Advancing Quantitative Risk Analysis for Critical Water Infrastructure

by

Thomas Ying-Jeh Chen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2019

Doctoral Committee:

Professor Seth D. Guikema, Chair
Associate Professor Eunshin Byon
Professor Glen T. Daigger
Professor Yafeng Yin

Thomas Ying-Jeh Chen

tyjchen@umich.edu

ORCID iD: 0000-0001-6698-7547

ACKNOWLEDGEMENTS

First and foremost I would like to thank my advisor, Dr. Seth Guikema, for the opportunity to pursue my doctorate studies under his guidance. Throughout my time at the University of Michigan, I have learned a tremendous amount both in terms of subject matter and research skill. I am grateful for the support Dr. Seth Guikema has given me to pursue research endeavors that interest me, and also opening new doors of academic interests along the way.

Secondly, I would also like to thank all the other faculty members and colleagues who I have worked together with throughout my PhD: Dr. Terje Aven, Dr. Sara Shashaani, Dr. Pascal Van Hentenryck, Valerie Washington, Jared Beekman, Craig Daly, Connor Riley. I have learned so much from each collaborator and the research we've done was truly enriched as a result.

Thirdly, I would like to thank my dissertation committee members for their guidance: Dr. Eunshin Byon, Dr. Glen Daigger, Dr. Yafeng Yin.

I would also like to thank the members of my research group, in particular: Tom Logan, Anna White, Tim Williams, Elnaz Kabir, Chengwei Zhai, Caroline Johnson.

I want to also thank all the friends i've made over the past 4 years, in particular: Gian Garcia, Justin Haney, Adam Vandeusen, Sammi Meister, Lauren Biernacki, and Ciara Timban. Finally, I would like to thank my family: my sister Jennifer, my dad Steve, and my mom Yvonne for their unwavering support.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES

LIST OF TABLES

TABLES

LIST OF APPENDICES

APPENDIX

ABSTRACT

Critical infrastructure systems play a vital role in the supply of lifeline services to businesses and the wider public. It is of paramount importance for national security, public health, and economic prosperity that these critical structures function properly. Unfortunately, with respect to drinking water infrastructures in the US, much of the pipeline assets are nearing the end of their useful life and utilities are challenged with maintaining these systems with limited budgets and information.

Risk analysis is a useful decision making tool which can allow managers to better identify weaknesses, and aid better investment decisions regarding maintenance, inspection, and repair. The current practice for risk analysis and management of critical water systems falls short of the approaches preferred by risk researchers. The aim of this thesis is to advance to practice and theory. This involves the evaluation of existing methods as well as the incorporation of modern analytical tools to fundamentally advance the state of practice. This thesis first critically analyzes a popular risk analysis standard (J100-10) to establish the knowledge gap between practice and theory in the water domain. Two quantitative methodologies are then explored: machine learning and mathematical optimization. The research here demonstrates how they can be integrated into a broader risk framework and used to improve assessments for water systems.

The work presented in this dissertation represents a significant contribution to the field of infrastructure risk and reliability analysis. While the domain application is specific to drinking water systems, the techniques can be applied for other types of networked infrastructures.

CHAPTER 1

Introduction

In many parts of the United States, drinking water infrastructure is nearing the end of its useful life and upgrades are needed to ensure the consistent delivery of complaint water with federal quality standards to end users [10]. It is estimated that over one trillion US dollars of capital investments are required for necessary upgrades to the nation's water infrastructure [19]. Furthermore, there are additional social and economic burdens when failures to the distribution system occur. These costs are often associated with public health risks [130], stoppage of service, and public disruption as a result of emergency repairs [205].

Despite drinking water systems being recognized as one of the most critical infrastructures [70], current practice for quantitative risk management of these systems falls short of the approaches preferred by risk researchers. Some examples include: 1) inappropriate metrics used to quantify risk may be simple to implement but can lead to misallocation of resources [60] and 2) the scope of most frameworks do not adequately address emerging threats to water systems [200]. The aim of the proposed research is to improve infrastructure risk analysis practice and theory. This involves the evaluation of existing methods as well as the incorporation of modern analytical tools to fundamentally advance the state of practice.

The research presented in this thesis aims to achieve the following goals:

1. Evaluate the J100-10 risk analysis standard against the state of the art.
2. Advance the literature on pipe break machine learning techniques to better forecast future failures.
3. Apply modern optimization tools to plan for inspection routes when having to account for limitations of inspection tools.

Together these goals combine to form a holistic body of work that represents a significant contribution to the risk and resilience literature for infrastructure systems. Section 1.3 discusses why these specific subjects are selected for this thesis.

1.1 Drinking Water Distribution System Overview

This section first provides some background on the schematic of drinking water distribution systems. In the US, water supply systems are usually owned and operated by local municipalities, but occasionally belong to private organizations. Figure 1.1 illustrates the basic process in how drinking water is extracted from its natural source, delivered to the end user, and transported back to the natural source.

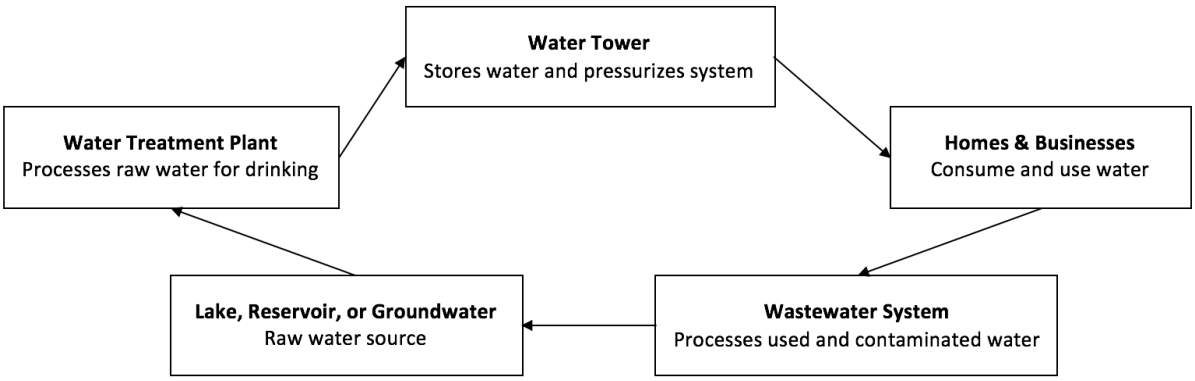


Figure 1.1: Schematic of Drinking Water System Cycle.

Large aqueducts or transmission mains deliver the water from a reservoir, lake, or underground aquifer to the treatment facility. There the water is processed and treated to meet chemical quality standards [10] such that it is fit for human consumption. Then the water is transported through transmissions mains either directly to both the end user as well as storage facilities scattered throughout the service area. In a supply network there needs to be sufficient positive pressure to ensure that water reaches all consumers. Pressure and flow is regulated by pumps and valves as well as natural elevation itself, where pressure is generated by having water flow from high to low points. Water towers provide storage capacity at raised elevations. When water is released, gravity provides additional pressurization for the delivery of water to consumers. Once homes and business have finished consumption, the wastewater system extracts and delivers the water back to a treatment site [192]. There the water is once again treated to meet quality standards such that it can be returned to the raw source (e.g. river, lake, ocean).

1.2 Research Motivation

This thesis focuses on advancing the risk analysis methodology of water distribution systems. Today, most Americans and American businesses get their water from one of approximately 51000 community water systems, which make up over 1 million miles of distribution pipe [39]. Many of these systems are small, serving populations of 500 or fewer people. Approximately 10000 large water systems located in densely populated areas serve around 82% of the total population. These vast and intricate systems that deliver the lifeline good of potable water are often taken for granted. However, increasing rates of failure due to aging combined with limited funding for asset renewal [38] has highlighted the importance for sound management of these buried infrastructures.

To provide a first inclination on the urgency of the matter, a report by the American Water

Works Association surveyed various types of buried pipeline assets and their corresponding design life [37]. Cast-iron distribution pipes laid in the late 1880s have an average lifespan of 120 years, while those laid in the 1920s which were constructed using different manufacturing techniques have a lifespan of 100 years. The pipes laid during the post-World War II economic boom were expected to have a useful life span of about 75 years. These values indicate that much of the underground pipeline network for potable water systems in the US will be due for replacement in the next 2 decades [54].

As a result, being able to accurately identify the assets that are most prone to failure, and making the right mitigation decisions on these risks, is of utmost importance for utility managers. However, a study by the EPA [38] points out that a simple prioritization based on material age would lead to a sub optimal investment strategy. The rate of deterioration of a water system is not a simple function of age but rather the cumulative effect of the internal and external forces acting on it. These include external and internal corrosion, hydraulic conditions inside the pipe, the soil properties surrounding the pipe, as well as the piping material itself. Therefore, in order for an effective use of limited capital resources, better frameworks for assessing system risk and decision making is needed.

The current state of practice for risk assessment of water infrastructure, and other infrastructure systems in general, lacks the sophistication found in literature [60]. In this thesis, I first critically evaluate the J100-10 [23] risk analysis standard which is widely used in the water industry. By highlighting its limitations and weaknesses, we can identify the current gap between practice and theory. While there are numerous critiques published in the literature regarding the J100-10 and other similar risk analysis frameworks, much of the assessments have solely focused on the individuals approaches rather than comparing multiple methods against one another [183]. The first part of the thesis aims to advance our understanding of some popular methods in the context of risk prioritization.

Secondly, modern mathematical techniques can lend themselves useful for addressing practical challenges in condition assessment and asset management. The rest of this thesis explores how statistical modeling and optimization tools can be applied to advance current practice of risk assessment. Therefore this work not only fills a research gap, but also makes advances towards introducing vital tools for practitioners which guide resource allocations to optimize risk reductions.

1.3 Project Objectives

The goal of this thesis is to enhance our understanding of risk assessment approaches in their utility for infrastructure asset management, as well as advancing the state of practice by examining the use of analytical tools to provide useful decision support. Table 1.1 below summarizes the intellectual

contribution and methodology focus for each of the remaining chapters of this dissertation.

Table 1.1: Intellectual Contribution and Methods Focus of Dissertation Chapters.

Chapter	Intellectual Contribution	Methodology Focus
2	Gap analysis between risk analysis literature and practice.	Foundations of the risk analysis field.
3, 4	Improved modeling for pipeline inspection planning.	Mathematical optimization.
5, 6	Extending the existing techniques for pipe break prediction.	Statistical modelling, supervised learning, unsupervised learning.

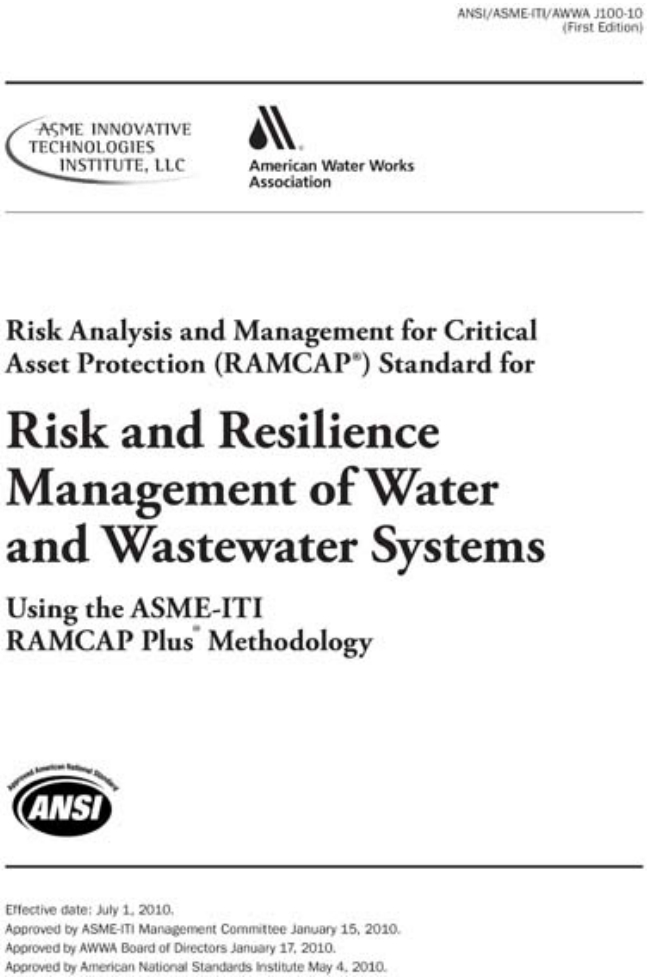
To help frame the research better, we first begin with an analysis on the current state of risk assessment protocols in the water industry. The goal of the thesis is to advance practice and theory for the risk and resilience assessments of water systems, a natural starting point is to first understand what is being done in currently practice and to establish a knowledge gap. Two methodological techniques are the focus for the remainder of the thesis: statistical learning, and mathematical optimization, each of them pertaining to specific processes in the risk assessment framework. The topic of pipe break prediction is selected because it is a critical component for any risk analysis for water systems due to pipeline failure. In order for utilities to better allocate their resources, they need to target the assets which are in poorest condition. The contributions here are to 1) address practical challenges when implementing these models, and 2) demonstrate an advancement in the modeling methodology which can improve performance. The problem of inspection planning is also selected because it is another key component of a risk assessment process. Robotic condition assessments are expensive, and utilities want to get the maximum return on investments for each inspection. This thesis aims to explore how the techniques of mathematical optimization can be used to help identify good routes for the use of robotic inspection tools.

1.3.1 Project 1: Review and Evaluation of the J100-10 Risk and Resilience Management Standard

The objective of the first project is to identify the gap between risk analysis practice and theory in the water infrastructure sector. To address the ever growing need of having a uniform and holistic risk assessment framework within the water industry, in 2010 the American Water Works Association published the "J100-10 Risk and Resilience Management of Water and Wastewater Systems" (referred to simply as J100-10) standard to be used by practitioners across the US [23]. This standard was modeled closely after the 2006 cross-sector infrastructure risk assessment framework

"Risk Analysis and Management for Critical Asset Protection" (or RAMCAP^{TM}) published by the Department of Homeland Security, but adapts the methods to be specific to the water sector.

In Chapter 2 of this thesis, I evaluate the methods listed in J100-10 against the current state of the art in the risk analysis literature. The analysis covers both conceptual and practical limitations that can lead to inadequate risk characterizations and misguide decision makers.


ANSI/ASME-ITI/AWWA J100-10
(First Edition)

ASME INNOVATIVE TECHNOLOGIES INSTITUTE, LLC

American Water Works Association

Risk Analysis and Management for Critical Asset Protection (RAMCAP®) Standard for

Risk and Resilience Management of Water and Wastewater Systems

Using the ASME-ITI RAMCAP Plus® Methodology

ANSI

Effective date: July 1, 2010.
Approved by ASME-ITI Management Committee January 15, 2010.
Approved by AWWA Board of Directors January 17, 2010.
Approved by American National Standards Institute May 4, 2010.

Figure 1.2: Front Cover of the J100-10 Risk Management Standard [23].

To my knowledge, no previous work has holistically reviewed an entire risk analysis standard against the foundations of the risk analysis field in this manner. The contribution here is to identify areas where risk analysis practice in the water infrastructure industry can be improved. By presenting this review of the J100-10 and highlighting of its main shortcomings we aim to establish the gap between practice and theory, and ultimately guide future improvements made to the standard.

1.3.2 Project 2: Routing Optimization of Robotic Pipeline Inspections

To aid the development of an effective asset management plan, inspection operations are often employed to gather information on the current condition of the system. Some commercially available products are shown in Figure 1.3.

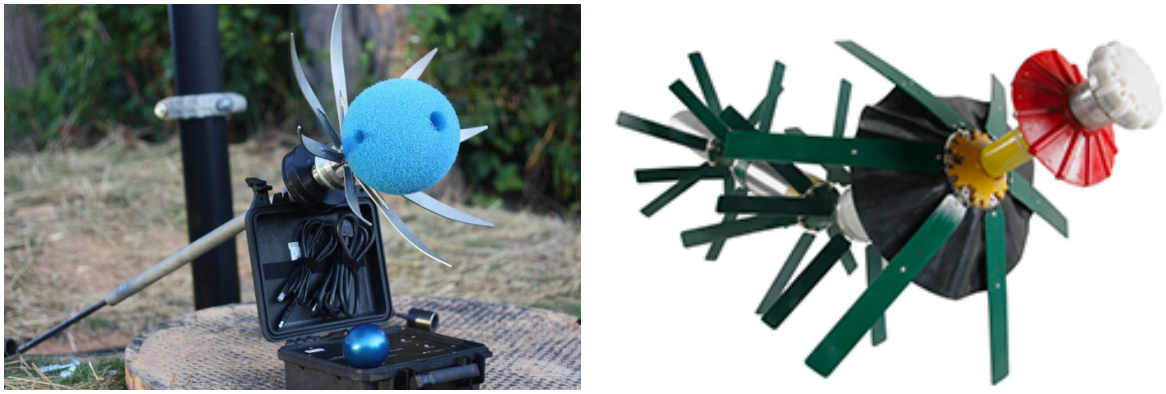


Figure 1.3: Commercial Robotic Pipeline Inspection Technologies developed by Pure Technologies U.S. [140]. Left: Smartball^{TM}, Right: PipeDiver^{TM}

Smartball^{TM} uses an acoustic technology to target leak detection, while PipeDiver^{TM} uses electromagnetic technology to scan for pipe wall defects [140]. These are two examples of the variety of inspection technologies utility managers could leverage, and having an effective plan for their deployment can increase the return on investments. However prioritizing inspections based solely on the risk of inspected assets, a widely popular approach in both the academic literature and practice, can lead to suboptimal results because the operational limits of the inspection technology must be accounted for [65]. These considerations can affect the quality of the collected data, and ultimately the return on investment for inspection expenditures.

The goal of this project is to present a mathematical optimization formulation for the planning of pipe inspection routes, and demonstrate how it can be solved in networks of various sizes and complexities. The work here is divided across two chapters. Chapter 3 introduces a general optimization framework for the problem, and explores the application of heuristic methods for identifying routes in both synthetic and real networks. Chapter 4 revisits the same problem and presents a full integer programming model, a variety of exact solution algorithms are demonstrated on a real network and their scalabilities are examined. The contribution of this project is to present both a heuristic and exact framework for the mathematical optimization of inspection routes while considering platform limitations. To my knowledge, no previous academic literature tackles the problem in this manner.

By presenting and comparing both exact and approximate solution methods, we aim to provide insight for decision makers on how mathematical optimization can provide useful decision support

for inspection planning.

1.3.3 Project 3: Pipe Break Machine Learning for Maintenance Prioritization

The failure of drinking water systems occurs under three broad categorizations: source water contamination, treatment deficiencies, and distribution network failure [158]. The focus of this project is the predictive modeling of distribution network failures, specifically water main breaks.

While there are many previous works on pipe break models [117, 169, 201], many of them are inapplicable simply because utilities do not have records of information regarding the pipelines themselves. The work is also divided into two chapters. Chapter 5 tackles the data scarcity problem many water utilities have [101, 43] by examining whether pipe break models built using only public data can accurately forecast future failures. The contribution here is to determine if statistical learning techniques can provide useful decision support for utility managers in identifying system vulnerabilities without any available system data. Chapter 6 expands the modeling literature by combining supervised and unsupervised modelling techniques, it explores if information about spatial clusters of pipe breaks can improve predictive accuracy. To my knowledge, no previous academic work on pipe break modeling has evaluated the combination of clustering and machine learning techniques for predictive accuracy.

Ultimately, the goal is to extend the literature on statistical modeling of pipeline failures, and demonstrate that better prioritization for renewal and replacement spending can be achieved.

CHAPTER 2

Review and Evaluation of the J100-10 Risk and Resilience Management Standard for Water and Wastewater Systems

Risk analysis standards are often employed to protect critical infrastructures which are vital to a nations security, economy, and safety of its citizens. We present an analysis framework for evaluating such standards and apply it to the J100-10 risk analysis standard for water and wastewater systems. In doing so, we identify gaps between practices recommended in the standard and the state of the art. While individual processes found within infrastructure risk analysis standards have been evaluated in the past, we present a foundational review and focus specifically on water systems. By highlighting both the conceptual shortcomings and practical limitations, we aim to prioritize the shortcomings needed to be addressed. Key findings from this study include: 1) risk definitions fail to address notions of uncertainty, 2) the sole use "worst reasonable case" assumptions can lead to mischaracterizations of risk, 3) analysis of risk and resilience at the threat-asset resolutions ignores dependencies within the system, and 4) stakeholders values need to be assessed when balancing the tradeoffs between risk reduction and resilience enhancement.

Keywords: Drinking Water Distribution System, Asset Management, Risk Analysis and Management

Note: The research presented in this chapter has been accepted for publication at the Journal of Risk Analysis, acceptance date on Oct 14, 2019. Co-authors: Valerie Nicole Washington, Seth David Guikema, Terje Aven.

2.1 Introduction

2.1.1 Background

Following the attacks of September 11, 2001, the federal government recognized the need to define and prioritize the requirements for protecting the nation's infrastructure [23]. As a result, the Homeland Security Act of 2002 [59] prescribed a cross-sector risk assessment plan to identify vulnerabilities for all critical infrastructure and key resources (CIKR) and define a framework to prioritize defense resource allocation. As defined in the National Infrastructure Protection Plan (NIPP) of 2009 [70], CIKRs include energy, water (drinking and waste), transportation, communications, and government facilities.

The potential importance of a uniform risk analysis procedure was recognized when the White House recruited the American Society of Mechanical Engineers (ASME) to develop a procedure applicable across different types of infrastructure [23]. The goal was that common terminology, metrics, and methodology would facilitate comparisons within and across CIKR sectors, and support decision making for risk reduction investments. In 2006, ASME released the specifications for Risk Analysis and Management for Critical Asset Protection (RAMCAP^{TM}), which serves as the basis for J100-10 [23]. RAMCAP^{TM} defines a seven-step process (discussed in Section 2.1.2) to assess risk and resilience for a given asset and to prioritize countermeasures.

RAMCAP^{TM} outlines three major objectives [21]: 1) to define a common framework for owners and operators of critical infrastructure to assess consequences and vulnerabilities relating to terrorist attacks on their assets and systems, 2) to provide guidance on methods that can be used to assess and evaluate risk through this framework, and 3) to provide an efficient and consistent mechanism to report risk information to the U.S. Department of Homeland Security (DHS).

The American Water Works Association (AWWA) adopted the RAMCAP^{TM} seven-step framework to create a water and wastewater sector specific risk analysis standard, and in 2010 published the J100-10 standard for Risk and Resilience Management of Water and Wastewater Systems [23]. While RAMCAP^{TM} and J100-10 were initially developed with the intent of analyzing risks associated with terrorist attacks [21], subsequent updates expanded the analysis breadth to include a variety of threats (e.g. natural hazards, dependency, and proximity threats). Beyond allowing utility operators to systematically assess risk, J100-10 provides methods to evaluate options for improving weaknesses in water and wastewater systems [23]. The aim is to prioritize the actions that better mitigate risks and can lead to more resilient critical infrastructure.

We use the term risk analysis in this chapter as it is defined in the Society of Risk Analysis (SRA) glossary [186]. Risk analysis is "a systematic process to comprehend the nature of risk and to express risk with the available knowledge". A fundamental principles document from SRA highlights some key criteria for a high quality risk analysis [187]: it needs to be reliable, valid,

and the decision maker needs to have confidence in the results. Reliable means that there is reproducibility in the process (encompassing analyst, methods, procedures etc.), and valid meaning there is success at characterizing the relevant risks. A key is that the degree of knowledge (or lack thereof) of the analyst is properly communicated to the decision maker. The ultimate goal is to inform and support decision making for risk management.

In this chapter, we provide an analysis framework for assessing risk analysis standards and present a holistic review of J100-10 to highlight its conceptual shortcomings and practical limitations. Our goal in this chapter is to begin a conversation about how to strengthen the J100-10 moving forward.

2.1.2 J100-10 Definitions

Two key components of a risk management standard are the definitions and the underlying conceptualizations of risk. Before proceeding further with our assessment, we include key definitions from J100-10 [23], which were adopted from RAMCAP^{TM}. The following definitions are taken verbatim from the standard, and a discussion on their sufficiency is presented in later sections. For ease of reading, we have eliminated block quotations.

Risk is "the potential for loss or harm due to the likelihood of an unwanted event and its adverse consequences" (page 18, J100-10 manual [23]). J100-10 uses the RAMCAP^{TM} approach to quantify risk using Equation (2.1) below [23]:

$$\text{Risk} = \text{Threat Likelihood} \times \text{Consequence} \times \text{Vulnerability} \tag{2.1}$$

Threat likelihood is "the probability that an undesired event will occur" (page 49, J100-10 manual [23]). With natural hazards, J100-10 states that this should be "the historical frequency of similar events, unless there is a belief that the future will differ from the past. With malevolent threats, the likelihood is a function of available intelligence, the objectives and capabilities of the adversary, and the attractiveness as a target" (page 49, J100-10 manual [23]).

Consequence is defined as "the immediate, short- and long-term effects of a malevolent attack or natural incident" (page 43, J100-10 manual [23]), which J100-10 specifies should be estimated exclusively on a "worst reasonable case basis" (page 8, J100-10 manual [23]). These effects include fatalities, injuries, and losses suffered by the owner of the asset and by the community served by that asset.

Vulnerability is "an inherent state of the system (e.g. physical, technical, organizational, cultural) that can be exploited by an adversary or impacted by a natural hazard to cause harm or damage" (page 49, J100-10 manual [23]). J100-10 specifies that vulnerability should be expressed as the likelihood of an event resulting in the estimated consequences, given that the event occurs.

There are various definitions of risk presented in the academic literature [186], a discussion on the sufficiency of the J100-10 definition in Equation (2.1) is presented in section 2.4.1.

Resilience is "the ability of an asset or system to withstand an attack or natural hazard without interruption of performing the asset or systems function or, if the function is interrupted, to restore the function rapidly" (page 19, J100-10 manual [23]). Resilience can be considered at the threat-asset level or at the system level. Asset-level resilience is defined on a scale such that lower values indicate greater resilience. It can be calculated using the following three metrics:

1. *Operational Resilience Metric (ORM)* measures the service denial due to a threat-asset pair, weighted by vulnerability and threat likelihood. It is calculated following equation (2.2) [23]:

$$\text{ORM} = \text{Duration} \times \text{Severity} \times \text{Vulnerability} \times \text{Threat Likelihood} \tag{2.2}$$

   Duration is the time, in days, of service denial and severity is the amount of service denied (in gallons of water per day).

2. *Owners Economic Resilience Metric (OERM)* converts ORM into a dollar value and characterizes the financial loss to the utility owner, and is calculated following equation (2.3) [23]:

$$\text{OERM} = \text{ORM} \times \text{Preincident Unit Price} \tag{2.3}$$

3. *Community Economic Resilience Metric* is the lost economic activity, in dollars, to the community served by the utility. Estimating these impacts requires a regional simulator and/or economic model to fully capture the direct and indirect effects.

An evaluation on the conceptualization of the J100-10 resilience definition (equation (2.2) and equation (2.3)) is presented in section 2.4.3.

2.1.3 J100-10 Risk Analysis Process

J100-10 outlines a seven-step risk analysis process, as shown in Figure 2.1 below.

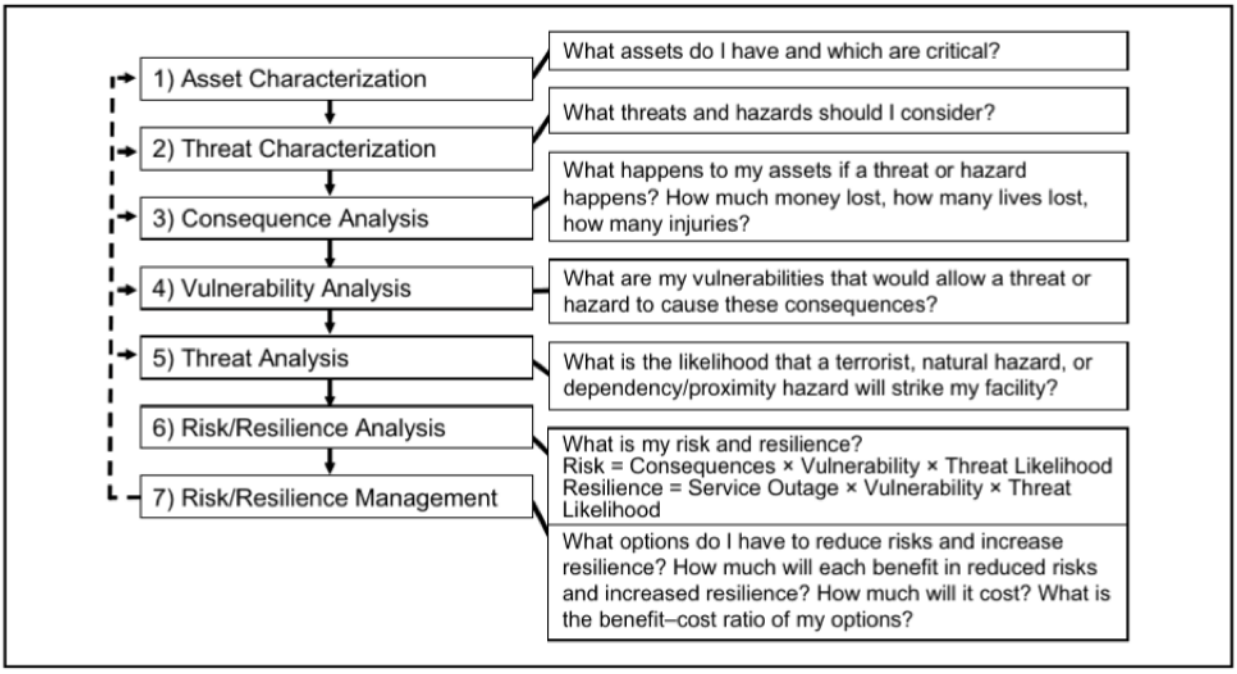


Figure 2.1: The adopted RAMCAP^{TM} process in the J100-10. Taken from the J100-10 Risk Management Standard [23].

Below we provide a brief description of each of the seven steps of the assessment methodology.

1. *Asset Characterization:* Identify the critical assets, which if compromised, would inhibit the organization from carrying out its mission or operational goals. Asset ranking can be used to prioritize components for analysis if the number is too large to include them all.

2. *Threat Characterization:* Identify and describe reference threats scenarios to estimate vulnerability and consequence. Reference categories include malevolent threats, natural hazards, and proximity and dependency threats. Additional threats can be added as long as they are used in the analysis of all assets under consideration.

3. *Consequence Analysis:* Identify and estimate the "worst reasonable consequence" generated by each threat-asset combination. Consequence metric categories include fatality count, serious injury count, financial loss to the owners, and economic losses to the community.

4. *Vulnerability Analysis:* Estimate the conditional likelihood that, given an adverse event occurs on the asset, the estimated consequences will occur. Some methods for estimating this value suggested by J100-10 include direct expert elicitation, path analysis, vulnerability logic diagrams, event trees, or a hybrid of these methods.

5. *Threat Assessment:* Estimate the probability that each of the identified threats will occur in a given time frame (typically one year). J100-10 provides guidance on how to estimate these values for different types of threats, e.g., an event tree based approach for malevolent threats, or using federal agency-specific resources for various natural hazards (e.g., the Federal Emergency Management Agency (FEMA) flood insurance rate maps, or the National Hurricane Center risk analysis program, HURISK).

6. *Risk and Resilience Assessment:* Use Equation (2.1) to calculate the risk metric, equation (2.2) and (2.3) to calculate the resilience metrics for each threat-asset pair. J100-10 outlines a utility resilience index (URI), which assesses the operational and financial capabilities of the utility to cope with various incidents that have the potential to disrupt service.

7. *Risk and Resilience Management:* Implement actions to achieve a level of acceptable risk and resilience at an acceptable cost. Benefit-cost analysis is useful for suggesting potential actions, e.g. new security countermeasures or consequence mitigation features. Benefits are calculated as the expected risk reduction or resilience increase and costs are defined in dollar units.

2.2 Literature Review

2.2.1 Standardized Risk Analysis Methods in the Water Sector

By one estimate, there are more than 250 critical infrastructure risk analysis methods [134]. Many of these methods have been used in other risk analysis standards to study water infrastructure prior to the development of RAMCAP^{TM} or J100-10. Three of these prior standards in particular have been widely documented and used [23]. They are 1) the Risk Assessment Methodology - Water (RAM-W^{TM}) [100] developed by Sandia National Laboratories, 2) the Scientech and PA Consulting Group Vulnerability Self-Assessment Tool (VSAT^{TM}) [99], and 3) the National Rural Water Association Security and Environmental Management System (SEMS^{TM}) [22]. RAM-W^{TM} was specifically developed to evaluate the risk of adversarial threats. It is a water sector-specific version of the RAMCAP^{TM} standard (see Section 2.1.3 for general seven-step approach) that focuses on risk quantification, while J100-10 analyzes both risk and resilience. VSAT^{TM} was originally intended for use by wastewater utilities, but was later adapted to include drinking water utilities. It uses a risk matrix, estimated as a combination of qualitative criticality and vulnerability ratings, to determine which assets need security improvements [13]. SEMS^{TM} was developed for small systems in rural areas. It uses a simple "yes" or "no" questionnaire to help owners of utilities identify vulnerabilities and improvement actions. While it does not describe any explicit quantification of

risk, SEMS^{TM} provides information about the operating conditions and asset status of the utility.

Following the release of RAMCAP^{TM}, the VSAT^{TM} and RAM-W^{TM} standards have been modified to be consistent with the RAMCAP^{TM} seven-step framework. SEMS^{TM} has been adapted to include questions that cover basic information required by RAMCAP^{TM} [23], such as certain security measures. Despite the wide variety of available assessment frameworks, we chose to evaluate J100-10 because it was the first standard to include both a wide range of risk sources and all types of water infrastructure in its analysis.

2.2.2 J100-10 and RAMCAP^{TM} Critiques

In this section, we review some of the previous critiques and contextualize them within our broader review of J100-10. The presented critiques of J100-10 have broader implications for the parent RAMCAP^{TM} standard. Because RAMCAP^{TM} serves as the foundation of J100-10, we include critiques of this standard as well.

While the J100-10 and RAMCAP^{TM} standards do not mandate that utilities report risk assessment results or implement countermeasures, some utilities have documented the use of the approach to guide decision making to improve facility security. A cross-infrastructure sector implementation is found in Krimgold (2012) [122], where the RAMCAP^{TM} methodology is implemented to analyze power, water, transport, and communications systems in an unnamed metropolitan region. This is done to better identify specific threats and their respective consequences across sectors. The study concludes that the RAMCAP^{TM} asset-level assessment provides useful guidance on defining risk through operational units, which assists in the prioritization of short- and long-term risk management goals.

Herrare et al. (2017) [93] examine an implementation of RAMCAP^{TM} to Colorado's transportation sector, which helped identify system vulnerabilities and assisted in supporting federal emergency response funding requests. The Department of Transportation favored the benefit-cost analysis within the risk and resilience management step used to evaluate multiple mitigation options since it provided a data-driven approach to support decision making.

An implementation specific to the water sector is found in Kerr et al. (2015) [109], which provides a case study from a utility in Peel, CA. In this study, the utility uses the J100-10 assessment method to develop a long-term strategy to manage and reduce risk through capital investment and operational planning. The authors find that using the J100-10 analysis framework gives the utility a more complete and unbiased understanding of the assets that are at highest risk, which allows for a clearer process for capital investment decision making. In addition, the risk and resilience management guidelines provide a framework for the continual review and revision of the analysis as mitigation plans are implemented.

A number of academic studies have critiqued the risk assessment methodology outlined in the RAMCAP^{TM} standard. High-level critiques include Cox (2008) [60], which emphasizes the shortcomings of the threat-vulnerability-consequence triplet definition of risk as well as the ordinal scales used in the RAMCAP^{TM} risk calculation. Some of the main limitations discussed by Cox (2008) [60] are that RAMCAP^{TM} fails to address possible correlations between the threat, vulnerability, and consequence components. Additionally, it does not account for non-additivity of risk when aggregating from the analysis level of threat-asset pairs to system-level risk estimates, the use of ordinal scoring values to calculate risk can lead to sub-optimal allocation of resources for implementing countermeasures, and notions of uncertainty related to the estimates of threats and consequences are not addressed in the analysis.

Burkhart (2015) [55] identifies consistency and scope problems in the J100-10 standard; for example, the utility is given the choice to analyze the resilience at either the asset or system level, but no guidance is provided on how to choose between the two resolutions. Furthermore, no concrete process is outlined for defining a single level of acceptable risk, especially if multiple decision makers are involved. As a more general critique of assessments using risk-based scoring methods for resource allocation, Cox (2009) [62] specifies that such an approach often fails to account for interdependencies and risk externalities (risk for parts of a system changes as countermeasures are added) among the considered threats.

Critiques of specific steps within the J100-10 process have been discussed in the academic literature. Cox (2008) [61] highlights the limitations of using risk matrices to drive prioritization decisions. Such use of risk matrix methods from RAMCAP^{TM} can be found in the asset characterization step, which is used to screen assets for analysis to reduce the scope of the risk assessment. The study argues that risk matrices often have poor risk resolution and errors in risk estimation, which can lead to suboptimal prioritization decisions.

Consequence estimation, as defined in the J100-10 and RAMCAP^{TM} standards, are based solely on a "worst reasonable case" [23] premise, the common thinking being that this results in a conservative (inflated) estimate of risk intended to add a factor of safety. A case study in off-sea oil drilling presented by Huage et al. (2014) [91] highlight the limitations of this approach. The authors explain that uncertainties related to characterizing extreme outcomes and their likelihoods can limit the usefulness of an assessment.

The threat analysis step in the RAMCAP^{TM} methodology defines 41 reference threats, which include terrorist threats, natural hazards, and dependency hazards. The J100-10 standard uses the same 41 reference threats and provides details for analyzing risk from these threats. However, White et al. (2016) [200] recognize the failure of this process to account for key emerging threats (climate change, aging infrastructure, and cyber attacks) and propose 13 additional reference threats to address these emerging issues. As a follow up study, White et al. (2016) [199] use

a simulated RAMCAP^{TM} model to analyze the performance under the proposed set of 54 threats.

The risk and resilience analysis step defines risk as the product of the consequence, vulnerability, and threat likelihood, which make up the triplet definition of risk. The shortcomings of this approach is well established in the risk science literature, where the main concern is that potentials for extreme outcomes are not properly reflected. Alternative and more general perspectives have been developed where risk captures the triplet events, consequences, and uncertainties, see SRA (2015) [186] and Aven 2012 [28], 2017 [32]. These perspectives build on Kaplan and Garrick (1981) [106] who refer to risk qualitatively as "uncertainty plus damage".

As shown above, there have been multiple case studies reported on the implementation of the J100-10 standard in the water and wastewater sector and of RAMCAP^{TM} in other infrastructure systems. There are also a number of studies by risk analysts highlighting the limitations of RAMCAP^{TM} and the methodologies it recommends for analyzing risk and resilience. These critiques have focused on specific issues within certain steps of the analysis. In the subsequent sections we will present a more comprehensive critique of the J100-10 assessment process as a whole.

2.3 Analysis Framework

Here we define our framework for evaluating the J100-10 standard. The approach can be implemented for a variety of risk analysis standards outside the water infrastructure domain. Based on the criteria for a risk analysis outlined in Section 2.1.1, we identify two questions of emphasis: 1) are risk and other key concepts (e.g. probability and resilience) being characterized adequately?, and 2) are the recommended procedures in line with the state-of-the-art in risk science?

As a result, in this research we conceptually compare J100-10 against the state of the art in risk science. We choose this approach because it focuses on the foundational issues of the risk analysis field and measures the process against these established principles. An alternative approach is to implement both J100-10 and a second risk analysis method and compare their outputs. This can be tricky because various assessments are beset by tradeoffs of completeness, consistency, and timeliness [199]. The development of a process to directly compare multiple frameworks is beyond the scope of our analysis and is left for future research.

Our analytical framework can be divided into two categories: conceptual and practical limitations. The former addresses the theoretical shortcomings. The latter addresses specific steps which could lead to poor risk characterizations. We primarily focus on the risk analysis portion of J100-10, but also discuss its guidelines for assessing resilience. We present our findings of the conceptual and practical limitations in Sections 2.4 and 2.5, respectively.

2.4 Conceptual Limitations

In the following section, we identify conceptual gaps related to definitions of key terms, how they are calculated and interpreted in the standard, and how they relate to the state of the art in the field of risk analysis.

2.4.1 Definitions of Risk

The operating risk definition in the J100-10 standard falls short because concepts of uncertainties are not included. J100-10 uses the expected consequences definition of risk, which is calculated as the product of the probability of a threat event, the conditional probability that the event will lead to the worst-case consequences, and the consequences themselves. As discussed in Section 2.2.2, this understanding of risk has severe limitations and its use can seriously mislead decision makers.

An analysis of the literature shows that there are multiple definitions of risk: some are broader, while others lead more naturally to quantifiable equations. By distinguishing between the concept of risk and how it is measured, a consensus can be reached on characteristics of risk, as shown by the Society for Risk Analysis Glossary (2015) [186]. Aven (2012) [28] discusses the issue and argues that a notion of uncertainty is required to capture the concept of risk. Analysts classify uncertainty in two ways [167]: 1) aleatory uncertainty, which reflects variation in populations and 2) epistemic uncertainty, which reflects lack of knowledge. The latter type of uncertainty is key to understanding and characterizing risk, while the former is used to build probabilistic models, when justified, and support the epistemic uncertainty characterizations. Understanding where sources of uncertainty lie can help utilities better interpret assessment results and guide management decisions to reduce uncertainty for future analyses. J100-10 does not attempt to address uncertainty in the analysis process, evidenced by the fact the word "uncertainty" does not appear anywhere in the standard. While there is debate regarding how uncertainties should be characterized and propagated in assessments, e.g., some arguing probabilities fully capture uncertainty [202] and others advocating for other methods [167, 94, 80], it is evident that the current J100-10 framework falls short because uncertainty is not addressed at all.

Including the concept of uncertainty in the definition of risk can improve the assessment framework of J100-10. The most common method is probabilistic risk assessments (PRA) [17], which uses probabilities as the sole measure of uncertainty. Flage et al. (2014) [80] and Shortridge et al. (2017) [183] outline a variety of other analysis methods, from simpler models that use qualitative assessments of uncertainty, to more sophisticated technical models (e.g. use of possibility bounds and evidence theory).

Another approach is to assess the underlying strength of knowledge when using probabilistic judgments, for example, in relation to expert opinions. Experts include utility operators and share-

holders, and they can be used to assess threat likelihoods and consequence measures when data is unavailable [23]. Typically, a stronger background knowledge is correlated with lower degrees of uncertainty. In performing this assessment, the uncertainty description becomes a function of their strength of background knowledge [20]. Askeland et al. (2017) [20] present a framework to evaluate strength of knowledge, categorizing it as "weak", "moderate", or "strong" based on five criteria: 1) experts understanding of the phenomena, 2) reliability and availability of data, 3) agreement among experts, 4) identification, documentation, and soundness of assumptions, and 5) evaluation of knowledge gaps and changes in knowledge over time. Aven et al. (2013) [25] present an alternative method for assessing strength of knowledge through assumption deviation risk scores. Assumption deviation risk is defined as "risk related to a deviation between what has been assumed and what actually occurs" [17]. To assess the risk, consideration is given to deviation probabilities, consequences of deviation, and related strength of knowledge judgments. Subsequent updates to the J100-10 standard can employ one or more of these methods or develop methods more suitable for application in the water industry.

2.4.2 Concepts of Probability

Probabilities are an integral part of the risk assessment process in J100-10. The standard defines probability on page 43 as follows:

"*A measure of the likelihood, degree of belief, frequency, or chance that a particular event will occur in a period of time (usually one year) or number of iterations or trials. This is usually expressed quantitatively as a value between 0 and 1, a range of values between 0 and 1, a distribution (density function), or the mean of such a distribution. Probability can also be expressed in qualitative terms, e.g. low, medium, or high, if there is a common understanding of the meaning of the qualitative terms*" [23].

The definition presented is unclear in two ways. First, there are multiple ways outlined to represent probabilities. For clear interpretation of results to drive decision making, it is vital to have a consistent probability representation. Second, how these probabilities should be interpreted is left ambiguous. Aven and Reniers (2013) [35] highlight the practical importance for decision makers to understand what the risk analysis is communicating. For this reason, a concise definition of probability and its interpretation is required. Many previous studies have discussed this issue at length, see for example: White et al. (2016) [200, 199], Aven and Reiners (2013) [35]. The body of work categorizes probability into two major schools of thought: frequentist and Bayesian.

The "frequentist" interpretation defines the probability of an event as the fraction of "successes" over a hypothetical infinite series of independent and identical trials. An asymptotic relationship is assumed where, as the number of trials increases, the fraction of successes will converge to the

"true" value (according to the law of large numbers), which is interpreted as the probability of the event. The true probability is in most cases, unknown and needs to be estimated. On the other hand, the "Bayesian" view defines probability as a measure of the assessor's degree of belief about the event. This numerical encoding of one's belief is always conditional on the assessor's knowledge base. Often, an example of drawing balls from an urn is used to provide an interpretation of the probabilities [35].

The J100-10 standard needs to be clear on which form of probability is used in each of the risk analysis steps because the two approaches can lead to different interpretations of the analysis, and ultimately lead to different actions in practice [35]. When a frequentist view is used, it is important that the historical records are representative of future scenarios. The uncertainties of the frequentist estimates also need to be addressed. Similarly, when a Bayesian probability is adopted, evaluating the analyst's strength of knowledge on the matter is critical to understanding the usefulness of the assessment. Furthermore, communicating this knowledge level is essential for the accurate interpretation of a Bayesian probability. This results in the need to see beyond just the numerical value. An assessment process is required to evaluate the strength of knowledge as well, where a high strength of subject knowledge gives the analysis more authority and vice versa [32, 29].

The J100-10 standard gives some flexibility for the analysts to decide which type of probability they wish to use (see page 29 of the J100-10 standard for eliciting probabilities for proximity and dependency hazards). Making the different types of probability clear and how they are to be interpreted can help the analyst choose the more suitable method depending on data availability and their strength of knowledge on the system.

While the J100-10 standard deals with threats from many different sources, a particular emphasis is misplaced on terrorism risk, as evidenced by 31 of the 41 reference hazards being malevolent threats. J100-10 acknowledges that a true terrorism threat likelihood estimation is beyond the scope of most water sector risk analysis [23], but suggests that estimating a proxy for this value can provide useful information for decision making. Equation (2.1) indicates that determining the annual likelihood of attack and the conditional likelihood of certain outcomes given an attack are key components of quantifying terrorism risk.

However, there is debate in the risk analysis literature regarding whether assigning static probabilities is even feasible. One side (see [62, 34, 36, 52]) argues that the intelligent nature of the adversary makes assigning meaningful and useful probabilities problematic if not impossible. Bayesian probabilities of attack can be elicited through experts, but are misleading because the defender and attacker act on different knowledge bases. Others argue that employing a game theoretic approach [177, 176, 165], which requires some simplifying assumptions on the adversary, provides a foundation from which probabilities can be assigned. Unfortunately these basic assumptions are rarely met in practice and renders the method deeply flawed. For example, there

is not common knowledge between all actors, nor do the attackers always behave rationally.

J100-10 takes a more simplistic approach for estimating static probabilities, adopting a method developed by Risk Management Solutions, LLC. The process is outlined in a RAND Corporation report [71] and detailed in Appendix F of J100-10 [23]. The method characterizes attack probability as the product of six values: 1) the likelihood an attack will occur, 2) the likelihood the attack will occur in a given metro area, 3) the likelihood water infrastructure will be targeted for attack, 4) the likelihood a subclass of facilities will be selected out of all water infrastructure (e.g. reservoirs, treatment plants, etc.), 5) the likelihood of a certain facility being targeted, and finally 6) the likelihood of the specific threat-asset pair being chosen.

Determining the likelihoods at each step uses a mixture of both frequentist and Bayesian perspectives. The approach J100-10 adopts is a Bayesian driven analysis when eliciting probabilities of attack for a metro region (step 2) and for a specific threat-asset pair (step 6). It is important that an appropriate elicitation from subject experts include consideration of adversary intent, capabilities, and options. In contrast, a frequentist approach is used when estimating the likelihood of which facility type (e.g. reservoir or pump station) and which specific site will be selected for attack. Because of the deep uncertainty surrounding intelligent adversaries, we argue that the J100-10 approach in trying to capture likelihoods of terrorism attack in a single value is inadequate and misleading as the process assumptions, the adequacy of historical data, and the strength of the assessor's knowledge all need to be communicated.

2.4.3 Evaluation of Resilience

While we focus our analysis on the risk analysis portion of J100-10, resilience is also an integral part of the decision making process in J100-10. Here we highlight some limitations regarding how resilience is evaluated.

There are various definitions of resilience across different disciplines. SRA defines resilience as the "ability of a system to sustain or restore its basic functionality following a risk source or an event" [186]. This is in line with the popular engineering (in particular infrastructure) view that conceptualizes resilience as the ability to "bounce back" following shocks [63]. Other characterizations of resilience, particularly in the social sciences, focus more on the capacity for adaptive learning and change following events [63].

A literature review by Hosseini et al. [95] highlights two key attributes for characterizing engineering resilience: 1) the system's preparedness to absorb disruptions to performance, and 2) the ability for performance recovery. To this end, the definition provided by J100-10 (see section 2.1.2) is in line with the engineering state of the art. However, the approaches J100-10 provides for characterizing resilience are too narrow. The Operational Resilience Metric (ORM) metric

in Equation (2.2) quantifies the expected amount of service denial because of a lost asset, and the Owner's Economic Resilience Metric (OERM) in Equation (2.3) measures the dollar value of this loss to the utility. These metrics are not adequate reflections of system resilience but rather measures of consequence, and using them as characterizations of resilience can seriously misguide the decision maker.

Since J100-10 is specific to water infrastructure, the key function for utilities to sustain or recover is the ability to meet demand for clean water and to prevent wastewater overflow. The temporal and dynamic aspects of service recovery is crucial for determining resilience [6, 89] but is completely omitted in J100-10. J100-10 instructs that individual component resilience be quantified using Equations (2.2) and (2.3); however, this notion has been thoroughly discredited in the literature. Park et al. (2013) [164] argue that the nonlinear and self-organizing features in complex systems makes resilience impossible to measure when solely focusing on individual assets. Rather an emphasis should be placed on the performance of the entire system as a whole.

Some alternative assessments of resilience which J100-10 can apply are presented here. Two survey-based methods for measuring system-wide resilience are provided in Shirali et al. (2013) [182] and Cutter et al. (2008) [64]. In both case studies, the authors worked with domain experts to characterize indicators of resilience (e.g. redundancy, robustness) and developed specific criteria to identify whether an organization met these indicators. Examples of quantitative methods for evaluating resilience involve stochastic simulation and optimization. In simulation driven methods [3, 188], infrastructure models are subjected to hypothetical hazards and key performance indicators (e.g. percentage of on-time deliveries for supply chains) are tracked. Optimization modelling [5, 78], in contrast, aims to estimate least cost recovery or best-case performance for a system after damage.

The above examples analyze resilience in relation to well-defined objectives and disruptions. Haimes (2009) [89] argues that resilience should be further expanded as the performance of a system can be different for different types of shocks (e.g. natural hazards vs intentional attacks). To address this issue, Aven (2017) [31] argues that risk and resilience assessments can be coupled together for a more complete analysis.

Finally, the notion of community resilience in J100-10 only references the economic impacts of hazards, ignoring the multi-faceted aspects of community resilience and the need for all attributes to be adequately captured in an analysis, as highlighted in Koloui et al. (2017) [120]. These multi-faceted aspects include physical, environmental, financial, and social impacts.

2.5 Practical Limitations

Here, we discuss some of the practical limitations of the J100-10 assessment framework. One such limitation is that the employed methods can lead to inaccurate representations of risk. Other limitations involve cases of ambiguity as a result of how key metrics are estimated and interpreted.

2.5.1 Use of Work Case Scenarios

As discussed in Section 2.2.2, relying exclusively on worst-case assumptions in performing risk analysis can result in misleading conclusions. Even if there is certainty on the most extreme consequence, the sole analysis on worst-case outcomes will always lead to mischaracterizations of risk because all other possible outcomes are excluded from the analysis. Consider for example, the threat-asset pair summarized in Table 2.1.

Table 2.1: Summary of Example Threat-Asset Pair 1 with Divergent Outcomes. Risk calculated using Equation (2.1). *Worst case only risk.

Scenario	Threat	Consequence	Vulnerability	Risk
1-1	0.1	10000	0.001	1*
1-2	0.1	500	0.049	2.45
1-3	0.1	100	0.950	9.5
			Expected Value	1 + 2.45 + 9.5 = 12.95

For the same threat event, which has probability 0.1 of occurrence, there are three possible outcome scenarios with varying likelihoods. This is shown by the different consequence values and their associated vulnerabilities. A worst-case-only analysis would conclude that the associated risk is 1 (based on scenario 1-1). However, if the other two outcome scenarios are taken into account, the expected value is 12.95. In comparison, consider the threat-asset pair shown in Table 2.1. For the same threat with likelihood 0.1, there are two possible consequence scenarios. A worst-case-only analysis would determine that the associated risk for this example is 0.4 (under scenario 2-1). However, the expected value of risk, which considers both outcomes weighted by their respective likelihoods, is 50.3.

Table 2.2: Summary of Example Threat-Asset Pair 2 with Divergent Outcomes. Risk calculated using Equation (2.1). *Worst case only risk.

Scenario	Threat	Consequence	Vulnerability	Risk
2-1	0.1	2000	0.002	0.4*
2-2	0.1	500	0.998	49.9
			Expected Value	0.4 + 49.9 = 50.3

These examples serve as simple illustrations as to why a full representation of all consequence scenarios is needed for an accurate representation of risk. Both example threat-asset pairs have high worst-case consequences with low associated vulnerabilities, which lead to very similar risk scoring (1 and 0.4 respectively). Taking a worst-case-only approach would lead risk analysts to conclude that both threat-asset pairs are subject to the same level of risk as measured by equation (2.1). Worst-case scenarios alone, however, do not accurately represent the risk of the threat-asset pairs. In both examples, the worst-case scenarios are also the least likely to occur. After considering the other possible scenarios, the resulting risk calculations again using equation (2.1) (12.95 and 50.3 respectively) show that the second example is clearly the riskier threat-asset pair, with close to four times the risk value. The assumption here is that the expected value is an adequate risk measure, which is a very questionable assumption. This clear distinction in the risk description is overlooked when a worst-case-only basis is used.

A worst-case-only approach is quite popular in other domains beyond critical infrastructure analysis (e.g. financial [215] and environmental risk assessments [97, 107]). The limitations of using conservative "worst case" methods have been thoroughly discussed and criticized in the literature [166] and the reader is referred Aven (2016) [30] for an expanded discussion. Considering the full range of possible outcomes and their consequences in the analysis will lead to more informative descriptions of risk. In addition to the probabilistic characterizations, judgments of the strength of knowledge supporting these should be included as highlighted in Section 2.4.2.

An alternate characterization of risk is to present information on the underlying consequence distributions, for example, showing the 25^{th}, 50^{th}, and 75^{th} percentiles as well as the expected and worst-case scenarios. A common and more complete probabilistic representation in the risk analysis literature is the use of F-N types of curves, discussed in Aven (2013) [26], which plot all possible consequence values against their respective inverse cumulative probabilities, i.e. the probabilities for events leading to at least N units of loss (e.g., fatalities).

2.5.2 Defining and Estimating Consequences

It is important to display a full range of consequence scenarios for risk estimations. The J100-10 framework defines four baseline metrics for measuring consequence. These are 1) number of fatalities, 2) number of serious injuries, 3) financial loss to utility owners, and 4) economic losses to the community. The standard suggests that other facets of consequence, such as degradation in public confidence and environmental impacts, can also be included if the analyst deems necessary. Detailed calculations using simulation and economic models or direct estimation by qualified experts are acceptable methods of determining consequences according to J100-10.

The risk valuation in Equation (2.1) requires a single value for the consequence metric. However it is unclear how, or even if, an analyst should aggregate across metrics. For example, no guidance is offered for combining the metric estimates of 10 deaths, 5 injuries, $5 million in financial losses to the utility, and $15 million in economic losses to the serviced community. Aggregating across different metrics is poor practice because no utility will view each category the same. A decision maker is likely to have different acceptable outcomes across the varying metrics. For example, health and safety violations are not acceptable and must be avoided at all cost, but once these are met the financial risks are material in the decision making framework. This process becomes more difficult when qualitative assessments of consequences are also considered.

There are a number of ways to encode consequences into a single metric. One method is to monetize fatalities and injuries to provide a common unit of measure to sum consequences from each category. A similar approach is to normalize each metric into an ordinal scale (e.g. 1-10) and sum the normalizations. J100-10 provides a 0-10 consequence scale for each category [23] which the analyst can opt to use. This approach makes an implicit assumption about the inherent value of different consequence outcomes, and disagreements about these valuations may arise when multiple decision makers are involved. For example, according to the J100-10, one fatality is equal to $1 million in economic losses to either the utility or the community. These assumptions need to be made explicit to the decision maker and J100-10 does not provide any direction on doing so.

Additional outcome aggregating methods are also presented in the risk analysis literature. The field of decision analysis supports the use of multi-attribute utility theory (MAUT) to encode a variety of decision maker preferences into a numerical value, and has been demonstrated in many engineering risk assessments [151, 150, 51]. Ayyub (2014) [40] introduces other methods for assessing consequences and severities, including cause-consequence (CS) diagrams and total economic valuation (TEV). CS diagrams use a tree representation of multiple consequence categories (e.g. fatalities, economic costs) and assess their respective severities using logic diagrams. These severities are combined additively in an ordinal scale. TEV uses willingness to pay or accept methods to estimate the market value, measured in dollars, of lost goods and services.

While there is a host of processes for combining consequence metrics into a single value, it

is unclear what this single value represents. In making this calculation, the system operator must make assumptions regarding the value of consequences to other stakeholders, and in doing so, the utility imposes its own value structure on these stakeholders. According to Arrow's impossibility theorem [18], it is generally impossible for any analyst to accurately encompass each stakeholder's diverse preferences under a set of numerical weights. Survey methods are available as a foundation to begin the analysis of contrasting value judgements, but they require time and resources that the utility may not be willing to commit.

Therefore, in some situations utilities may find it beneficial to keep the consequence categories disaggregated. While this can lead to a less quantifiable measure of risk (i.e. Equation (2.1) can no longer be applied), more information can be communicated in the assessment results. Lundberg and Willis (2019) [142] present one approach for carrying out risk assessments while dealing with non-aggregate outcomes. The authors use a survey-based method to identify a ranking of consequences attributes. This information allows the analyst to prioritize one category over another. Kabir et al. (2018) [102] presents a quantitative Bayesian network model for modeling consequences due to infrastructure failures. The model disaggregates outcomes based on health and safety, environmental, societal, and economic impacts. Expert judgement is used to define the dependencies between various outcome measures.

2.5.3 Analysis Resolution of Threat-Asset Pairs

An accurate estimation of the consequences of a hazard on complex systems requires the analysis of multiple components together and the consideration of their interdependencies. Consequently, analyzing risk and resilience only at the threat-asset pair resolution overlooks the dependency between components [6].

This integrated relationship between assets can be illustrated through a simple example. A reliability block diagram (RBD) is a visual method that describes how individual components contribute to the overall functioning of a complex system [26]. Here, the functioning or success of the system is defined as the extent to which it can carry out its mission. In the case of water systems, this involves the adequate delivery of clean drinking water to end users. Each block in a diagram represents a system function, which can correspond to individual components of the system, e.g., treatment plant or storage tank, that can fail with a given probability upon an incident hazard. Blocks can be connected in parallel or series; parallel paths introduce redundancy into the system, where all blocks within a parallel block must fail before the network fails. On the other hand, any failure to a single block in a group of blocks connected in series will cause system failure.

Figure 2.2 illustrates a simple system with three components, represented by blocks A, B, and

C. Component A is connected in series to a parallel set of components B and C. This means failures to A alone, or B and C together, or to all three components can lead to system failure. Risk analysis of this system at the threat-asset level involves evaluating the consequences of failure when only A, B, or C fails individually. The redundancy relationship between B and C is not captured in the analysis at this resolution. A consequence estimate on the failure of asset B assuming asset C is functional may only include costs of damage repair; however, if asset C also fails, the consequence involves repairs to both components as well as economic losses due to service interruption.

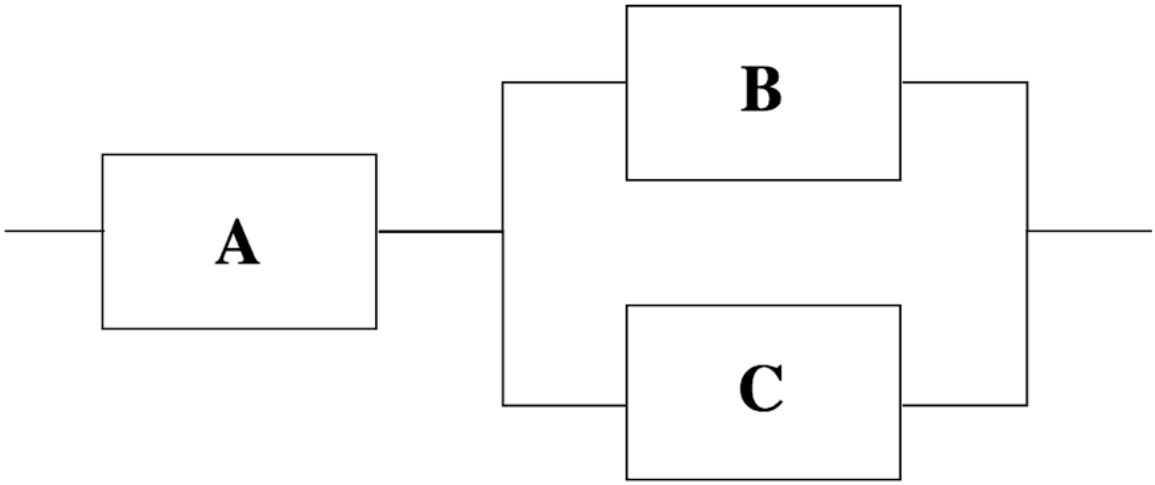


Figure 2.2: Reiliability Block Diagram of Example System.

The simplifying example above serves to illustrate that an accurate assessment of threat consequences requires information from multiple components of the system, and examining risk at the asset levels overlooks this relationship by requiring the analyst to make implicit assumptions about the condition of other components. The assessment can be improved where joint impacts, particularly cases where consequences of failures to a group of assets will exceed the sum of consequences from individual failures itself, are captured.

Aside from reliability block diagrams, graph theory (or network theory) is another method researchers have used to study the system-wide impacts related to individual component failure (see [209, 210, 9, 127]). In these network models, infrastructure components are represented through a series of arcs and nodes [73]. Each node represents a demand point, storage site, treatment site, or generation facility. Arcs represent distribution assets (e.g. wire cables for power systems, pipelines for water and gas networks). These studies have aimed to examine which network metrics (betweenness, centrality, etc.) are most useful in providing an accurate characterization of network resilience. Papers by Alderson et al. [6, 4] emphasize the use of physical infrastructure models rather than simple topological representations to provide the most accurate reflections of

network performance. LaRocca et al. (2015) [128] compared a range of topological metrics and physical models to measure power system performance, and found that combining graph theory with physical flow models provided the most accurate insights.

2.5.4 Threats Defined

Another issue of implementation is the limited scope of the 41 reference threats listed in Figure 2.3.

Hazard Type	Hazard Description			
Natural	N(H) Hurricanes	N(E) Earthquakes	N(T) Tornadoes	N(F) Floods
	N(W) Wildfire		N(I) Ice storms	
Dependency & Proximity	D(U) Loss of Utilities	D(S) Loss of Suppliers	D(E) Loss of Employees	D(C)Loss of Customers
	D(T) Loss of Transportation		D(P) Proximity to other targets	
Product Contamination	C(C) Chemical	C(R) Radionuclide	C(B) Biotoxin	C(P) Pathogen
	C(W) Weaponization of water disposal system			
Sabotage	S(PI) Physical–Insider	S(PU) Physical–Outsider	S(CI) Cyber–Insider	S(CU) Cyber–Outsider
Theft or Diversion	T(PI) Physical–Insider	T(PU) Physical–Outsider	T(CI) Cyber–Insider	T(CU) Cyber–Outsider
Attack: Marine	(M1) Small Boat	(M2) Fast Boat	(M3) Barge	(M4) Ocean Ship
Attack: Aircraft	(A1) Helicopter	(A2) Small Plane	(A3) Medium, Regional Jet	(A4) Long-Flight Jet
Attack: Automotive	(V1) Car	(V2) Van	(V3) Midsize Truck	(V4) Large Truck (18 Wheeler)
Attack: Assault Team	(AT1) 1 Assailant	(AT2) 2-4 Assailants	(AT3) 5-8 Assailants	(AT4) 9-16 Assailants

Figure 2.3: RAMCAP^{TM} Reference Hazards used in the J100-10. Figure taken from the J100-10 Risk Management Standard [23].

The RAMCAP^{TM} framework, which the J100-10 standard is based on, was originally developed to deal with terrorism threats, and 31 out of the 41 reference threats deal with malevolent threats. As a result, the analysis scope can be biased towards this single threat category. This can lead to a suboptimal allocation of resources to countermeasures that are dedicated to increasing the physical security of the system at the expense of hardening the system against (arguably) more frequent natural hazards. For example, a countermeasure, such as adding more security personnel, can decrease the risk for many of the 31 reference terrorist threats. Because of the large overlap in

the types of threats and how to defend against them, implementing mitigation options for one of these threats also serves to mitigate several other threats. As a result, the estimated net benefit of counter-terrorism defenses will be over inflated.

On the other hand, countermeasures for natural hazards tend to be more specific to the threat, e.g., installing flood-walls around coastal treatment plants to reduce flood damage. The limited overlap in affected threats from these countermeasures can lead to lower net benefits after summing over all threat-asset pairs. This shows that the J100-10 reference threat set typically biases the user to allocate resources to defend against terrorist threats over other hazard categories. For some general guidance on how to use cost-benefit type analysis, see Aven (2017) [24] and Ale et al. (2018) [7].

As noted in Section 2.2.2, two studies presented by White et al. [200, 199] argue that the operating 41 reference threats do not adequately address the emerging threats of climate change, aging infrastructure, and cybersecurity. While J100-10 allows analysts to include additional threats, it lacks guidance in how to define events that encompass these emerging threats and how to calculate the respective threat likelihoods. Furthermore, the subjectivity involved in adding more events can lead to inconsistencies when different analysts are performing the risk assessments.

2.5.5 Risk versus Resilience Tradeoff

In Steps 6 and 7 of the J100-10 methodology, risk and resilience are calculated, countermeasures are defined, and resources are allocated based on cost-benefit analysis. However there is ambiguity in choosing how to allocate these resources based on the different metrics. Step 7 (risk and resilience management) specifies that utilities need to define what acceptable levels of risk and resilience are, and implement countermeasures to meet these pre-defined thresholds.

As defined by J100-10, resilience and risk are two different outcomes. When dealing with various outcomes, an analyst must work with the stakeholders to elicit the value of resilience enhancement versus risk reduction. Decision makers need to understand the tradeoffs between the risk and resilience objectives in order for the assessment to be actionable. Unfortunately, the importance of eliciting these value judgements is omitted from J100-10.

There is, however, a strong argument in the risk research community that the separation between risk and resilience is artificial and that the risk concept should cover resilience [33]. This is because any actions performed to affect one will also affect the other: reductions in risk will also increase resilience, and vice versa. Aven (2017) [31] argues that assessments are more effective when the two outcomes are considered together, rather than treated as separated objectives.

As it currently stands, J100-10 is too vague in its definition of the relationship between risk and resilience. Improvements to the standard can either solely focus on risk, and target reductions

in risk, or integrate risk and resilience together for a more holistic assessment.

2.6 Discussion

In this study, we performed a comprehensive review of the risk and resilience assessment framework J100-10, a certified standard adopted by the water and wastewater industry. The framework adopts the seven-step methodology outlined in RAMCAP^{TM}, which applies to multiple sectors of critical infrastructure and key resources. Our analysis examined both conceptual limitations within the standard and practical issues with carrying out the risk and resilience assessment processes.

The main conceptual shortcomings are 1) the exclusion of notions of uncertainty when defining risk, 2) a clear definition for probability and how to interpret the values is not presented, and 3) resilience measures are too narrow. In particular, the differences between frequentist and Bayesian probability needs to be highlighted, and the conceptualization used needs to be communicated in the final analysis results. Our key findings on the practical limitations relate to the mischaracterization of risk, the biased emphasis placed on malevolent threats, and the general ambiguity in defining and comparing key metrics.

When calculating risk, using only a worst-case assumption of the associated consequences without considering the full range of possible outcome scenarios will result in a poor risk characterization. Furthermore, risk and resilience analysis at the resolution of individual threat-asset pairs ignores key dependencies between assets in connected systems. This resolution can lead to risk judgments that are too low in cases where combined consequences of hazards on multiple assets at a time will be far greater than the sum of the individual parts.

On the same note of accurately representing consequences, the standard uses four key metrics: fatalities, injuries, and economic losses to both the utility and community. Additional qualitative evaluations of consequence can also be included. The J100-10 standard does not provide adequate guidance on how to bring these four metrics, measured in different units, and other qualitative aspects of consequence, together into a single consequence value. This ambiguity can lead to inconsistencies in the risk analysis process.

The J100-10 defines 41 reference threats as part of the assessment, 31 of which are related to malevolent threats. The disproportionate representation of risk related to one category of threat can lead to biased conclusions about inflated benefits gained from counter-terrorism defenses. It is important for resulting updates of the J100-10 and RAMCAP^{TM} standard to account for any overlap when weighing the tradeoff between countermeasures designed to address malevolent threats versus natural hazards versus proximity and dependency hazards.

Lastly, the J100-10 standard needs to provide guidance on how to best assess stakeholder values relating to the two goals of risk reduction and resilience enhancement. This is critical for using

the J100-10 in an effective decision making context. The vagueness of the current standard can also introduce arbitrariness and inconsistencies, with potential for poor investments of available resources.

The shortcomings summarized above can assist with prioritization in redrafts of the standard by highlighting areas that need to be addressed. By closing the gap between the standard's methods and those that are the state of the art in the risk analysis literature, more informed risk-driven decisions can be made to better protect the nation's critical lifeline infrastructure.

Acknowledgements

We thank the University of Michigan for funding this research. The opinions and views expressed are those of the researchers and do not necessarily reflect those of the sponsors.

CHAPTER 3

Optimal Pipe Inspection Paths Considering Inspection Tool Limitations

The inspection of deteriorating water distribution pipes is an important process for utilities. It helps them gain a better understanding of the condition of their buried conveyance systems and aids better decision making for risk-based asset management. In-pipe continuous inspection tools provide high resolution and accurate data, but they have seen relatively limited use due to cost and operational constraints. To facilitate-cost efficient deployment of these technologies and maximal information gain, a process that finds high risk pipes to inspect while accounting for the limitations of the tools at hand is needed. This chapter shows how to incorporate these considerations within an optimization formulation, and examines the use of Evolutionary Programming, Simulated Annealing, and Greedy Search heuristics to identify inspection paths. Case studies performed on both synthetic and real world networks demonstrate that Evolutionary Programs are the most effective. While only three factors are used to characterize tool limitations, the method presented in this chapter can be extended to include technology-specific complexities in real world applications.

Keywords: Drinking Water Distribution System, Condition Assessment, Asset Management

Note: The research presented in this chapter has been published in the Journal of Reliability Engineering and System Safety. Citation: Thomas Y.J. Chen, Seth D. Guikema, Craig M. Daly. Optimal Pipe Inspection Paths Considering Inspection Tool Limitations. *Reliability Engineering and System Safety*, 181:156-166, 2019.

Preamble

The research in this chapter presents work that was done between 2016 - 2018. The contribution of this chapter is significant because it establishes the problem of inspection routing with consideration to tool limitations. The goal of the chapter is to highlight why platform considerations need to be taken into account when planning for inspections, and demonstrates how to solve it using heuristic solutions. From our review, prior to the publication of this chapter, no work in the academic literature addresses this issue. Chapter 4 was done sequentially (between 2018 - 2019) and explores the same problem with more mathematical sophistication, but this chapter first establishes the initial formulation of the routing problem.

3.1 Introduction

Drinking water distribution systems are one of the most critical infrastructures [70], however many utilities are challenged with managing these aging networks with insufficient budgets and limited availability of information [19]. Experts estimate that on average, utilities in the US lose over 14% of their treated drinking water daily due to leaky distribution pipes, and about $500 billion is required to address the replacement needs of these decaying assets over the next 40 years [39]. A proactive management framework is needed to reduce the spending on emergency repairs [161]. Furthermore, pipe breaks are also associated with public health risks [46, 184, 110], and can impact public confidence in the utility.

While the number of breaks and their associated costs can vary, pipe failures and leaks are a substantial problem in many countries. Kettler and Coulter [110] surveys failures rates on pipes of different age and size in 4 cities across the US and Canada, they found that failures can range from 1.05 to 0.05 breaks per km per year based on how pipes are categorized. These failures are expensive, a study from the Water Research Foundation reports that direct costs from pipeline breakages can range from $5,000 to $250,000 [204].

To aid the development of an effective asset management plan, inspection operations are often employed to gather information on the current condition of the system. As outlined in Roman and Pellegrino [173], besides obtaining information on pipe health and condition, other benefits of using inspection robotics include: 1) removing humans from potentially hazardous work situations, 2) allowing for inspection of inaccessible areas and, depending on the tool, 3) providing on-line inspection without stoppage of pipe operation.

In order for utilities to better address their most pressing liabilities, the planning of these inspections is critical. The highest risk pipes need to be prioritized so more information can be obtained to guide decision making on how to mitigate the risks of failure (e.g. replacement, repair,

leave alone).

Tur and Garthwaite [194] summarizes some current inspection technologies and Daly et al. [65] provides an overview of data collection techniques associated with these tools. This includes but are not limited to: targeted testing and continuous testing. Targeted testing extracts data at discrete points within the network, however obtaining an accurate condition characterization for an entire pipeline is difficult due to the variability of structural damage along a length of pipe [65]. In contrast, continuous testing obtains information along the entire length of the pipeline. Due to the continuous and high resolution data available with this method, it is typically preferred for inspecting critical mains (typically with diameters larger than 12 in.). Despite the breadth of conditional assessment tools available for deployment, Tur and Garthwaite [194] reports that due to economic, regulatory, operational, and physical constraints, typical inspection capabilities for these assets are limited. It is industry standard that a utility invests in inspections of approximately 2% of system length annually.

While the focus of this research is the development of an optimization framework for continuous inspection routes, the same technique can also be applied to identify contiguous regions of pipe for discrete inspections.

Prioritizing inspections based solely on a risk-based ranking of assets can lead to suboptimal results because operational limitations of the tools must be accounted for when planning for deployment. These constraints can affect both the accuracy and quality of the data obtained as well as the cost associated with inspection [47]. Often these constraints are related to the physical characteristics of the pipe segment. For example, sensors must be calibrated for specific materials of pipe (e.g. cast iron, concrete), this is a particular concern for free floating devices which collect inspection data while moving along the flow of water. These devices cannot be reconfigured and recalibrated on the fly, so multiple runs along the identified path are needed to obtain complete data in paths with multiple pipe materials. This can greatly increase the cost of inspection as the number of material changes go up [159].

In 2010 the National Academy of Corrosion Engineers (NACE) published a standard practice guideline for in-line pipe inspections which highlights many operational issues when planning for inspection deployments [159]. Some of them include 1) limiting sharp turns (90-degree bends) and/or valves which can increase the likelihood of tool damage, and 2) ensuring the flow rate and water pressure inside the pipe meet the specifications of the robotic sensor.

Another important consideration is pipe diameter. Electromagnetic devices which measure wall thickness require the sensor be placed a certain distance from the wall. A path that includes high variability in pipe diameter will require multiple recalibrations of the sensor offset, which will result in signal disruptions along the inspection and disjointed data collection.

The contribution of this study is to demonstrate how to incorporate the limitations of assessment

technologies within an optimization framework when planning for inspections. This is overlooked in popular practice where only the criticality of the pipeline assets is considered. To the authors knowledge, no past research has tackled inspection path planning in this fashion. This chapter aims to fill this gap by presenting a general optimization formulation, and comparing the performance of three solution algorithms applied to synthetic and real-world distribution networks. The integer programming formulation presented is simplified, the derivation of the full formulation (and the corresponding algorithms to solve it) is beyond the scope of this study and is left for future work.

The solution algorithms compared in this chapter are: Greedy Search Heuristic, Simulated Annealing, and Evolutionary Programming. We chose these methods because they provide a straightforward implementation in a network-based problem, as well as presenting a range of derivative-free approaches (the underlying optimization problem is integer) with varying degrees of computational complexity. Other optimization methods that have been studied to model the optimal design and control of water systems, such as ant colony optimization [213, 138] and particle swarm [156, 75], could potentially also be applied in a similar manner but are not included in this study.

The rest of this chapter will be organized as follows: a review on inspection technologies and planning is in Section 3.2, the simplified optimization formulation is presented in Section 3.3, its implementation using various networks and the aforementioned solution methods is covered in Section 3.4, performance results are reported in Section 3.5, followed by a discussion of their implications in Section 3.6.

3.2 Literature review and background

There are two fundamental types of continuous inspection platforms, tethered and untethered [159]. Tethered platforms include robotic crawlers which can be operated from above ground and free floating tools that move independently with the flow of liquid. Tethers provide a physical link between the operator and the tool, and serves to provide power supply and communications to the robotic units. Data is collected continuously during an inspection and typically stored on computers and not within the tools themselves. In comparison, untethered tools are autonomous units that contain their own power supply and data storage. These tools are inserted into a pipe and move with the flow of liquid to a point of extraction where data is then downloaded for analysis. While movement of these tools cannot be controlled, they are often tracked from above ground using sensors attached to the pipeline at defined intervals along the inspection path, typically at valves or other locations readily accessible from the surface. If an untethered tool becomes damaged or stuck during an inspection the recovery efforts are more complex and potentially costlier.

Both tethered and untethered tools can be outfitted with a variety of sensors. The two main types of sensors utilized for pressure pipe inspection are electromagnetic and acoustic, though

sonar and laser sensors are common too. Typically, acoustic sensors are used for leak detection while electromagnetic sensors are used to identify defects that risk perforating the pipe wall. Sonar is used to identify large defects in the pipeline as well as areas where debris and sediment build-up are present. Lasers have been used to measure ovality in pipelines which could be an indication of a structural condition.

A widely adopted method in practice for inspection planning is using a risk based ranking to prioritize assets, where the riskiest pipes are inspected first [145, 111]. This also serves to ensure that costly data collection techniques are targeted to the most at-risk assets. In this context, risk is typically defined as the expected consequences of an adverse event (specifically pipe failures). This is a popular definition used in many engineering applications [27] where risk scores are calculated as the product between the consequences, typically measured on a unit less scale (e.g. 0 - 1), and likelihood of a hazard. As part of the risk calculation, many studies have attempted the statistical modeling of pipe break probabilities, which range from simple parametric models [179, 15, 207], to more sophisticated non-parametric approaches [81, 56]. Recent advances have also focused on the consequence aspect of risk. A Water Research Foundation report [205] summarizes a comprehensive survey of pipe break records focused on understanding key drivers of failure consequences to both the operators and the customer.

Effective asset management not only includes the accurate modeling of risk, but also the development of an efficient operation plan. Researchers have worked to implement algorithms for obtaining the best scheduling of repair and maintenance tasks. Cost minimization was often the prevailing objective, from which a variety of methods have been developed (see [87, 12, 50, 49, 114]). For example, Dandy and Engelhardt (2001) [67] uses the Evolutionary Programming algorithm to determine a cost minimizing 5 year rehabilitation and replacement schedule for an Australian municipality. This was extended in Dandy and Engelhardt (2006) [66] which formulated the same problem in a multi-objective framework to capture a variety of shareholder interests.

Reducing the time taken for individual inspections is also a key consideration for cost efficiency, Lu et al. [141] and Kawaguchi et al. [108] solves for specific inspection routes by likening the task to a travelling salesman problem, which lends itself to a host of solution methods documented in Laporte [126]. However the practical application of these results are limited because, as pointed out in Tur and Garthwaite [194], it is not economically feasible for an entire system to be inspected by urban utilities over a limited time horizon.

Many previous studies have also presented optimization models for risk based inspection and maintenance (see [111, 190, 185]). In these models the risk considered pertains only to the distribution assets. This is an important omission (see Section 3.1) that affects both data quality and associated deployment costs. From the authors review, no mathematical programs have been developed for finding optimal inspection routes which facilitates maximal information gain by

considering both 1) the sparing use of the tools and 2) the operational limits of the technology.

3.3 Optimization formulation

Let N be the number of discrete pipe segments in a given distribution network. Let X define a vector of length N to index each individual segment and to represent the inspection decision, where each element is a binary variable corresponding whether a pipe is chosen for inspection or not. Algebraically this can be shown as X $\in \mathbb{R}^N$, X_i = 1 if segment i is inspected, X_i = 0 otherwise, for all i in 1 to N. Similarly vectors R and L, both of length N, can also be defined to represent the risk score and length of each individual pipe segment where R, L $\in \mathbb{R}^N$.

In order for an inspection path to be feasible, it must satisfy some physical constraints. This includes: 1) it has to be fully contiguous where all the selected segments connect, 2) the path must be a simple path where no pipe junction is traversed more than once, this implies no looping occurs or branching in the path can occur. Let $\mathbb{I}$ define an indicator function which is used to check for the feasibility of any candidate solution X, where $\mathbb{I}$(X) = 1 if conditions above are satisfied and 0 otherwise.

This research will consider only pipe feature changes of material and diameter as tool-limiting factors, however technology specific complexity (e.g. penalizing sharp turns for tethered tools) can be added to the formulation in more advanced settings. Feature-change counter functions C_m(X) and C_d(X) are defined to return the number of material and diameter changes respectively along any feasible inspection path X. Furthermore, each change of pipe material and diameter will have a constant penalty P_m and P_d associated with it.

These penalties are not parameters of an optimization algorithm, rather a reflection of the limitations of inspection tool as they traverse non-ideal conditions inside a pipe. In practice, the value of these coefficients should be defined by engineers with experiences using inspections platforms to accurately reflect the penalties associated with encountering pipe property changes along an inspection route.

Bringing the above information all together, the optimization model is as follows:

$$\max Z = R \cdot X^T - P_m C_m(X) - P_d C_d(X) \tag{3.1a}$$

Subject to:

$$L \cdot X^T \leq D \tag{3.1b}$$

$$\mathbb{I}(X) = 1 \tag{3.1c}$$

Where equation (3.1a) characterizes the objective function of searching for a path to include high-risk pipe segments, while also being penalized for high variability in physical features along the path. Equation (3.1b) specifies the distance constraint of the selected assets, which must be less than some quantity D, this reflects the limited inspection capabilities for a utility. Equation (3.1c) enforces that only physically feasible paths be considered as part of the candidate solutions. For this work we only consider the selection of one optimal path under the budget constraint. Due to the large set up costs for running an inspection (e.g. road closures, digging down to the pipe, crew dispatch), utilities in practice typically plan for a single inspection route at a time. The exploration of optimizing inspections where multiple paths can be considered is left for future research.

3.4 Case study networks, risk modeling, and solution algorithms

This section describes the methodology used for this study, including how synthetic networks were developed, how a real network was applied, and how a simple risk model was implemented to characterize asset condition. In most engineering applications, risk analysis considers both the probability and consequence of failure [26]. Since it is beyond the focus of the study to apply a sophisticated modeling of risk, an age-based pipe break likelihood model from Pelletier et al. [168] is adopted instead, though more advanced risk models could be used instead. Finally, the solution methods used to solve the formulation presented in Section 3.3 is outlined.

3.4.1 Networks

A popular network theoretic representation of a water distribution system is through a system of connected/disconnected arcs and nodes [73, 209]. Each node representing a water source, a connecting valve, a storage or treatment facility, or end user; while arcs represent pipeline segments. Directed edges are used to represent flow direction within pipes. We use this arc/node representation of a water network in our optimization formulation and solution methods. While it is possible to explore other abstractions of the physical network, it is beyond the scope of the research and can be examined in future work.

Since this research deals with incorporating tool limits for inspection route optimization problems, we include both synthetic and real networks in our benchmark trials. Two synthetic networks are used in this study, one with a highly regularized structure and another with a more realistic/sparse structure. As a test bed, a square Grid network (10 edges per side) was developed in the R statistical software [44], where every node is connected to all of its adjacent nodes (see Figure 3.1).

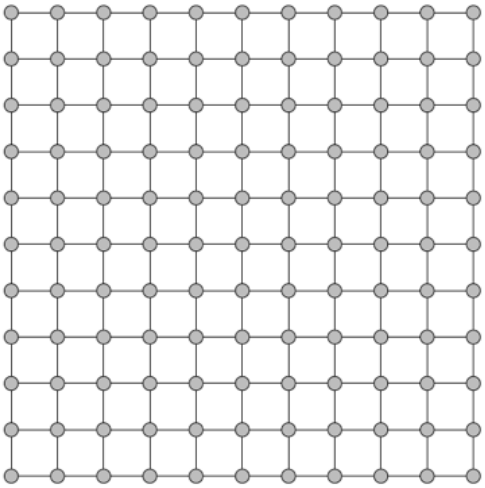


Figure 3.1: Layout of Grid Network

The other, more realistic, synthetic network used was an open sourced virtual pipe network named Micropolis, presented in Brumbelow et al. [53]. Developed in EPANET and ArcGIS (ESRI, 2011), the virtual city represents a water system of a small rural town of approximately 5000 residents (see Figure 3.2).

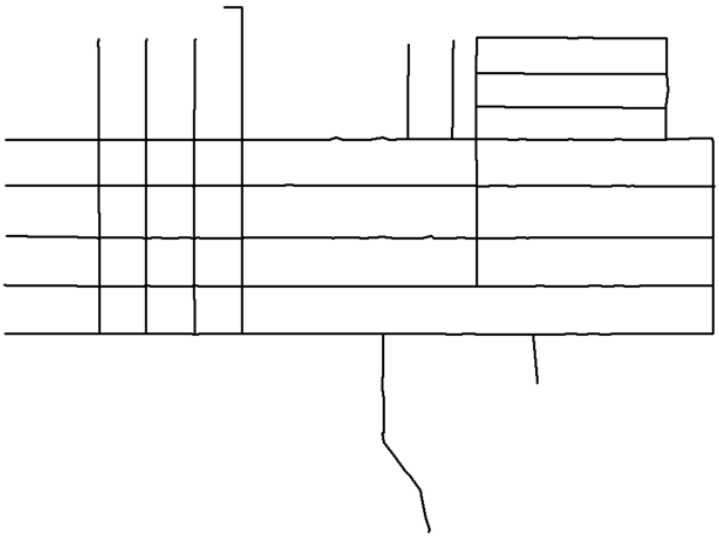


Figure 3.2: Layout of Micropolis Network

Due to the synthetic nature of the Grid network which lacks any pipe attribute information, physical features had to be assigned to each arc, as well as a randomized probability of failure to characterize risk. Each pipe segment (arc) was randomly assigned one of 3 classes of material and diameter. The actual type of material or diameter is irrelevant as the objective function in

equation (3.1a) will only consider the change in features along an identified path. 5 random grid networks are generated to test the optimization algorithms.

The other synthetic system we explore, Micropolis [53], contains a more realistic structure for a water network. It contains pipes segments from 5 different classes of diameter and 3 different classes of material. Micropolis has been used as a test bed in previous research [214, 136] studying water distribution systems, demonstrating that it is an effective tool to simulate realistic networks when data is unavailable. The Micropolis dataset was visualized using the mapping software ArcGIS (ESRI, 2011), shown in Figure 3.2.

Finally, to determine the efficacy of the solution algorithms in realistic settings, we move away from synthetic systems and use a subsection of the water distribution system (WDS) from Ann Arbor, Michigan. The network is also mapped in ArcGIS and displayed in Figure 3.3. The distribution systems contain pipe segments of 12 different diameter sizes and 10 different material designs.

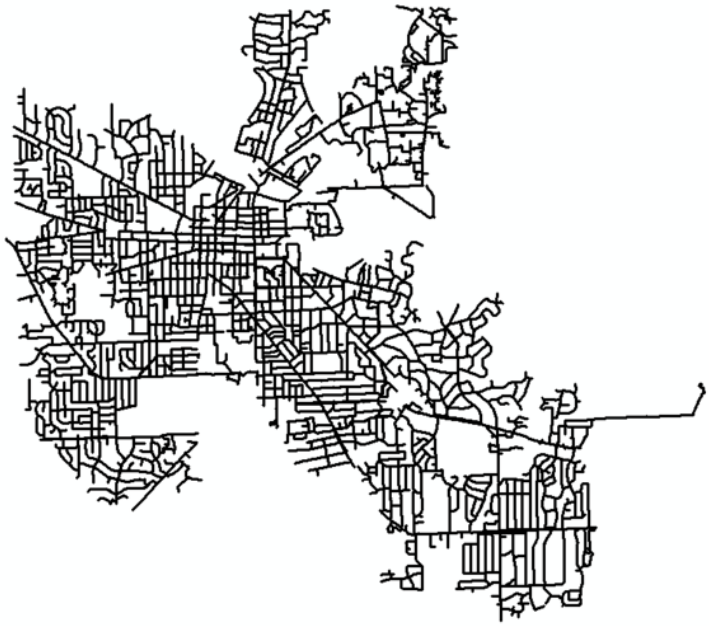


Figure 3.3: Layout of the City of Ann Arbor Water Distribution System

Table 3.1 summarizes some characteristics of each of the examined systems.

Table 3.1: Summary of case study networks

Characteristic	Square Grid	Micropolis	Ann Arbor WDS
Pipe Network Length (mi)	NA	32	242
Number of Segments	220	567	13058
Age Information	No	Yes	Yes
Year of Installation of First Pipe	NA	1910	1880
Diameter Information	Synthetic	Yes	Yes
Material Information	Synthetic	Yes	Yes

3.4.2 Probability of failure risk model

In order to characterize the physical condition of the pipe segments in the networks, an age-based pipe break likelihood model is implemented where data related to pipe age is available (the Micropolis and Ann Arbor systems). It is beyond the scope of this research to explore a sophisticated modeling of risk, our goal here is to have a method to assign rewards for inspecting each pipeline asset. A failure probability model is convenient for this framework since higher likelihoods represent riskier assets and the probability values can be used as the reward coefficients in equation (3.1a). To similarly bound risk values between 0 and 1 for the Grid network, a uniform distribution is used to assign pipe break likelihood risk scores to each arc.

The probability model we use is taken from a case study reported by Pelletier et al. [168]. A Weibull distribution is fit against historical failure data from a municipality in Gatineau, Canada. A Weibull distribution is characterized by two parameters κ and ρ, it is associated with the time to first failure from initial installation [174]. The authors found a better statistical fit of the failure data is achieved by discretizing the pipe system in two based on installation year, before and after 1960, and fitting two separate distributions. Historical failure data is required to fit this distribution, but since failure data is not available for both Micropolis and Ann Arbor, we will use the reported parameters from Pelletier et al. [168] to derive a risk model for our benchmark trials.

A hazard function which corresponds to the annual probability of failure can be derived from a Weibull distribution, and is presented in equation (3.2) below.

$$\lambda(t) = \kappa\rho(\kappa t)^{\rho-1} \tag{3.2}$$

It is assumed that each pipe has not experienced a break up to the year 2017 (t is the pipe age at 2017), thus λ(t) is the probability of failure at the year 2017. Table 3.2 presents the κ and ρ parameter set used to model the Micropolis and Ann Arbor systems.

Table 3.2: Hazard function parameters of Pipe Break Likelihood Risk Model, used for Ann Arbor and Micropolis System. Model parameters taken from Pelletier et al. [168]

Parameter	Pipes Installed Before 1960	Pipes Installed at 1960 or After
κ	0.022	0.029
ρ	2.725	2.172

Figure 3.4 shows the distributions of failure likelihoods among the pipe segments in the Micropolis and Ann Arbor Systems using the aforementioned risk model. By observation, the failure probabilities obtained from equation (3.2) all fall below 0.5 in both networks. There are three large peaks in the failure density distribution for the Micropolis system since there are only three unique installation dates, and age was the only determining factor in probability estimation.

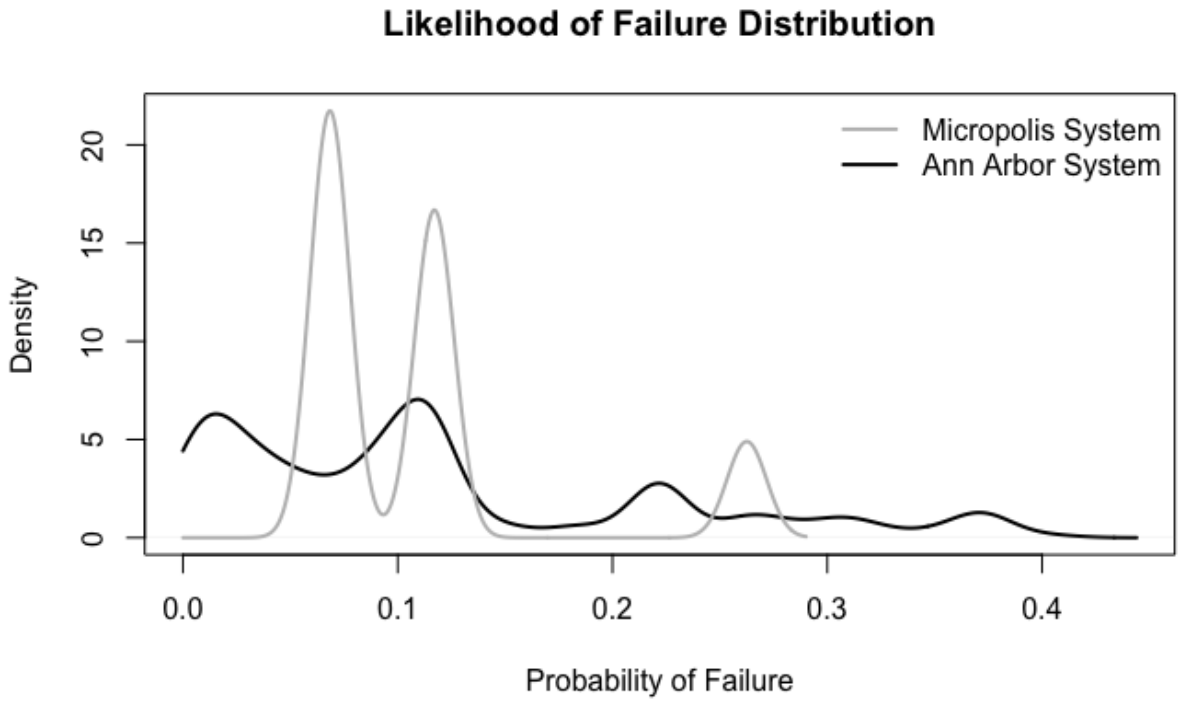


Figure 3.4: Likelihood of Failure Distribution for Micropolis and Ann Arbor Networks, evaluated using Risk Equation (3.2).

In contrast, there is a wider spread in distribution for the Ann Arbor system due to the larger pipe age heterogeneity in the dataset. In contrast, the probability of failure was uniformly assigned in each of the randomly generated grid networks and should have even greater variance.

3.4.3 Optimization algorithms

The optimization algorithms used (Evolutionary Program, Simulated Annealing, Greedy Search) for this research were implemented using the statistical language R. The following subsection will outline each of the respective methods, but some optimization model parameters and formulation coefficients must be first defined.

Since each individual failure likelihood/inspection reward value is bounded between 0 and 1, a 0.33 penalty coefficient (corresponding to P_m and P_d in equation (3.1a)) was chosen to sufficiently impact the value of the inspection path. As discussed in section 3.1, in practice the value of these cost coefficients should be defined by experts to best reflect the penalties associated with encountering non-ideal conditions along an inspection route. However, since this work is exploratory in nature, we select 0.33 since it sufficiently impacts the value of each pipeline asset.

The selected cost coefficients P_m and P_d assume that a change in pipe material and diameter has an equal penalty. However, in practical applications where certain features are more crucial to the effective operation of the inspection tool (e.g. limiting sharp turns is more critical than homogenous pipe features for tethered tools), differing penalizing factors can be included. Furthermore, a 2% of total system length constraint (value D in constraint (3.1b)) is defined to reflect the limited inspection capabilities of utilities. This is a typical value seen in industry practice, both for capital planning and budgeting purposes. This length limit can be adjusted to reflect different levels of budgeting and regulatory limits on a case-by-case basis.

3.4.3.1 Greedy search

This approach involves an iterative process of selecting the locally optimal solution and taking steps in that direction within the feasible region [74]. The implementation of the algorithm starts by randomly selecting an arc to initialize a candidate path, followed by an exhaustive search to enumerate all feasible paths that can be taken from current path that adds on N more segments. After all feasible paths are generated, the objective function (3.1a) is used to evaluate them and the highest scoring one is selected. From the selected path only the single arc that directly extends out of the current candidate path's end is added, now the path is one arc longer. This process is repeated until the distance limit is met. To avoid the algorithm reaching locally optimal paths, a 4-step (N = 4) Greedy Search was used in this study. Figure 3.5 shows a flow chart summarizing each step of the algorithm.

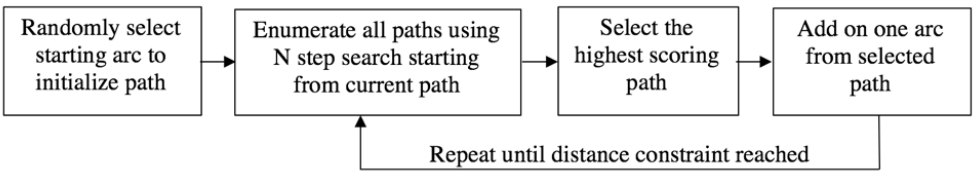


Figure 3.5: Schematic for Greedy Search

The greedy search method works well because there is a natural implementation in a link/node graph setting, in particular the exhaustive N path enumeration can be performed using standard network traversal algorithms [2]. The efficacy of the greedy search can vary depending on the randomly selected arc in the first step. To boost the search capacity, a M-random start version of the heuristic is also implemented. The algorithm is simply run M number of times, each with a random starting point, and the best performing solution is chosen at the end. M is set to 50 in the corresponding case studies.

3.4.3.2 Simulated annealing

Introduced in Kirkpatric et al. [115], the Simulated Annealing algorithm involves the iterative comparison of neighboring solutions until a set of terminating conditions are met. It can be summarized as an iterative 3-step process: 1) generating a neighboring path to the current solution, 2) comparing path scores between the 2 neighbors, and 3) selecting from the pair a candidate path for the next iteration.

The algorithm starts by taking an initial solution, which is a randomly selected feasible path, and compares it against a neighboring path using the objective function (3.1a). A neighboring solution is generated by replacing a selected arc from either end of the current path with an unselected one while maintaining feasibility. If the neighboring path scores higher than the original, it is selected for the next iteration. However, if the neighbor scores lower, it is selected with probability P instead, otherwise the original path is kept. Allowing for the initial rejection of a better performing neighbor potentially avoids having the algorithm get stuck local optimas.

Following the formulation of the Simulated Annealing algorithm from [115], the selection probability P is calculated following equation (3.3) below:

$$P = e^{\frac{Z_{neighbor} - Z_{current}}{T}} \tag{3.3}$$

Where $Z_{neighbor}$ and $Z_{current}$ are the respective objective function values of the two paths under consideration, and T is an iteration-step dependent multiplier. T is set to 1 initially and is reduced by a factor of 0.9 with each iteration. In this equation, when the difference in performance is small,

the likelihood of accepting the worse solution is higher, this likelihood drops as the difference increases or as more iterations have been completed.

A flowchart of the algorithm steps is shown in Figure 3.6. Through a number of test trials, we found that over each network the path typically converges within 2000 iterations. Hence, the simulated annealing algorithm is run for 2000 steps in each of the following trials, this is defined as the stopping criterion for the method.

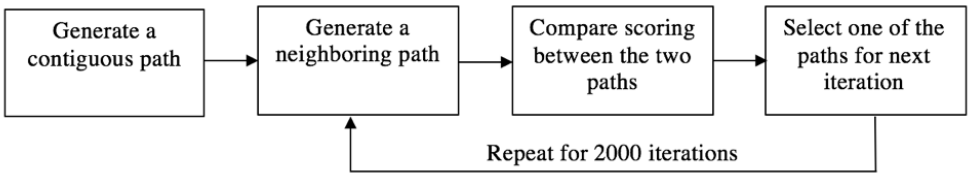


Figure 3.6: Schematic for Simulated Annealing

Like the greedy search, the simulated annealing algorithm also works well for path finding problems because of its natural implementation in a graph setting. The algorithm involves the iterative movement between neighboring solutions, and in a link/node representation of a path this is simply the swapping of one node and edge off the end of a path for another node/edge combination.

Similar to the Greedy Search, the performance of the Simulated Annealing can also vary depending on the location of the initializing path. Thus a M-random start version of the heuristic is also applied where we run M trials of the method and choose the best solution after all runs of the algorithm are completed. M is set to 50 in the corresponding case studies.

3.4.3.3 Evolutionary program

Evolutionary Programs [206] are a subclass of evolutionary methods that have been often used in water resource planning [67, 83]. Unlike the Greedy Search and Simulated Annealing, which only selects one solution to move between iterations, the Evolutionary Program will consider a set of candidate solutions to increase search capacity.

The Evolutionary Program is initialized by first generating a group of N feasible paths, followed by an iterative 4-step process: 1) computation of the performance measure of each solution using objective function (3.1a) and the subsequent ranking of solutions, 2) ranking-based selection (with replacement) of N candidate paths, 3) randomization of the selected paths and computation of their performance, 4) choose the N best performing paths from the combined selected and altered paths and define it as the new solution set. A new set of solutions are generated at each iteration of this four step process until a stopping criterion is met.

In each step, the algorithm attempts to identify the best performing paths and include them in the next iteration set. These solutions are chosen using linear-based rank selection, where the probability of selection for each candidate is associated with its relative performance amongst others. Better solutions have a higher probability of selection. This selection method is chosen over a roulette based approach, where solutions are selected with probability proportional to its objective function value, to avoid premature convergence at local optima's [152]. Once a solution is identified, the solution is randomly altered to expand the number of candidate paths considered and avoid getting trapped in local optimums.

In the classic Evolutionary Program approach outlined in Yu and Gen [206], the path randomization step is known as the mutation function. Similar to the other 2 methods presented above, the Evolutionary Program also has a natural implementation for path identification in a link/node graph setting. In this research, the mutation function is implemented by randomly choosing and removing up to half of the segments from either end of the selected path, and adding back a different series of segments until the length limit is met to generate a different path. The two paths are neighbors to each other since they share at least half of their comprised pipe segments.

Figure 3.7 shows a flow chart schematic of the algorithm. In the case studies a total of 100 sets is generated (we also found that the best path typically converges within 100 steps of the method), and the size of each solution set is 50, the same number as M in the M-start versions of the Greedy Search and Simulated Annealing. The best performing path from the final set is selected as the final solution. Since the Evolutionary Program itself is already a group based search, a 50-start version of the algorithm is not implemented.

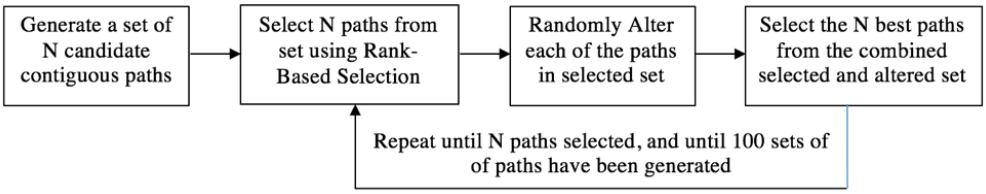


Figure 3.7: Schematic for Evolutionary Program

3.5 Algorithm testing and results

As discussed in the previous section, there are a total of 7 networks under consideration in this study: 5 randomized grid networks, the Micropolis system, and the Ann Arbor system. For each of the 7 systems, the optimization algorithms were applied in 50 trials and the objective function value of the resulting inspection path was recorded. The rest of this section will present a summary of the solution results and a discussion of their significance, example routes identified by the various

solution methods are shown in the appendix.

Since the length of individual pipe segments can vary greatly relative to the length limit of the inspection path (2% of the total system) and across different systems, the value of the objective function (3.1a) can be greatly impacted due to the number of segments comprising a path. It is possible to implement network preprocessing methods beforehand to standardize the segment lengths, but this is beyond the scope of the study. To maintain generalizability, our analysis will focus on the relative performance of each algorithm within a given network rather than across networks.

Table 3.3 shows the mean and standard deviation of the path scores for each of the generated Grid networks, which has randomized risk scores and physical properties. Across all five networks there is a distinct hierarchy. On average: the 50 start Simulated Annealing performs the best, followed by the 50 start Greedy Search, then the Evolutionary Program, then finally the single runs of the Simulated Annealing and Greedy Search. A similar ordering is also observed for the standard deviation associated with the path values, the 50 start Simulated Annealing has the lowest standard deviation which suggests the algorithm is much more consistent in identifying good solutions.

Table 3.3: Objective function mean and standard deviation of identified solutions from 5 randomly generated grid networks. Higher scores indicate better paths.

	Greedy Search	50 Start Greedy Search	Simulated Annealing	50 Start Simulated Annealing	Evolutionary Program
	Sample mean over 50 trials (sample standard deviation)				
Network 1	3.75 (0.61)	5.17 (0.28)	4.94 (0.40)	5.45 (0.09)	5.10 (0.22)
Network 2	3.28 (0.60)	4.61 (0.16)	3.98 (0.41)	4.52 (0.01)	4.38 (0.21)
Network 3	3.25 (0.73)	4.37 (0.13)	4.45 (0.37)	4.71 (0.01)	4.65 (0.17)
Network 4	3.32 (0.63)	4.71 (0.23)	4.20 (0.38)	4.73 (0.01)	4.46 (0.22)
Network 5	3.24 (0.62)	4.65 (0.22)	4.27 (0.36)	4.70 (0.10)	4.47 (0.19)

The distinction between the heightened effectiveness of ensemble-based searches over single searches is clear. Barring computational limitations, running the same heuristic multiple times and picking the best solution can only improve the performance over a single run unless the global optimal is found on the first trial. The average improvement in mean scores between Greedy Search to the 50 run version across the 5 networks is 39.6%, while the improvement for the Simulated Annealing is much smaller at 10.5%. These results suggest that when using less computationally expensive algorithms, the expected improvement in running the algorithm a large number of times is greater than when using more exhaustive approaches.

Even though the Evolutionary Program is the most computationally expensive algorithm used in the benchmark trials, it falls short of the 50 start Simulated Annealing. Possible reasons for

this could be related to the size of the solution set within the Evolutionary Program, increasing the number of paths compared could expand the search capability of the method. Furthermore, in highly connected networks such as the Grid, a Simulated Annealing approach may be better suited to explore the feasible region since it has the freedom to move around the entire network, and is less likely to be hindered by topological bottlenecks.

Table 3.4 shows the mean and standard deviation of the solution values for optimization trials in the Micropolis and Ann Arbor networks. For the micropolis system, the ordering of performance for the algorithms is as follows: the Evolutionary Program performs best (highest average score) and is most consistent (lowest standard deviation), followed by the 50 start Simulated Annealing, then the 50 start Greedy Search, and finally the single run versions of the Simulated Annealing and Greedy Search.

Table 3.4: Objective function average and standard deviation of identified solutions. Higher scores indicate better paths.

	Greedy Search	50 Start Greedy Search	Simulated Annealing	50 Start Simulated Annealing	Evolution Program
	Sample mean over 50 trials (sample standard deviation)				
Micropolis	0.85 (0.79)	3.12 (0.81)	1.78 (0.94)	3.58 (0.49)	4.36 (0.34)
Ann Arbor	7.85 (10.55)	33.49 (3.21)	6.74 (6.65)	26.45 (3.18)	44.08 (2.25)

This relationship is also seen in Figure 3.8, where the probability density distribution of the solution values are plotted, the distribution associated with the paths found using the Evolutionary Program has the sharpest peak. It is seen that the distributions of both the 50 start algorithms have overlap with that of the Evolutionary Program and is far better than their single run counterparts. However, this overlap is small, with most of the distribution of the 50 start methods falling below the Evolutionary Programs, indicating that while all ensemble based searches are capable of identifying good inspection paths, the Evolutionary Program is much more consistent and effective.

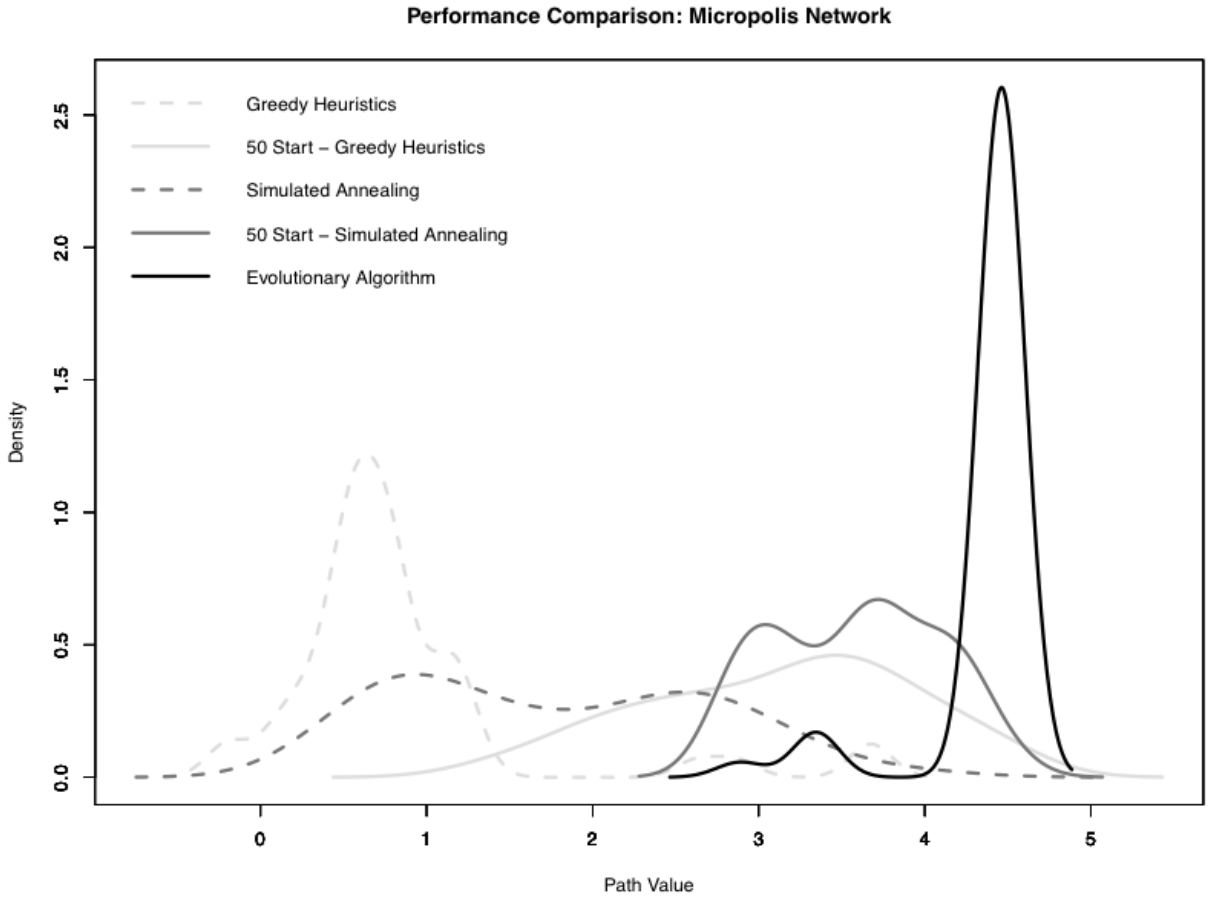


Figure 3.8: Density distribution of path value from 50 trials of each optimization algorithms on Micropolis network.

In contrast to the results seen from the randomized grids, the Evolutionary Program outperforms the 50 start Simulated Annealing, where average score of the former is approximately 22% greater than the latter. This gap in performance is much larger than that observed in the randomized grids, where the 50 start Simulated Annealing outperforms the Evolutionary Program by less than 5% on average. One possible reason for this discrepancy is due to the way the algorithms search the network. Simulated Annealing relies on moving between neighbors after each iteration, this means network topology and the ease of which regions of high value paths can be accessed from other parts of the system is vital to the algorithms effectiveness. On the other hand, the Evolutionary Program does not rely on moving between neighbors, instead it directly moves groups of solutions to regions where good solutions are found. While network topology is still an important property that determines how well the mutation function can explore the surrounding region of a given path, it is less likely to get stuck at poor solutions due to topological constraints. Thus the disconnected network structure of Micropolis means the Evolutionary Program is more adequate

to find good inspection paths

For the Ann Arbor system, the ordering of performance for the algorithms is similar to the Micropolis case: again referring to Table 3.4, the Evolutionary Program performs best (highest average score) and is most consistent (lowest standard deviation), followed by the 50 start Greedy Search, then the 50 start Simulated Annealing, and finally the single run versions of the Simulated Annealing and Greedy Search. Figure 3.9 shows the probability density plots of the respective scores over the 50 trials.

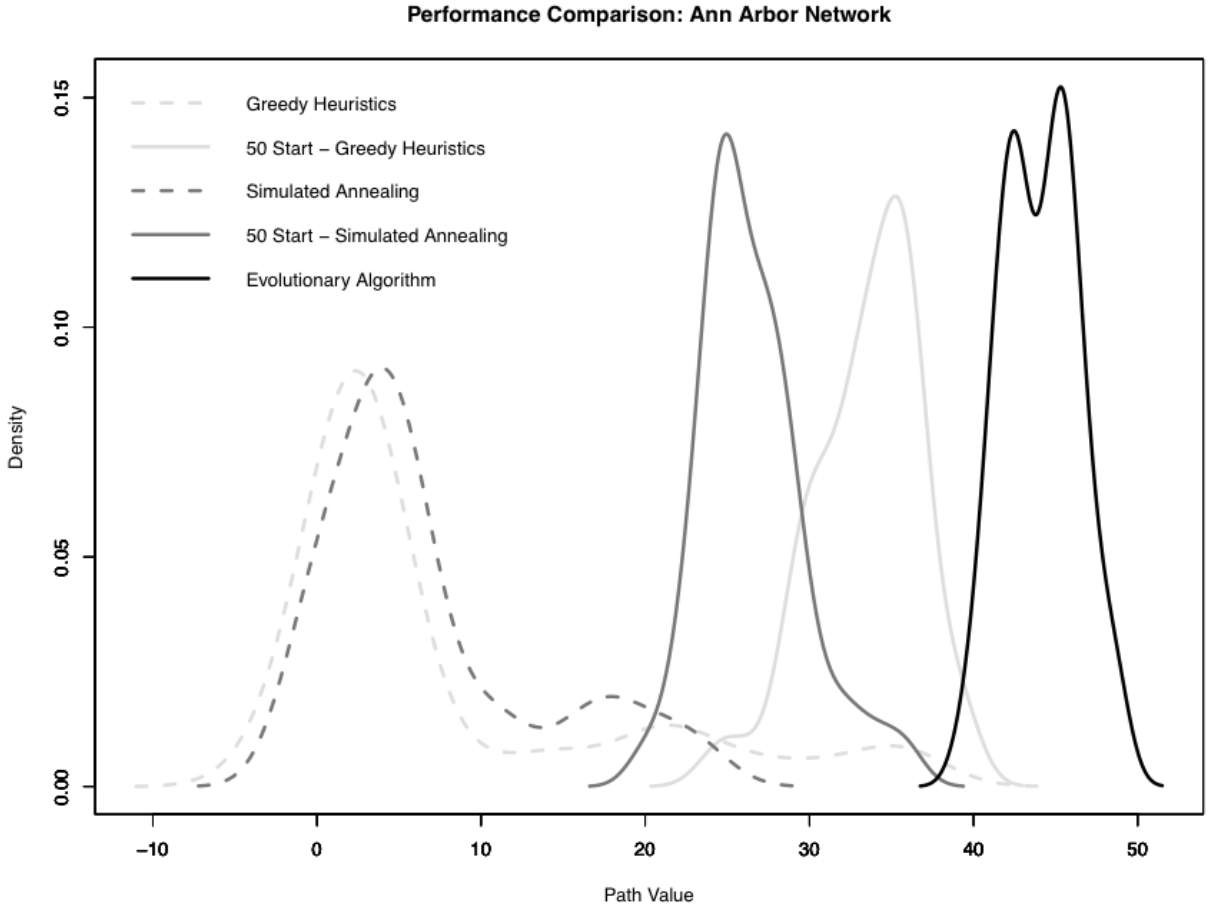


Figure 3.9: Density distribution of path value from 50 trials of each optimization algorithms on Ann Arbor network.

It is evident that gap between the Evolutionary Program performance and that of any other approach is far greater here than in previous examples. The average path score obtained using the Evolutionary Program is 33% higher than the 50 start Greedy Search, the second best approach. Furthermore, there is almost no overlap between the Evolutionary Program distribution and that of any other approach. This indicates that in real networks that are both larger (26 times larger than

Micropolis and 60 times larger than the grid in terms the number of pipe segments) and much more disconnected, using a heuristic that is less reliant on network topology is vital for identifying good inspection paths.

An interesting observation seen in the Ann Arbor case study is the reversal between the performances of the 50 start Greedy Search and 50 start Simulated Annealing. In all the previous cases the Simulated Annealing, the more computationally intensive algorithm, performed better than the Greedy Search. The opposite is seen in the Ann Arbor system. A hypothesis for this observation is due to the disconnected layout of the Ann Arbor system, which contains more "dead ends" compared to Micropolis. As a result, the Simulated Annealing may be more prone to getting trapped, and leaving these dead ends becomes difficult due to the disconnected layout.

3.5.1 Optimality gap comparison

Due to the homogenous length of the pipe segments in the grid, every solution path will always contain the same number segments (in our case, 10), this is in contrast to the Micropolis and Ann Arbor network solutions where the number of segments comprising a path is highly dependent on their individual lengths. Thus for the grid example, the value of a global optimum (10) can be obtained by setting the probability of failure of each segment to 1 and augmenting the path such that there is no pipe feature change along its length.

To further compare the effectiveness of the heuristic algorithms, an optimality gap estimator was implemented where another 50 random grids was generated. Each grid was induced with a global optimum by selecting a random path and setting all comprising arcs with risk value 1 and homogenous physical properties. This guarantees us that there exists a globally optimal path in the network. For each of the augmented grids, the heuristics were run and the ratio of the scores between the identified path and the optimal path was recorded in Table 3.5.

Table 3.5: The average ratio, over 50 trials, between the value of a path identified by a heuristic algorithm and induced global optimum.

	Greedy Search	50 Start Greedy Search	Simulated Annealing	50 Start Simulated Annealing	Evolution Program
Ratio	0.597	0.970	0.921	1.00	0.993

It was found that in the presence of a global optima, all approaches except the single run Greedy Search performed very well by averaging over 92% when comparing objective function values to the optimal. The 50 start simulated Annealing in particular never failed to find the global optimal, while the Evolutionary Program was almost near optimal (99.3% average value of the optimal). However, since all the ensemble based methods performed near-optimally (comparing 97% for the

50 start Greedy Search to 99% from the Evolutionary Program), it did not provide more meaningful insight into helping differentiate the different ensemble-based optimization approaches.

3.6 Discussion

It is shown from Section 3.5, that as the size of the system grows and the complexity of the network layout increases, more computationally intensive methods like the Evolutionary Program are more effective at finding good paths than less refined approaches. While the presented methods were chosen because they provided a straightforward implementation in a network setting, future research can aim to adapt other methods for comparison.

To demonstrate the importance of considering tool constraints when planning for inspections, the Evolutionary Program is applied again to identify an inspection path in both the Micropolis and Ann Arbor systems, except this time, the algorithm will the applied once with the original objective function in equation (3.1a), and another with the tool-related penalty terms $\mathrm{P}_m\mathrm{C}_m(\mathrm{X})$ and $\mathrm{P}_d\mathrm{C}_d(\mathrm{X})$ removed. The latter case will serve to show the difference when the only consideration is to find a path with the riskiest pipes and tool limitations are disregarded.

In the Micropolis example, when pipe feature change penalties are included, the obtained solution has objective value 2.89 with no pipe or diameter changes. In contrast, when the feature change penalties are removed, the resulting path has value 3.88 and 5 total changes in pipe property. While the second solution includes riskier pipes for inspection, the 3 changes in material and 2 changes in diameter along that path may lead to greater risk of deployment of the tools.

The difference in the resulting paths is further emphasized in the Ann Arbor example. When the feature change penalties are included, the obtained solution has value 43.07 with 8 pipe diameter changes and 8 material changes. On the other hand, without the feature change penalties, the resulting path has value 48.88 and more than double the number of feature changes (20 diameter changes and 15 material changes respectively). Again, while the second path includes more high risk assets, the lack of consideration for the tool limitations can lead to a suboptimal inspection due to greater risk of operational issues during inspections. For example, if a free floating device which cannot be recalibrated instantaneously was used, multiple runs of the tool along the identified path would be required, each time calibrating to a unique combination of pipe material and diameter. This increases operational costs and risk of damage to the technology, the first path on the other hand would be preferred since it presents a less risky deployment while also including critical pipes.

How we characterized the tool limits in this study is through a simple penalty parameter which only takes into account pipe material and diameter and assumes an equal cost for each respective change. This may not always be the case, since costs could grow in a non-linear fashion. This can

be modeled as part of more sophisticated formulations in future research. Further considerations for tools include limiting the turns in a path as some tools operate better traveling down straight lines [159]. The direction of flow as well as the flow rate of water in the segments is also an important factor in evaluating candidate inspection paths not considered in this study. In subsequent modeling efforts, added complexities that better reflect real world operations can allow decision makers to gain more useful insight from the model results.

In summary, while only simplifying assumptions are included, it is shown that the Evolutionary Program applied in more realistic networks can aid the formulation of better inspection processes. A better inspection process will lead to collection of better information regarding the condition of transmission and distribution system assets. Having clearer knowledge on where the system is in good condition while knowing where system issues need addressing will lead to a more resilient water distribution network; protected against future failures leading to a reduction in related health concerns and minimizing costs associated with pipe breaks.

3.7 Conclusion

In this research, a general optimization framework is presented for identifying inspection paths within drinking water networks. The formulation aims to select the most critical segments for assessment, while also accounting for the operations limits of the technology at hand. While the given examples relate to water distribution systems specifically, the same formulation can be used to plan for inspections of all types of infrastructure networks.

Examples of applying the optimization formulation are shown using synthetic and real networks. It is shown that as the scale of the network grows beyond the computational limits of exact methods, and complexity of its layout increases, more advanced heuristics are required to identify good paths.

Future research should aim to explore other optimization algorithms to compare effectiveness and efficiency, to extend the current formulation to include both targeted and continuous inspection tools, and to model tool-specific considerations when evaluating identified paths.

Acknowledgements

We thank the University of Michigan and NSF (grant number 1621116) for funding this research. The opinions and views expressed are those of the researchers and do not necessarily reflect those of the sponsors.

CHAPTER 4

Optimizing Inspection Routes in Pipeline Networks

Maintaining an aging network is a challenge for many water utilities due to limited budgets and uncertainty surrounding the physical condition of buried pipeline assets. The deployment of robotic inspections provides high quality data, but these platforms have limited use due to cost and operational constraints. To facilitate cost-efficient inspections, operators need to identify high-risk assets while accounting for the effectiveness of the tools at hand. This chapter addresses inspection planning with the goal of finding an optimal route considering tool limitations. An exact integer programming formulation is presented where only three factors are used to characterize tool constraints. Two classes of solution methods are explored: 1) tree based searches, and 2) integer programming. This chapter demonstrates how each method can be used to identify optimal inspection paths within a real water distribution system. Empirical trials suggest that tree-based search methods are the most efficient when the path limit is short, but do not scale well when the path length increases. In contrast, integer-programming methods are more effective for longer path lengths but have scalability issues for large network sizes. Data preprocessing, where the input network size is reduced, can provide large computational time reductions while returning near-optimal solutions.

Keywords: Decision Support Systems, Risk Analysis, Routing, Drinking Water Distribution Systems, Asset Management.

Note: The research presented in this chapter has been published at the Journal of Reliability Engineering and System Safety. Citation: Thomas Y.J. Chen, Connor T. Riley, Pascal Van Hentenryck, Seth D. Guikema. *Reliability Engineering and System Safety* 2019.

doi: https://doi.org.10.1016/j.ress.2019.106700

4.1 Introduction

There are over 150,000 public water systems in the USA, approximately 51,000 of these are community water systems that serve an estimated 90% of the total population [76]. These large networks (the average utility owns over 1600 miles of pipe [194]) are recognized as one of the most critical infrastructures [70], but many operating utilities are challenged with maintaining them with insufficient budgets and limited information [19]. Leaky distribution pipes account for approximately 14% of treated drinking water loss. These breaks, which occur as a result of aging, can also impact water quality for the consumer (inadequate disinfectant residual, low pressure, etc.) [131] with potential health impacts [184]. Experts estimate over $500 billion in capital investments are required to address the replacement needs of these decaying assets over the next 40 years [39].

For utilities, having an effective asset management plan can lead to higher returns on capital spending by targeting the highest risk assets for renewal [147]. This can be a difficult task when dealing with buried infrastructure because there is limited and uncertain information surrounding their current physical condition [19]. Robotic inspections are often deployed as a means for collecting real-time information on pipe health and condition. Other benefits of using these platforms include: 1) removing humans from hazardous work situations, 2) allowing for inspection of inaccessible areas and, depending on the tool, 3) providing online inspection without stoppage of pipe operation [173]. Tur and Gathwaite [194] summarize some current inspection technologies and Daly et al. [65] provide an overview of the data collection techniques associated with these tools. Despite the continuous high quality data available with the use of robotic inspections, economic and regulatory constraints have limited their widespread use [160] .

In order to facilitate higher returns on spending for inspections and asset renewal, identifying regions of high failure risk is critical. Prioritizing inspections based solely on a risk-based ranking of assets is popular in both practice and previous research (see [87, 157, 145]), but can lead to suboptimal results. The sensors used by the tool as well as the physical properties of the pipeline can affect the quality of data obtained and must be accounted for when designing inspection routes.

This research tackles the inspection routing problem such that tool limitations are also considered. We specifically target the potential for signal losses as a result of sensor calibration between varying pipe material and size [159]. It is possible to extend the formulation to model other limitations within an inspection routing framework. We present an optimization model for identifying a continuous route for inspection and explore the effectiveness of five different solution algorithms. The aim is to examine and discuss both the scalability and practicality of these methods for real-world systems. To broaden the related research on inspection planning, the model formulation will include two important considerations: 1) the inclusion of high-risk pipe segments along the selected path, and 2) a penalty to reflect the limits of the technology platform to effectively collect

high quality data along the identified path.

The same problem is first discussed in Chen et al. (2018) [58], where a general routing framework is presented and solved using heuristics. The authors focused on the importance of considering tool limits rather than obtaining paths which are provably global optimums. The main contributions of this work can be summarized as follows.

1. We propose an exact optimization formulation for finding optimal pipeline inspection routes while also considering platform limitations. These are critical considerations that must be taken into account when planning robotic inspections [58]. From our review of the literature regarding risk based inspections, we find no previous work that presents an exact formulation of this problem.
2. We present several complementary methods to obtain such optimal paths that are applicable to real networks. The comparison of different solution methods is informative for evaluating how different features of the problem (e.g. available budget, network size) impact the most appropriate method for finding optimal paths efficiently.

The methods presented here can certainly be improved, but they highlight the power of optimization to address practical problems in inspection planning. Together these two contributions advance the modeling approach for inspection planning as better paths can now be found. This research also highlights the strengths and weaknesses of different optimization techniques when applied for this task.

The rest of this chapter will be organized as follows: a review of inspection technologies and planning is in section 4.2. Section 4.3 specifies the routing problem and presents the optimization formulation. Section 4.4 covers the solution methods, as well as their implementations on a real network. Empirical results are presented and discussed in section 4.5 and we conclude our findings in section 4.6.

4.2 Literature Review and Background

The condition assessment of pipelines is an useful aid for decision making regarding repair, rehabilitation, and replacement. It is typically used to mitigate costs of failure by determining potential failure locations in advance [137]. A popular method for assessment is through the use of inspection robotics, where inferential indicators of stresses and defects can be identified [147].

Continuous inspections are a means for collecting information along an entire length of pipe. This is in contrast to spot or discrete inspections where only select locations are targeted (e.g. every 10 meters along the pipeline) [65]. A 2013 EPA report summarizes a number of continuous

inspection platforms which are currently on the market and includes a field demonstration of their capabilities [160]. The most commonly attached device is a CCTV camera which would allow video to be captured inside the pipe. On top of that, many other sensors can also be outfitted to the inspection robot, mostly classified as electromagnetic or acoustic [137]. Acoustic sensors are geared towards leak detection while electromagnetic sensors are used to identify defects on the pipe surface. Other less commonly used platforms include sonar and lasers, typically used to locate debris and sediment build-up.

As mentioned in Section 4.1, both the academic and industry literature which deals with inspection planning aims to prioritize assets by risk (see [145] and [159] for examples). This ensures that costly data collection techniques are targeted to the most vulnerable assets, and many previous studies have presented mathematical models for this task (see [111, 190, 185]). The academic literature presents various applications of risk based inspection planning on engineered systems such as offshore structures [155], oil rigs [45], ships [72, 139], and aircrafts [208]. These models take in a user-defined estimate of asset risk and determines which assets to inspect and how often. Typically, either a budget is specified or the objective function minimizes the cost associated with carrying out these inspections.

For example, Li et al. (2012) [135] formulates an hierarchal integer program to determine the minimum number of locations needed such that there can be complete coverage of the pipeline environment using CCTV cameras. Lu et al. (1999) [141] solves for specific routes by likening the task as a traveling salesman problem (TSP), except each location may be visited more than once for complete inspection. A TSP formulation can lend itself useful to a host of solution methods, some of which are explored and discussed in Laporte (1992) [126]. However routing problems are NP complete [141] which means an efficient (non exponential) solution method is not currently known.

Works by Straub et al. (2005 [190], 2006 [189]) and Luque et al. (2019) [143] aim to find optimal replacement and inspection policies for multi-component systems (e.g. roadways or pipelines) using bayesian methods. Each of these studies report that the large number of assets consisting a infrastructure network make identifying optimal inspection policies with exact methods computationally infeasible. A method to integrate risk based inspections into a wider decision making framework is presented in Khan et al. (2004) [112]. Here expert knowledge is combined with risk assessment outputs using multi-attribute decision-making tools. To address uncertainties related to the structural deterioration rates, works by Kallen and Noortwijk (2005) [104] and Memarzadeh and Pozzi (2016) [149] propose stochastic models to identify inspection policies under uncertainty.

A case study that tackles optimal inspection scheduling is found in Dandy and Engelhardt (2001) [67]. The authors demonstrate the use of genetic algorithms to formulate a cost-minimizing five year rehabilitation and inspection plan for an Australian municipality. This work is extended

in Dandy and Engelhardt (2006) [66] where multiple shareholder interests are captured in the optimization (e.g. least cost, maximum system reliability). Besides only considering asset risk in the planning stage, these studies also formulate the problem to provide complete coverage of the pipeline network. The practical implication of the findings become limited because most utilities do not have the budget to inspect their entire system over a limited time horizon. On top of economic constraints, regulatory and operational considerations limit inspection, replacement, and repair capabilities to approximately 2% of the system length for most municipalities [58].

In a best-practice guideline published by the National Association of Corrosion Engineers (NACE) [159] for robotic inline inspections, a number of operational issues are listed which must be accounted in the planning stage. These concerns are associated with physical properties inside the pipe which may affect the quality of the data collected. Mazumder et al. (2018) [147] identifies some state of the art inspection technologies and highlights their limitations. For example: the effectiveness of laser scanners depend of the pipe surface roughness and color, signals from impact echos can be affected by the presence of embedded items inside the pipe, and acoustic methods may not provide accurate readings for plastic pipes.

It is important that limitations must be acknowledged in the planning of robotic inspections. The goal of the routing optimization is to identify the optimal path which maximizes the risk of inspected assets, while being penalized for pipe properties which can lead to signal loss. To the authors knowledge, no past research has presented an exact mathematical model which tackles inspection path planning in this fashion.

4.3 Problem Definition and Optimization Formulation

In this section we will define the inspection routing problem and present it's mathematical formulation. Only three factors pertaining to the limitations of an inspection platform are addressed in the optimization, however it is possible to extend the general approach presented below to model a variety of technology-specific considerations.

4.3.1 Model Specification

The routing optimization presented here is similar to the prize collecting traveling salesman problem [42]. The goal is to find the maximum value path where only a subset of the pipes can be traversed. Due to physical constraints on the water flow, the solution must be an elementary path, i.e., a path that visits a junction at most once. Indeed, traversing a cycle would often require the real-time operation of valves, which most utilities are not capable of. Furthermore, passing the same junction twice with tethered platforms increases the risk of having the tool being stuck.

A limited inspection budget is specified as a limit on the total length of the path. The rewards for the designated route are associated with the pipe risk (edges), and the penalties are assigned based on properties which negatively impact inspection readings. We focus on adjacent pipes of different material and diameter. Many robotic sensors need a certain liftoff from the wall surface [137] for accurate readings and require different calibrations based on material type (cast iron, steel, etc.) and pipe diameter. When tools move between adjacent pipe with different properties, sensors need to be recalibrated and no inspection readings are taken during this process. This is a particular concern for free floating devices which collect inspection data while moving along the flow of water [137]. Condition data is lost along the length of pipe that is traversed while the tool is adjusting to different pipe properties [58].

While our model only addresses material and diameter changes, it can be extended to include other technology specific penalties. For example: avoiding sharp bends for tethered tools, avoiding non-metallic pipe for electromagnetic tools, etc. To the authors knowledge, no past research has presented an exact mathematical model which tackles inspection path planning in this fashion. We only consider the section of a single inspection route in our optimization. Due to the large set up costs for performing a single inspection (e.g. crew dispatch, road closures, excavation equipment to insert and retrieve tool), most utilities in practice only plan for single deployments at a time. It is possible to optimize for the planning of multiple routes constrained under one budget, e.g. identify multiple paths where the sum of length is less than 1 miles rather than just a single path less than 1 mile. Instead we focus the formulation and the exploration of techniques to finding one path, variants of this problem is left for future research.

4.3.2 Mathematical Formulation

We start by first defining an algebraic representation of a water distribution system. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph, where $\mathcal{V}$ represents the set of pipe junctions and $\mathcal{E}$ represents the set of all pipe segments. Let each vertex in the network be indexed by $i \in \mathcal{V}$, and each edge be represented by the pair of vertices $(i, j) \in \mathcal{E}$. This is a popular abstraction of a water system which has been applied in various research ([209, 8]). Each vertex can represent a water source, a connecting valve, a storage/treatment facility, or a pipe junction. Each edge represents a pipeline segment, and for simplicity, we assume that these are undirected edges.

Let R_{ij} and L_{ij} represent the risk score and length (in miles) of each pipe segment. Note that each edge $(i, j) \in \mathcal{E}$ also has a corresponding pipe material and diameter. Additionally, let s and t be the indexing of pseudo source and sink vertices, any feasible path will start and end at these 2 vertices. The goal of the mathematical program is to select a series of edges (i, j) which forms the highest-value elementary path, bound by the total distance (D_L, D_U).

We define $\mathcal{D}_i$ as the set of indexes j for all vertices adjacent to vertex $i \in \mathcal{V}$. Let all pairs of adjacent edges be indexed by the triplet (i, j, k), where j is the connecting vertex. The set $\mathcal{T}$ represents all pairs of adjacent edges in $\mathcal{G}$. From the network data, the following information can be extracted: M_{ijk} equals 1 if a material change occurs between the pair of adjacent pipes $(i, j, k) \in \mathcal{T}$, 0 otherwise. D_{ijk} equals 1 if a diameter change occurs between the pair of adjacent pipes $(i, j, k) \in \mathcal{T}$, 0 otherwise. Finally, let P_M and P_D be the path penalty factors associated with a material and diameter change along a selected pair of adjacent pipes $(i, j, k) \in \mathcal{T}$.

The decision variables for this problem are as follows:

- X_{ij} equals 1 if the edge (i, j) is selected for inspection, 0 otherwise. $\forall (i, j) \in \mathcal{E}$.
- Y_{ijk} equals 1 if adjacent edges X_{ij} and X_{jk}, are both selected for inspection, 0 otherwise. $\forall (i, j, k) \in \mathcal{T}$
- U_i is the vertex labelling variable used for subtour elimination $\forall i \in \mathcal{N}$. We use the Miller-Tucker-Zemlin (MTZ) formulation [153] of subtour elimination constraint set.

Below is the full integer programming formulation.

$$\max \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{D}_i} R_{ij} X_{ij} - P_M \sum_{(i,j,k) \in \mathcal{T}} M_{ijk} Y_{ijk} - P_D \sum_{(i,j,k) \in \mathcal{T}} D_{ijk} Y_{ijk} \tag{4.1a}$$

Subject to:

$$D_L \leq \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{D}_i} X_{ij} L_{ij} \leq D_U \tag{4.1b}$$

$$\sum_{i \in \mathcal{V}} X_{si} = \sum_{i \in \mathcal{V}} X_{it} = 1 \tag{4.1c}$$

$$\sum_{j \in \mathcal{D}_i} X_{ji} + X_{si} \leq 1, \forall i \in \mathcal{V} \tag{4.1d}$$

$$X_{ij} + X_{ji} \leq 1, \forall (i, j) \in \mathcal{E} \tag{4.1e}$$

$$\sum_{j \in \mathcal{D}_i} X_{ji} + X_{si} = \sum_{k \in \mathcal{D}_i} X_{ik} + X_{it}, \forall i \in \mathcal{V} \tag{4.1f}$$

$$Y_{ijk} \leq X_{ij}, \forall (i, j, k) \in \mathcal{T} \tag{4.1g}$$

$$Y_{ijk} \leq X_{jk}, \forall (i, j, k) \in \mathcal{T} \tag{4.1h}$$

$$Y_{ijk} \geq X_{ij} + X_{jk} - 1, \forall (i, j, k) \in \mathcal{T} \tag{4.1i}$$

$$U_i - U_j + 1 \leq (\mid \mathcal{V} \mid -1)(1 - X_{ij}), \forall (i, j) \in \mathcal{E} \tag{4.1j}$$

$$U_s = 1 \tag{4.1k}$$

$$U_t = \mid \mathcal{V} \mid +2 \tag{4.1l}$$

$$X_{ij} \in \{0,1\}, \forall (i,j) \in \mathcal{E} \tag{4.1m}$$

$$Y_{ijk} \in \{0,1\}, \forall (i,j,k) \in \mathcal{T} \tag{4.1n}$$

$$U_i \in \{2, ..., \mid \mathcal{N} \mid +1\}, \forall i \in \mathcal{V} \tag{4.1o}$$

The objective function (4.1a) is the sum of risk scores along the selected edges, minus the total pipe junction transfer costs that are incurred due to material and diameter changes. We will assume that the rewards and penalties grow in a linear fashion. These penalties are defined as the total number of pipe feature changes that occur along the path, multiplied by the corresponding penalty factor.

Constraints (4.1b) impose that the selected path be bound by the distance limit (D_L, D_U). Constraints (4.1c) - (4.1f) are flow balance constraints which impose that the selected solution must be an elementary path (no repeated vertices): (4.1c) enforces a unit flow at the source and sink vertices, (4.1d) and (4.1e) prevent branching or looping by restricting the flow entering each vertex, and (4.1f) is the set of flow balance constraints. Constraints (4.1g) - (4.1i) imposes the relationship between the edge selection variable X_{ij} and the edge pair selection indicator Y_{ijk}: Y_{ijk} is 1 if and only if both X_{ij} and X_{jk} are 1.

Since feasible solutions can only include elementary paths, by definition no sub-tours are allowed. Constraints (4.1j) - (4.1l) handle sub-tour elimination following the MTZ vertex labelling formulation [153] by imposing a strict ordering on the vertex labels. Any feasible solution will first consist of an edge out the source X_{si}, followed by a series of connected edges (X_{ij}, X_{jk}, etc.), then an edge into the sink X_{kt}. For any feasible path, constraint (4.1j) imposes that the labels U_i for all intermediate vertices must be strictly increasing when ordered from source to sink. This prevents subtours since (4.1j) will be violated if a path loops back on itself. Constraints (4.1k) and (4.1l) gives the source the smallest label and the sink the largest label. Finally, constraints (4.1m) - (4.1o) define the domain of each decision variable.

One other practical application of this research is to identify inspection paths under uncertain conditions. No model is perfect, and the risk science literature argues single metrics paint an incomplete characterization of risk [79] because uncertainties and assumptions behind these estimates are not communicated. One way to elicit uncertainties is to quantify risk as a set of possible values rather than a single number. To integrate the optimization of inspections into this broader risk definition, model (4.1) can be solved across different reward values (R_{ij}) and penalty coefficients (M_{ijk}, D_{ijk}) and the different paths are recorded. Finding optimal paths under varying inputs represent uncertainties in the risk analysis and tool performance. Instead of identifying a single solution, the decision maker can now find a group routes, each optimal under different assumptions. This allows for the comparison of different inspection investments against each other,

and ultimately enhances the utility of the optimization model as a decision support tool.

4.4 Methodology

This section describes the methodology used for this research. This includes the extraction of network information from a real water distribution system and the application of a risk model to characterize asset condition. It also includes a description of the five solution methods used to solve the formulation presented in Model (4.1).

4.4.1 Network Test Case

We partnered with the local utility in Ann Arbor, MI, to obtain a spatial file of the city's water distribution system. Figure 4.1 below depicts the full system layout.

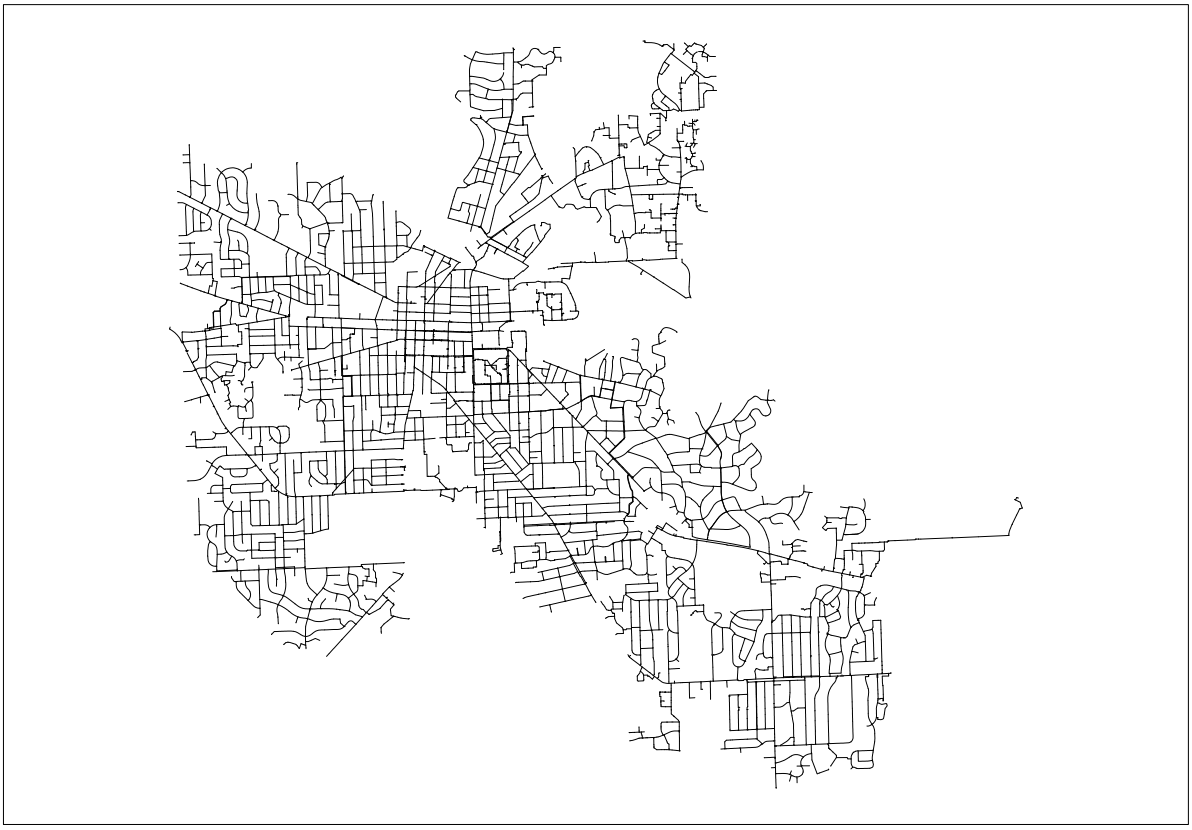


Figure 4.1: City of Ann Arbor Water Pipe Network.

The Ann Arbor network contains pipe of 9 different material classes (predominantly Cast Iron and Ductile Iron) and 12 different diameter sizes ranging from 2 - 24 inches. There are a total

of 12,538 unique pipe junctions (the set $\mathcal{V}$), and 13,058 unique pipe segments (the set $\mathcal{E}$) in the network, accounting for a total of 242.26 miles of total pipe.

From the spatial file we can extract the following information: the length of each pipe segment L_{ij} $(i,j) \in \mathcal{E}$, the set of all adjacent pipes $(i,j,k) \in \mathcal{T}$, and the presence of material and diameter changes between each pair of adjacent pipe M_{ijk} & D_{ijk}, $(i,j,k) \in \mathcal{T}$. Note that in the optimization formulation the specific material or diameter does not matter (e.g. 12 in. steel or 4 in. PVC), we only care about the change in these properties between adjacent pipes.

4.4.2 Network Test Case with Data Preprocessing

Each method we investigate is guaranteed to find the optimal inspection path, the more interesting discussion is on the efficiency and scalability of these methods. The size of the optimization problem can be characterized by the size of the input network and the length of the allowable path. In the link-node network representation, size does not pertain to the total mileage of pipe but rather the total number of edges $|\mathcal{E}|$ and vertices $|\mathcal{V}|$.

Beyond directly working with the original spatial data itself, we also implement a network preprocessing step which can reduce the cardinality of the edge and vertex set. The preprocessing involves two steps: 1) merging adjacent edges which are the same pipe type, and 2) removal of short edges. The first involves identifying all adjacent pipes which have the same material and diameter and merging them into a single pipe. Note we merge pipes when there are strictly two which meet at a junction. If there are three or more pipes meeting at an intersection we cannot merge. Rewards for the set of new psuedo-edges is the sum of the constituent edge rewards R_{ij}. Only material and diameter is considered for when checking "identical pipes" since they are the ones used in the optimization model. The second step simply removes pipes which are three feet or less in length, this corresponds to less than 0.1% of the total network length. The reduction in the input network after preprocessing the data is summarized in table 4.1 below.

Table 4.1: Change in Optimization Model Input after Data Preprocessing.

	Original Network	Network with Edge Aggregation and Removal	Reduction
$\|\mathcal{E}\|$	13058	4565	65.04%
$\|\mathcal{V}\|$	12538	4584	63.44%

Note that it is possible that the optimal solution is changed when joining and removing pipes because feasible solutions are removed. Part of the research goal is to also explore the tradeoff between reducing the computational burden by preprocessing the network and the resulting solution quality. Besides exploring the change in solution time as a function of the size of the input network, we also adjust the length of the allowable inspection path. For each of the two graphs we also use

the following path distance limits: $D_L = 0.00$ mi., $D_U = \{0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00\}$ mi. The two mile limit is used to reflect the limited inspection budgets, single inspection routes in practice typically do not extend beyond this limit [160].

4.4.3 Probability of Failure Risk Model

In order to characterize the physical condition of the pipe segments in the full Ann Arbor system as well as the preprocessed network, an age-based pipe-break likelihood of failure model is implemented. The risk score assigns the rewards (the R_{ij} coefficients) for inspecting each asset.

The reward for inspecting each asset is set to the probability of failure. Implementing a more complicated risk function would lead to better inspection paths but is beyond the scope of this research. We focus the work on the model formulation, and the subsequent analysis on the performance of the optimization algorithms.

We use a model published in the literature. Similar to the methods presented in Chen et al. (2018) [58], we use a case study by Genevieve et al. (2013) [168] where a Weibull distribution is fit against historical failure data from a Canadian utility. A Weibull distribution is characterized by two parameters, κ and ρ. It models the time to first failure from initial installation [174]. The authors [168] found a better statistical fit of the failure data is achieved by discretizing the pipe system into two classes based on installation year, before and after 1960, and fitting two separate distributions.

A hazard function, $\lambda(\theta)$, which corresponds to the annual probability of failure can be derived from a Weibull distribution.

$$\lambda(\theta) = R_{ij} = \kappa\rho(\kappa\theta)^{\rho-1} \tag{4.2}$$

A Weibull distribution is characterized by two parameters, κ and ρ. Where θ is the age of the pipe segment in years, κ (units year^{-1}) is the scale parameter of the distribution, and ρ (unitless) is the shape parameter. ρ values greater than 1 signify that failure rates of a pipeline will increase over time Table 4.2 below represents the κ and ρ parameters used for the test networks. We apply equation 4.2 to all edges $(i, j) \in \mathcal{E}$ in the network to compute the inspection rewards R_{ij}.

Table 4.2: Hazard Function Parameters of Pipe Break Likelihood Risk Model, taken from Genevieve et al. (2013) [168].

Parameter	Pipes Installed Before 1960	Pipes Installed at 1960 or After
κ	0.022	0.029
ρ	2.725	2.172

4.4.4 Solution Methods

We implement five algorithms to solve the optimization formulation presented in Model 4.1: integer programming branch and bound (IP), constraint generation (CG), depth first search (DFS), breadth first search (BFS), and a pruned depth first search (DFS$_{pruned}$). These algorithms are categorized as either: 1) tree traversals or 2) integer programs.

We ran our experiments on a machine with 32 cores running at 2.6GHz. For the integer-programming methods, the optimization model is written in Python3 using the "gurobipy" package and then solved using Gurobi (written in C++). The tree-based methods are implemented and solved entirely in Python3. Once again, the goal of this research is not to compare different methods, but to show their practicability, strengths, and weaknesses of different. Finding the best optimization model is beyond the scope of this research.

4.4.4.1 Tree Search

We first examine two exhaustive search techniques: breadth first search (BFS) and depth first search (DFS). Both enumerate all feasible paths starting from each vertex $v \in \mathcal{V}$ in parallel. See references by Ahuja et al. (1993) [2] for a full description of both graph traversal algorithms. For our implementation, the starting vertices v are sorted based on their geographical location and processed in order of increasing latitude and longitude.

The DFS was also tested with pruning, we call this the DFS$_{pruned}$ method. The algorithm relies on a depth-first traversal over the network, but computes a heuristic at every step to evaluate an upper-bound objective for the current path. The traversal along a branch of the DFS is terminated if this upper bound is less that the best solution currently found. This allows us to find the optimal solution without fathoming every feasible path, effectively pruning the size of the search space.

The method also processes each vertex $v \in \mathcal{V}$ in order based on latitude and longitude. To prune paths for a particular start vertex v_s, we record all edges contained within a radius of length D_U from v_s and sort these edges based on unit reward L. L is defined as the ratio between reward R_{ij} and length L_{ij}. We sort based on unit reward since the pipe segments lengths L_{ij} are non-homogenous in the given data. The DFS traversal algorithm is again executed in parallel. At each node of the DFS, the subpath contains a series of edges E and has certain distance from the limit $(D_U - d_E)$. The length of the current path is d_E, the current value is R_c, and a heuristic $h(n)$ for the best additional value that can be achieved in $(D_U - d_E)$. To calculate $h(n)$, the rewards of the edges from the set $L \setminus E$ are summed starting from the edge with the best unit reward, until the distance of these edges is equal to $(D_U - d_E)$. At every node n, if $R_c + h(n)$ is less than the value of the best path currently found in the search, then n is not expanded further.

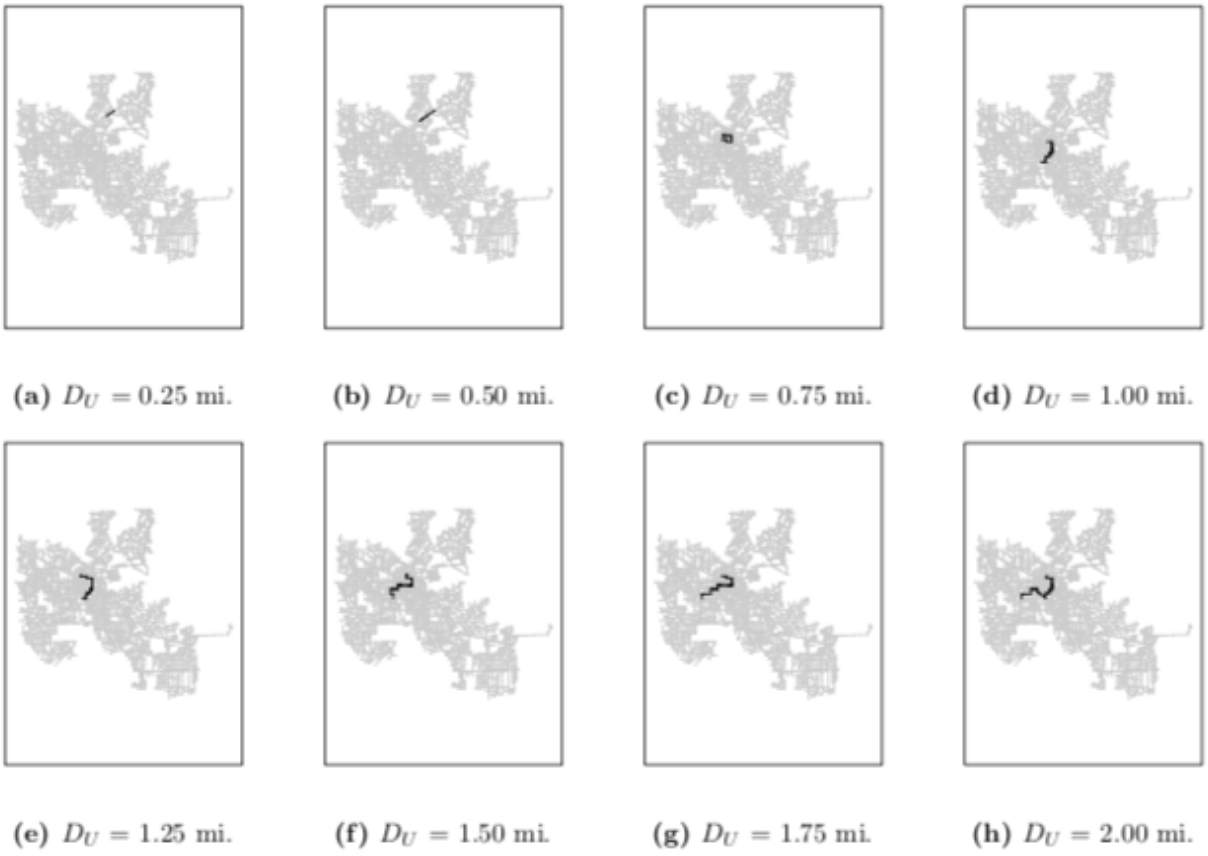
4.4.4.2 Integer Programming

Model (4.1) was solved using Gurobi [98] both using the full formulation (IP) and using a constraint generation (CG) approach. The Gurobi solver uses a branch and bound method [129] to identify the optimal solution.

In constraint generation, constraints (4.1j) - (4.1l) (the subtour elimination constraints) are removed from the model and the relaxed problem in solved. We remove these constraints because they give the master problem a weak linear relaxation, making it more computationally inefficient. For all subtours $s \in S$ found by the relaxation, add the constraint $\sum_{(i,j)\in S} X_{ij} \leq |\mathcal{S}| - 1$ and resolve the integer program with these added constraints. Here we are specifically targeting the identified subtours and removing them from feasibility. By repeating this process until no subtours are present in the solution of the relaxation, we are guaranteed a solution that is also optimal to the master problem in Model (4.1).

4.5 Results and Discussion

We solved for the optimal path within the Ann Arbor system with and without data preprocessing, each with 8 different limits on the allowable length D_U. The reward coefficients R_{ij} for each pipe segment is defined using equation (4.2), and the penalty coefficient due to material and diameter changes (P_M and P_D) is set to 1. Since all rewards are bound between 0 and 1, we select a penalty of 1 to sufficiently negate the value of any selected pipe if a penalty is incurred when traversing it. In practice the value of these penalties should be defined by experts with experience using these inspection platforms [58] to better reflect the losses incurred due to signal disruptions. The figure below identifies the solution path in the unprocessed system over each D_U increment. We set a limit of 20,000 seconds (5.5 hours) in each trial. For the runs which completed within the allotted time, each method returned the identical solution.



(a) $D_U = 0.25$ mi. **(b)** $D_U = 0.50$ mi. **(c)** $D_U = 0.75$ mi. **(d)** $D_U = 1.00$ mi.

(e) $D_U = 1.25$ mi. **(f)** $D_U = 1.50$ mi. **(g)** $D_U = 1.75$ mi. **(h)** $D_U = 2.00$ mi.

Figure 4.2: Optimal Inspection Paths for Unprocessed Ann Arbor Water Pipe Network.

Five experiments were run for each algorithm/D_U/network combination. In table 4.3 below we present the value of the optimal path and the average computation time required to obtain the solution. The fastest average time over each D_U is in bold, the results displayed are only for the network before any preprocessing is applied. Since the exact solution time can vary across machines and trials due to a number of factors (e.g. background processes, CPU and memory capacity etc.), we focus our analyses on the solution times relative to each other rather than the raw number.

Table 4.3: Network without PreProcessing - D_U vs Path Value and Solution Time (sec.).

D_U	Objective Value	IP	CG	DFS	BFS	DFS$_{pruned}$
0.25	308.78	81.24	68.33	**12.13**	12.46	41.19
0.50	521.90	2008.72	3854.97	**30.50**	31.20	357.35
0.75	757.57	1090.10	5737.12	221.18	**210.89**	1513.82
1.00	994.72	**1230.94**	5781.23	1629.61	1677.13	1981.49
1.25	1227.48	**1371.22**	6690.16	>20000	>20000	2166.43
1.50	1448.59	**1540.87**	9886.01	>20000	>20000	>20000
1.75	1694.01	**1113.81**	5993.20	>20000	>20000	>20000
2.00	1934.82	**1744.07**	4441.38	>20000	>20000	>20000

The results show that, when $0.25 \leq D_U \leq 0.75$, the DFS is the most efficient and, when $1.00 \leq D_U \leq 2.00$, the IP branch and bound is best. The IP and CG branch and bound are only methods able to find the optimal solution within 5.5 hours over all of the D_U increments, suggesting they are relatively insensitive to D_U. The CG is slower than the IP in every instance except when D_U = 0.25 mi. A potential cause for the generally slower processing time is due to the increased number of subproblems it is required to solve: over the D_U increments the CG solves an average of 7 relaxations before obtaining the optimal solution.

We find that the number of subproblems solved before reaching the optimal solution is positively correlated to the overall solution time. The solution times for the integer programming methods are not strictly increasing with D_U, unlike the graph traversals. It is known that the branch and bound algorithm scales exponentially with the number of decision variables and constraints [163]. This implies that the efficiency of the integer programming approach is dependent on the size of the input network. In order to formulate Model (4.1), having more edges (pipes) increases the number of decision variables and more nodes (pipe junctions) increases the number of constraints.

When the allowable path lengths are small (0.25 - 0.75 miles) the DFS and BFS exhaustive searches are by far the fastest, the solution time being orders of magnitude less than the others. The DFS and BFS are identical algorithms with respect to the number of steps taken (search tree size) [2], any difference in solution times are due to implementation details. However the exhaustive search does not scale well. With $D_U \leq 1$ mi., a solution can be found within 1 hour, but with $D_U > 1$ mi. the methods take longer than 5.5 hours.

The DFS$_{pruned}$ is slower than both of the exhaustive search methods when $D_U \leq 1$ mi. However it manages to return a solution when $D_U = 1.25$ mi. whereas the exhaustive searches cannot. The reason for the slower computation time at the start is due to the extra computation from having to solve a heuristic (see section 4.4.1) at every step. The results imply that when D_U is small, a

full enumeration of all feasible paths is preferable over using a heuristic to reduce the number of fathomed solutions. However when D_U exceeds a certain threshold, the reduction in the search space is enough such that pruning becomes more effective.

We next examine the effects that data preprocessing has on solution quality and computation time. Table 4.4 below presents the results after preprocessing steps are applied (joining edges with same properties and deleting edges shorter than 3 ft.). The new solution times are reported, along with the percentage reduction in the new objective value. Again, we will focus our discussion on the relative solution times of each method.

Table 4.4: Network with Preprocessing - Solution Time (sec.) and % Reduction in Solution Value.

D_U	Objective Value Change	IP	CG	DFS	BFS	DFS$_{pruned}$
0.25	-12.22%	41.26	14.26	4.62	**4.14**	7.91
0.50	-1.07%	72.41	410.90	4.70	**4.67**	15.46
0.75	-0.00%	31.93	53.44	**5.07**	5.32	41.69
1.00	-2.29%	22.51	76.95	18.43	**17.66**	78.66
1.25	-0.42%	**61.20**	159.32	83.67	80.40	103.47
1.50	-0.11%	**20.36**	54.86	333.06	330.86	201.01
1.75	-0.07%	**63.70**	132.41	1226.43	1210.02	225.64
2.00	-0.06%	62.53	**49.08**	4234.59	4306.73	259.33

Unlike the results without network preprocessing, all the tree search instances were able to find the optimal path within 5.5 hours. In most cases the decrease in solution quality (objective function value) was also minimal. The objective value reduced by less than 2.5% in all cases except at D_U = 0.25 mi., where the objective value reduced by 12%. The reduction in solutions times for the DFS/BFS ranged between 62-99% and increased with D_U. We again note that the DFS and BFS solution times are near identical, and any differences are due to implementation details.

The average reduction in solution times for the DFS_{Pruned} was about 83%. For the IP, the average solution time reduced by 50% in the smallest case (D_U = 0.25 mi.) but had at least a 94% reduction in all other instances. Reductions in solution times were also similar for the CG, averaging at about 95% across all instances. These results indicate that preprocessing the data can boost computational efficiency while returning near-optimal solutions.

When examining which methods are the fastest in relation to D_U, a similar pattern from the unprocessed network case is observed. For shorter paths ($D_U \leq 1.00$), the exhaustive DFS/BFS searches are the fastest. Meanwhile, for longer paths ($D_U \geq 1.25$), an integer programming branch and bound approach becomes preferable. An interesting observation is that the CG is the fastest in the largest instance (D_U = 2 mi.).

Similar trends are also revealed when comparing the exhaustive DFS/BFS against the DFS$_{Pruned}$.

At shorter path lengths, the full enumeration approach is faster. In contrast, at longer path lengths ($D_U \geq 1.50$) we see that the pruned search becomes preferable. This indicates that even for the processed network, involving additional computational costs to reduce the search space becomes favorable after a certain length threshold.

To further compare the BFS/DFS against the DFS$_{Pruned}$, we graph the size of the search tree for each method. The size of the search tree is the total number of steps taken by the algorithm to find the optimal path. Figure 4.3 shows the growth in the complexity of the search as D_U is scaled, in both the unprocessed and preprocessed network.

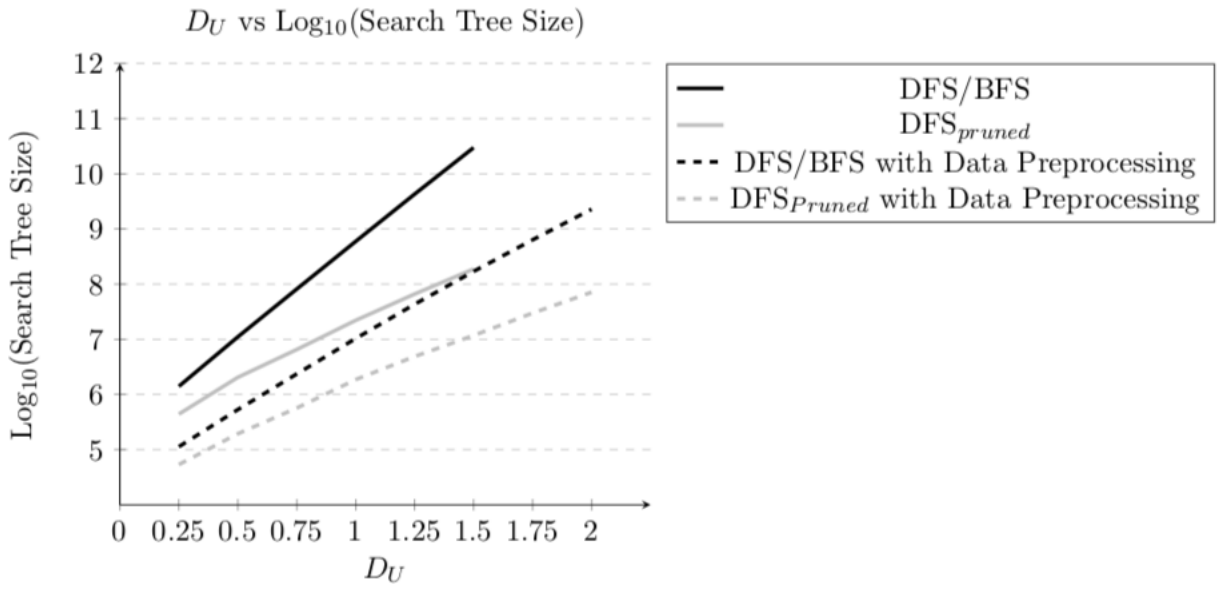


Figure 4.3: Exponential Growth in Tree Search Complexity as D_U Increases.

All the lines in figure 4.3 are roughly linear in the log scale. This shows that the size of the search space grows exponentially with D_U, both with and without network preprocessing. While the DFS and BFS complexities are polynomial (order 2) [2] in the which they search a network, the exponential growth in complexity with input size indicates that these methods scale poorly with path length.

Table 4.4 shows that DFS$_{pruned}$ is slower than an exhaustive search until a certain threshold in path length is reached. Figure 4.3 shows the corresponding reduction in the search space the algorithm produced. For the DFS$_{pruned}$ with and without network preprocessing, there is a $>88\%$ reduction in the search tree across all instances. However the exponential growth in the input, as evidenced by the grey lines in figure 4.3, show that even the pruned searches scale poorly with D_U.

Comparing solution times reported in tables 4.3 and 4.4, the data shows that preprocessing can provide a large boost in the computational efficiency. Furthermore the resulting paths are near-

optimal, the average reduction in the solution value is 2%. However, branch and bound and tree searches still may not scale well for larger networks and longer paths. This may be a problem for utilities with a larger pipeline network and a larger inspection budget.

The efficiency of DFS$_{Pruned}$ is in part dependent on the quality on the best solution found at any time. Currently the first solution is the first DFS path found and is updated throughout the traversal. If a high quality solution is instead used as the first path, a larger reduction of the search tree can be achieved. To provide a upper bound for the optimal efficiency of DFS$_{Pruned}$, we rerun the analysis where the known optimal path is used as the starting solution. This guarantees us a maximal pruning of the search tree in both the processed and unprocessed network. The size of the search trees is again plotted below for the original and unprocessed network.

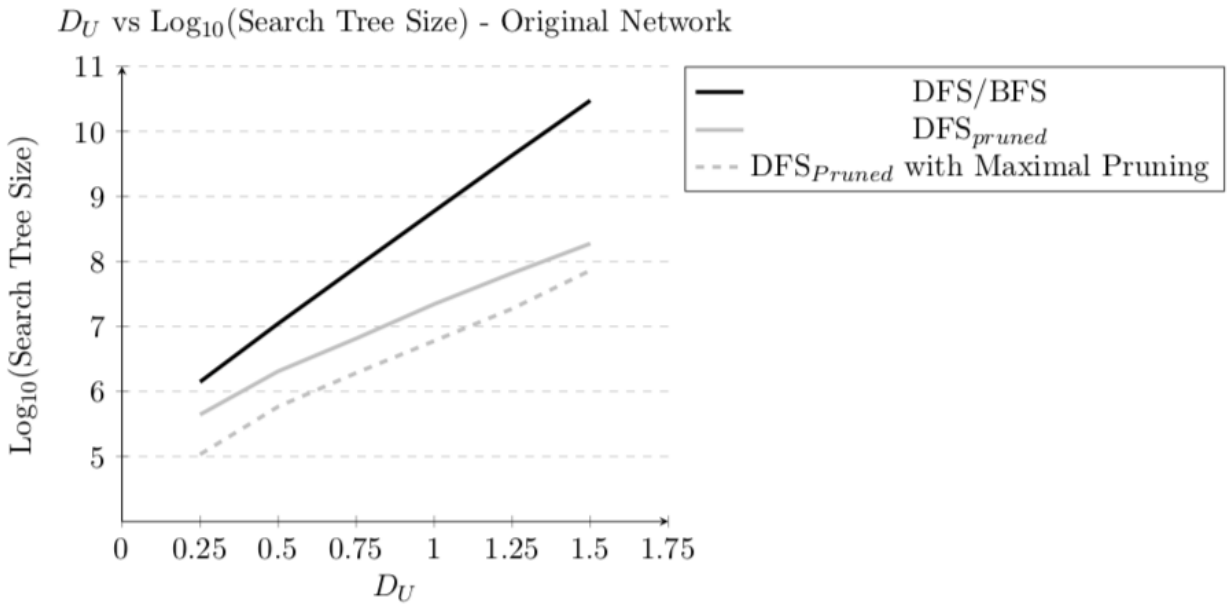


Figure 4.4: Tree Search Complexity Comparison, Original Network.

The average reduction in the search tree size when using the optimal solution is 70% from the original DFS$_{pruned}$, and 97% from the exhaustive searches. Below is the search tree size plots for the instance with data preprocessing.

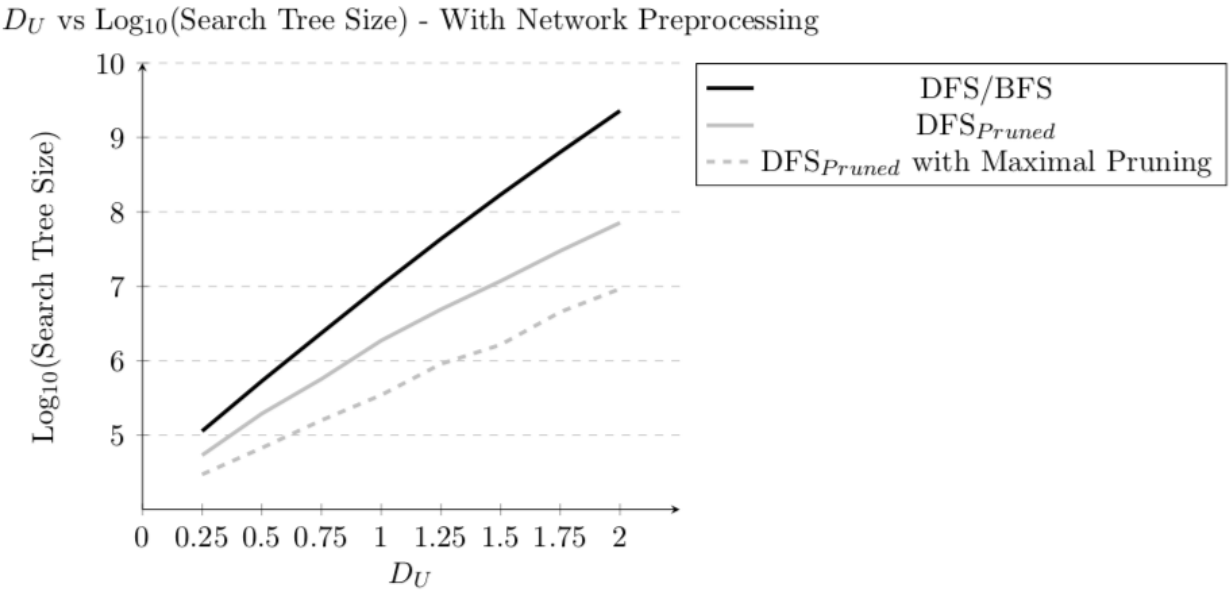


Figure 4.5: Tree Search Complexity Comparison, Network with Data Preprocessing.

The average reduction in the search tree size when using the optimal solution is 75% from the original DFS_{pruned}, and 93% from the exhaustive searches. The results show that in both the cases with and without network preprocessing, the reduction in the search space is similar when using the pruned search. However there is still room for additional pruning, and in turn improved computational efficiency. This is suggested in figures 4.4 and 4.5 by the gap between the solid grey and dashed lines.

The current implementation of DFS_{Pruned} processes each network vertex v in order based on longitude and latitude. The pruning capability of relies on the best solution found at each stage of the algorithm. Therefore, if good paths are found earlier during the traversal, the efficiency can be increased. Chen et al. [58] explores some heuristic algorithms for this task and can provide a good starting point. In this setting, instead of using the first DFS path out of the first vertex as the initial best solution, scan over the entire network and use a high quality heuristic solution instead. Exploring the possibility of combining heuristics with the current methods, as well as other classes of solution algorithms is left for future research.

4.6 Conclusion

In this research, we tackled the problem of optimal routing for in-pipe robotic inspections. The work extended the previous literature where the optimization will take into account both: 1) the limited capabilities to use these tools, and 2) the physical pipe properties down a given route which may limit the effectiveness of the tools. While the focus is on water main inspections, the methods

presented here can be extended for inspection planning in other types of networked infrastructure (e.g. power networks, gas networks, road networks).

An exact integer programming formulation was presented, along with five solution algorithms. The methods fall under 2 categories: integer programming and graph search. Other classes of methods may also work (for example, dynamic programming) but is beyond scope and is left for future work. We demonstrate the application of these algorithms using the water pipe network in Ann Arbor, Michigan. Empirical trials suggest that tree based searches scale poorly with the allowable path length, while integer programming methods are less sensitive. On the other hand, integer-programming-based methods do not scale well with the number of unique pipes and pipe junctions. We also show that graph preprocessing by 1) reducing the number of unique edges, and 2) ignoring edges below a length threshold, can provide a large boost in computation time while returning near-optimal solution paths. Using a heuristic to identify a good path first can further boost the efficiency of the pruned graph search.

Extensions to this research can tackle the planning of multiple paths at once, as well as modeling tool-specific considerations beyond just pipe property (material and size) changes. Using more advanced pipe failure risk models can identify better paths and enhance the usefulness of the optimization as a decision support tool. Examining the use of other risk frameworks in the inspection planning context, in particular how to best address uncertainties, is left for future work.

Acknowledgements

We thank the University of Michigan for funding this research. The opinions and views expressed are those of the researchers and do not necessarily reflect those of the sponsors.

CHAPTER 5

Statistical Modeling in the Absence of System Specific Data: Exploratory Empirical Analysis for Prediction of Water Main Breaks

The replacement of deteriorating distribution pipes is an important process for water utilities. It helps reduce capital spending on water main breaks and improves customer satisfaction. To assist with the development of an effective renewal plan, statistical models which forecast future breakage rates have been used to guide planning for asset management. However, this process is difficult for older utilities which lack readily available pipe network data. We examine whether accurate and useful predictive models can be built in the absence of pipe-feature data. Using the historical break record from a Mid-Atlantic utility, two datasets at different spatial scales are created using publicly available demographic and environmental information. Empirical results suggest that while accuracy suffers from the lack of pipe-level details, it is still possible to create a model which provides useful information for prioritization of high-risk regions for management.

Keywords: Drinking Water Distribution System, Replacement and Rehabilitation Planning, Asset Management, Statistical Modeling, Risk Prioritization

Note: The research presented in this chapter has been published in the Journal of Infrastructure Systems. Citation: Thomas Y.J. Chen, Jared A. Beekman, Seth D. Guikema, Sara Shashaani. Statistical Modeling in the Absence of System Specific Data: Exploratory Empirical Analysis for Prediction of Water Main Breaks. *Journal of Infrastructure Systems*, 25(2):04019009, 2019.

5.1 Introduction

The management of an aging water distribution system is important to water utilities due to the vital societal and economic impacts that are incurred when service is not provided [77]. However, management of these systems can be challenging due to limited availability of information and uncertainty regarding the physical condition of the pipes [19]. Aging, corrosion, and other environmental factors play a role in the deterioration of water mains [119] which can lead to leaks and breaks. Pipe breaks can pose public health dangers [184] by contaminating the distribution systems, and often lead to costly repair operations [196] (Walski and Pelliccia 1982). A survey by the Water Research Foundation estimates the average main break costs $42000 [205].

Being able to forecast which regions of the distribution network are at highest likelihood of failure can help utilities with efficient inspection and repair decisions [116]. However, this can be challenging for some utilities due to uncertain and limited information of system-specific features [19]. Utilities such as these do not have a digitized map of the system, nor do they have records of basic pipe characteristics such as material, size, and age. This lack of geospatial data is surprisingly common in the US, especially among small to mid-size utilities. As a result, these utilities often plan their asset management in an ad-hoc fashion such as relying on expert judgment or prioritizing assets based on past failure trends.

To address this problem, this research aims to develop statistical models which forecast future breakage probability using public demographic and environmental information as proxies for pipe-level data. Using models such as these would potentially allow utilities to identify where vulnerabilities lie in the system.

Traditional modeling approaches may fail to predict accurately when there is a significant imbalance in the response variable (one class largely outnumbers the other class). This is common in segment-level pipe break data because pipe breaks occur infrequently at the level of individual segments. The proposed method in this study uses sampling methods outlined in He and Garcia (2009) [92] to achieve better balance in the training dataset.

Because managing pipe failures is a binary problem in practice that identifies whether or not the onset of a single failure of a distribution asset would require full repair operations, the modeling of pipe breaks is often framed as a binary classification. In this setting, the outcome is either no failure (a negative response) or at least one failure (a positive response) [168, 201, 207]. Thus, this research will explore the use of binary classification models to accurately predict the onset of pipe failures. The contribution here is exploring the use of only publicly available data in building predictive models, and determining whether these methods can be useful for utilities challenged by the lack of system level information.

Building models while lacking important explanatory variables raises the possibility of omit-

ted variable bias. Omitted variable bias [132] is a model misspecification issue that arises when variables (e.g. pipe-level data) highly correlated to the response (pipe break likelihood) are not available. There are statistical methods to account for the model bias that arises in this situation, such as multi-level modeling [148]. However, as discussed in Gelman et al. (2014) [84] and Allison (1999) [11], these control methods are developed for obtaining causal inference (i.e. controlling the error in the model training), rather than attaining optimal predictive accuracy.

Since the focus of this research is on the utility of statistical models in practice, predictive accuracy will be the primary measure of model performance. For completeness, multi-level models will be included for comparison but will be evaluated solely on predictive performance. Additionally, variable statistical significance and partial dependency analysis will not be specifically presented when comparing different predictive models because they evaluate in-sample fit rather than out of sample predictive accuracy. We will, however, estimate the statistical significance of our variables and their effects on the response within a logistic regression setting. Having knowledge on which variables the models found most useful can guide a utility's future efforts when collecting pipe level data.

We will assess model performance within 2 datasets at varying spatial resolutions (road level and census tract level) by carrying out using 2 types of holdouts tests: 1) a random cross-validation where the dataset is split randomly into training and validation sets, and 2) a temporal holdout where data from 51 of the 52 total months are used for model training and the resulting model is validated against the remaining month. In the random holdouts, we assess predictive performance using common classification metrics, while in the temporal holdouts we determine whether statistical models can prioritize high-risk assets and regions better than a method commonly adopted in practice [172].

5.2 Literature Review

Numerous previous studies are available concerning drinking water distribution pipe break modeling and these works can be summarized under two categories: physical-based and statistical-based models. Rajani and Kleiner (2001) [169] provide an overview of the physical/ mechanical models developed to understand the structural performance of water mains. These methods address the components of the physical processes that lead to pipe breakage, such as corrosion [170] and stress/strain within the pipe itself [144]. For example, one study [113] focuses on the relationship between corrosion pits and the structural resistance of steel pipes. The analysis uses empirical data where steel pipes of varying size and depths of corrosion were tested until failure. From this, an analytical model was developed to quantify the pressure at which a pipe with a given corrosion pit depth would fail.

In a companion paper, Kleiner and Rajani (2001) [117] provide an overview of statistical-based models by summarizing a large bodywork aimed at quantifying structural deterioration of water mains through the analysis of historical performance data. For example, one study [110] used condition and age data on pipes in Winnipeg, Manitoba over a 10 year period to fit a linear regression model to analyze the failure rates across different materials over time.

Statistical models can be more flexible than physical models since they can be applied with various types of input data, while physical-based models often require information that is not readily available or difficult to obtain. Statistical-based models for drinking water distribution pipe breaks are either inferential or predictive [90]. Inferential models aim to evaluate data as it is presented, formulating trends based on statistical measures such correlation and covariance. The goal of inferential models is to improve understanding, not to make accurate predictions. Examples are Kettler and Goulter (1986) [110] and Shi et al. (2013) [181], both using spatial and temporal clustering. However, Pelletier et al. (2003) [168] note that this type of statistical approach requires an explicit set of pipe characteristics that are often not available, making them difficult to apply.

On the other hand, many predictive statistical-based models for analyzing pipe breaks have been developed in previous works with varying degrees of model complexity. Two popular modeling tools proposed in literature includes 1) survival analysis [193] which estimates the time until next failure at the asset level, and 2) regression modeling which estimates the likelihood or number of failures at the next time frame [81, 110]. Since many water utilities only recently started recording pipe breaks, survival analysis models are less useful since they rely on having large amounts of historical information. Linear regression models were studied in Kettler and Goulter (1986) [110], and as a follow up Andreou (1986) [14] uses proportional hazards models and Generalized Linear models to achieve better predictive accuracy. Yamijala et al. (2009) [207] compared a range of regression models (time linear models, time exponential models, and logistic regression) and found that although the logistic regression model performed the best, none of them achieved particularly high predictive accuracy when used to rank assets according to risk.

To assess models with non-linear structures, Francis et al. (2014) [81] used Bayesian Belief Networks to construct a knowledge model for pipe breaks. To explore other non-additive methods, this research uses tree-based models explored in Chen et al. (2017) [56] to predict pipe break probabilities. These models include Classification Trees [133], Random Forests [48] and Boosted Trees [82], each of which will be discussed more in the "Data and Methods" section.

In recent statistical modeling studies, the topic of imbalanced learning [191] has drawn much attention. It focuses on the observed phenomenon, as seen in pipe failure history, where standard classification models do not provide high predictive accuracy when the instances of one response significantly outnumber those of the other. He and Garcia (2009) [92] offer a number of approaches to tackle the class imbalance problem, from working at the data level by changing the class dis-

tribution using sampling techniques, to the algorithmic level where the modeling approaches are adjusted. In modeling water main failures, Wang et al. (2013) [197] have shown that while sampling techniques did not lead to better accuracy, they did offer efficiency improvements since the training data can be reduced in size. For this study, a hybrid under-sampling and oversampling technique was used on the training data to increase balance between records of pipe breaks and no breaks.

5.3 Data and Methods

This section outlines the data collection and modeling methodology used in this chapter, including how the break data was spatially aggregated and processed, as well as an overview of all the statistical models used and the binary classification metrics used to evaluate them.

5.3.1 Pipe Break and Environmental Data

We partnered with a utility based in the mid-Atlantic region to obtain information on pipe break dates and locations between 2010 and 2014 (52 months). We aggregated this information to a monthly temporal scale and combined it with public demographic and environmental data to generate 2 datasets at different spatial resolutions: one at the road segment level and the other at the census tract level. We use the streets in the study area to proxy the location of individual pipe segments and the census tracts to proxy different geographic zones in a distribution system.

The partnering utility does not have a spatial dataset of their system, meaning no information (including location) on the distribution pipes is available. As a result, a publicly available (U.S. Census Bureau) spatial file of the area's roads was used as a network proxy, while census tracts were used to divide the service area into small regions (average 0.9 square miles). Failure history, climate data, and demographic data were collected and used as predictors. Table 5.1 describes each of the 24 predictors, their label in the model, and respected sources.

Table 5.1: Summary of dataset used for classification modeling.

Variable Name	Variable Description
BREAK	Binary Classification Response.
BREAK_$HISTORY$	Number of Breaks on road segment/census tract over past 6 months.
AA_$TOTAL$	Number of African American households in census tract.
AA_$PERCENT$	Percentage of African American households in census tract.
ASIAN_$TOTAL$	Number of Asian households in census tract.
ASIAN_$PERCENT$	Percentage of Asian households in census tract.
AGE_2005	Number of households built later than 2005 in census tract.
AGE_2000	Number of households built between 2000 and 2005 in census tract.
AGE_1990	Number of households built between 1990 and 2000 in census tract.
AGE_1980	Number of households built between 1980 and 1990 in census tract.
AGE_1970	Number of households built between 1970 and 1980 in census tract.
AGE_1960	Number of households built between 1960 and 1970 in census tract.
AGE_1950	Number of households built between 1950 and 1960 in census tract.
AGE_1940	Number of households built between 1940 and 1950 in census tract.
AGE_$LESS$_1940	Number of households built earlier than 1940 in census tract.
MEDIAN_$INCOME$	Median Income of households in census tract.
LAND_USE	Major land use identifier of road segment or census tract.
SOIL_$CLASS$	The highest level in soil taxonomy.
BOTTOM_$DEPTH$	Depth to bedrock below a given road segment or census tract.
CLAY_$PERCENT$	Percentage of soil that is clay.
COR_$CONCRETE$	Concrete corrosivity rating of soil: high, moderate, low, none.
COR_$STEEL$	Steel corrosivity rating of soil: high, moderate, low, none.
RUNOFF	Runoff potential rating of soil: high, moderate, low, none.
FROST_$ACTION$	Susceptibility rating of soil for frost heaving: high, moderate, low, none.
FREEZE	Number of days in month that go below freezing (32F).
TEMP_MIN	Minimum monthly temperature (F).
PRECIP	Average monthly precipitation (in.).
TEMP_$STDEV$	Standard deviation of daily TMIN over a month.

Using ArcGIS (ESRI), the environmental and demographic variables presented in Table 5.1 are expressed both spatially and temporally.

Information at both the road segment and census tract level was generated using a spatial overlay between database layers and the network file. The location of each pipe break record was geocoded and spatially referenced to a corresponding road and census tract. Monthly climate data

in the study period was obtained from a nearby NOAA weather station and assumed homogenous over the entire region. Figure 5.1 shows a geographic overlay between pipe break locations and both the road and census tract layers. The approximate distribution of the breakage data is as follows: 99.8% negative responses (no breaks) 0.2% positive responses (at least 1 break) at the road level, and 85.4% negative responses and 14.6% positive responses at the census tract level.

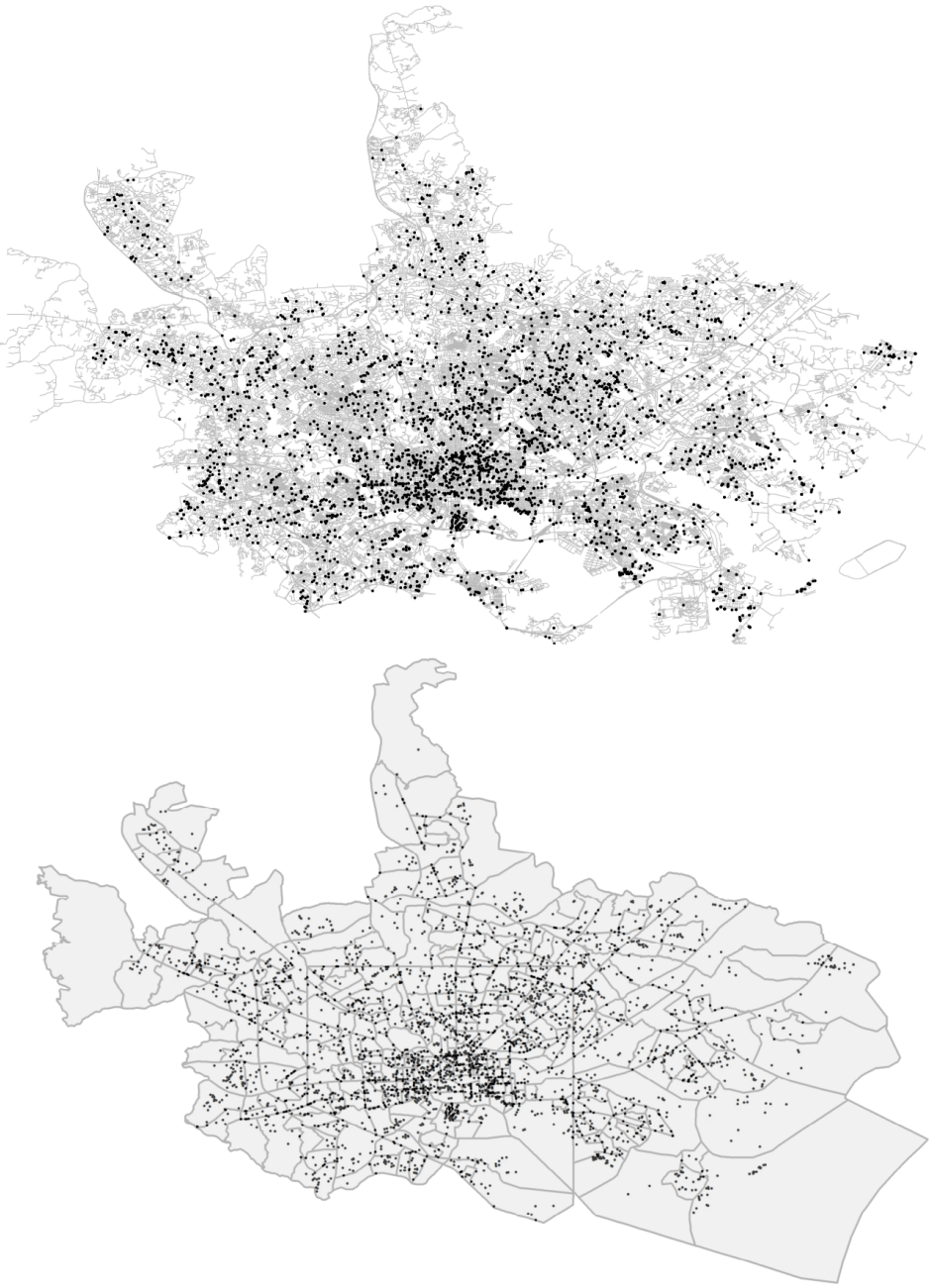


Figure 5.1: Location of pipe break records overlaid with road and census tract layers.

5.3.2 Binary Classification Models

We choose 7 different types of classification models to determine the likelihood of breaks, which is expressed both at the road segment and at the census tract resolution. Below is a summary of them. A full reference for all methods can be found in Hastie et al. (2009) [90].

1. Generalized Linear Model: A linear combination of the explanatory variables, fitted to model the log odds ratio, which is the log ratio between the probability of experiencing at least one break and the probability of no breaks.

$$\text{Log Odds Ratio} = \ln\left(\frac{\text{Pr(BREAK = 1)}}{\text{1 - Pr(BREAK = 1)}}\right) \tag{5.1}$$

   This ratio is a measure of association between a particular set of variables and the likelihood of the positive response 'at least one break' occurring. A larger log odds ratio is associated with a greater likelihood of the response.

2. The Generalized Additive Model: A linear combination of smoothed functions for each explanatory variable, fitted to model the log odds ratio.

3. The Classification Tree: A recursive partitioning technique that iteratively partitions data. Each split is selected to minimize impurity in the resulting subsamples, which is a measure of the inhomogeneity of data points from different classes within a region.

4. The Random Forest: An ensemble of Classification Trees trained with bootstrapped replications of the original data with a randomized subset of explanatory variables used for each splitting node. The Classification Trees grouped together are approximately uncorrelated conditioned on the original data and can reduce variance.

5. The Boosted Trees: An ensemble of Classification Trees trained sequentially, with each tree capturing the error of the set trained before it, thus reducing bias.

6. The Mixed Effects Model: Similar to a Generalized Linear Model in structure, but includes random Gaussian variables to control for missing variables [85], trained using maximum likelihood estimation. Two structures of Mixed Effects Models are used: one that assumes a uniquely parameterized normal variable for every observation (random intercept), and another that groups observations based on time and location by assuming the normal variable for rows from the same month and same road segment/census tract have the same mean and variance (random slopes).

Since we are trying to estimate pipe failures without any pipe information, the motivation for implementing a mixed effects model is to account for model biases which may arise [198]. Mixed effects models introduce hierarchy in the data where a given variable will have a different affect on the outcome based on different possible groupings. If the effects between groups are drastically different it may indicate that there are important features missing. We note that the most intuitive grouping would involve pipe level features (material, age, size) which are not available, but we still explore how random effects modeling techniques can be applied in this setting. It can benefit a utility to invest in the collection of pipe feature data since it allows for the use of multilevel models to explore group-level effects.

For both of the Mixed Effects Models structures, we train a model using all variables listed in Table 5.1 and another with a feature selection step to remove highly correlated variables. We estimate the variable inflation factor (VIF) for each feature to check for multicollinearity and remove those with VIF's greater than 10, we refer to this as before feature selection.

For all other model structures besides the Mixed Effects, 4 distinct models are built. One using all the explanatory variables listed in Table 5.1 and another which includes after feature selection, then another pair with and without after feature selection; only first we perform before feature selection and remove features with high multicollinearity (VIF $\geq$ 10) beforehand. The multicollinearity check is done to estimate model performance when only linearly independent features are included. After feature selection is implemented because not all variables may have a statistically significant dependence to the response, and including them can lead to model overfitting. These insignificant variables can differ depending on the structural form of the model (e.g. linear or non-linear), and identifying these variables and removing them can result in models with better out-of-sample predictive accuracy.

For linear models (Generalized Linear model, Generalized Additive model), the stepwise feature selection method [105] is used to iteratively remove and add variables based on Akaike Information Criterion until all the remaining ones have statistically significant relationship to the response at the 0.05 level. Finally, for the Random Forest and Boosted Trees method, the importance of each variable is evaluated by training the model with and without it. The change in in-sample accuracy with the variable excluded is calculated and variables are ranked according to the magnitude of that change. The top 10 most important variables (ones that lead to the largest decrease in accuracy when omitted) are kept and the second model is built off of them.

5.3.3 Predictive Performance Metrics

One common criterion to evaluate binary classification models is overall accuracy, calculated as the fraction of total observations that are correctly classified. However, since the negative responses

(no failures) in both datasets outnumber the positive ones, using overall accuracy would skew the measure heavily towards the negative responses. Thus overall accuracy is not used as an evaluation metric as it would provide misleading information about the model performance. Furthermore, utility managers would gain more insight from the analysis if the accuracy of the positive and negative responses are evaluated separately. Thus, the following measures are used.

1. Brier Score (BS): A proper score function (minimized for a perfect model) for binary classification to measure the accuracy of probabilistic predictions with the following formulation.

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2 \tag{5.2}$$

   Where N is the total number of observations, O_i is the i'th event outcome (0 or 1 in a binary setting), and P_i is the probability that the i'th observation is positive ('at least one break' in this study).

2. True Positive Rate (TPR): The fraction of observations that fall in the positive class (at least one break) which are correctly classified by the model.

$$\text{TPR} = \frac{\text{Correctly classified observations with BREAK = 1}}{\text{Total number of observations with BREAK = 1}} \tag{5.3}$$

3. False Positive Rate (FPR): The fraction of observations that fall in the negative class (no breaks) which are incorrectly classified by the model.

$$\text{FPR} = \frac{\text{Incorrectly classified observations with BREAK = 0}}{\text{Total number of observations with BREAK = 0}} \tag{5.4}$$

4. Positive Predicted Value (PPV): The fraction of observations that are predicted positive class by the model which are correctly identified.

$$\text{PPV} = \frac{\text{Correctly classified observations with BREAK = 1}}{\text{Total number of observations predicted with BREAK = 1}} \tag{5.5}$$

5. Negative Predicted Value (NPV): The fraction of observations that are predicted negative class by the model that are correctly identified.

$$\text{NPV} = \frac{\text{Correctly classified observations with BREAK = 0}}{\text{Total number of observations predicted with BREAK = 0}} \tag{5.6}$$

5.4 Random Holdout Results and Discussion

For both the datasets at the road segment level and at the census tract level, the predictive accuracy of each model was tested using holdout cross-validation repeated 100 times. To ensure there are enough positive instances to build the models, in each trial we divide the full dataset into 2 according to the response and combine a randomly selected 75% of the positive class and negative class observations to form the training data, the remaining instances are left as the validation data. These models are built using the training data and tested on the unbalanced validation set by taking this unseen data and making predictions on their outcome with each of the respective methods. Predictive accuracy can be measured by comparing the predictions against the actual response: BREAK = 0 or BREAK = 1. Each of the performance measures listed in the "Predictive Performance Measures" section is recorded over the 100 trails.

Since the focus of the research is on the predictive accuracy of the methods, information related to goodness of in-sample fit is not presented (e.g. variable statistical significance, R^2). While feature selection involves in-sample fit by using variable significance, it is mainly done to avoid overfitting against the training data in order to improve the out-of-sample accuracy.

To balance the training data, random over- and under-sampling was performed in each holdout to boost the occurrence of positive responses to 50% (from 0.2% in the original road level dataset, and from 14.6% in the census tract level dataset). This was done by splitting the dataset into the two corresponding classes, applying random sampling without replacement to the majority classes to reduce their prevalence, and applying random sampling with replacement to the minority class to increase its prevalence. In the remainder of this section, the results obtained from the random cross-validation trials between the two datasets are presented and discussed.

Table 5.2 and 5.3 show the mean and standard deviation for each classification metric over the 100 holdouts in the road level and census tract level datasets respectively. They summarize the results from all the model forms, including 1) models trained using all available features, 2) models with features selected through feature selection, 3) models trained with all linearly independent features, and 4) models trained with features selected through feature selection starting from linearly independent features.

Table 5.2: Performance summary of binary classification models in the road level, sample mean and standard deviation reported over 100 trails. [a]Models where feature selection was used.

Model	BS	TPR	FPR	PPV	NPV
GLM	$1.76\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$7.35\text{x}10^{-1}$ ($7\text{x}10^{-3}$)	$4.22\text{x}10^{-1}$ ($3\text{x}10^{-3}$)	$6.76\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($3\text{x}10^{-5}$)
	$1.90\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$7.07\text{x}10^{-1}$ ($9\text{x}10^{-3}$)	$4.67\text{x}10^{-1}$ ($4\text{x}10^{-3}$)	$6.40\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
GLM[a]	$1.76\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$7.3\text{x}10^{-1}$ ($1\text{x}10^{-2}$)	$4.2\text{x}10^{-1}$ ($1\text{x}10^{-2}$)	$6.87\text{x}10^{-3}$ ($4\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
	$1.90\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$7.0\text{x}10^{-1}$ ($1\text{x}10^{-2}$)	$4.6\text{x}10^{-1}$ ($1\text{x}10^{-2}$)	$6.53\text{x}10^{-3}$ ($4\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
GAM	$1.70\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$7.45\text{x}10^{-1}$ ($7\text{x}10^{-3}$)	$4.18\text{x}10^{-1}$ ($3\text{x}10^{-3}$)	$6.94\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
	$1.81\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$7.20\text{x}10^{-1}$ ($8\text{x}10^{-3}$)	$4.51\text{x}10^{-1}$ ($3\text{x}10^{-3}$)	$6.74\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
GAM[a]	$1.83\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$7.08\text{x}10^{-1}$ ($7\text{x}10^{-3}$)	$4.35\text{x}10^{-1}$ ($2\text{x}10^{-3}$)	$7.00\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
	$1.82\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$7.08\text{x}10^{-1}$ ($8\text{x}10^{-3}$)	$4.36\text{x}10^{-1}$ ($2\text{x}10^{-3}$)	$6.93\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
CART	$1.86\text{x}10^{-1}$ ($2\text{x}10^{-3}$)	$6.0\text{x}10^{-1}$ ($2\text{x}10^{-2}$)	$3.8\text{x}10^{-1}$ ($2\text{x}10^{-2}$)	$8.50\text{x}10^{-3}$ ($5\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
	$1.94\text{x}10^{-1}$ ($3\text{x}10^{-3}$)	$5.3\text{x}10^{-1}$ ($5\text{x}10^{-2}$)	$3.5\text{x}10^{-1}$ ($6\text{x}10^{-2}$)	$9.42\text{x}10^{-3}$ ($2\text{x}10^{-3}$)	$9.99\text{x}10^{-1}$ ($1\text{x}10^{-5}$)
RF	$6.09\text{x}10^{-2}$ ($8\text{x}10^{-3}$)	$4.4\text{x}10^{-1}$ ($1\text{x}10^{-2}$)	$1.37\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$9.35\text{x}10^{-3}$ ($4\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
	$7.44\text{x}10^{-2}$ ($1\text{x}10^{-3}$)	$4.3\text{x}10^{-1}$ ($1\text{x}10^{-2}$)	$1.59\text{x}10^{-1}$ ($2\text{x}10^{-3}$)	$7.49\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
RF[a]	$1.4\text{x}10^{-1}$ ($3\text{x}10^{-2}$)	$6.3\text{x}10^{-1}$ ($6\text{x}10^{-2}$)	$2.4\text{x}10^{-1}$ ($2\text{x}10^{-2}$)	$7.11\text{x}10^{-3}$ ($1\text{x}10^{-3}$)	$9.99\text{x}10^{-1}$ ($1\text{x}10^{-5}$)
	$1.3\text{x}10^{-1}$ ($2\text{x}10^{-2}$)	$4.8\text{x}10^{-1}$ ($4\text{x}10^{-2}$)	$2.1\text{x}10^{-1}$ ($2\text{x}10^{-2}$)	$5.20\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($7\text{x}10^{-5}$)
GBM	$1.63\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$7.3\text{x}10^{-1}$ ($1\text{x}10^{-2}$)	$3.8\text{x}10^{-1}$ ($1\text{x}10^{-2}$)	$8.00\text{x}10^{-3}$ ($4\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($3\text{x}10^{-5}$)
	$1.72\text{x}10^{1}$ ($1\text{x}10^{-3}$)	$7.1\text{x}10^{-1}$ ($1\text{x}10^{-2}$)	$3.9\text{x}10^{-1}$ ($2\text{x}10^{-2}$)	$7.77\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
GBM[a]	$1.67\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$6.8\text{x}10^{-1}$ ($1\text{x}10^{-2}$)	$3.4\text{x}10^{-1}$ ($1\text{x}10^{-2}$)	$9.56\text{x}10^{-3}$ ($6\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
	$1.74\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$6.4\text{x}10^{-1}$ ($2\text{x}10^{-2}$)	$3.3\text{x}10^{-1}$ ($2\text{x}10^{-2}$)	$9.75\text{x}10^{-3}$ ($7\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
GLMM1	$1.76\text{x}10^{-1}$ ($1\text{x}10^{-3}$)	$7.35\text{x}10^{-1}$ ($7\text{x}10^{-3}$)	$4.22\text{x}10^{-1}$ ($3\text{x}10^{-3}$)	$6.76\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($3\text{x}10^{-5}$)
	$1.90\text{x}10^{1}$ ($1\text{x}10^{-3}$)	$7.07\text{x}10^{-1}$ ($9\text{x}10^{-3}$)	$4.67\text{x}10^{-1}$ ($4\text{x}10^{-3}$)	$6.40\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($4\text{x}10^{-5}$)
GLMM2	$4.16\text{x}10^{-2}$ ($8\text{x}10^{-4}$)	$8.31\text{x}10^{-1}$ ($8\text{x}10^{-3}$)	$4.93\text{x}10^{-1}$ ($9\text{x}10^{-4}$)	$6.47\text{x}10^{-3}$ ($3\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($3\text{x}10^{-5}$)
	$3.98\text{x}10^{-2}$ ($6\text{x}10^{-3}$)	$8.13\text{x}10^{-1}$ ($8\text{x}10^{-3}$)	$5.17\text{x}10^{-1}$ ($2\text{x}10^{-3}$)	$6.22\text{x}10^{-3}$ ($2\text{x}10^{-4}$)	$9.99\text{x}10^{-1}$ ($3\text{x}10^{-5}$)

Table 5.3: Performance summary of binary classification models in the census tract level, sample mean and standard deviation reported over 100 trails. [a]Models where feature selection was used.

Model	BS	TPR	FPR	PPV	NPV
GLM	0.237 (0.002)	0.600 (0.008)	0.483 (0.002)	0.203 (0.007)	0.892 (0.005)
	0.237 (0.001)	0.600 (0.009)	0.482 (0.002)	0.203 (0.007)	0.892 (0.006)
GLM[a]	0.237 (0.002)	0.600 (0.009)	0.483 (0.002)	0.203 (0.007)	0.892 (0.005)
	0.237 (0.001)	0.600 (0.009)	0.482 (0.002)	0.202 (0.007)	0.892 (0.006)
GAM	0.227 (0.002)	0.637 (0.009)	0.477 (0.002)	0.223 (0.008)	0.910 (0.004)
	0.227 (0.002)	0.637 (0.008)	0.477 (0.002)	0.223 (0.008)	0.910 (0.004)
GAM[a]	0.228 (0.002)	0.637 (0.008)	0.477 (0.002)	0.221 (0.008)	0.910 (0.004)
	0.228 (0.002)	0.634 (0.009)	0.477 (0.002)	0.221 (0.008)	0.910 (0.004)
CART	0.239 (0.002)	0.63 (0.4)	0.57 (0.3)	0.176 (0.007)	0.914 (0.004)
	0.239 (0.002)	0.63 (0.4)	0.57 (0.3)	0.175 (0.007)	0.914 (0.004)
RF	0.157 (0.003)	0.428 (0.009)	0.306 (0.005)	0.25 (0.01)	0.889 (0.004)
	0.156 (0.003)	0.427 (0.009)	0.306 (0.005)	0.26 (0.02)	0.890 (0.004)
RF[a]	0.145 (0.002)	0.42 (0.01)	0.300 (0.005)	0.27 (0.02)	0.892 (0.004)
	0.145 (0.002)	0.425 (0.009)	0.301 (0.005)	0.27 (0.02)	0.892 (0.004)
GBM	0.218 (0.002)	0.658 (0.008)	0.471 (0.002)	0.236 (0.008)	0.920 (0.004)
	0.218 (0.002)	0.657 (0.009)	0.472 (0.002)	0.236 (0.008)	0.920 (0.004)
GBM[a]	0.219 (0.002)	0.654 (0.008)	0.472 (0.002)	0.234 (0.008)	0.919 (0.004)
	0.219 (0.002)	0.654 (0.009)	0.472 (0.002)	0.234 (0.008)	0.919 (0.004)
GLMM1	0.237 (0.002)	0.600 (0.009)	0.483 (0.002)	0.203 (0.007)	0.892 (0.004)
	0.237 (0.001)	0.600 (0.009)	0.483 (0.002)	0.202 (0.007)	0.892 (0.006)
GLMM2	0.216 (0.003)	0.663 (0.008)	0.472 (0.002)	0.239 (0.009)	0.922 (0.004)
	0.216 (0.003)	0.664 (0.008)	0.472 (0.002)	0.239 (0.008)	0.922 (0.004)

For each model, first 2 values are results without Multi-collinear Feature Removal: Mean over 100 trials (standard deviation); second 2 values are results with Multi-collinear Feature Removal: Mean over 100 trials (standard deviation). Model abbreviations: GLM: Generalized Linear Models. GAM: Generalized Additive Models. CART: Classification Tree. RF: Random Forest. GBM: Boosted Trees. GLMM1: Mixed Effects Models with Random Intercepts. GLMM2: Mixed Effects Models with Random Slopes.

We observe that in Table 5.3, the use of only linearly independent features had nearly no effect on predictive performance. This is evidenced by the fact that the average accuracy metrics (and the standard deviation) were identical whether the feature removal step was used or not. In contrast, removing multi-collinear features in the road level data appears to worsen performance in the road level results in Table 5.3. With the variable removal step applied the following trend is seen across

most model structures: True Positive Rates are lower, False Positive Rates are higher, and Positive Predicted Values are lower. The rest of this section will only target the results when all available features (including those linearly dependent) are used.

In the road level results, every model has near perfect negative predictive values, meaning that when the models assign a negative prediction to an observation, it is right 99.9% of the time. Whereas in the census tract level, while all models still perform well according to this measure (between 89% and 92%), there is a drop off in accuracy as spatial resolution increases and there is a better balance between the 2 classes. The reason why all methods perform well in this regard is intuitive: since the testing dataset is heavily skewed with negative outcomes (99.8% in the road data, 85.4% in the tract data), these observations are easy to predict correctly. This suggests that a more meaningful comparison between the models involves metrics related to the accuracy of the positive responses (BREAK = 1), and comparing multiple metrics in a holistic fashion.

From the road level results in Table 5.2 all model structures except the Random Forest without feature selection have an average False Positive Rate above 20%, while in the tract level results in Table 5.3 all model structures have an average above 30% False Positive Rates. Since the positive responses occur infrequently (0.2% and 14.6% respectively), this indicates that almost all the models are substantially over-predicting failures. As a result of over-prediction, the rate at which a positive prediction is correct, namely the positive predictive value, is expected to be very low. While all the models have a Positive Predicted Value less than 1% in the road level results, the Boosted Trees and Random Forest performed best in this aspect (0.956% and 0.935% respectively). These measures are much higher in the census tract level results; all models have Positive Predicted Values between 17.6% and 27%. The Random Forest with feature selection was the best in this aspect.

As the spatial resolution between the datasets is increased, there is a better balance between the two classes. These results point to an interesting finding: while a better balance can lead to some loss of accuracy in predicting the "no failure" responses, there is a large improvement for predicting the "failure" cases, as evidenced by the significant difference in Positive Predicted Values between the two tables.

The Brier Score, measuring the mean square error of the predicted likelihoods of failure, indicates that there are 3 highly accurate models in the road level results: The Random Forest using all covariates, and the two Mixed Effects Models. The average Brier Scores from these methods are smaller than the others by at least an order of 10. We note that the Random Forest model also had the lowest average False Positive Rate, while the Mixed Effects Models had the lowest average Positive Predicted Values. This shows there is agreement between the different metrics when identifying accurate models.

In contrast, the Brier Score values from the census tract results do not show a single model

outperforming the others by such a large margin. The two Random Forest models are the best, achieving an average Brier Score approximately 35% lower than the others. Again, there is agreement across different measures as the Random Forests also had the best False Positive Rate and Positive Predicted Value results.

The True Positive Rates equal the proportion of observations that are positive responses which are correctly identified. When observing the road level results from the linear models, Boosted Trees, and Mixed Effects Models, they all seem to perform well (all above 70%). However, this statistic is misleading due to over-prediction of positive responses. For example, the Mixed Effects Model with random slopes has the highest True Positive Rate at 83.1%, however the False Positive Rate of 49.3% indicates that this model is assigning roughly half of the observations as positive predictions. Similarly, the Mixed Effects Model with random slopes in the census tract results had the highest True Positive Rates (66.3%), but its False Positive Rate is also comparably high (47.2%). Given the low occurrence rate of the positive class, it is likely a large portion of them will be identified by simply assigning positive predictions to a large number of observations. As a result, while the True Positive Rate provides some information on a models ability to identify correctly the positive instances, it must be considered alongside other metrics for correct interpretation.

As evidenced by the similar performance measures with and without feature selection, it is found that the predictive performance of the Generalized Linear model, Generalized Additive model, and the Classification Tree is not very sensitive to feature selection. The results from the road level dataset suggest that using feature selection on the Random Forest leads to a model that is more likely to over-predict. The increase in the Brier Score by almost 4 times indicates that there is more bias. Over-prediction is evidenced by the increased False Positive Rate, lower positive predictive value, and increased True Positive Rate. On the other hand, when applied to the census tract level dataset, feature selection for the Random Forest appears to marginally improve performance on average but leads to higher variance especially for the road level data.

The performance of the Boosted Trees in the census tract level data did not change significantly with feature selection but did lead to improved performance in the road level results. While the Brier Score stays roughly the same, the False Positive Rate decreases and the Positive Predicted Value increases, suggesting that the new model is less likely to over-predict and positive predictions are more likely to be correct. However, the tradeoff for lowering the rate of over-prediction is that fewer true positive outcomes are identified, as evidenced by the lower True Positive Rate.

In the road level results, the Mixed Effects Models appear to over-predict the most amongst all the models. The highest False Positive Rates, True Positive Rates, and the lowest Positive Predicted Values all belong to the two models. Random slopes and intercepts were used as controlling parameters to account for omitted variable bias, but the training process of these models is meant for improving causal inference of the in-sample data. While the Brier Scores of the models are

the lowest by more than a magnitude of 10, this can be a misleading indication of good predictive performance when considering the other metrics. On the other hand, the Mixed Effects Models performed much better in the census tract level results. In particular, the random slopes model performed relatively well based on these metrics: average True Positive Rate, Positive Predicted Value, and Negative Predicted Value. However, both random slope and intercept models do have high False Positive Rates as well, indicating over-prediction.

As highlighted in the introduction section, while we do not discuss variable statistical significance as it pertains to out-of-sample accuracy; we can use it to provide insight on variables that provide the greatest explanatory power. As part of an extended analysis we use both the road level and census tract level dataset to train a Generalized Linear model. Note that we first balanced the response distribution through resampling and perform before feature selection to remove multicollinear features. The summary of variable statistical significance is summarized in Table 5.4 and Table 5.5, which shows both the coefficient value as well as its p-value from the logistic regression, lower p-values indicate that a variable is statistically significant.

Table 5.4: Generalized Linear Model Feature Statistical Significance for Road Level Data.

Covariate	Regression coefficient	p-value
BREAK_$HISTORY$	2.707273	0.000000
LAND_$USE(WATER)$	1.182370	0.000000
AA_$PERCENT$	0.581550	0.000000
LAND_$USE(Developed)$	0.207457	0.001545
SOIL_$CLASS(Inceptisols)$	0.203086	0.000015
FREEZE	0.047318	0.000000
T_$STDEV$	0.013281	0.017823
TMIN	0.008200	0.000000
CLAY_$PERCENT$	0.007532	0.000000
AGE_$LESS$_1940	0.001044	0.000000
AGE_1960	0.000732	0.000000
AGE_1970	0.000266	0.000007
AGE_1980	0.000084	0.081396
MEDIAN_$INCOME$	0.000001	0.087193
AGE_2005	-0.000139	0.299151
AGE_1950	-0.000141	0.000060
ASIAN_$TOTAL$	-0.000185	0.263724
AGE_1940	-0.000368	0.000000
AA_$TOTAL$	-0.000395	0.000000
AGE_1990	-0.000752	0.000000
AGE_2000	-0.001281	0.000000
BOTTOM_$DEPTH$	-0.001440	0.000000
SOIL_$CLASS(Entisols)$	-0.069317	0.008203
SOIL_$CLASS(Alfisols)$	-0.127571	0.000002
SOIL_$CLASS(Ultisols)$	-0.177209	0.000000
PRCP	-0.472325	0.000001
(Intercept)	-1.377525	0.000000
ASIAN_$PERCENT$	-2.152039	0.000000

Table 5.5: Generalized Linear Model Feature Statistical Significance for Census Tract Level Data.

Covariate	Regression coefficient	p-value
ASIAN_$PERCENT$	1.439946	0.082844
LAND_$USE(INDUSTIAL)$	0.874150	0.000000
PRCP	0.749920	0.000000
LAND_$USE(ResourceConservation)$	0.565614	0.000006
LAND_$USE(MixedUse)$	0.564866	0.000016
LAND_$USE(Residential)$	0.415841	0.000155
AA_$PERCENT$	0.331253	0.001341
BREAK_$HISTORY$	0.188556	0.000000
T_$STDEV$	0.047920	0.000000
CLAY_$TOTAL$	0.001112	0.749617
AGE_2005	0.000895	0.000089
AGE_1960	0.000751	0.000000
AGE_$LESS$_1940	0.000471	0.000000
AGE_1980	0.000340	0.000035
AGE_1970	0.000308	0.001955
AGE_1940	0.000261	0.002288
AGE_1950	0.000252	0.000033
MEDIAN_$INCOME$	0.000189	0.089160
AGE_1990	0.000185	0.028243
AGE_2000	0.000039	0.764792
AA_$TOTAL$	-0.000331	0.000000
ASIAN_$TOTAL$	-0.001780	0.000105
TMIN	-0.003170	0.078820
FREEZE	-0.009900	0.144116
BOTTOM_$DEPTH$	-0.009900	0.000026
(Intercept)	-1.545253	0.000000

For both logistic regressions, before feature selection was implemented to remove highly correlated variables. Any feature with a variable inflation factor of larger than 10 was removed. Between the logistic regressions using both datasets, the historical failure count and the number of old households (AGE_$LESS$_1940) are identified as one of the most important features. This indicates that historical failure activity and pipe age (older houses typically are connected to older pipes) are potential drivers of failure.

After that there is variability as to which variables are significant. For the road level data, demographic information such as the percentage of Asian American or African American house-

holds are highly significant, while in the census tract level data land use is much more important. Environmental conditions pertaining to soil conditions are also more significant in the road level data. It is possible that operating at a finer spatial scale allows the relationships between the soil data and the road data to be more important in the model.

Beyond using the statistical models only to forecast regions with high breakage likelihood, identifying which features are significant can further help utilities by pointing out the effects of each individual variable on breakage likelihood. In both Table 5.4 and 5.5, the coefficients for percentage of break history is positive, indicating that a high number of historical breaks are associated with higher breakage likelihoods. Similarly, the coefficient for the number of older households is also positive but much smaller in magnitude, suggesting while there is a positive correlation between the variables the effect of this variable on the response is weaker.

We acknowledge that our analyses of feature importance are confined strictly to the linear logistic regression realm, it is possible there are significant nonlinear relationships in the data which are not detected and can be left for exploration in future work.

In summary, we show through the random holdout results that predictive accuracy suffers when faced with the lack of system specific data. There is also a tradeoff between spatial precision and model accuracy. When analyzing the pipe break records at the road level, the imbalance between the positive and negative classes causes all the models to over-predict. While better accuracy can be achieved by aggregating to larger geographic regions such as census tracts (average 0.9 square mile in area), these classifiers become less useful for guiding asset management because they cannot point to where specifically in a census tract a pipe break will occur. This indicates that in order for utilities to achieve correct forecasts on the future condition of their system, they must collect accurate and precise data on the current system at a fine spatial scale.

5.5 Temporal Holdout Results and Discussion

In the previous section, we acknowledge that predictive accuracy is limited in the absence of system specific data. Here we aim to test whether statistical models built without system data, despite their limited accuracy, can still provide useful insight to guide asset management. Utilities without system level data often rely solely on historical break records to plan their maintenance activities. Regions or pipe segments are sorted based on how often they previously experienced failures, and the ones with the most failures are inspected first. Our goal here is to determine if a sorting based on the predictions of a statistical model can achieve better accuracy than one based on historical failure rates. Temporal holdouts are used to compare the models against the history-based sorting. This section will present the evaluation criteria, show the results of this analysis, and discuss them.

As discussed briefly in the introduction, in each trial the data is divided into two sets: the

training set which consists of 51 of the 52 months worth of data, and the remainder month being the validation set. We repeat this holdout process iteratively for all 52 months. Over- and under-sampling is again used to boost the occurrence of the positive class (breaks) to 50% in both the road and census tract level training data. Each model is built using this balanced training set and predictions are made on the unbalanced validation data. Similar to the previous section, we implement 4 different models for each unique structure: 1) using all features, 2) using after feature selection, 3) using only linearly independent features (before feature selection), 4) using variable selection on linearly independent features (both before and after feature selection).

Each model is a binary classifier that provides the predicted likelihood of failure, which we can use to sort our observations (highest to lowest) in the validation set. Since each model produces a different set of predictions, each resulting ranking will also be different. For example, the Boosted Trees ranking will sort the validation set highest to lowest based on its predictions for probability of break, while the Random Forest ranking will do the same based on its different set of predictions.

We also use the break history, which is included in the analysis as an explanatory variable, to sort the observations in the validation data. Here the roads or census tracts with the most historical failures are ranked higher. This is what the utilities typically do to decide where to perform inspection and maintenance.

Since the true response is known, we can compare in a pairwise fashion the accuracy of a statistical models ranking against the historically based rank at different cutoff points. For example, if we set the cutoff to 50 observations and wish to analyze the Classification Tree, we will compare the top 50 observations from 1) the ranking based off the Classification Tree predictions and 2) the ranking based off the historical failure rates. Out of these two sets of 50 observations, if the Classification Tree rank has more or equal to the number of breaks as found in the historical rank, we can conclude that it has achieved a non-dominated sorting. This indicates that the Classification Tree is at least as good, if not better, as the history driven ranking.

We perform this pairwise comparison between every model and the history-based rank for each of the 52 temporal holdout trials at the following defined cutoffs: 10, 20, 30, 40, 50. While there are many more roads and census tracts than 50 in the datasets, we choose these cutoffs because they reflect a realistic application of these models. Utilities have limited resources to spend on maintenance, so in a given time frame they can only focus on the top-ranked regions or roads for management. In Tables 5.6 and 5.7, we report the fraction of trials in which the model achieved a non-dominated ranking of observations and the best performing models at each cutoff. There does not seem to be a noticeable effect of feature selection on the prioritization accuracy of the models. In some cases models incorporating feature selection leads to better a sorting of the observations while in others it does not.

Table 5.6: Temporal holdout results for the road level data. [a]Models in which feature selection is used.

Model	Top 10	Top 20	Top 30	Top 40	Top 50
GLM	0.827, 0.855	0.673, 0.808	0.692, 0.808	0.788, 0.788	0.788, 0.827
GLM[a]	0.827, 0.855	0.673, 0.808	0.692, 0.808	0.788, 0.827	0.788, 0.827
GAM	0.827, 0.577	0.673, 0.615	0.673, 0.538	0.788, 0.558	0.788, 0.577
GAM[a]	0.827, 0.538	0.712, 0.596	0.673, 0.538	0.750, 0.538	0.808, 0.538
CART	0.442, 0.442	0.346, 0.346	0.308, 0.308	0.269, 0.269	0.308, 0.308
RF	0.538, 0.500	0.423, 0.346	0.423, 0.327	0.346, 0.231	0.346, 0.192
RF[a]	0.346, 0.423	0.288, 0.250	0.231, 0.192	0.154, 0.154	0.115, 0.192
GBM	0.462, 0.442	0.327, 0.327	0.250, 0.327	0.346, 0.288	0.346, 0.346
GBM[a]	0.462, 0.481	0.308, 0.385	0.308, 0.308	0.269, 0.327	0.211, 0.404
GLMM1	0.769, 0.865	0.673, 0.788	0.673, 0.808	0.731, 0.769	0.750, 0.846
GLMM2	0.596, 0.538	0.577, 0.558	0.519, 0.500	0.462, 0.442	0.442, 0.442
Best Model	GLM	GLM	GLM/GLMM1	GLM/GAM	GLM

Table 5.7: Temporal holdout results for the census tract level data. [a]Models in which feature selection is used.

Model	Top 10	Top 20	Top 30	Top 40	Top 50
GLM	0.634, 0.577	0.808, 0.769	0.731, 0.692	0.558, 0.596	0.635, 0.634
GLM[a]	0.654, 0.615	0.769, 0.750	0.692, 0.731	0.577, 0.596	0.635, 0.635
GAM	0.538, 0.654	0.788, 0.673	0.712, 0.673	0.712, 0.750	0.692, 0.712
GAM[a]	0.615, 0.462	0.788, 0.692	0.654, 0.673	0.712, 0.673	0.712, 0.635
CART	0.385, 0.385	0.327, 0.308	0.288, 0.269	0.250, 0.250	0.308, 0.288
RF	0.481, 0.519	0.558, 0.442	0.462, 0.327	0.442, 0.327	0.519, 0.404
RF[a]	0.500, 0.500	0.385, 0.442	0.403, 0.385	0.346, 0.288	0.403, 0.385
GBM	0.558, 0.634	0.750, 0.769	0.731, 0.750	0.731, 0.712	0.808, 0.692
GBM[a]	0.653, 0.673	0.692, 0.635	0.692, 0.654	0.692, 0.692	0.693, 0.731
GLMM1	0.615, 0.654	0.731, 0.750	0.731, 0.654	0.538, 0.519	0.635, 0.596
GLMM2	0.673, 0.538	0.712, 0.692	0.731, 0.673	0.692, 0.596	0.712, 0.635
Best Model	GBM[a]	GLM	GBM	GBM	GBM

For each model, first value is without Multicollinear Feature Removal, second value is with Multicollinear Feature Removal. Model abbreviations: GLM: Generalized Linear Models. GAM: Generalized Additive Models. CART: Classification Tree. RF: Random Forest. GBM: Boosted Trees. GLMM1: Mixed Effects Models with Random Intercepts. GLMM2: Mixed Effects Models with Random Slopes.

Table 5.6 reports the relative performance of the statistical learning based model sorting against the history-based sorting when analyzed at the road segment level. The Generalized Linear model using only linearly independent features and the Generalized Additive models using all available features are consistently two of the best performing models across all cutoff points. The Mixed Effects model with random intercepts also performed comparably well. At the cutoff of 10 observations, the 2 models produced a non-dominated ranking relative to the history-based method in 88.5% of the trials. Similarly, in cutoffs of 20, 40 and 50, the model was non-dominated in more than 78% of the trials. At the cutoff 30, the Mixed Intercepts Models (random slopes) performed equally well as the linear and additive models. These high percentages indicate that the statistical models are more likely to prioritize high probability of failure roads effectively.

Table 5.7 reports the relative performance of the statistical learning based models against the history-based sorting when analyzed at the census tract spatial resolution. Unlike the results in the road level case, the Boosted Trees was the best performing model in this instance. Between the cutoffs of 10, the Boosted Trees had above 65% of trials where it produced a non-dominated ranking compared to the historical-based rank. Between cutoff points 30 to 50, the Boosted Trees was also the best model, with at least 73% of the trails performing at least as good as the history rank. This suggests that when using statistical models for prioritization at the census tract level, it could be beneficial to use multiple methods together for better accuracy.

Outside of the cutoff point 10, the best models perform at least as well as the history-based rank in a high fraction of the holdout trials, all above 71%. Indicating that while a history-based rank can sort a small number of very high-risk regions well, it becomes less accurate as it has to handle more observations and that using statistical learning based models can guide decision making better.

We demonstrate through our results in Table 5.6 and Table 5.7 that despite the limited classification accuracy of the statistical methods, an accurate prioritization of high-risk assets can be achieved based off the predicted probabilities of failure. This is the case in both the road and census tract level analysis when compared against the common practice of sorting by historical failure rates. In the road level results, we find that the Generalized Linear and Generalized Additive models achieve the most accurate sorting. While in the census tract level results we find that while a history based rank can initially be more useful, the Generalized Additive models and Boosted Trees combined will eventually outperform it. This indicates that these statistical models can be useful for utilities which lack readily available data, and can serve as a viable alternative to their current methods.

Our conclusions are based off empirical findings that are specific to the datasets we use, but the goal of this chapter was to demonstrate the utility of statistical models when faced with the data challenge commonly faced by water utilities. While our results show that there is benefit in

adopting statistical models for prediction, we do acknowledge that the effectiveness of the models can be improved with data specific to the distribution system and utilities should invest in the collection of them.

5.6 Conclusion

In this research, the challenge for water utilities in having to manage their distribution network without any readily available pipe level information is addressed. We examine whether accurate and useful predictive models can be built in the complete absence of pipe-feature data. It is found that when evaluating the models in a binary classification context, the predictive accuracy is low, and the model suffer from the lack of asset level information. Future directions for this study are to incorporate more advanced modeling techniques to handle class imbalance, which can improve accuracy.

Despite the limited accuracy of the models, models without detailed pipe-level data are able to better prioritize the high-risk assets when compared to a historical failure rate based ranking, which is a prioritization method commonly adopted in practice. Hence, these models can be useful to aid inspection and maintenance planning.

Acknowledgements

The authors would like to thank the University of Michigan for funding this research. The opinions and views expressed are those of the researchers and do not necessarily reflect those of the sponsors.

CHAPTER 6

Prediction of Water Main Failures with the Spatial Clustering of Breaks

Due to limited budgets and an aging system, infrastructure managers have increasingly sought cost-effective means to evaluate asset condition. Better information on the physical health of the infrastructure can help achieve higher returns on investments in replacement and repair spending. This is a particular challenge for water distribution systems due to the vast amount of buried and unseen pipelines. While robotic inspections can provide high quality data, they can often be cost prohibitive. An alternative method is to perform a desktop analysis by using past performance information to estimate current pipeline conditions. A spatial clustering of pipe breaks fits well into a wider asset management framework with the aim of identifying regions with abnormally high failure rates. The information about spatial clusters identified using historical breaks, if and where they exist, can potentially improve predictions on the location of future breaks. In this research, we present three algorithms (poisson based, density based, and locally weighted density based) for scanning and clustering pipe break data and demonstrate their application on a real pipeline network. We also explore whether the use of spatial clusters as an explanatory variable can improve the accuracy of pipe break machine learning models. Empirical findings show that the locally weighted density scan provides the greatest precision for finding high breakage zones. The application of these clusters generally improves the performance of predictive models by helping them prioritize high risk pipes with greater accuracy.

Keywords: Infrastructure Resilience, Water Distribution Systems, Spatial Clustering, Statistical Modeling, Pipe Break Prediction.

Note: The research presented in this chapter is under first round review at the Journal of Reliability Engineering and System Safety, submission date on May 29, 2019. Co-authors: Seth David Guikema.

6.1 Introduction

In many parts of the United States, drinking water infrastructure is nearing the end of its useful life and upgrades are needed to ensure the consistent delivery of safe water to end users [10]. A 2017 report published by the American Society of Civil Engineers (ASCE) estimates that over one trillion US dollars of capital investments are required for necessary upgrades to the nation's water infrastructure [19]. Despite being recognized as one of the most critical infrastructures [70], leaks and breaks in the distribution system are occurring more frequently as the pipelines degrade [54]. An estimate from ASCE reports that over 240,000 water main breaks occur across the country per year [19]. This accounts for about 14% of treated water lost throughout the distribution system, often termed as non-revenue water [154]. On top of the economic burden caused by the aging infrastructure (each break costs about $42,000 [205]), researchers have also linked pipe breaks to compromised water quality and health risks [76, 184].

Spending on infrastructure repair and replacement is needed to ensure drinking water infrastructures can continue to function well [41, 178]. Because many utilities operate on a limited budget [58], having an effective asset management framework is critical. This ensures that capital investments are targeted to the most vulnerable regions and risks can appropriately mitigated. However, a common challenge that prohibits many US utilities from formulating asset managements plans is the lack of readily available system data [57, 101, 146]. This means that basic pipeline information (material, age, size) can be missing from digitized databases. As a result, many utility managers rely on their expert judgement instead to estimate pipeline condition and to plan maintenance works [154].

Spatial models can be useful for helping managers identify high risk zones in the network. They are a cheaper alternative to robotic inspections since they only require information regarding the failure locations. Spatial models also have less data requirements; utilities do not need to invest in collecting missing pipeline information [147]. The key idea is to monitor the location of failures over an extended period of time and aim to identify spatial clusters of breaks. A cluster here is defined as a contiguous collection of pipes with an anomalously high failure rate compared to regions not in clusters [69]. Having knowledge on whether clusters exist, and if so where they are located, is useful information for utility operators. It can act as a potential indicator of system distress. For utilities which lack comprehensive pipeline data these clusters can be directly used to inform capital spending directed at these high risk zones. On the other hand, utilities which have pipeline information can use this information to improve the accuracy of pipe break forecasting models [68, 69].

The first goal of this research is to explore the effectiveness of three different spatial clustering algorithms in finding high breakage zones. The second goal is to determine whether the use

of spatial clusters derived from breaks in the past can assist in predicting the location of future breaks. To the authors knowledge, no past research has jointly compared clustering approaches for grouping regions within a network and consequently explored their application in pipe break prediction models.

We partnered with a utility in the Midwest to implement our approach in a real water distribution system. We will first compare the following clustering approaches: a poisson based model, a density based approached, and a locally weighted density approach. As an extension from earlier work, we will focus on clustering the pipes on which the breaks occur rather than grouping the breaks themselves. Then we will take the best performing algorithm and test whether the inclusion of clusters as an explanatory variable can improve machine learning pipe break prediction models. The aim is to demonstrate that high precision clusters (as many breaks captured in as little of the network as possible) can provide useful information to better predict future pipeline failures.

6.2 Related Research

The related works for this research deal with: 1) spatial scan statistics and their application for analyzing critical infrastructure, and 2) the statistical modeling of water main breaks. In this section we will review the methods which were explored as well as their relevance to this chapter.

6.2.1 Pipe Break Clustering

Spatial scan and cluster methods were developed and frequently used in the analysis of epidemiological data [124, 96]. Researchers have used the statistical tools to identify regions (zip codes, counties) which are likely to be experiencing a disease outbreak. There are two major components to this process. The first involves identifying candidate regions using a search window, defined by a particular size and shape (e.g. circular window with 500m radius). The second is the choice of probability model to evaluate the number of events observed within a given space. The observed events are assumed to follow a parametric distribution (e.g. Poisson, Bernoulli) [123] and a hypothesis test is carried out to determine the statistical significance of the observations.

Past works have compared the effectiveness of various non-compact shapes (oval, rectangular, etc.) [125]. Other researchers have explored non-parametric methods to better capture the incidence of clusters [175], these methods only compare relative intensity of events and could be more flexible [121]. Case studies in the water infrastructure domain are found in Goulter and Kazemi (1998) [86] and Shi et al. (2013) [181]. The researchers used clustering methods to identify regions prone to main breaks and to analyze spatial factors unique to these areas.

These works discussed above have relied on the use of euclidean search spaces for the iden-

tification of clusters. This may not apply well when analyzing water networks since the search space is now constrained to a 2-dimension planar graph [211]. As a result, the methods above need to be adapted to scan over planar graphs when applied to networked infrastructure (power, water, gas). Shi and Janeja (2009) [180] presents an algorithm that relies on connectivity measures for efficiently scanning a linear graph for potential clusters. Yiu and Mamoulis (2004) [211] present network based formulations of partition based and hierarchical based clustering methods for road networks. A domain specific case study for water systems in presented in De Oliveira et al. (2011) [69], where a Poisson model based scan and clustering algorithm is presented. The methodology also controls for pipeline features such as age and material which may affect breakage rates. The same authors present a network implementation of the density based clustering algorithm OPTICS [16] in a related paper [68].

6.2.2 Pipe Break Prediction

Previous work on the statistical prediction of pipe breaks can be described as one of two categories: physical based and statistical based models [169]. Physical based methods aim to characterize the structural performance of the water mains subject to environmental loadings. Some examples include: modeling the thinning of steel pipe walls due to corrosion [171], and describing the mechanics of in-pipe stress due to external frost loading [118] and temperature changes [88]. A common challenge for this approach is the requirement of empirical data that is not readily available or difficult to obtain [57].

Statistical based methods can be more flexible since they can be applied with different types of input data. The main goal is to quantify structural deterioration by analyzing past performance data [117]. Two popular tools used in the literature are survival analysis and regression modeling. Survival analysis estimates the time until the next pipe break [193], and regression modeling estimates the likelihood or number of breaks in a time window (e.g. month, year) [56]. Survival analysis requires utilities have well documented records on the time and location of pipe breaks, and are less useful since many utility operators only recently started recording breaks [144]. Yamijala et al. (2009) [81] compared a range of regression models (time linear, time exponential, logistic regression) and found that none of them achieved particularly high accuracy due to the low occurrence rate of breaks. Chen et al. (2019) [57] explored the effectiveness of using only environmental data (soil conditions, weather data) to model breaks, and showed that these models can still be useful in prioritizing high risk regions.

A case study presented by Wood and Lence (2009) [203] demonstrated that predictive models built on basic pipeline attributes (age, diameter, material) can be useful for guiding pipe replacement and identifying key factors affecting breaks. Previous work has shown that categorizing

observations based on common pipeline features could lead to better model accuracy [103, 212]. Researchers have further suggested the inclusion spatial clusters as an explanatory variable could potentially improve regression models [69, 68]. To the authors knowledge, no previous work has provided a framework for comparing the precision of clustering algorithms for networked infrastructure and determined their usefulness in forecasting future breaks.

6.3 Data and Methods

As stated in the introduction, the first goal of this project is to compare three different clustering methods in their ability to capture pipe break zones in a pipe network with high precision. High precision here is defined as capturing many breaks while only classifying a small portion of network as spatial clusters. The second goal is to take the best performing cluster approach and determine whether machine learning pipe break models can be enhanced with spatial clusters as a variable. This section will outline the distribution network data from the partnering utility, as well as the clustering and machine learning approaches we explored.

6.3.1 Pipeline Network and Break History

We need two types of data to carry out the proposed research. The first are the water distribution network data, preferably in a digitized map format that includes physical attributes of each individual pipeline. The second are the pipe failure records themselves, these can be recorded either in text format (reporting the location and time of each break) or also in a digitized map.

We partnered with a mid-sized utility in the midwest which serves a population of approximately 100,000. The distribution system contains 423 miles of total pipe, with 11430 unique junctions and 12092 unique segments. The utility has had 755 pipe breaks (as recorded from repair work orders) spanning from 2008 to 2017, Figure 6.1 maps the distribution system with all break records overlaid. The digitized map contains pipe segment attributes and operating information such as age, material, diameter, average flow, and average pressure. We will use this dataset to implement and evaluate the quality of both the clustering algorithms and the subsequent predictive models.

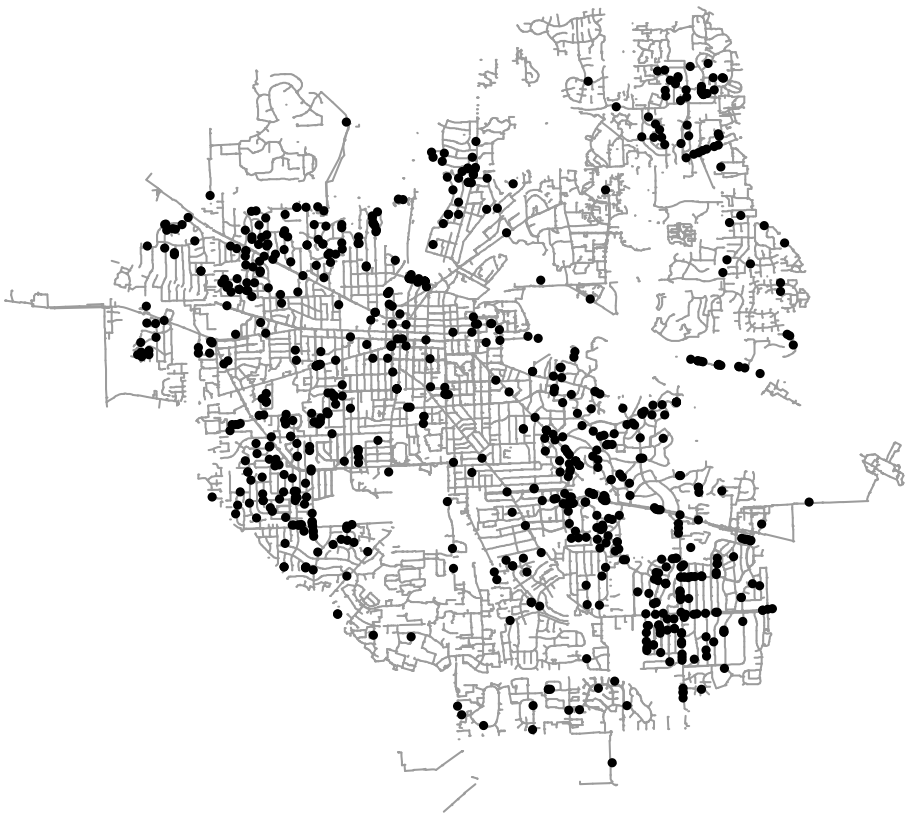


Figure 6.1: Network Layout with Breaks from May 2008 - Apr 2017.

6.3.2 Clustering Methods

Three algorithms are implemented and compared: a poisson based approach and two density based methods. We will outline each of them respectively.

We first start by defining an algebraic representation of a water distribution system. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represent an undirected graph, where $\mathcal{V}$ represents the set of pipe junctions and $\mathcal{E}$ represents the set of all pipe segments. Each recorded main break is referenced to its nearest pipe junction, let these points be denoted by $\mathcal{F}$ ($\mathcal{F} \in \mathcal{V}$). The aim of the clustering algorithm is to identify a set of discrete subgraphs $\mathcal{C}$ where the edges contained in $\mathcal{C}$ have a higher breakage rate than the ones not ($\mathcal{E} \setminus \mathcal{C}$).

6.3.2.1 DBSCAN

The algorithm recursively checks for each failure node in $\mathcal{F}$ whether a minimum count of breaks $\mathcal{B}$ exist within a threshold network distance $\mathcal{D}$. Two failure nodes are considered in close proximity to

each other if the minimum network distance between them is less than $\mathcal{D}$. If such condition is met, all edges contained within $\mathcal{D}$ from the root node are marked as a cluster, otherwise the algorithm moves onto the next failure node. Length constrained graph traversals such as the depth-first search [2] can be used to identify both the cluster condition and the set of contained edges.

The following parameter combinations are implemented for comparison: $\mathcal{B}$: {5, 8, 10, 12, 15} and $\mathcal{D}$: {0.125, 0.250, 0.375, 0.500} mi.

6.3.2.2 Locally Weighted DBSCAN

A potential flaw in the DBSCAN method is the reliance on global density parameters: $\mathcal{B}$ and $\mathcal{D}$. This assumes that every cluster in the search region can be characterized by the same properties: a minimum number of points contained within some radius. As a result, clusters with lower density may be missed and high density clusters may be overestimated. To more effectively reveal clusters across different regions of the data, different local densities may by needed [16].

OPTICS gets around this issue by detecting clusters of varying densities [16]. A networked implementation of this method is presented in De Oliveria et al. (2011) [69]. However since the output of the OPTICS method is a node grouping rather than a desired edge grouping, we instead adapt the existing DBSCAN method. A length constrained breadth-first graph traversal is performed out of each failure node in $\mathcal{F}$. To ensure that the search radius is strictly increasing, the closest unmarked point is traversed in each step. If the minimum break count $\mathcal{B}$ is met before the threshold distance $\mathcal{D}$, terminate the search early and return the traversed edges as a cluster.

The following parameter combinations are implemented for comparison: $\mathcal{B}$: {5, 8, 10, 12, 15} and $\mathcal{D}$: {0.125, 0.25, 0.375 ,0.500} mi.

6.3.2.3 Poisson Based Model

This approach is adapted from the existing literature, and reimplemented for our dataset. De Oliveira et al. [69] presents a framework for scanning a planar network and identifying hot zones with high events rates. This method assumes that the underlying break process (X) follows a poisson distribution. To account for potentially varying break distributions due to different pipeline features, we define the material-specific and diameter-specific breakage processes by the poisson random variables X_m and X_d.

$$X_m \sim Poisson(\lambda_m) \tag{6.1}$$

$$X_d \sim Poisson(\lambda_d) \tag{6.2}$$

Where λ_m and λ_d are the total number of breaks divided by the total length of pipe (in feet) considering only specific material and diameter classes. We perform a length constrained depth-first traversal out of each breakage node $\mathcal{F}$ to identify all edges contained within the distance $\mathcal{D}$. Any connected set of edges which contain at least $\mathcal{B}$ number of failures are considered as a candidate cluster. In each candidate, we use the parametric distributions X_m, and X_d to compute the expected number of breaks in the subgraph and compare it with the observed break count. Hypothesis testing is then used to calculate the statistical significance of the observation relative to the assumed distributions. Let O be the observed number of breaks in a given candidate cluster. The statistical significance of the O when evaluated against both poisson distributions (X_m and X_d) can be evaluated using the following p-values.

$$p_m = P(O \geq X_m) \tag{6.3}$$

$$p_d = P(O \geq X_d) \tag{6.4}$$

The regions are considered statistically significant if p_m and p_d are both less than 5%. This implies that under the poisson assumption there is a less than 5% chance the O breaks occur in the region randomly, thus we label all edges in the region as clusters. The following parameter combinations are implemented for comparison: $\mathcal{B}$: $\{5, 8, 10, 12, 15\}$ and $\mathcal{D}$: $\{0.125, 0.250, 0.375, 0.500\}$ mi.

6.3.3 Pipe Break Machine Learning Models

After comparing the performance between the three clustering algorithms, we take the best approach and subsequently explore whether including clusters in a pipe break machine learning model can improve accuracy. Below we will outline the dataset and evaluation framework we implemented, as well as the pipe break prediction models explored.

6.3.3.1 Dataset and Holdout Framework

The digitized dataset of the distribution network also comes with basic pipe attributes and information on operating conditions. We can also use the known location of each pipeline to infer key environmental variables. Using the mapping software ArcGIS, we performed a spatial analysis between the pipe locations and environmental information. We also gathered annual summaries of climatological data over the study area to infer if weather patterns after pipe break. Below we list the additional attributes collected and their respective sources. We associate this data with each individual pipe segment $\mathcal{E}$.

- Proximity to Major Roads: modeled as a binary variable, 1 if the pipe segment lies within 100 ft. of a highway, 0 otherwise. Obtained from the U.S. Census Bureau [195].
- Land Use: modeled as binary variable, 1 if pipe segment is underneath residential land, 0 otherwise. Obtained from the U.S. Census Bureau [195].
- Soil Conditions: corrosivity ratings for steel and concrete, runoff potential, frost heaving potential. Obtained from the U.S. Department of Agriculture [162].
- Annual Climate Information: total precipitation, days with temperature below freezing, days with temperature below 0F, days with more than 1 in. of precipitation, days with more than 1 in. of snowfall. Obtained from the National Oceanic Atmospheric Association [1].

We model pipe breaks at the annual scale: the number of breaks per pipe segment per year. As mentioned in previous sections the break records span between 2008 - 2017. The accuracy of each prediction model is evaluated in a holdout setting where a portion of the data is used for model training and predictions are made against previously unseen records. The predictions are then compared against known break responses and accuracy is summarized. To do this, the break records are divided into annual windows spanning from May to April. We select these windows such that each winter season is fully contained in an annual timeframe. Table 6.1 summarizes the full dataset used.

Table 6.1: Summary of dataset used for Regression Modeling.

Variable Name	Variable Description
Breaks (Response)	Number of Breaks on Pipe Segment in a given annual window.
Cluster	Indicator Variable. 1 if segment is in high breakage cluster, 1 otherwise.
Material	Pipe Material. Classified as Cast Iron or Other.
Diameter	Pipe Diameter (inches).
Length	Pipe Length in feet.
Age	Pipe Age (years).
PSI	Average operating pressure in pipe (pounds per square inch).
GPM	Average flow in pipe (gallons per minute).
HEAD	Average hydraulic head inside pipe (meters).
RoadProx	Indicator Variable. 1 if segment is within 200 ft. of major highway, 0 otherwise.
CorConrete	Soil susceptibility of corrosion to concrete. Classified as Low, High.
CorSteel	Soil susceptibility of corrosion to steel. Classified as Low, High.
Runoff	Runoff potential class for soil. Classified as Low, Medium, High.
FrostAct	Soil susceptibility to frost heaving. Classified as Low, Medium, High.
PRCP	Annual cumulative rainfall in annual window (inches.)
Days00	Days in annual window below 0 Fahrenheit.
Days32	Days in annual window below 0 Freezing (32 Fahrenheit).
PRCP01	Days in annual window with more than 1 inch of rainfall.
SNOW01	Days in annual window with more than 1 inch of snowfall.

Our final dataset includes break responses from 4 of the 10 available years: 2013 - 2014, 2014 - 2015, 2015 - 2016, and 2016 - 2017. For each time frame, we use break records from the preceding five years to classify high breakage clusters. As a result, breaks from 2008 - 2012 cannot be used as a response. A total of four holdout trials are performed where each individual year is left out for validation and the models are built with the remaining three. This allows us to determine whether clusters built from the past are indicative of future breaks. Table 6.2 summarizes the response and cluster information, this includes the response time frames as well as the time frames used to obtain historical clusters.

Table 6.2: Summary of Time Frames in the Pipe Break Dataset.

Year	**Response Window**	**Years Used for Cluster Training**
1	May 2013 - Apr 2014	May 2008 - Apr 2013
2	May 2014 - Apr 2015	May 2009 - Apr 2014
3	May 2015 - Apr 2016	May 2010 - Apr 2015
4	May 2016 - Apr 2017	May 2011 - Apr 2016

6.3.3.2 Regression Models

We chose five different regression models to estimate the number of breaks at each individual pipe segment. A summary of the approaches is presented below, we refer the reader to Hastie at al. (2009) [90] for a full description.

1. Generalized Linear Models (GLM): Because pipe breaks are discrete (zero, one, two breaks), we assume the response distribution is poisson. This model is a linear combination of the explanatory variables, fitted to the log of the response (number of breaks per segment).
2. Generalized Additive Models (GAM): We also assume the response distribution is poisson. This approach is also a linear combination of each explanatory variable, except here each continuous feature is modeled with smoothed splines to potentially capture non-linear relationships.
3. Regression Tree (RT): An iterative partitioning technique. Each split is selected to minimize the in-sample error of the resulting subspaces.
4. Random Forest (RF): An ensemble of regression trees, each built with bootstrap replicates of the original dataset. This allows for model variance to be reduced. Only a random subset of the explanatory variables are used at each partitioning in a tree. The average over all trees is taken as the final prediction.
5. Gradient Boosted Trees (GBT): An ensemble of regression trees, each trained sequentially. The training samples and the output trees are reweighed after each iteration based on in-sample error, this technique is shown to reduce bias. The final prediction is a weighted average over each regressor.

These five models are selected because they span a wide range in terms of model complexity and parametric assumptions. We also recognize that this is not an exhaustive list. The aim of the research is not to find the best predictive model, but rather to explore the effect of spatial clusters on model accuracy. It is possible other regression techniques can achieve higher accuracy, and finding these models is left for future research.

6.4 Clustering Results

The three clustering algorithms (DBSCAN, Locally Weighted DBSCAN, Poisson Based) were implemented in Python 3. There are many proposed metrics to quantify the effectiveness of identified clusters [96]. Some rely on statistical measures such as Moran's I or Tango's statistics, while others use likelihood based hypothesis tests. These measures were developed for euclidean search

spaces so they do not apply to this research. We will instead use the following domain-specific measures for evaluating failures on infrastructure: 1) proportion of breaks included in cluster, 2) proportional length of system in cluster. High precision clusters are desirable and these regions will have a high proportion of breaks included within a limited portion of the system length.

Table 6.3 reports the tuning results from DBSCAN.

Table 6.3: DBSCAN Clustering Results: Cluster Break Capture %, Cluster Length Capture %$\mathcal{B}$ is the break threshold and $\mathcal{D}$ is the search radius.

$\mathcal{D}$	$\mathcal{B}$ = **5**	$\mathcal{B}$ = **8**	$\mathcal{B}$ = **10**	$\mathcal{B}$ = **12**	$\mathcal{B}$ = **15**
0.125 mi.	37.2%, 6.3%	16.2%, 1.9%	8.5%, 0.9%	4.4%, 0.4%	2.1%, 0.1%
0.250 mi.	72.6%, 29.6%	50.1%, 16.7%	38.1%, 12.1%	31.4%, 9.1%	18.4%, 4.7%
0.375 mi.	85.7%, 52.6%	80.8%, 44.5%	68.1%, 35.4%	62.8%, 31.1%	57.1%, 25.6%
0.500 mi.	89.5%, 62.3%	86.9%, 59.1%	81.9%, 55.3%	78.2%, 50.0%	73.4%, 45.2%

In each cell the within-cluster break and length proportion capture are reported. The resulting outputs differ based on the input parameter combinations $\mathcal{B}$ and $\mathcal{D}$. Some patterns can be seen in the table. When $\mathcal{B}$ is held constant, both the proportion of break and network length contained within the cluster increases with $\mathcal{D}$. This is an intuitive result, for the same break threshold we expect to have more breaks identified and more pipe segments included with a larger search radius. Conversely, when $\mathcal{D}$ is held constant, both the proportion of break and network length contained within the cluster decreases with $\mathcal{B}$. This is also an intuitive result, we expect to have fewer breaks and pipe segments identified when we increase the break threshold for cluster consideration.

Table 6.4 reports the tuning results from the locally weighted DBSCAN.

Table 6.4: Locally Weighted DBSCAN Clustering Results: Cluster Break Capture %, Cluster Length Capture %. $\mathcal{B}$ is the break threshold and $\mathcal{D}$ is the search radius.

$\mathcal{D}$	$\mathcal{B}$ = **5**	$\mathcal{B}$ = **8**	$\mathcal{B}$ = **10**	$\mathcal{B}$ = **12**	$\mathcal{B}$ = **15**
0.125 mi.	36.0%, 5.3%	16.0%, 1.7%	8.5%, 0.7%	4.4%, 0.4%	2.1%, 0.1%
0.250 mi.	67.8%, 22.0%	46.6%, 12.8%	35.9%, 9.4%	29.8%, 7.3%	17.1%, 4.1%
0.375 mi.	81.1%, 41.5%	77.6%, 38.0%	63.6%, 29.5%	59.6%, 26.7%	52.3%, 22.9%
0.500 mi.	85.6%, 48.7%	85.0%, 52.6%	79.5%, 49.4%	74.0%, 45.3%	69.9%, 40.6%

Again, in each cell both the within-cluster break and length capture are reported as a result of parameter inputs $\mathcal{B}$ and $\mathcal{D}$. It is noted when comparing instances of the same input parameters against the DBSCAN, the break and length capture in the locally weighted case is always less. This is because the local weighted version terminates the graph search early once the break threshold is met, while the full DBSCAN will continue until the distance limit is met.

The empirical results in Tables 6.3 and 6.4 suggest that the locally weighted DBSCAN results in higher precision clusters. For example: in the instance where $\mathcal{B}$ = 0.5 mi. and $\mathcal{D}$ = 5, the locally weighted DBSCAN captures only 4% fewer breaks in more than 10% less of the total system length. This suggests that while the extended search means the DBSCAN can capture more breaks, the tradeoff is the lowering of the resulting cluster precision.

Clustering results for the Poisson model are reported in Table 6.5, the 5% significance level is selected for any candidate region to be considered a cluster.

Table 6.5: Poisson Model Clustering Results, for Significant Clusters at the 5% Level: Cluster Break Capture %, Cluster Length Capture %. $\mathcal{B}$ is the break threshold and $\mathcal{D}$ is the search radius.

$\mathcal{D}$	$\mathcal{B}$ **= 5**	$\mathcal{B}$ **= 8**	$\mathcal{B}$ **= 10**	$\mathcal{B}$ **= 12**	$\mathcal{B}$ **= 15**
0.125 mi.	26.8%, 3.8%	15.0%, 1.8%	8.0%, 0.8%	4.4%, 0.4%	2.1%, 0.1%
0.250 mi.	37.4%, 10.1%	36.7%, 10.2%	33.5%, 9.5%	30.6%, 8.4%	18.4%, 4.6%
0.375 mi.	49.3%, 17.6%	51.1%, 18.7%	50.6%, 19.1%	49.5%, 18.6%	48.1%, 18.3%
0.500 mi.	47.2%, 20.8%	51.0%, 22.5%	45.3%, 20.8%	45.2%, 20.5%	52.6%, 23.5%

This approach is the much more conservative compared to the DBSCAN methods in terms of classifying pipes as clusters. For the same input parameters ($\mathcal{D}$, $\mathcal{B}$) the break capture and length capture for the Poisson approach is always the lowest. There is also no discernible pattern related to the parameters, possibly due to the significance tests required for any candidate space to be included as cluster. Empirical results from the last two rows ($\mathcal{D} \geq 0.375$ mi.) suggests the algorithm becomes insensitive to the $\mathcal{B}$ parameter after a certain threshold. This is evidenced by the fact that all entries in the same row are very similar.

When selecting the best cluster, we ignore instances where the resulting clusters cover more than 50% of the network length. This is because these results are not high precision. Of the remaining trials where fewer than half of the network is identified, we select the case where the highest proportion of breaks are captured. The locally weighted DBSCAN is selected as the best clustering approach out of the three, specifically the instance where the input search radius ($\mathcal{D}$) is 0.5 miles and the break threshold ($\mathcal{B}$) is 5 breaks. 85.6% of all the training breaks are contained within 48.7% of the system length, Figure 6.2 shows the identified clusters overlaid with the break points. In the subsequent machine learning trials, this algorithm is used to train the clustering indicator variables.

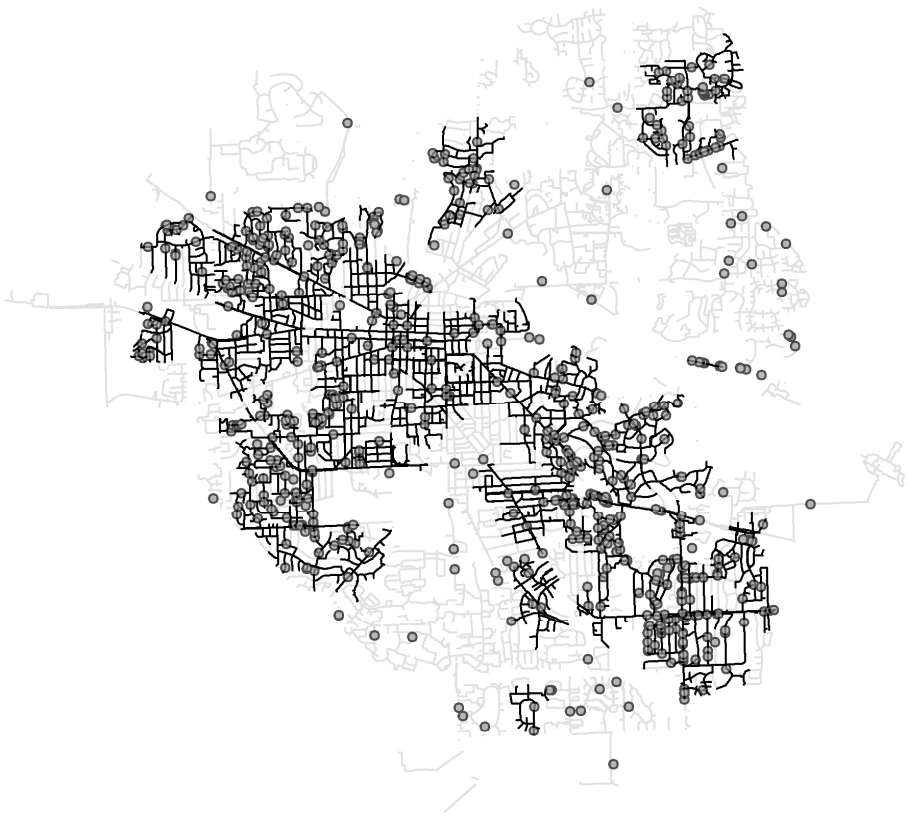


Figure 6.2: Cluster output from Locally Weighted DBSCAN with Parameters $\mathcal{D}$ = 0.5 mi. and $\mathcal{B}$ = 5.

6.5 Pipe Break Machine Learning Evaluation and Results

As shown in Table 6.2, the pipe break dataset was divided into 4 annual time frames between May and April: 2013 - 2014, 2014 - 2015, 2015 - 2016, and 2016 - 2017. Cluster indicators for each window were generated with the locally weighted DBSCAN method using break data from the preceding five years. This is a binary vector where 1 indicates that a pipe segment belongs in a high breakage cluster and 0 otherwise.

Four holdout trials were performed where data from three years were used for model building, and remaining year was used for model validation. In the methods section above, we noted that five regression models were explored (GLM, GAM, RT, RF, and GBT). For each model structure, one with-cluster version is trained where the cluster indicator is included in the training data and another without-cluster version where the indicator is removed. The goal of this section is to compare the accuracy between the with-cluster and without-cluster models.

6.5.1 Evaluation Criteria

One common criteria to evaluate regression models is mean squared error. However, since the number of instances with zero breaks so heavily outnumber non-zero breaks in this dataset (99.4% of all observations), using mean squared error would heavily skew the accuracy evaluation towards the zero break instances [57]. In practice, utilities are more interested in the higher risk pipes that are predicted to have the most breaks. This is in part due to limited budgets available where only the highest priority assets can be addressed in a given year. As a result, a more useful evaluation criteria should reward models that better separate high breakage pipes.

A natural approach is to form a ranking based on the predictions of each regression model and evaluate whether observations with breaks are ranked higher than those without. For any holdout year and regression model, let $\widehat{Y}_i$ be the predicted number of breaks for the i'th pipe segment ($\forall$ i $\in \mathcal{E}$). Sort each of the observations in $\mathcal{E}$ high to low based on $\widehat{Y}_i$. Let $Y*$ be the real break observation vector when sorted based on the $\widehat{Y}_i$ predictions. A break capture vector can be generated by computing the cumulative sum at each index along the sorted $Y*$ vector. Divide the cumulative sum vector by the total number of breaks, now the vector is scaled between 0 - 1 and reflects the fraction of breaks captured instead than the raw number. A length capture vector can similarly be calculated by the cumulative sum of each pipe length along the same sorted vector. Divide the length capture by the total length of the entire system to normalize between 0 - 1. Let $\mathbb{B}$ and $\mathbb{L}$ be the rank ordered break and length capture vectors. Note that both start at 0 and end at 1, $\mathbb{L}$ is strictly increasing and $\mathbb{B}$ is strictly non-decreasing.

The $\mathbb{B}$ and $\mathbb{L}$ vectors characterize how well a set of predictions prioritize high risk breaks. For example, suppose the 100th index along the break capture vector $\mathbb{B}$ is 0.4 and is 0.3 along the length capture vector $\mathbb{L}$. This means that at the top 100 ranked pipe segments, 40% of all breaks in the validation data are captured within 30% of the system length. Plotting $\mathbb{B}$ against $\mathbb{L}$ and taking the area under the curve will give us a single metric that captures prioritization accuracy. This metric is similarly scaled between 0 - 1 and models that produce more accurate rankings will have higher area under the curve.

6.5.2 Holdout Trials

In all four trials the resulting regression tree only had one level, meaning the final prediction only took one of two possible values. This shows that the model is under-fitting the data, and is typically associated with high bias (error) [90]. Since the predictions only take two unique values, they are also not useful in doing any prioritization and are discarded from further analysis. Only results from the remaining four models are presented and discussed below.

Figure 6.3 below shows the rank ordered plots of $\mathbb{B}$ against $\mathbb{L}$ for the holdout year May 16 -

Apr 17.

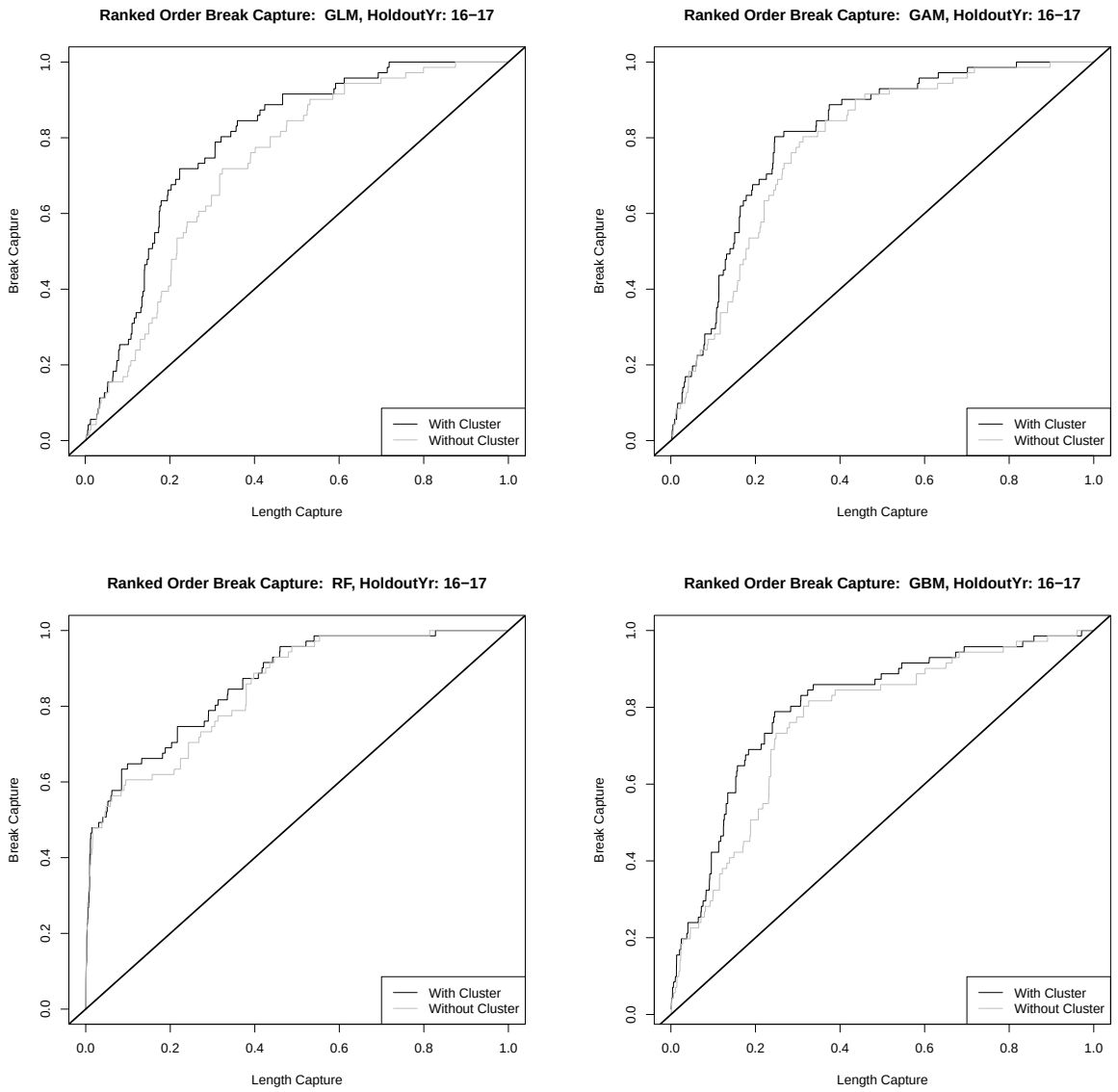


Figure 6.3: Holdout Results for May 16 - Apr 17. Rank Ordered Plot, Break Proportion Capture vs. Length Proportion Capture.

Each subplot indicates results for a single model structure, the black line shows the performance of the with-cluster model and the gray line shows the performance of the without-cluster model. Results for the other three holdout years are included in the appendix. In each of the four models the black lines are higher than the gray lines. This indicates a better prioritization of high breakage pipes is achieved when cluster indicators are included. The gap between the two lines is smallest in the random forest, suggesting that while the with-cluster model still improves the accuracy, the improvement is smaller than in the other models.

The area under the curve metrics across all validation trials are reported in Table 6.6.

Table 6.6: Area Under the Ranked Ordered Curve: With Cluster Indicator, Without Cluster Indicator.

Holdout Year	GLM	GAM	RF	GBT
May 13 - Apr 14	**0.653**, 0.640	**0.669**, 0.668	0.647, **0.677**	0.606, **0.627**
May 14 - Apr 15	**0.735**, 0.676	**0.776**, 0.740	**0.755**, 0.706	**0.732**, 0.717
May 15 - Apr 16	**0.765**, 0.727	**0.794**, 0.767	**0.887**, 0.881	**0.803**, 0.756
May 16 - Apr 17	**0.790**, 0.727	**0.810**, 0.779	**0.863**, 0.845	**0.804**, 0.763

In each cell, results from the with-cluster and without-cluster models are shown and the higher value is indicated in bold. Empirical results show that the with-cluster model produced a better ranking in all trials except two: the RF and GBT in the first holdout year. In three out of the four years the random forest model had the highest ranking metric, and across all four years a with-cluster model was the overall best. The average improvement in the area under curve metric from including the cluster indicator is 6.2% in the GLM and 3.1% in the GAM. Omitting the 2 trials where the cluster led to a worse model, the average improvement in the area under curve metric was 3.3% and 4.6% in the RF and GBT respectively.

Because utilities only have the resources to address to highest priority needs, an evaluation metric that reflects accuracy at the only the top ranked assets is a more useful. We focus on the accuracy of the top ranked 20% of system length. The area under the ranked-ordered curve can be adapted such that only the first portion of the curve $\mathbb{L} \leq 0.2$ is considered. Table 6.7 shows the adjusted results, with-cluster and without-cluster area under curve metrics are reported side by side in each cell.

Table 6.7: Area Under the Ranked Ordered Curve at the top 20%: With Cluster Indicator, Without Cluster Indicator.

Holdout Year	GLM	GAM	RF	GBT
May 13 - Apr 14	**0.0336**, 0.0299	**0.0414**, 0.0393	0.0510, **0.0595**	**0.0428**, 0.0353
May 14 - Apr 15	**0.0432**, 0.0330	**0.0567**, 0.0484	0.0493, **0.0496**	**0.0524**, 0.0469
May 15 - Apr 16	**0.0558**, 0.0473	**0.0684**, 0.0641	**0.1319**, 0.1296	0.0655, **0.0664**
May 16 - Apr 17	**0.0605**, 0.0405	**0.0703**, 0.0567	**0.1173**, 0.1109	**0.0808**, 0.0609

The with-cluster models always outperform the without-cluster models in the GLM and GAM instances. The average improvement from adding the cluster indicator is 27.7% and 13.3% respectively between the two models. The results between the RF and GBT are more similar, the with-cluster models produced a better ranking in just 5 of the 8 trials. In three out of the four holdout years, a with-cluster model produced the best overall ranking.

The empirical results suggest that inclusion of the cluster indicator variable generally improves the ranking accuracy of a predictive model. Similar findings are derived when comparing all instances and only the top ranked 20% of pipe segments. For simpler models like the GLM and GAM, both with strong parametric assumptions, having the cluster indicator always improved the model. In contrast the cluster indicator usually improved the RT and GBT, but there are a number of cases (5 out of 16) where it did not. It is possible with more complex models like the RT and GBT that are able to capture a variety of non-linear relationships, the cluster indicator does not provide as great discriminatory power between high and low breakage pipes.

These findings show that including high breakage clusters in a machine learning model can help better prioritize high risk pipes. Specifically, the ranked ordered sorting of observations based on the predicted number of breaks are generally more accurate. Spatial clusters always improve simpler models, whereas for more complex models they still generally help. Our results suggest that spatial clusters trained from historical break events are indeed useful for predicting future breaks. We acknowledge that our findings here are based only on empirical results and pertain specifically to our dataset. It is possible that the utility of using spatial clusters will change in a different test case, this is left for future exploration.

6.6 Conclusion

In this research we explored the effectiveness of three different spatial clustering algorithms to identify high breakage zones within a pipeline network. We adapted existing methods presented in the literature for applications in planar graphs, and these methods can be adapted for use in other infrastructure domains. Empirical results from a real water network shows that a locally weighted DBSCAN algorithm produces the highest precision clusters. This is where the most breaks are captured within a limited total mileage of pipe.

We also examined whether the inclusion of spatial clusters as an explanatory variable can improve pipe break predictions. Since in practice utilities can only afford to target the highest priority assets for replacement/rehabilitation, overall accuracy measures are not particularly useful. Instead we adapt our evaluation of criteria based on the ability of predictive models to achieve an accurate sorting of observations, where the best models will capture more breaks the highest ranked instances.

We performed holdout trials where clusters were trained from the preceding five years and used in conjunction with spatial pipeline data for predict future breaks. Another set of models were trained without the clustering information for comparison. Our experiments show that including spatial clusters generally leads to improved accuracy, and the best overall model always had clustering information. This shows that combining two different statistical techniques: clustering and

predictive modeling, can potentially lead to better identification of high risk critical infrastructure. Giving decision makers better information to make the right capital investment decision.

Acknowledgements

We thank the University of Michigan and the National Science Foundation (NSF, grant number 1621116) for funding this research. The opinions and views expressed are those of the researchers and do not necessarily reflect those of the sponsors.

CHAPTER 7

Conclusion

The research presented in this thesis forms a body of work that critically analyzes the current state of risk analysis and management for drinking water infrastructures. I aim to demonstrate how risk assessments in this domain can be improved such that better decision support can be provided for the protection of these critical systems. This dissertation sits at the intersection of the fields of risk analysis, civil and environmental engineering, statistics, optimization, and decision support. The works presented borrows from each of these respective domains to provide new insight to the complex problem of managing the aging water infrastructure in the US. The intellectual contributions of each individual chapter are summarized in section 7.1, and possible directions for future work are discussed in section 7.2.

7.1 Summary of Contributions

Chapter 2 provides context for why risk assessment practices in the water infrastructure domain need advancing. More specifically, it demonstrates why there is a need for the research in this thesis. I aim to establish the existing knowledge gap between risk analysis practice and theory by critically analyzing a popular J100-10 standard [23] published by the American Water Works Association (AWWA). Through a holistic comparison between the recommended practices within the standard and the foundational concepts of the risk analysis field, conceptual and theoretical limitations can be identified. We find that key concepts of risk and resilience are inadequately conceptualized, and methods for quantifying these metrics can lead to misrepresentations which could ultimately misguide the decision maker.

The insights gained in this chapter pertain only to the J100-10 application of risk analysis for the water infrastructure domain. I acknowledge there are numerous other risk assessment frameworks presented in both the academic and professional literature. The point is to begin a conversation on how this popular standard can be improved moving forward, and to highlight the gap between theory and practice. Advancements in future editions of the J100-10 that fall closer in line with the state of the art can lead to better practice of risk assessments in the water infrastructure domain.

The rest of the dissertation examines two specific steps in the asset management of water infrastructure: pipe rehabilitation planning and robotic inspection planning, and demonstrates the application of quantitative tools to improve the decision making process. These topics were selected because they are two critical steps in any risk assessment process for water systems.

Chapters 3 and 4 explores the mathematical optimization of inspection path planning. Utility operators need to manage a vast network of underground pipelines, but only have limited budgets for their inspection. For higher returns on investment, identifying an inspection path that maximizes the traversal of high risk pipelines can assist better decision making for risk mitigation (e.g.

replacement versus repair). There are a variety of previous research that have examined this optimization problem, and have presented a multitude of formulations and algorithms for this task. A key omission from previous works is the limitations of the robotic inspection tools themselves are not included in the modeling framework. For different robotic sensors, specific properties of the inspection path can impact the quality of the inspection reading and lower its effectiveness for decision making.

The contribution of chapter 3 is twofold, the first is to highlight how omitting platform limitations can lead to lower quality inspections. The second, a general framework for incorporating tool limitations into a mathematical optimization framework is presented and applied for both real and synthetic networks. The main contribution here is that we highlight the importance of tool limitations when planning for inspections, something that is omitting in the risk and reliability literature prior to publishing this chapter in 2018. Three heuristic optimization algorithms are applied and their effectiveness is explored.

Chapter 4 extends this work and presents an exact integer program formulation for the routing problem. The model is a variant on the prize collecting traveling salesman problem, where the goal is to find a maximum value subpath within a large network. Material and diameter changes between adjacent pipes, which impact sensor readings, are included as penalties to the objective. Five different solution algorithms (variants on integer programming branch and bound, and variants on tree search) are demonstrated for a mid-sized utility in the US, and their scalabilities are explored. The contribution here is the mathematical model itself, as well as the assessment of algorithm scalability for this routing problem. Together these two chapters help improve the state of risk analysis practice in the water domain where utilities can better allocate inspection investments to better manage their critical infrastructures.

Chapters 5 and 6 examines the statistical modeling of pipe breaks and its application for replacement planning. Typical water systems in the US have been installed over 50 years ago and have limited availability of real time condition data. To better allocate replacement and repair spending, utilities need to know which assets are in poorest condition. Statistical models of pipe break predictions are a useful decision support tool because they help utilities identify potential areas of greatest risk for infrastructure failure. This information can help utility operators proactively manage their assets by targeting high risk areas before failures occur.

Many small to mid-sized utilities have no records of any pipeline information at all, and this challenge is addressed in chapter 5. The contribution is to demonstrate whether accurate and useful models can be built in the absence of any pipe-specific features. We partnered with a utility in the Mid-Atlantic and obtained the date and time of pipe repair work orders as the break response. Publicly available environmental and demographic information are used as regressors, and a variety of machine learning models are trained. We find that accuracy suffers when key pipeline infor-

mation such as diameter, material, and age are omitted. However when compared to prioritizing assets solely based on failure history, a common practice amongst utilities, a better sorting of high risk assets can be achieved when using the statistical models. The results show that utilities with no digitized network information can still leverage public data to create statistical models that are useful for replacement planning. This work also demonstrates that useful decision support can be obtained despite limited accuracy with these methods, and can drive data collection efforts in the future.

Chapter 6 aims to extend the methodology in the pipe break modeling literature. The contribution is to explore if information about the spatial clustering of high breakage density zones, if and where they exist, can improve machine learning models when used as a regressor. From my review, no previous work has combined an unsupervised learning technique with supervised methods to improve predictive accuracy in this domain. We partnered with a mid-sized utility in the Mid-West and aplied three clustering algorithms for the identification of pipe break clusters. Machine learning models for the estimation of pipe break count are then trained with and without clustering information. The goal is to determine if the models having the clustering data will outperform those without, and if so, by how much. Findings from this research show that models with clustering information included tend to outperform those without. This indicates that the two-tiered framework which combines unsupervised and supervised techniques can produce better performing models. The magnitude of this improvement depends on the complexity of the model, simpler models will experience larger improvements over more complex models. Practical uses of this research aim to guide the modeling approach used by utilities to obtain more accurate break forecasts.

7.2 Future Research Directions

One extension from the critical review of J100-10 is to establish an alternate assessment procedure that addresses all of the identified flaws. Another direction for future research is developing a framework for an implementation-driven comparison between various risk assessment methods. The methods for a critical review of a risk analysis standard, such as that presented in Chapter 2 and be applied to domains beyond water infrastructure. The presented approach focuses on the foundational issues of the risk analysis field. An alternate approach is to implement a variety of risk analysis standards and compare the assessment results. A framework for doing so is currently unknown due to challenges of differing scope and focus across standards.

The inspection routing optimization work in Chapters 3 and 4 incorporates a constant penalty factor for the number of pipe feature changes occurring along an identified route. Future work could expand on the modeling framework to incorporate a variety of technology-specific consid-

erations. Extensions can also explore how to optimize the planning for multiple tools, whether it involve different inspection robots or multiple allowable runs of the same tool. The added complexity of optimizing for an ensemble of paths, rather than just a single path can lead to useful and interesting research.

Chapter 6 identifies a clustering algorithm for finding high breakage spatial clusters, and demonstrates how to incorporate this information as a regressor in a statistical model. However the focus is purely on the improvement of statistical accuracy for future break prediction, and does not explore the patterns between pipes within and outside these identified clusters. Perhaps there is a certain class of pipe material that is predominantly included in these clusters, or a certain installation age that of those pipes. This type of analysis can provide better understanding on the underlying mechanisms that are driving high breakage zones. The information can in turn be used in conjunction with failure forecasts to better inform pipe rehabilitation and replacement decisions.

The statistical modeling methods presented in Chapters 5 and 6 can be extended by incorporating representations of uncertainty in the output. Beyond just likelihood of an event (pipe break), uncertainty is a big part of an adequate risk representation [28]. Current model outputs only provide point predictions on either the probability of failure or the expected number of failures per asset. How to illicit uncertainty in these predictions, and how to incorporate them into a wider decision making framework, is currently not well understood within the infrastructure planning realm.

APPENDIX A

Chapter 3 Appendix - Example Optimal Paths

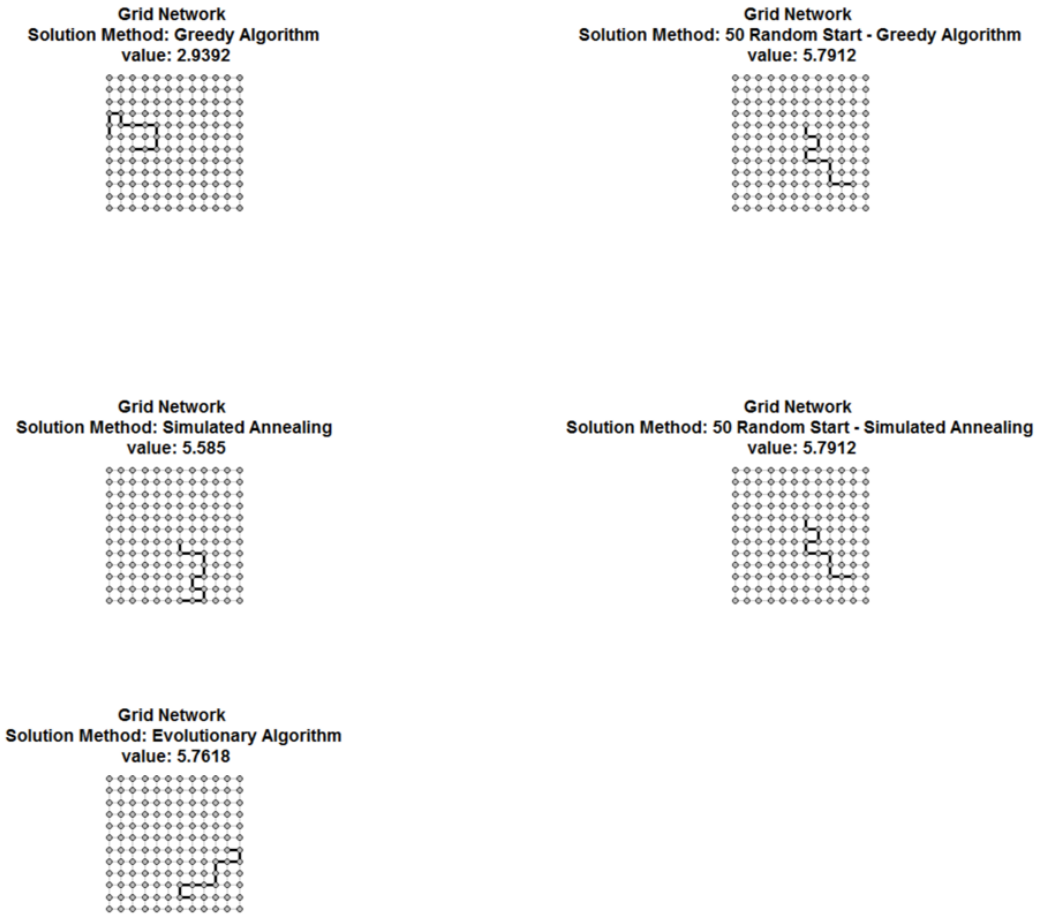


Figure A.1: Example Solution Paths in Random Grid Network.

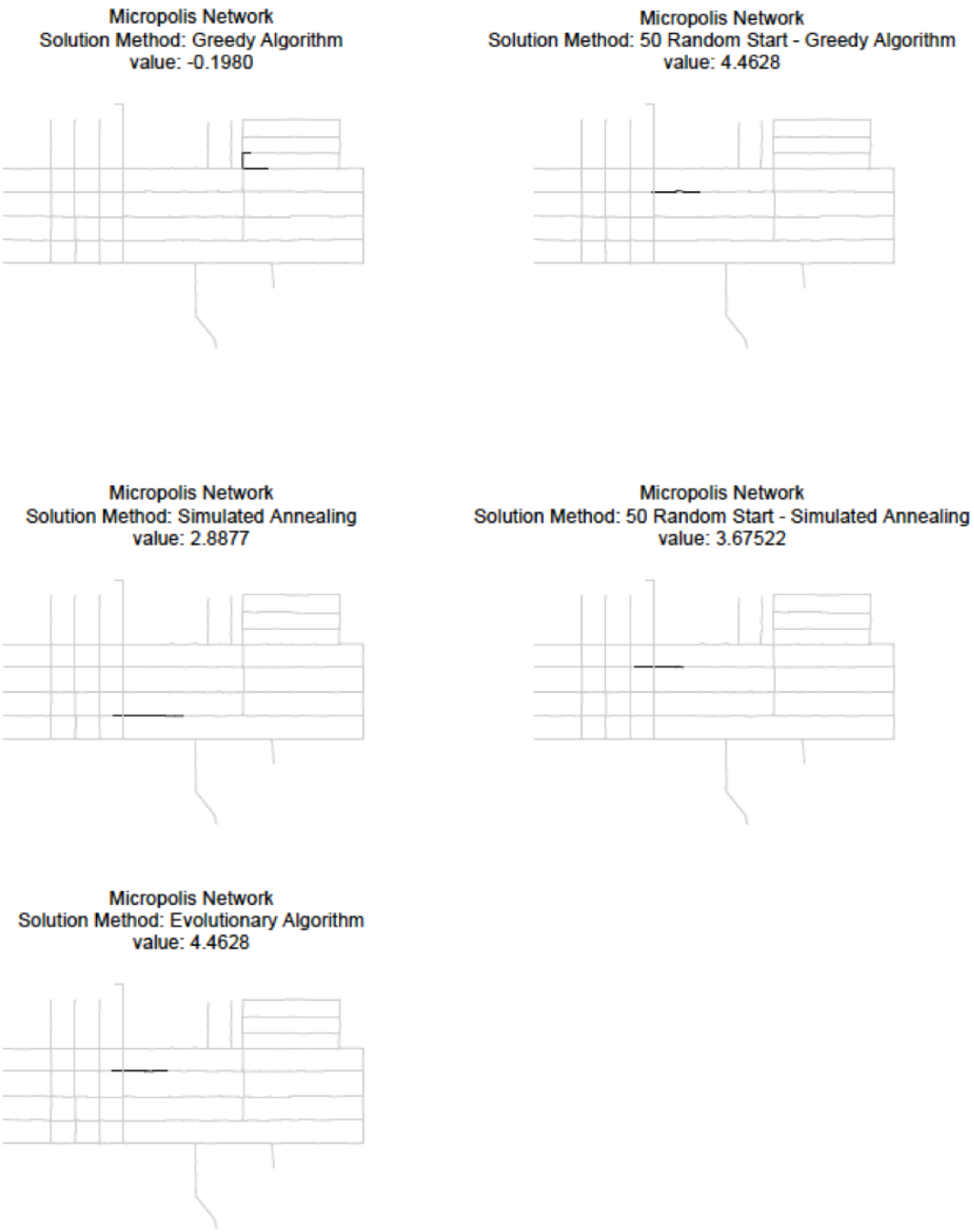


Figure A.2: Example Solution Paths in Micropolis Network.

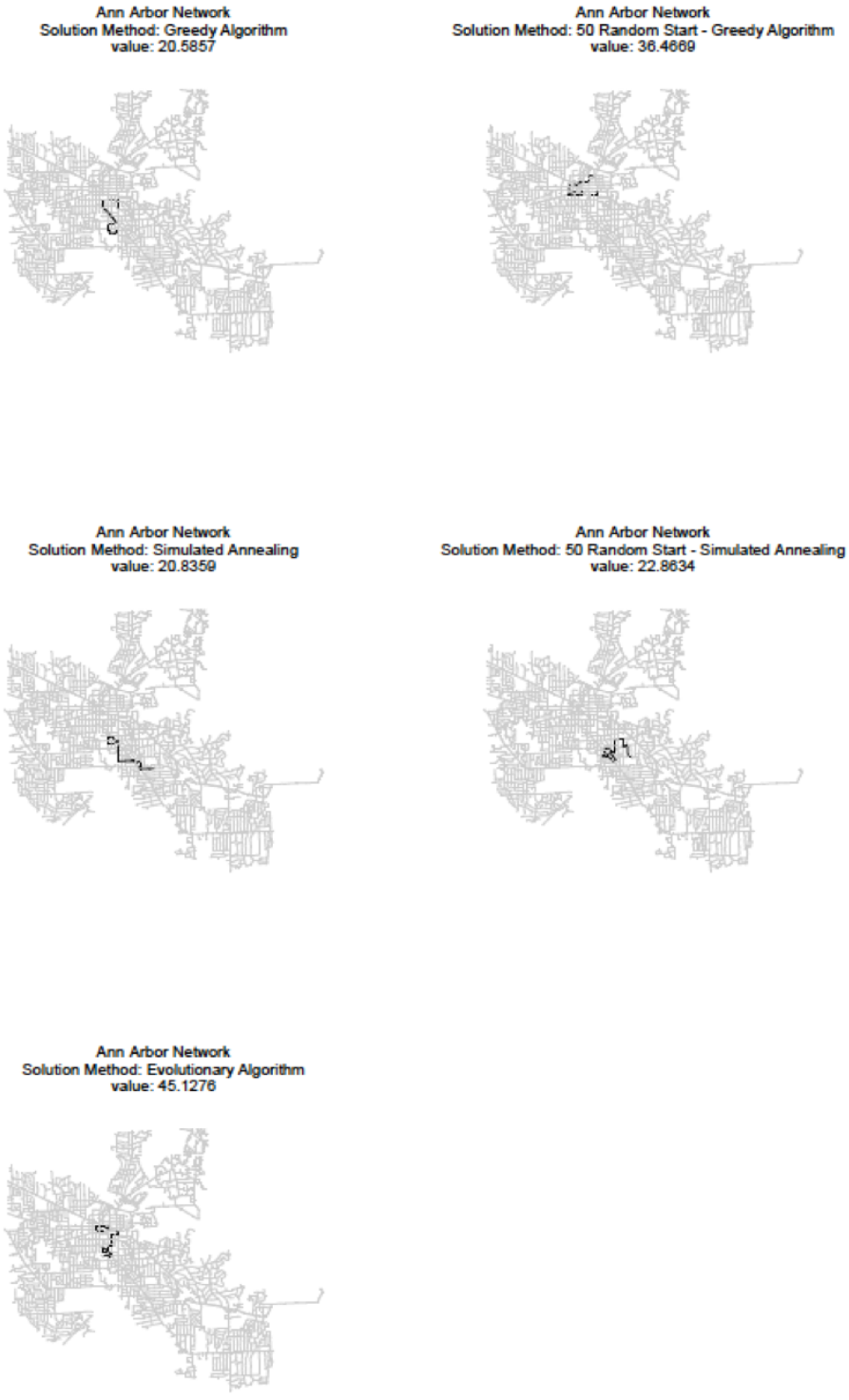


Figure A.3: Example Solution Paths in Ann Arbor Network.

APPENDIX B

Chapter 6 Appendix - Rank Ordered Plots

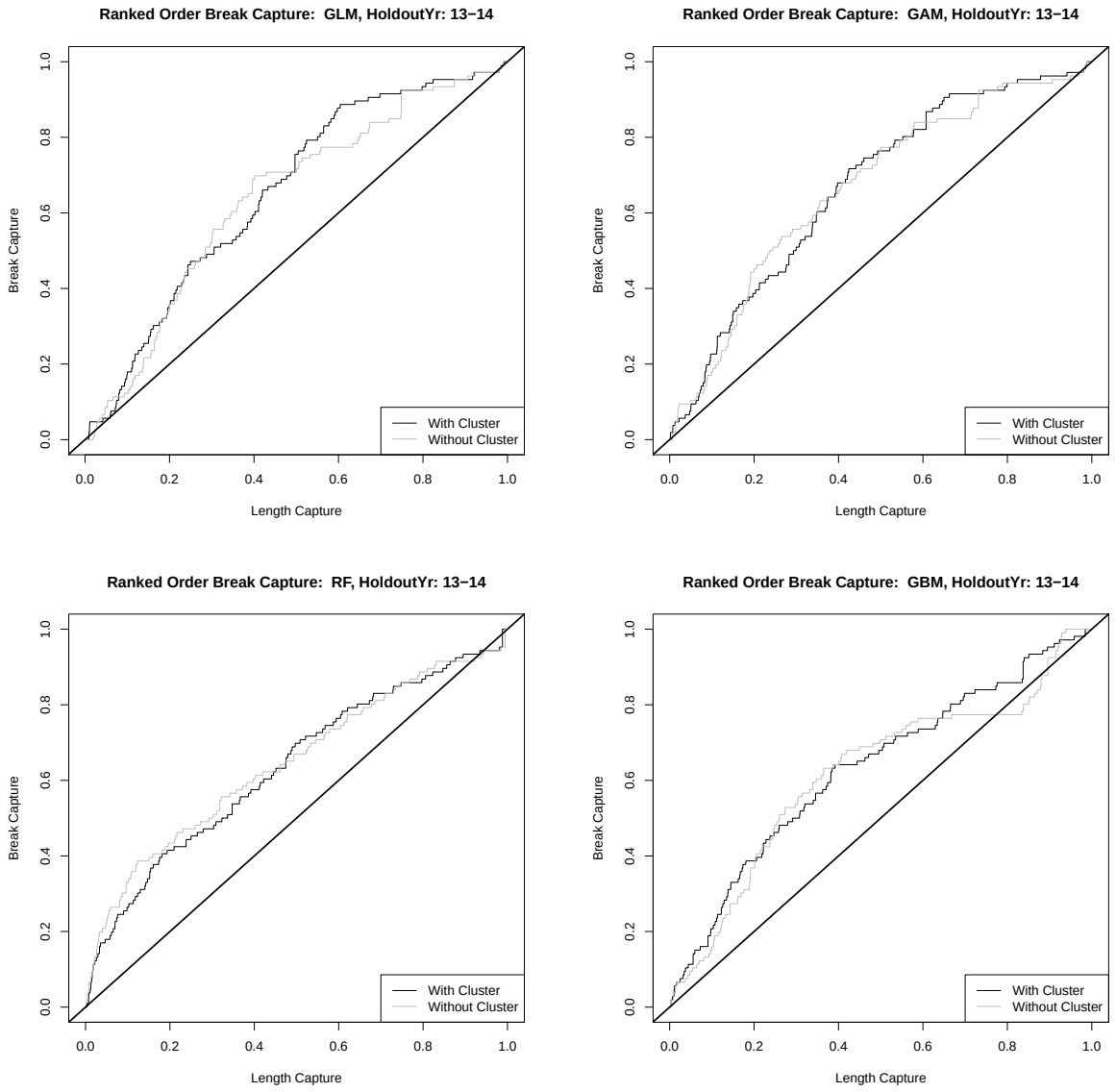


Figure B.1: Holdout Results for May 13 - Apr 14. Rank Ordered Plot, Break Proportion Capture vs. Length Proportion Capture.

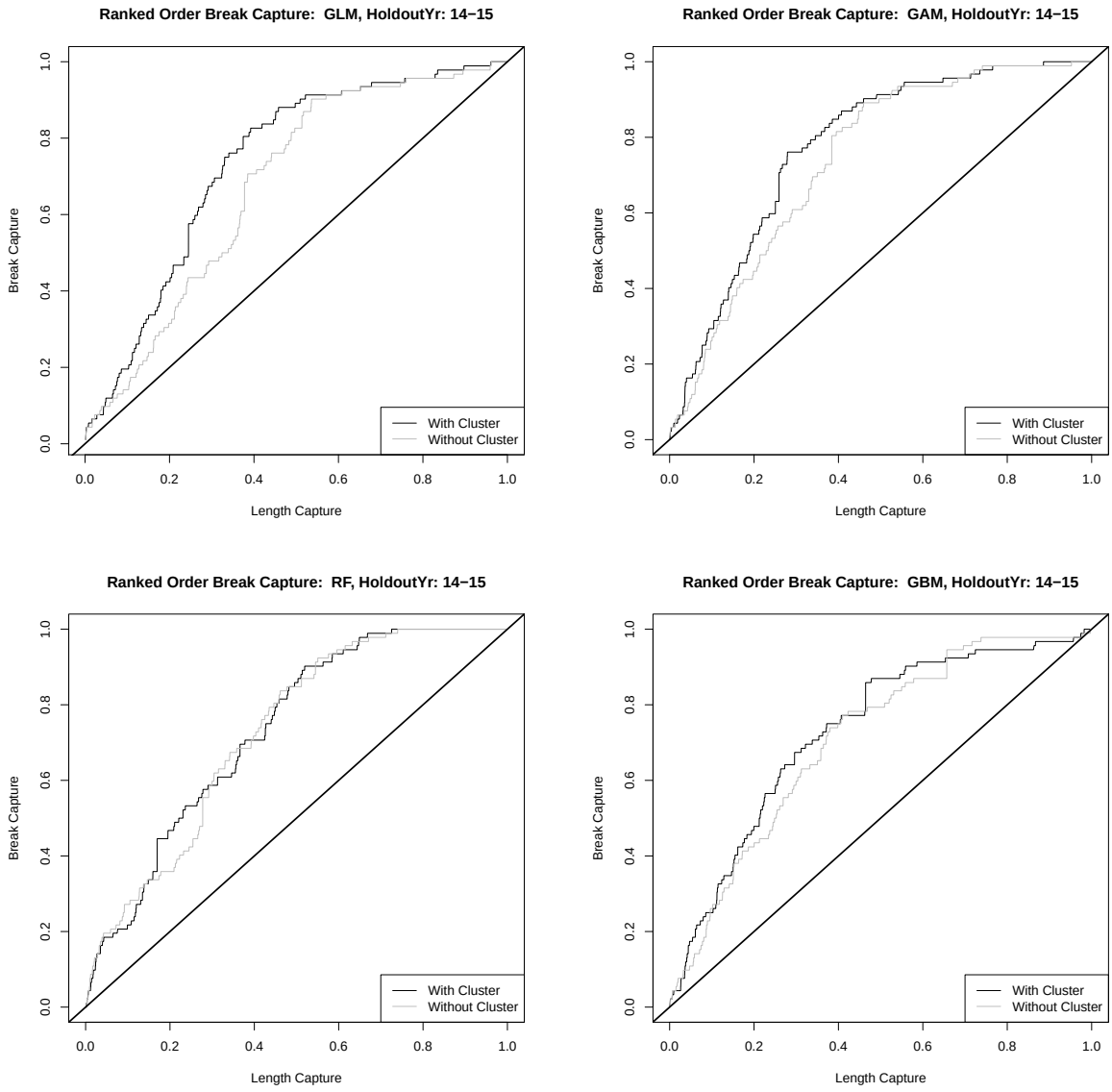


Figure B.2: Holdout Results for May 14 - Apr 15. Rank Ordered Plot, Break Proportion Capture vs. Length Proportion Capture.

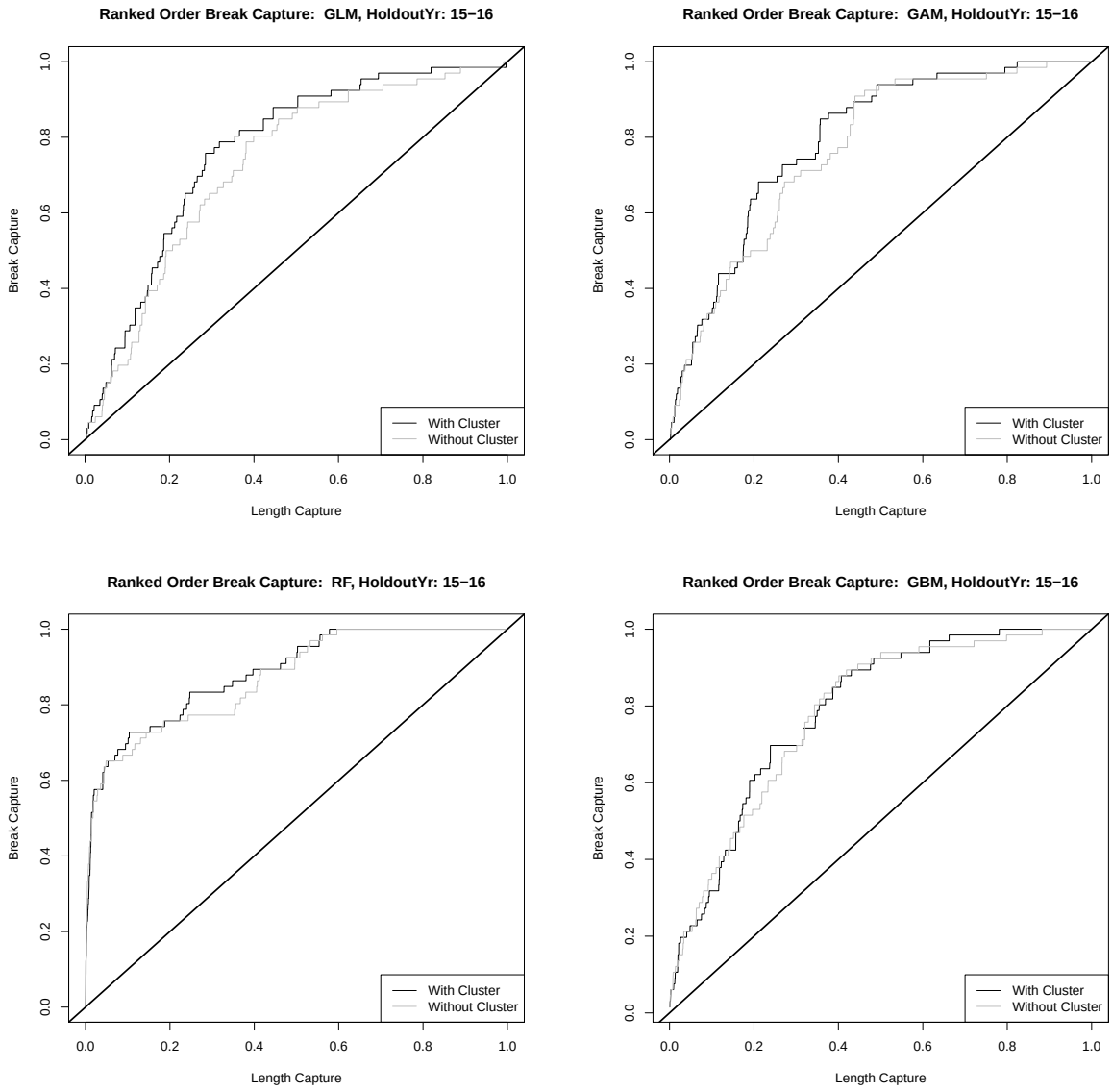


Figure B.3: Holdout Results for May 15 - Apr 16. Rank Ordered Plot, Break Proportion Capture vs. Length Proportion Capture.

BIBLIOGRAPHY

[1] NOAA (National Oceanic Administration) and Atmospheric. Local climatological data (LCD) dataset documentation. Technical report, NOAA, Silver Spring, MD, 2010.

[2] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, New Jersey, 1993.

[3] Pavel Albores and Duncan Shaw. Government preparedness: Using simulation to prepare for a terrorist attack. *Computers and Operations Research*, 35(6):1924–1943, 2008.

[4] David L. Alderson, Gerald G. Brown, Matthew W. Carlyle, and Louis Anthony Cox. Sometimes There Is No "Most-Vital" Arc: Assessing and Improving the Operational Resilience of Systems. *Military Operations Research*, 18(1):21–37, 2013.

[5] David L. Alderson, Gerald G. Brown, and W. Matthew Carlyle. *Assessing and Improving Operational Resilience of Critical Infrastructures and Other Systems*. 2015.

[6] David L Alderson, Gerald G Brown, and W Matthew Carlyle. Operational Models of Infrastructure Resilience. *Risk Analysis*, 35(4):562–586, 2015.

[7] B.J.M. Ale, D.N.D. Hartford, and D.H. Slater. The practical value of a life : priceless, or a CBA calculation? *Medical Research Archives*, 6(3):1–12, 2018.

[8] Mohammed J.F. Alenazi and James P.G. Sterbenz. Comprehensive comparison and accuracy of graph metrics in predicting network resilience. In *2015 11th International Conference on the Design of Reliable Communication Networks, DRCN 2015*, pages 157–164, 2015.

[9] Mohammed J.F. Alenazi and James P.G. Sterbenz. Evaluation and comparison of several graph robustness metrics to improve network resilience. In *Proceedings of 2015 7th International Workshop on Reliable Networks Design and Modeling, RNDM 2015*, pages 7–13, 2015.

[10] Maura Allaire, Haowei Wu, and Upmanu Lall. National trends in drinking water quality violations. *Proceedings of the National Academy of Sciences*, 115(9):2078–2083, 2018.

[11] Paul Allison. *Multiple Regression: A Primer*. Pine Fore Press, 1999.

[12] Stefano Alvisi and Marco Franchini. Multiobjective Optimization of Rehabilitation and Leakage Detection Scheduling in Water Distribution Systems. *Journal of Water Resources Planning and Management*, 135(6):426–439, 2009.

[13] Sandra F. Amass. *The Science of Homeland Security, Volume 1*. Purdue University Press, Wast Lafayette, Ind., 2006.

[14] Stefanos A. Andreou. *Predictive Models for Pipe Break Failures and Their Implications on Maintenance Planning Strategies for Deteriorating Water Distribution Systems*. PhD thesis, Massachusetts Institute of Technology, 1986.

[15] Stefanos A. Andreou, David H. Marks, and Robert M. Clark. A New Methodology for Modelling Break Failure Patterns in Deteriorating Water Distribution Systems: Theory. *Advances in Water Resources*, 10(1):2–10, 1987.

[16] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod Record*, 28(2):49–60, 1999.

[17] George E Apostolakis. How Useful Is Quantitative Risk Assessment? *Risk Analysis*, 24(3):515–520, 2004.

[18] Kenneth J. Arrow. A Difficulty in the Concept of Social Welfare. *The Journal of Political Economy*, 58(4):328–346, 1950.

[19] (American Society of Civil Engineers) ASCE. 2017 Infrastructure Report Card. Technical report, American Society of Civil Engineers, Reston, VA, 2017.

[20] Tore Askeland, Roger Flage, and Terje Aven. Moving beyond probabilities Strength of knowledge characterisations applied to security. *Reliability Engineering and System Safety*, 159:196–205, 2017.

[21] ASME-ITI LLC. RAMCAP EXECUTIVE SUMMARY RAMCAP A 7 Step Approach. pages 1–8, 2005.

[22] National Rural Water Association. Security Vulnerability Guide for Small Drinking Water Systems Serving Populations Between 3300 and 10000. Technical report, National Rural Water Association, 2002.

[23] The American Water Works Association. J100-10 Risk and Resilience Management of Water and Wastewater Systems. Technical report, 2010.

[24] Terge Aven. On some foundational issues related to cost-benefit and risk. *International Journal of Business Continuity and Risk Management*, 7(3):182–191, 2017.

[25] Terge Aven, Piero Baraldi, Flage Roger, and Enrico Zio. *Uncertainty in risk assessment: the representation and treatment of uncertainties by probabilistic and non-probabilistic methods*. 2013.

[26] Terje Aven. *Foundations of Risk Analysis. A Knowledge and Decision-Oriented Pespective*. John Wiley & Sons, Ltd, 2003.

[27] Terje Aven. On How To Define, Understand and Describe Risk. *Reliability Engineering and System Safety*, 95(6):623–631, 2010.

[28] Terje Aven. The risk concept-historical and recent development trends. *Reliability Engineering and System Safety*, 99:33–44, 2012.

[29] Terje Aven. Practical implications of the new risk perspectives. *Reliability Engineering and System Safety*, 115:136–145, 2013.

[30] Terje Aven. On the use of conservatism in risk assessments. *Reliability Engineering and System Safety*, 146:33–38, 2016.

[31] Terje Aven. How some types of risk assessments can support resilience analysis and management. *Reliability Engineering and System Safety*, 167(August 2016):536–543, 2017.

[32] Terje Aven. Improving risk characterisations in practical situations by highlighting knowledge aspects, with applications to risk matrices. *Reliability Engineering and System Safety*, 167:42–48, 2017.

[33] Terje Aven. The Call for a Shift from Risk to Resilience: What Does it Mean? *Risk Analysis*, (2009), 2018.

[34] Terje Aven and Seth Guikema. On the Concept and Definition of Terrorism Risk. *Risk Analysis*, 35(12):2162–2171, 2015.

[35] Terje Aven and Genserik Reniers. How to define and interpret a probability in a risk and safety setting. *Safety Science*, 51(1):223–231, 2013.

[36] Terje Aven and Ortwin Renn. The Role of Quantitative Risk Assessments for Characterizing Risk and Uncertainty and Delineating Appropriate Risk Management Options, with Special Emphasis on Terrorism Risk. *Risk Analysis*, 29(4):587–600, 2009.

[37] (American Water Works Association) AWWA. Dawn of the Replacement Era - Reinvesting in Drinking Water Infrastructure. Technical report, American Water Works Association, 2001.

[38] (American Water Works Association) AWWA. Deteriorating Buried Infrastructure Management Challenges and Strategies. Technical report, American Water Works Association, 2002.

[39] (American Water Works Association) AWWA. Buried No Longer: Confronting America's Water Infrastructure Challenge. Technical report, American Water Works Association, Denver, CO, 2011.

[40] Bilal M. Ayyub. *Risk Analysis in Engineering and Economics*. CRC Press, 2 edition, 2014.

[41] Gregory M. Baird. A game plan for aging water infrastructure. *Journal - American Water Works Association*, 102(4):74–82, 2010.

[42] Egon Balas. The prize collecting traveling salesman problem. *Networks*, 19(6):621–636, 1989.

[43] Ngandu Balekelayi and Solomon Tesfamariam. Statistical Inference of Sewer Pipe Deterioration Using Bayesian Geoadditive Regression Model. *Journal of Infrastructure Systems*, 25(3):04019021, 2019.

[44] Richard A. Becker, John M. Chambers, and Allan R. Wilks. *The New S Language*, volume 1. 1988.

[45] Massimo Bertolini, Maurizio Bevilacqua, Filippo E. Ciarapica, and G. Giacchetta. Development of Risk-Based Inspection and Maintenance procedures for an oil refinery. *Journal of Loss Prevention in the Process Industries*, 22(2):244–253, 2009.

[46] Marie Claude Besner, Michèle Prévost, and Stig Regli. Assessing the public health risk of microbial intrusion events in distribution systems: Conceptual model, available data, and challenges. *Water Research*, 45(3):961–979, 2011.

[47] Nur Afiqah Binti Haji Yahya, Negin Ashrafi, and Ali Hussein Humod. Development and Adaptability of In-Pipe Inspection Robots. *IOSR Journal of Mechanical and Civil Engineering*, 11(4):01–08, 2014.

[48] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[49] E Downey Brill, Shoou-Yuh Chang, and Lewis D Hopkins. Modeling to Generate Alternatives: The HSJ Approach and an Illustration Using a Problem in Land Use Planning. *Management Science*, 28(3):221–235, 1982.

[50] E.D. Brill, S.Y. Chang, and L.D. Hopkins. Use of Mathematical Models to Generate Alternative Solutions to Water Resources Planning Problems. *Water Resources*, 18(1):58–64, 1982.

[51] A. J. Brito and A. T. de Almeida. Multi-attribute risk assessment for risk ranking of natural gas pipelines. *Reliability Engineering and System Safety*, 94(2):187–198, 2009.

[52] Gerald G. Brown and Louis Anthony Tony Cox. How Probabilistic Risk Assessment Can Mislead Terrorism Risk Analysts. *Risk Analysis*, 31(2):196–204, 2011.

[53] Kelly Brumbelow, Jacob Torres, Seth Guikema, Elizabeth Bristow, and Lufthansa Kanta. Virtual Cities for Water Distribution and Infrastructure System Research. *World Environmental and Water Resources Congress 2007*, pages 1–7, 2007.

[54] Patricia Buckley, Lester Gunnion, and Will Sarni. The Aging Water Infrastructure: Out of sight, out of mind? Technical Report March, Deloitte University Press, 2016.

[55] Aaron Burkhart. *Lifeline Infrastructure Risk Analysis Application*. PhD thesis, University of Colorado at Colorado Springs, 2015.

[56] Thomas Y.J. Chen, Jared A. Beekman, and Seth D. Guikema. Drinking Water Distribution Systems Asset Management: Statistical Modelling of Pipe Breaks. In *Condition Assessment, Surveying, and Geomatics - Proceedings of Sessions of the Pipelines 2017 Conference*, pages 173–186, 2017.

[57] Thomas Y.J. Chen, Jared A. Beekman, Seth D. Guikema, and Sara Shashaani. Statistical Modeling in Absence of System Specific Data: Exploratory Empirical Analysis for Prediction of Water Main Breaks. *Journal of Infrastructure Systems*, 25(2):04019009, 2019.

[58] Thomas Y.J. Chen, Seth D. Guikema, and Craig M. Daly. Optimal Pipe Inspection Paths Considering Inspection Tool Limitations. *Reliability Engineering & System Safety*, 181:156–166, 2019.

[59] United State Congress. Homeland Secuirty Act of 2002, 2002.

[60] Louis Anthony Cox. Some Limitations of Risk = Threat Vulnerability Consequence for Risk Analysis of Terrorist Attacks. *Risk Analysis*, 28(6):1749–1761, 2008.

[61] Louis Anthony Cox. What's Wrong with Risk Matrices? *Risk Analysis*, 28(2):497–512, 2008.

[62] Louis Anthony Cox. Improving Risk-Based Decision Making for Terrorism Applications. *Risk Analysis*, 29(3):336–341, 2009.

[63] Susan L. Cutter. Resilience to What? Resilience for Whom? *Geographical Journal*, 182(2):110–113, 2016.

[64] Susan L. Cutter, Lindsey Barnes, Melissa Berry, Christopher Burton, Elijah Evans, Eric Tate, and Jennifer Webb. A place-based model for understanding community resilience to natural disasters. *Global Environmental Change*, 18(4):598–606, 2008.

[65] Craig Daly, Chongyang Kate Zhao, Minh Smith, and Gert Van Der Walt. Not All Data Is Created Equal: Impact on Decision Making. In *Pipelines 2016*, pages 490–505, 2016.

[66] G. C. Dandy and M. O. Engelhardt. Multi-Objective Trade-Offs between Cost and Reliability in the Replacement of Water Mains. *Journal of Water Resources Planning and Management*, 132(2):79–88, 2006.

[67] G.C. Dandy and M. Engelhardt. Optimal Scheduling of Water Pipe Replacement Using Genetic Algorithms. *Journal of Water Resources Planning and Management*, 127(4):214–223, 2001.

[68] Daniel P. De Oliveira, James H. Garrett, and Lucio Soibelman. A density-based spatial clustering approach for defining local indicators of drinking water distribution pipe breakage. *Advanced Engineering Informatics*, 25(2):380–389, 2011.

[69] Daniel P. De Oliveira, Daniel B. Neill, James H. Garrett, and Lucio Soibelman. Detection of Patterns in Water Distribution Pipe Breakage Using Spatial Scan Statistics for Point Events in a Physical Network. *Journal of Computing in Civil Engineering*, 25(1):21–30, 2011.

[70] (Department of Homeland Security) DHS. National Infrastructure Protection Plan. Technical report, Department of Homeland Security, Washington, DC, 2009.

[71] Lloyd Dixon, Robert J. Lempert, T. LaTournette, and Robert T. Reville. The Federal Role in Terrorism Insurance: Evaluating Alternatives in an Uncertain World. Technical report, 2007.

[72] You Dong and Dan M. Frangopol. Risk-informed life-cycle optimum inspection and maintenance of ship structures considering corrosion and fatigue. *Ocean Engineering*, 101:161–171, 2015.

[73] Sarah Dunn, Gaihua Fu, Sean Wilkinson, and Richard Dawson. Network theory for infrastructure systems modelling. In *Proceedings of the ICE - Engineering Sustainability*, volume 166, pages 281–292, 2013.

[74] Jack Edmonds. Matroids and the Greedy Algorithm. *Mathematical Programming*, 1(1):127–136, 1971.

[75] Muzaffar M. Eusuff and Kevin E. Lansey. Optimization of Water Distribution Network Design Using the Shuffled Frog Leaping Algorithm. *Journal of Water Resources Planning and Management*, 129(3):210–225, 2003.

[76] Natalie G. Exum, Elin Betanzo, Kellogg J. Schwab, Thomas Y.J. Chen, Seth D. Guikema, and David P.E. Harvey. Extreme Precipitation, Public Health Emergencies, and Safe Drinking Water in the USA. *Current Environmental Health Reports*, pages 1–11, 2018.

[77] B.C. Ezell. Infrastructure Vulnerability Assessment Model (I-VAM). *Risk Analysis*, 27(3):571–583, 2007.

[78] Reza Faturechi, Eyal Levenberg, and Elise Miller-Hooks. Evaluating and optimizing resilience of airport pavement networks. *Computers and Operations Research*, 43:335–348, 2014.

[79] Roger Flage and Terje Aven. Expressing and Communicating Uncertainty in Relation to Quantitative Risk Analysis. *Safety and Reliability: Methodology and Applications*, 2(13):9–18, 2009.

[80] Roger Flage, Terje Aven, Enrico Zio, and Piero Baraldi. Concerns, Challenges, and Directions of Development for the Issue of Representing Uncertainty in Risk Assessment. *Risk Analysis*, 34(7):1196–1207, 2014.

[81] Royce A. Francis, Seth D. Guikema, and Lucas Henneman. Bayesian Belief Networks for predicting drinking water distribution system pipe breaks. *Reliability Engineering and System Safety*, 130:1–11, 2014.

[82] Jerome H. Freidman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[83] T. Fwa, W. Chan, and C. Tan. Genetic-Algorithm Programming of Road Maintenance and Rehabilitation. *Journal of Transportation Engineering*, 122(3):246–253, 1996.

[84] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. Taylor & Francis, 2014.

[85] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

[86] Ian C. Goulter and Ahad Kazemi. Spatial and Temporal Groupings of Water Main Pipe Breakage in Winnipeg. *Canadian Journal of Civil Engineering*, 15(1):91–97, 1988.

[87] Vincent M. Guillaumot, Pablo L. Durango-Cohen, and Samer M. Madanat. Adaptive Optimization of Infrastructure Maintenance and Inspection Decisions under Performance Model Uncertainty. *Journal of Infrastructure Systems*, 9(4):133–139, 2003.

[88] Ahmad Habibian. Effect of Temperature Changes on Water-Main Breaks. *Journal of Infrastructure Systems*, 120(2):312–321, 1994.

[89] Yacov Y. Haimes. On the Definition of Resilience in Systems. *Risk Analysis*, 29(4):498–501, 2009.

[90] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 2009.

[91] Kjellrun Hiss Hauge, Anne Blanchard, Gisle Andersen, Ragnhild Boland, Bjørn Einar, Daniel Howell, Sonnich Meier, Erik Olsen, and Frode Vikebø. Inadequate risk assessments A study on worst-case scenarios related to petroleum exploitation in the Lofoten area. *Marine Policy*, 44:82–89, 2014.

[92] Haibo He and Edwardo A. Garcia. Learning From Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[93] Elizabeth Kemp Herrera, Aimee Flannery, and Michael Krimmer. Risk and Resilience Analysis for Highway Assets. *Transportation Research Record: Journal of the Transportation Research Board*, 2604(1):1–8, 2017.

[94] F. Owen Hoffman and Jana S. Hammonds. Propagation of Uncertainty in Risk Assessments: The Need to Distinguish Between Uncertainty Due to Lack of Knowledge and Uncertainty Due to Variability. *Risk Analysis*, 14(5):707–712, 1994.

[95] Seyedmohsen Hosseini, Kash Barker, and Jose E. Ramirez-Marquez. A review of definitions and measures of system resilience. *Reliability Engineering and System Safety*, 145:47–61, 2016.

[96] Lan Huang, Linda W. Pickle, and Barnali Das. Evaluating Spatial Methods for Investigating Global Clustering and Cluster Detection of Cancer Cases. *Statistics in Medicene*, 27(25):5111–5142, 2009.

[97] Marijke Huysman, Tamas Madarasz, and Alain Dassargues. Risk Assessment of Groundwater Pollution Using Sensitivity Analysis and Worst Case Scenario Analysis. *Environmental Geology*, 50(2):180–193, 2006.

[98] Gurobi Optimization Inc. Gurobi Optimizer Reference Manual, 2014.

[99] SCIENTECH Inc. VSAT User's Manual (Vulnerability Self-Assessment Tool). Technical report, Association of Metropoligy Seweages Agencies, 2002.

[100] Calvin D. Jaeger, Michael M. Hightower, and Teresa Torres. Evolution of Sandia's Risk Assessment Methodology for Water and Wastewater Utilities (RAM-W). In *World Environmental and Water Resources Congress 2010*, pages 3804–3010, 2010.

[101] Lindsay Jenkins, Sanjiv Gokhale, and Mark McDonald. Comparison of Pipeline Failure Prediction Models for Water Distribution Networks with Uncertain and Limited Data. *Journal of Pipeline Systems Engineering and Practice*, 6(2):04014012, 2015.

[102] Golam Kabir, Ngandu B. S. Balek, and Solomon Tesfamariam. Consequence-based framework for buried infrastructure systems: A Bayesian belief network model. *Reliability Engineering and System Safety*, 180:290–301, 2018.

[103] Konstantinos Kakoudakis, Kourosh Behzadian, Raziyeh Farmani, and David Butler. Pipeline failure prediction in water distribution networks using evolutionary polynomial regression combined with K -means clustering. *Urban Water Journal*, 14(7):737–742, 2017.

[104] Maarten J. Kallen and Jan M. Van Noortwijk. Optimal maintenance decisions under imperfect inspection. *Reliability Engineering and System Safety*, 90(2-3):177–185, 2005.

[105] Yutaka Kano and Akira Harada. Stepwise variable selection in factor analysis. *Psychometrika*, 65(1):7–22, 2000.

[106] Stanley Kaplan and B John Garrick. On The Quantitative Definition of Risk. *Risk Analysis*, 1(1):11–27, 1981.

[107] Matthias Karl, Richard F Wright, Tore F Berglen, and Bruce Denby. Worst Case Scenario Study to Assess the Environmental Impact of Amine Emissions from a CO2 Capture Plant. *International Journal of Greenhouse Gas Control*, 5(3):439–447, 2011.

[108] Y. Kawaguchi, Yun-Hui Liu Yun-Hui Liu, T. Tsubouchi, and S. Arimoto. An Efficient Algorithm Of Path Planning For An Internal Gas Pipe Inspection Robot. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2:1155–1160, 1992.

[109] David J. Kerr, Amanvir Singh, and Imran Motala. Understanding Risk and Resilience to Better Manage Water Transmission Systems. In *Pipelines 2015*, number 20, pages 1772–1785, 2015.

[110] A.J. Kettler and C. Goulter. An Analysis of Pipe Breakage in Urban Water Distribution Networks. *Canadian Journal of Civil Engineering*, 12(1):286–293, 1986.

[111] Faisal I. Khan and Mahmoud M. Haddara. Risk-based maintenance (RBM): A quantitative approach for maintenance/inspection scheduling and planning. *Journal of Loss Prevention in the Process Industries*, 16(6):561–573, 2003.

[112] Faisal I. Khan, Rahed Sadiq, and Mohamed M. Haddara. Risk-based inspection and maintenance (RBIM) Multi-attribute Decision-making with Aggregative Risk Analysis. *Process Safety and Environmental Protection*, 82(6):398–411, 2004.

[113] J.F. Kiefner and P.H. Vieth. A Modified Criterion for Evaluating the Remaining Strength of Corroded Pipe. Technical report, Battelle Columbus Div., OH (USA), 1989.

[114] J Kim, C. Baek, D. Jo, E. Kim, and M. Park. Optimal planning model for rehabilitation of water networks. *Water Science and Technology*, 4(3):133–148, 2004.

[115] S. Kirkpatric, C.D. Gelatt, and M.P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[116] Y. Kleiner and Balvant Rajani. Forecasting Variations and Trends in Water-Main Breaks. *Journal of Infrastructure Systems*, 8(4):122–131, 2002.

[117] Yehuda Kleiner and Balvant Rajani. Comprehensive Review of Structure Deterioration of Water Mains: Statistical Models. *Urban Water*, 3(3):151–164, 2001.

[118] Yehuda Kleiner and Balvant Rajani. Forecasting Variations and Trends in Water-Main Breaks. *Journal of Infrastructure Systems*, 8(4):122–131, 2002.

[119] Yehuda Kleiner and Balvant Rajani. Comparison of four models to rank failure likelihood of individual pipes. *Journal of Hydroinformatics*, 14(3):659, 2012.

[120] Maria Koliou, John W. van de Lindt, Therese P. McAllister, Bruce R. Ellingwood, Maria Dillard, and Harvey Cutler. State of the research in community resilience: progress and challenges. *Sustainable and Resilient Infrastructure*, pages 1–21, 2017.

[121] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.

[122] Frederick Krimgold. Regional Resilience and Security for Critical Infrastructure. In *Comparative Analysis of Technological and Intelligent Terrorism Impacts on Complex Technical Systems*, pages 61–68. 2012.

[123] Martin Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496, 1997.

[124] Martin Kulldorff, Richard Heffernan, Jessica Hartman, Renato Assunção, and Farzad Mostashari. A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine*, 2(3):0216–0224, 2005.

[125] Martin Kulldorff and Neville Nagarwalla. Spatial disease clusters Detection and Inference. *Statistics in Medicene*, 14:799–810, 1995.

[126] Gilbert Laporte. The Traveling Salesman Problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, 59:231–247, 1992.

[127] Sarah Larocca and Seth Guikema. A survey of network theoretic approaches for risk analysis of complex infrastructure systems. In *Vulnerability, Uncertainty, and Risk: Analysis, Modeling, and Management*, pages 155–162, 2011.

[128] Sarah Larocca, Jonas Johansson, Henrik Hassel, and Seth Guikema. Topological Performance Measures as Surrogates for Physical Flow Models for Risk and Vulnerability Analysis for Electric Power Systems. *Risk Analysis*, 35(4):608–623, 2015.

[129] Eugene L. Lawler and David E. Wood. Branch-and-bound methods: A Survey. *Operations Research*, 14(4):699–719, 1966.

[130] Mark W Lechevallier, Richard W Gullick, Mohammad R Karim, Melinda Friedman, and James E Funk. The potential for health risks from intrusion of contaminants into the distribution system from pressure transients. Technical Report 1, 2003.

[131] E J Lee and K J Schwab. Deficiencies in drinking water distribution systems in developing countries. *J Water Health*, 3(2):109–127, 2005.

[132] Lung Fei Lee. Specification error in multinomial logit models. Analysis of the omitted variable bias. *Journal of Econometrics*, 20(2):197–209, 1982.

[133] Roger J Lewis, D Ph, and West Carson Street. An introduction to classification and regression tree (CART) analysis. *2000 Annual Meeting of the Society for Academic Emergency Medicine*, (310):14, 2000.

[134] Ted G. Lewis, Rudolph P. Darken, Thomas Mackin, and Donald Dudenhoeffer. Model-based risk analysis for critical infrastructures. *WIT Transactions on State-of-the-Art in Science and Engineering*, 54:3–19, 2012.

[135] Xin Li, Wuyi Yu, Xiao Lin, and S. S. Iyengar. On optimizing autonomous pipeline inspection. *IEEE Transactions on Robotics*, 28(1):223–233, 2012.

[136] Li Liu, S Ranji Ranjithan, and G Mahinthakumar. Contamination Source Identification in Water Distribution Systems Using an Adaptive Dynamic Optimization Procedure. *Journal of Water Resources Planning and Management*, 137(April):183–192, 2011.

[137] Zheng Liu and Yehuda Kleiner. State of the art review of inspection technologies for condition assessment of water pipes. *Measurement: Journal of the International Measurement Confederation*, 46(1):1–15, 2013.

[138] Manuel López-Ibáñez, T Devi Prasad, and Ben Paechter. Ant Colony Optimization for Optimal Control of Pumps in Water Distribution Networks. *J. Water Resour. Plan. Manag.*, 134(4):337–346, 2008.

[139] Inge Lotsberg, Gudfinnur Sigurdsson, and Per Terje Wold. Probabilistic Inspection Planning of the Asgard A FPSO Hull with Respect to Fatigure. *Journal of Offshore Mechanics and Artic Engineering*, 122(2):134–140, 2000.

[140] Pure Technologies U.S. Ltd. Asset Management for Sewer Collection Systems. Technical report, Pure Technologies U.S. Ltd.

[141] C. G. Lu, D. Morton, M. H. Wu, and P. Myler. Genetic algorithm modelling and solution of inspection path planning on a coordinate measuring machine (CMM). *International Journal of Advanced Manufacturing Technology*, 15(6):409–416, 1999.

[142] Russell Lundberg and Henry H. Willis. Examining the effectiveness of risk elicitations: comparing a deliberative risk ranking to a nationally representative survey on homeland security risk. *Journal of Risk Research*, pages 1–15, 2019.

[143] Jesus Luque and Daniel Straub. Risk-based optimal inspection strategies for structural systems using dynamic Bayesian networks. *Structural Safety*, 76(June 2017):68–80, 2019.

[144] A. Mailhot, G. Pelletier, J.F. Noël, and J.P. Villeneuve. Modeling the evolution of the structural state of water pipe networks with brief recorded pipe break histories: Methodology and application. *Water Resources Research*, 36(10):3053–3062, 2000.

[145] A. Mancuso, M. Compare, A. Salo, E. Zio, and T. Laakso. Risk-based optimization of pipe inspections in large underground networks with imprecise information. *Reliability Engineering and System Safety*, 152:228–238, 2016.

[146] John C. Matthews, Ariamalar Selvakumar, and Wendy Condit. Current and Emerging Water Main Renewal Technologies. *Journal of Infrastructure Systems*, 19(2):231–241, 2012.

[147] Ram K. Mazumder, Abdullahi M. Salman, Yue Li, and Xiong Yu. Performance Evaluation of Water Distribution Systems and Asset Management. *Journal of Infrastructure Systems*, 24(3):03118001, 2018.

[148] Charles E. McCullouch. *Generalized Linear Mixed Models*, volume 7. Institute of Mathematical Statistics, 2003.

[149] Milad Memarzadeh and Matteo Pozzi. Integrated Inspection Scheduling and Maintenance Planning for Infrastructure Systems. *Computer-Aided Civil and Infrastructure Engineering*, 31(6):403–415, 2016.

[150] Miley W. Merkhofer and Ralph L. Keeney. A Multiattribute Utility Analysis of Alternative Sites for the Disposal of Nuclear Waste, 1987.

[151] David Michaud and G.E. George E Apostolakis. Methodology for Ranking the Elements of Water-Supply Networks. *Journal of Infrastructure Systems*, 12(4):230–242, 2006.

[152] B.L. Brad L Miller and David E D.E. Goldberg. Genetic Algorithms, Tournament Selection, and the Effects of Noise. *Complex Systems*, 9(3):193–212, 1995.

[153] C. E. Miller, A. W. Tucker, and R. A. Zemlin. Integer Programming Formulation of Traveling Salesman Problems. *Journal of the Association for Computing Machinery*, 7(4):326–329, 1960.

[154] Anu Mittal and Mark Gaffigan. ENERGY-WATER NEXUS: Amount of Energy Needed to Supply, Use, and Treat Water Is Location-Specific and Can Be Reduced by Certain Technologies and Approaches. Technical report, United States Government Accountability Office, Washington, DC, 2011.

[155] Torgier Moan. Reliability-based management of inspection, maintenance and repair of offshore structures. *Structure and Infrastructure Engineering*, 1(1):33–62, 2005.

[156] Idel Montalvo, Joaquin Izquierdo, Rafael Perez, and Michael M. Tung. Particle Swarm Optimization applied to the design of water supply systems. *Computers and Mathematics with Applications*, 56(3):769–776, 2008.

[157] G. Morcous and Z. Lounis. Maintenance optimization of infrastructure networks using genetic algorithms. *Automation in Construction*, 14(1):129–142, 2005.

[158] N. A. Moreira and M. Bondelind. Safe drinking water and waterborne outbreaks. *Journal of Water and Health*, 15(1):83–96, 2017.

[159] NACE International. Standard practice: In-line inspection of pipelines. Technical Report 21094, National Association of Corrosion Engineers, Houston, TX, 2010.

[160] Bruce Nestleroth, Stephanie Flamberg, Vivek Lal, Wendy Condit, John Matthews, Abraham Chen, and Lili Wang. Field Demonstration of Innovative Condition Assessment Technologies for Water Mains : Acoustic Pipe Wall Assessment , Internal Inspection , and External Inspection. Technical Report July, United States EPA, 2013.

[161] Karin Nygard, Erik Wahl, Truls Krogh, Odd Atle Tveit, Erik Bøhleng, Aage Tverdal, and Preben Aavitsland. Breaks and maintenance work in the water distribution systems and gastrointestinal illness: A cohort study. *International Journal of Epidemiology*, 36(4):873–880, 2007.

[162] USDA (United States Department of Agriculture). Soil survey geographic databse (SSURGO) data packing and use. Technical Report November, USDA, Washington, DC, 2012.

[163] Christos H. Papadimitriou and Kenneth Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Courier Corporation, 1998.

[164] J. Park, T.P. Seager, P. S. C. Rao, M. Convertino, and I. Linkov. Integrating Risk and Resilience Approaches to Catastrophe Management in Engineering Systems. *Risk Analysis*, 33(3):356–367, 2013.

[165] Elisabeth Pate-Cornell and Seth Guikema. Probabilistic Modeling of Terrorist Threats: A Systems Analysis Approach to Setting Priorities Among Countermeasures. *Military Operations Research*, 7(4):5–20, 2002.

[166] M Elizabeth Paté-Cornell. Conditional uncertainty analysis and implications for decision making: The case of WIPP. *Risk Analysis*, 19(5):995–1002, 1999.

[167] M.Elisabeth Paté-Cornell. Uncertainties in Risk Analysis: Six Levels of Treatment. *Reliability Engineering & System Safety*, 54(2):95–111, 1996.

[168] Geneviev Pelletier, Alain Mailhot, and Jean-pierre Villeneuve. Modeling Water Pipe Breaks Three Case Studies. *Journal of Water Resources Planning and Management*, 129(2):115–123, 2003.

[169] B Rajani and Y. Kleiner. Comprehensive Review of Structural Deterioration of Water Mains: Physically Based Models. *Urban Water*, 3(3):117–190, 2001.

[170] Balvant Rajani and Jon Makar. A Methodology to Estimate Remaining Service Life of Grey Cast Iron Water Mains. *Canadian Journal of Civil Engineering*, 27(6):1259–1272, 2000.

[171] Balvant Rajani and Jon Makar. A methodology to estimate remaining service life of grey cast iron water mains. *Canadian Journal of Civil Engineering*, 27(6):1259–1272, 2000.

[172] Peter D. Rogers and Neil S. Grigg. Failure Assessment Modeling to Prioritize Water Pipe Renewal: Two Case Studies. *Journal of Infrastructure Systems*, 15(3):162–171, 2009.

[173] Harry T. Roman and Bruce A. Pellegrino. Pipe Crawling Inspection Robots: An Overview. *IEEE Transactions on Energy Conversion*, 8(3):576–583, 1993.

[174] Sheldon Ross. *A First Course in Probability*. Number 8. Pearson Education Limited, 2010.

[175] Jorg Sander, Martin Ester, Hans P. Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, 2(2):169–194, 1998.

[176] Todd Sandler and Walter Enders. An Economic Perspective of Transnational Terrorism. *European Journal of Political Economy*, 20(2):301–316, 2004.

[177] Todd Sandler and Daniel G. Arce M. Terrorism & Game Theory. *Simulation & Gaming*, 34(3):319–337, 2003.

[178] Ariamalar Selvakumar and Anthony N. Tafuri. Rehabilitation of Aging Water Infrastructure Systems: Key Challenges and Issues. *Journal of Infrastructure Systems*, 18(3):202–209, 2012.

[179] U. Shamir and C. D D Howard. An analytic approach to scheduling pipe replacement. *Am. Water Works Assoc. J.*, 71(5 , May 1979):248–258, 1979.

[180] Lei Shi and Vandana P. Janeja. Anomalous window discovery through scan statistics for linear intersecting paths (SSLIP). In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 767–776, 2009.

[181] Wen-Zhong Shi, An-Shu Zhang, and On-Ki Ho. Spatial analysis of water mains failure clusters and factors: a Hong Kong case study. *Annals of GIS*, 19(2):89–97, 2013.

[182] Gh A. Shirali, I. Mohammadfam, and V. Ebrahimipour. A new method for quantitative assessment of resilience engineering by PCA and NT approach: A case study in a process industry. *Reliability Engineering and System Safety*, 119:88–94, 2013.

[183] Julie Shortridge, Terje Aven, and Seth Guikema. Risk assessment under deep uncertainty: A methodological comparison. *Reliability Engineering and System Safety*, 159:12–23, 2017.

[184] Julie E Shortridge and Seth D Guikema. Public health and pipe breaks in water distribution systems : Analysis with internet search volume as a proxy. *Water Research*, 53:26–34, 2014.

[185] Maneesh Singh and Tore Markeset. A methodology for risk-based inspection planning of oil and gas pipes based on fuzzy logic framework. *Engineering Failure Analysis*, 16(7):2098–2113, 2009.

[186] Society of Risk Analysis (SRA). Society of Risk Analysis Glossary. Technical report, Society of Risk Analysis, McLean, VA, 2015.

[187] Society of Risk Analysis (SRA). Society for Risk Analysis: Fundamental Principles. Technical Report August, Society of Risk Analysis, McLean, VA, 2018.

[188] Virginia L.M. Spiegler, Mohamed M. Naim, and Joakim Wikner. A control engineering approach to the assessment of supply chain resilience. *International Journal of Production Research*, 50(21):6162–6187, 2012.

[189] Daniel Straub and Michael H. Faber. Computational aspects of risk-based inspection planning. *Computer-Aided Civil and Infrastructure Engineering*, 21(3):179–192, 2006.

[190] Daniel Straub and Michael Havbro Faber. Risk based inspection planning for structural systems. *Structural Safety*, 27(4):335–355, 2005.

[191] Y.M. Sun, A.K.C. Wong, and M.S. Kamel. Classification of Imbalanced Data : A Review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719, 2009.

[192] George Tchobanoglous, Franklin L. Burton, and H .David Stensel. *Wastewater Engineering: Treatment and Reuse*. Metcalf & Eddy, Inc, 4 edition, 1991.

[193] Stavroula Tsitsifli, Vasilis Kanakoudis, and Ioannis Bakouros. Pipe Networks Risk Assessment Based on Survival Analysis. *Water Resources Management*, 25(14):3729–3746, 2011.

[194] Josep M. Mirats Tur and William Garthwaite. Robotic devices for water main in-pipe inspection: A survey. *Journal of Field Robotics*, 27(4):491–508, jun 2010.

[195] U.S. Census Bureau. TIGER / Line Shapefiles Technical Documentation, 2012.

[196] Thomas M. Walski and Anthony Pelliccia. Economic Analysis of Water Main Breaks. *American Water Works Association*, 74(3):140–147, 1982.

[197] Rui Wang, Weishan Dong, Yu Wang, Ke Tang, and Xin Yao. Pipe Failure Prediction: A Data Mining Method. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 1208–1218, 2013.

[198] Simon P Washington, Matthew G Karlaftis, and Fred L Mannering. *Statistical and Econometric Methods for Transportation Data Analysis Library of Congress Cataloging-in-Publication Data*. 2003.

[199] Richard White, Aaron Burkhart, Terrance Boult, and Edward Chow. Towards a Comparable Cross-Sector Risk Analysis: RAMCAP Revisited. In *Critical Infrastructure Protection X*, pages 221–238, 2016.

[200] Richard White, Randy George, Terrance Boult, and C. Edward Chow. Apples to Apples: RAMCAP and Emerging Threats to Lifeline Infrastructure. *Homeland Security Affairs 12*, 1(1), 2016.

[201] D. Wilson, Y. Filion, and I. Moore. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. *Urban Water Journal*, 14(2):173–184, 2017.

[202] Robert L. Winkler. Uncertainty in probabilistic risk assessment. *Reliability Engineering & System Safety*, 54(2-3):127–132, 1996.

[203] Andrew Wood and Barbara J. Lence. Using Water Main Break Data to Improve Asset Management for Small and Medium Utilities: District of Maple Ridge, B.C. *Journal of Infrastructure Systems*, 15(2):111–119, 2009.

[204] (Water Research Foundation) WRF. What Pipe Breaks and Leaks Reveal About Pipe Health - Fact Sheet. Technical report, 2016.

[205] (Water Research Foundation) WRF. Managing Infrastructure Risk: The Consequence of Failure for Buried Assets. Technical report, Water Research Foundation, Denver, CO, 2017.

[206] Yu Xinjie and Gen Mitsuo. *Introduction to Evolutionary Algorithms*. Springer, 2010.

[207] Shridhar Yamijala, Seth D. Guikema, and Kelly Brumbelow. Statistical models for the analysis of water distribution system pipe break data. *Reliability Engineering and System Safety*, 94(2):282–293, 2009.

[208] J. N. Yang and W. J. Trapp. Inspection frequency optimization for aircraft structures based on reliability analysis. *Journal of Aircraft*, 12(5):494–496, 1975.

[209] A. Yazdani and P. Jeffrey. Applying network theory to quantify the redundancy and structural robustness of water distribution systems. *Journal of Water Resources Planning and Management*, 138(2):153–161, 2012.

[210] Alireza Yazdani and Paul Jeffrey. Water distribution system vulnerability analysis using weighted and directed network models. *Water Resources Research*, 48(6):1–10, 2012.

[211] Man L. Yiu and Nikos Mamoulis. Clustering Objects on a Spatial Network. *SIGMOD Conference*, pages 443–454, 2004.

[212] Hamed Zamenian, Juyeong Choi, and Seyed Amir Sadeghi. Systematic approach for asset management of urban water pipeline infrastructure systems. *Built Environmental Project and Asset Management*, 7(5):506–517, 2017.

[213] Aaron C. Zecchin, Holger R. Maier, Angus R. Simpson, Michael Leonard, and John B. Nixon. Ant Colony Optimization Applied to Water Distribution System Design: Comparative Study of Five Algorithms. *Journal of Water Resources Planning and Management*, 133(1):87–92, 2007.

[214] Emily M Zechman. Agent-Based Modeling to Simulate Contamination Events and Evaluate Threat Management Strategies in Water Distribution Systems. *Risk Analysis*, 31(5):758–772, 2011.

[215] Shushang Zhu and Masao Fukushima. Worst-Case Conditional Value-at-Risk with Application to Robust Portfolio Management. *Operations Research*, 57(5):1155–1168, 2009.