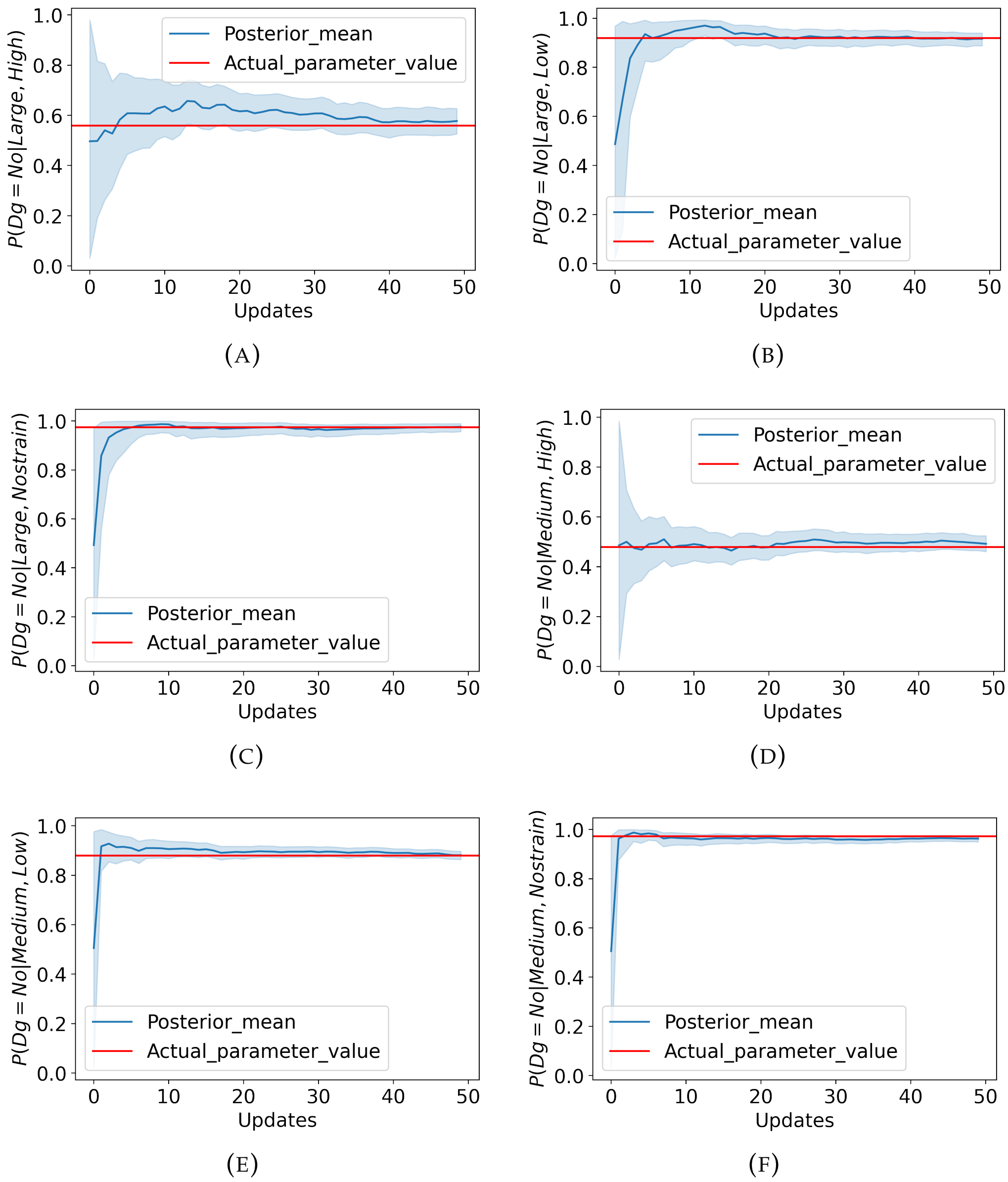


FIGURE A.9: Mean and 95% credible intervals for Corrosion CPT parameter posterior distributions based on 250 data updates

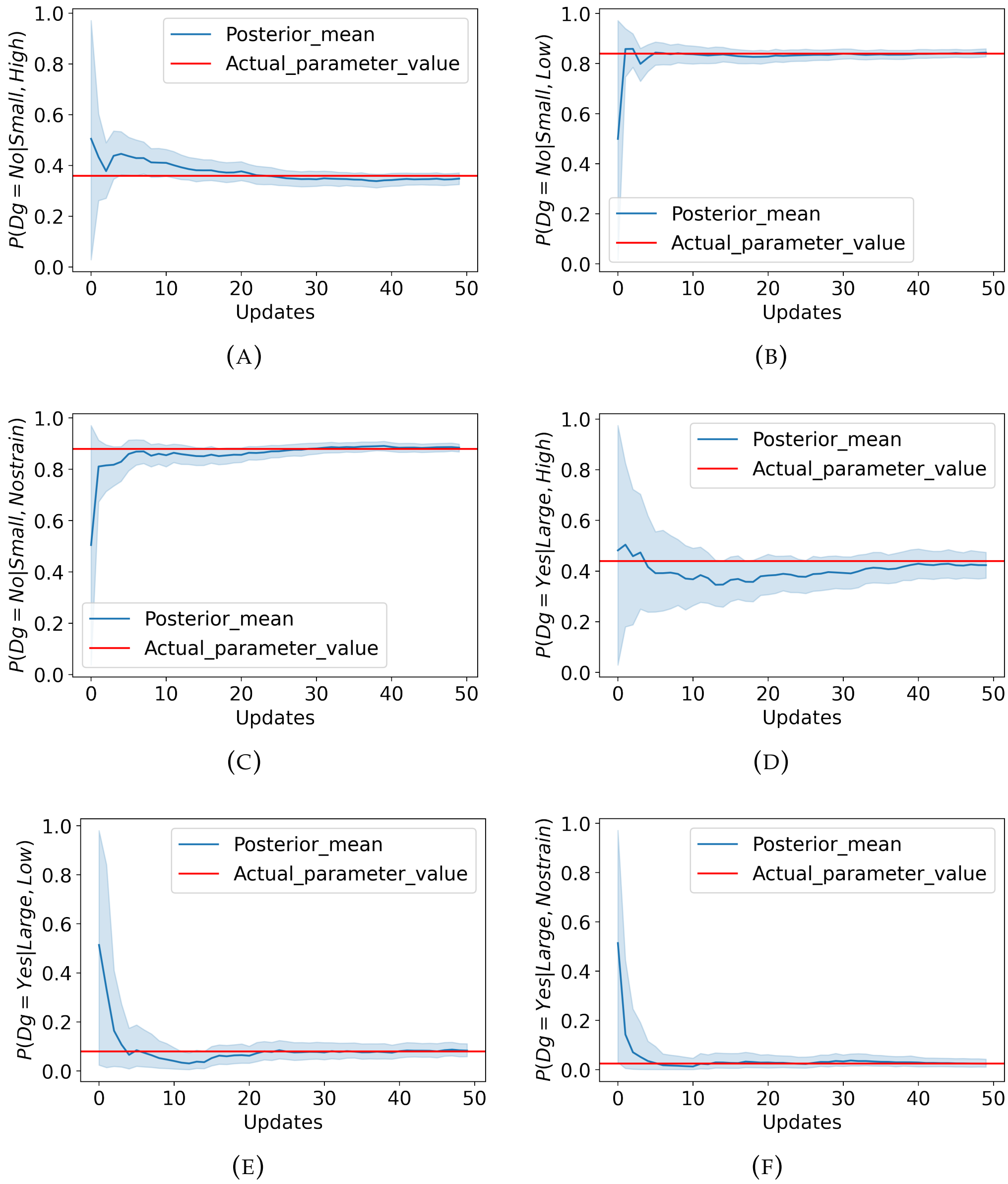
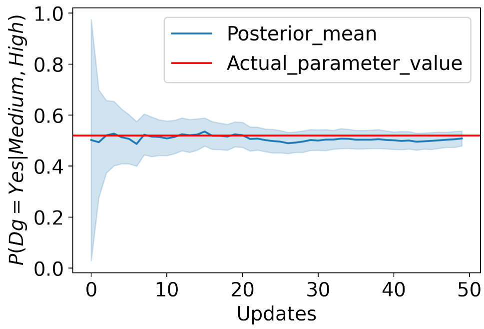
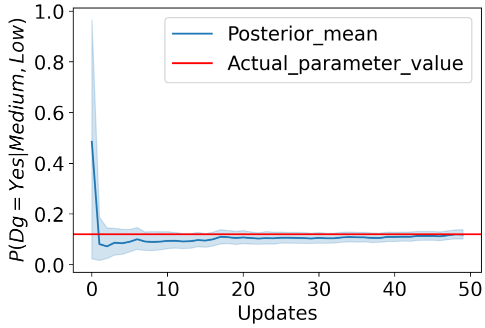


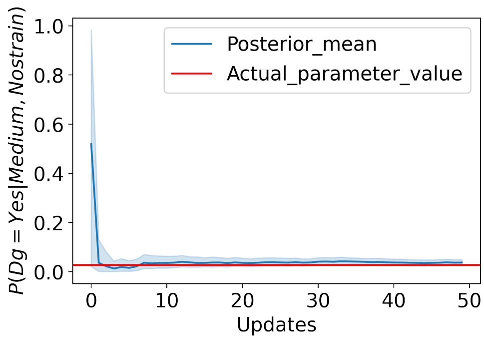
FIGURE A.10: Mean and 95% credible intervals for Damage CPT parameter posterior distributions based on 250 data updates. Plot continuation for Damage CPT parameters



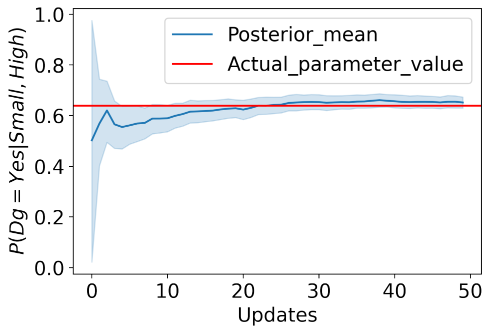
(A)



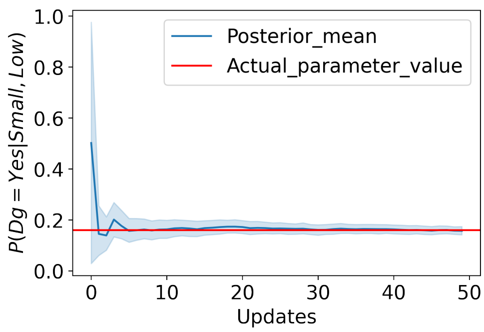
(B)



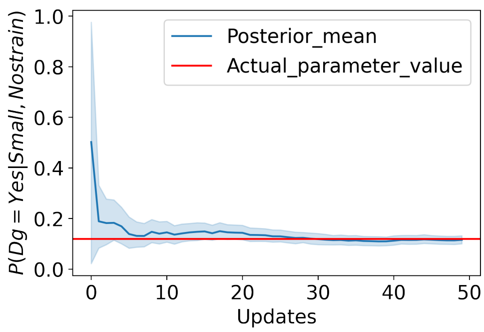
(C)



(D)

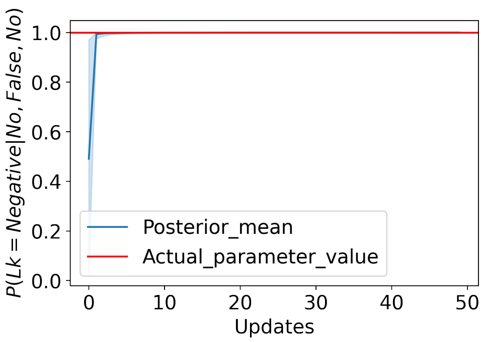


(E)

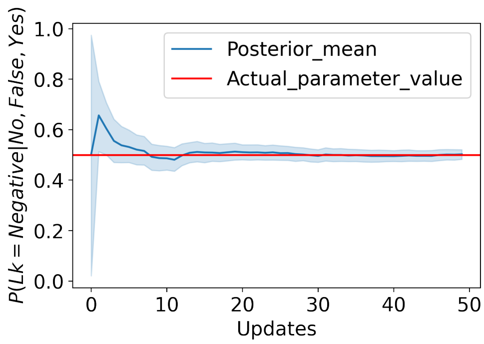


(F)

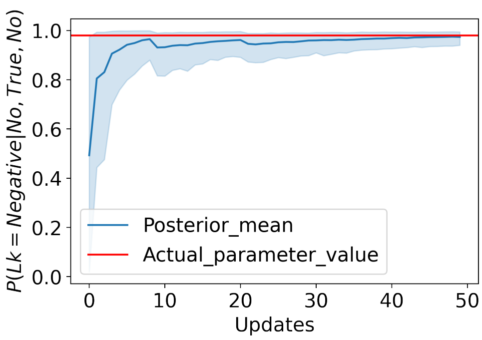
FIGURE A.11: Mean and 95% credible intervals for Damage CPT parameter posterior distributions based on 250 data updates. Plot continuation for Damage CPT parameters



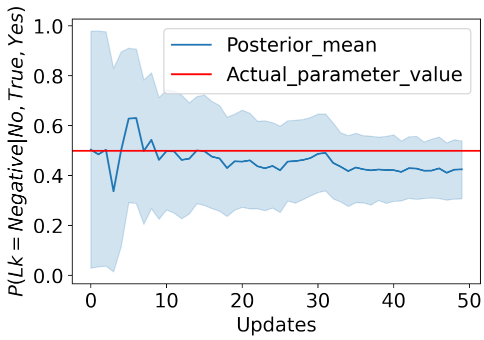
(A)



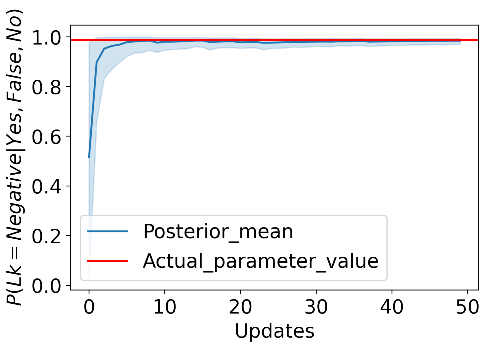
(B)



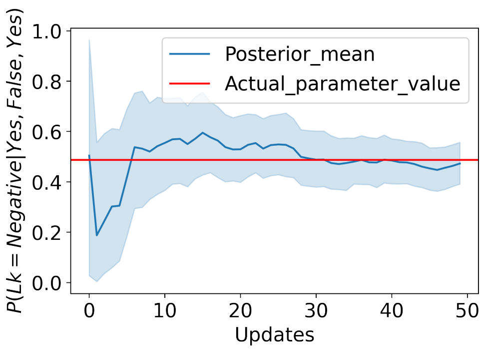
(C)



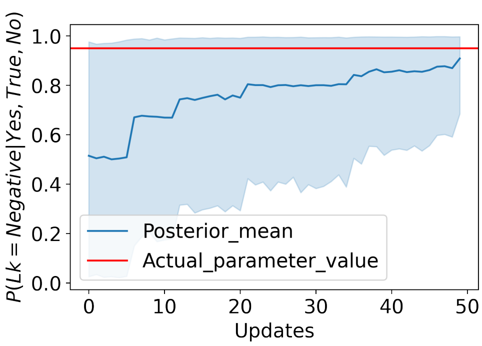
(D)



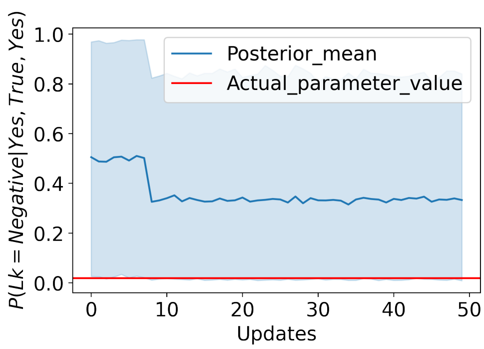
(E)



(F)

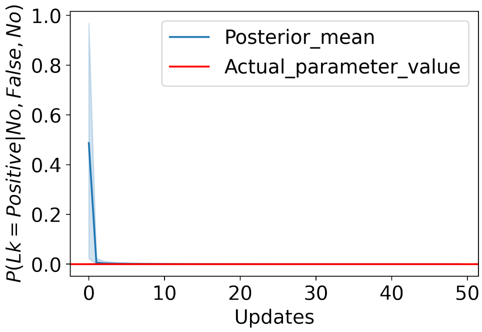


(G)

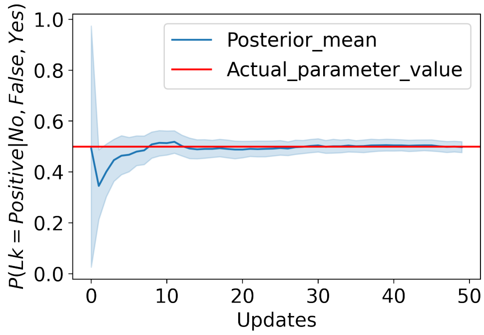


(H)

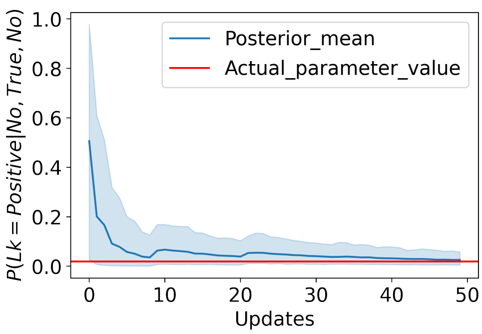
FIGURE A.12: Mean and 95% credible intervals for Corrosion CPT parameter posterior distributions based on 250 data updates



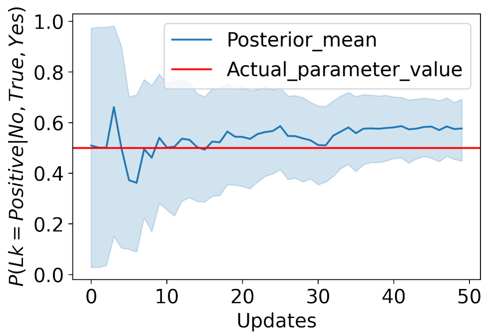
(A)



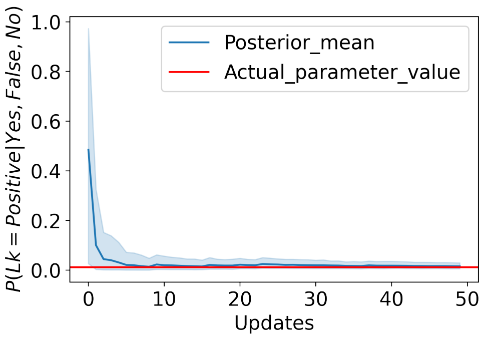
(B)



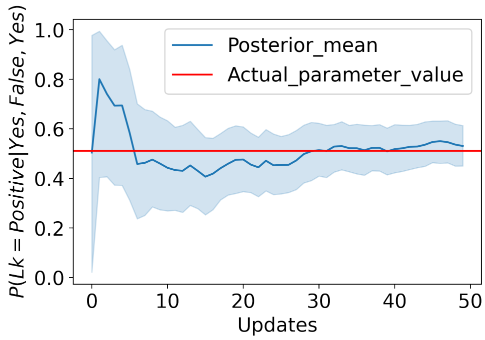
(C)



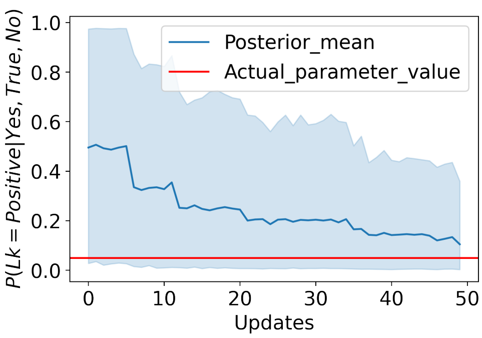
(D)



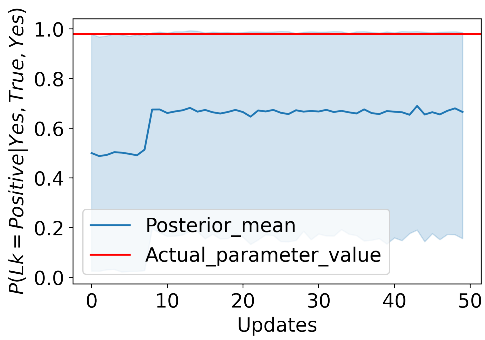
(E)



(F)



(G)



(H)

FIGURE A.13: Mean and 95% credible intervals for pipe leak status CPT parameter posterior distributions based on 250 data updates. Plot continuation for pipe leak status CPT parameters

Bibliography

Adedeji, K. B., Hamam, Y., Abe, B. T. and, Abu-Mahfouz, A. (2017). "Burst leakage-pressure dependency in water piping networks: Its impact on leak openings". *2017 IEEE AFRICON*. IEEE, pp. 1502–1507.

Asnaashari, A., McBean, E. A., Gharabaghi, B. and, Tutt, D. (2013). "Forecasting watermain failure using artificial neural network modelling". *Canadian Water Resources Journal* 38.1, pp. 24–33.

Ben-Gal, I. (2008). "Bayesian networks". *Encyclopedia of statistics in quality and reliability* 1.

Beretta, S., Castelli, M., Gonçalves, I., Henriques, R. and, Ramazzotti, D. (2018). "Learning the structure of Bayesian Networks: A quantitative assessment of the effect of different algorithmic schemes". *Complexity* 2018.

Boughorbel, S., Jarray, F. and, El-Anbari, M. (2017). "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". *PloS one* 12.6, e0177678.

Boulos, P. F. and, Aboujaoude, A. S. (2011). "Managing leaks using flow steptesting, network modeling, and field measurement". *Journal-American Water Works Association* 103.2, pp. 90–97.

Cao, Y. (2014). "Study of Four Types of Learning Bayesian Networks Cases". *Applied Mathematics & Information Sciences* 8.1, p. 379.

Demissie, G. A. (2017). "Prediction of water distribution system pipes and isolation valves failure using Bayesian models with the consideration of soil corrosion and climate change". PhD thesis. University of British Columbia.

Dighade, R., Kadu, M. and, Pande, A. (2014). "Challenges in water loss management of water distribution systems in developing countries". *International Journal of Innovative Research in Science, Engineering and Technology* 3.6, pp. 13838–13846.

Eyuboglu, S., Mahdi, H., Al-Shukri, H. and, Rock, L. (2003). "Detection of water leaks using ground penetrating radar". *Proceedings of the Third International Conference on Applied Geophysics, Orlando-FL*, pp. 8–12.

Farley, M., Wyeth, G., Md., B. Z., Ghazali, Istandar, A. and, Singh, S. (2010). *THE MANAGER'S NON-REVENUE WATER HANDBOOK FOR AFRICA A GUIDE TO UNDERSTANDING WATER LOSSES*.

Farmani, R., Kakoudakis, K., Behzadian, K. and, Butler, D. (2017). "Pipe failure prediction in water distribution systems considering static and dynamic factors". *Procedia Engineering* 186, pp. 117–126.

Ferguson, T. S. (2000). "Linear programming: A concise introduction".

Frauendorfer, R. and, Liemberger, R. (2010). *The issues and challenges of reducing non-revenue water*. Asian Development Bank.

Germanopoulos, G. (1985). "A technical note on the inclusion of pressure dependent demand and leakage terms in water supply network models". *Civil Engineering Systems* 2.3, pp. 171–179.

Giustolisi, O., Savic, D. and, Kapelan, Z. (2008). "Pressure-driven demand and leakage simulation for water distribution networks". *Journal of Hydraulic Engineering* 134.5, pp. 626–635.

Hunaidi, O., Wang, A., Bracken, M., Gambino, T. and, Fricke, C. (2004). "Acoustic methods for locating leaks in municipal water pipe networks". *International conference on water demand management*. Dead Sea Jordan, pp. 1–14.

Jafar, R., Shahrour, I. and, Juran, I. (2010). "Application of Artificial Neural Networks (ANN) to model the failure of urban water mains". *Mathematical and Computer Modelling* 51.9-10, pp. 1170–1180.

Kabir, G., Demissie, G., Sadiq, R. and, Tesfamariam, S. (2015). "Integrating failure prediction models for water mains: Bayesian belief network based data fusion". *Knowledge-Based Systems* 85, pp. 159–169.

Kabir, G., Tesfamariam, S., Francisque, A. and, Sadiq, R. (2015). "Evaluating risk of water mains failure using a Bayesian belief network model". *European Journal of Operational Research* 240.1, pp. 220–234.

Karkera, K. R. (2014). *Building probabilistic graphical models with Python*. Packt Publishing Ltd.

Kleiner, Y. and, Rajani, B. (2002). "Forecasting variations and trends in water-main breaks". *Journal of infrastructure systems* 8.4, pp. 122–131.

Korb, K. B. and, Nicholson, A. E. (2010). *Bayesian artificial intelligence*. CRC press.

Le Gat, Y. and, Eisenbeis, P. (2000). "Using maintenance records to forecast failures in water networks". *Urban water* 2.3, pp. 173–181.

Liemberger, R. and, Farley, M. (2004). "Developing a nonrevenue water reduction strategy Part 1: Investigating and assessing water losses". *Paper to IWA Congress*. Citeseer.

McKenzie, R. (1999). *Development of a standardised approach to evaluate burst and background losses in water distribution systems in South Africa*.

Ngubane, M. S. (2009). "Planning for recreational facilities and open spaces: a case study of Esikhawini Township at Umhlathuze Municipality, KwaZulu-Natal." PhD thesis.

Ogutu, A., Kogeda, O. P. and, Lall, M. (2016). "Decoding leakage tendencies of water pipelines in dolomitic land: a case study of the city of Tshwane". *Proceedings of The World Congress on Engineering and Computer Science*, pp. 931–936.

Ogutu, A. G., Kogeda, O. P. and, Lall, M. (2018). “A Probabilistic Assessment of Location Dependent Failure Trends in South African Water Distribution Networks”. *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Vol. 1.

Ogutu, G. A., Okuthe, P. K. and, Lall, M. (2017). “A Review of Probabilistic Modeling of Pipeline Leakage using Bayesian Networks”. *Journal of Engineering and Applied Sciences* 12.12, pp. 3163–3173.

Parvizsedghy, L., Gkountis, I., Senouci, A., Zayed, T., Alsharqawi, M., El Chanati, H., El-Abbasy, M. and, Mosleh, F. (2017). “Deterioration assessment models for water pipelines”. *International Journal of Civil and Environmental Engineering* 11.7, pp. 1013–1022.

Perdikou, S., Themistocleous, K., Agapiou, A. and, Hadjimitsis, D. G. (2014). “Introduction—The problem of water leakages”. *Integrated Use of Space, Geophysical and Hyperspectral Technologies Intended for Monitoring Water Leakages in Water Supply Networks*. IntechOpen.

Poulakis, Z., Valougeorgis, D. and, Papadimitriou, C. (2003). “Leakage detection in water pipe networks using a Bayesian probabilistic framework”. *Probabilistic Engineering Mechanics* 18.4, pp. 315–327.

Puust, R., Kapelan, Z., Savic, D. A. and, Koppel, T. (2010). “A review of methods for leakage management in pipe networks”. *Urban Water Journal* 7.1, pp. 25–45.

Saghi, H. and, Aval, A. A. (2015). “Effective factors in causing leakage in water supply systems and urban water distribution networks”. *American Journal of Civil Engineering* 3.2, p. 60.

Sharma, S. (2008). *Leakage management & control (An overview)*. URL: `http://www.switchurbanwater.eu/outputs/pdfs/GEN_PRS_Leakage_Management_and_Control_AC_Apr08.pdf` (visited on 03/14/2020).

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

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

South African Weather Service (2020). *Annual State of the Climate of South Africa 2019*. South African Weather Service.

Strijdom, L., Speight, V. and, Jacobs, H. E. (2017). "An assessment of sub-standard water pressure in South African potable distribution systems". *Journal of Water, Sanitation and Hygiene for Development* 7.4, pp. 557–567.

Uusitalo, L. (2007). "Advantages and challenges of Bayesian networks in environmental modelling". *Ecological modelling* 203.3-4, pp. 312–318.

Warming, L. (2019). *Fighting Non Revenue Water With Smart Metering*. URL: https://www.kamstrup.com/en-en/blog/world-water-day-2019 (visited on 03/01/2020).

Water-Operations-Partnerships (2011). *The state of African utilities: Performance assessment and benchmarking report*.

Water, U.-H. L. V., team, S. I., Water, N. and, team, S. C. (2012). *Utility Management Series for Small Towns: Leakage Control Manual*. Volume 5.

WHO (2001). *Leakage Management and Control - A Best Practice Training Manual*. (Visited on 03/07/2020).

WorldBank (2018). *Kosovo Water Security Outlook*. Word Bank.

WRDMAP (2010). *Active Leakage Control as a Key Component in Increasing Efficiency in Urban Water Supplies [Thematic Paper 3.3]*.

El-Zahab, S. and, Zayed, T. (2019). “Leak detection in water distribution networks: an introductory overview”. *Smart Water* 4.1, p. 5.