METHODOLOGY TO ENHANCE THE RELIABILITY OF DRINKING WATER PIPELINE PERFORMANCE ANALYSIS

PRUTHVI S. PATEL

THESIS SUBMITTED TO THE FACULTY OF VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

CIVIL ENGINEERING

SUNIL K. SINHA, CHAIR

ROBERTO T. LEON

LEE SEARS

MAY 8TH, 2018

BLACKSBURG, VIRGINIA

KEYWORDS: RESULT RELIABILITY, INFORMATION DOMAIN ANALYSIS, ENGINEERING DOMAIN ANALYSIS, MANAGEMENT DOMAIN ANALYSIS, LOCAL LEVEL ANALYSIS, REGIONAL LEVEL ANALYSIS, GLOBAL LEVEL ANALYSIS

METHODOLOGY TO ENHANCE THE RELIABILITY OF DRINKING WATER PIPELINE PERFORMANCE ANALYSIS

PRUTHVI S. PATEL

ABSTRACT

Currently, water utilities are facing monetary crises to maintain and expand services to meet the current as well as the future demands. Standard practice in pipeline infrastructure asset management is to collect data and predict the condition of pipelines using models and tools. Water utilities want to be proactive in fixing or replacing the pipes as fixing-when-it-fails ideology leads to increased cost and can affect environmental quality and societal health.

There is a number of modeling techniques available for assessing the condition of the pipelines, but there is a massive shortage of methods to check the reliability of the results obtained using different modeling techniques. It is mainly because of the limited data one utility collects and absence of piloting of these models at various water utilities.

In general, water utilities feel confident about their in-house condition prediction and failure models but are willing to utilize a reliable methodology which can overcome the issues related to the validation of the results. This study presents the methodology that can enhance the reliability of model results for water pipeline performance analysis which can be used to parallel the output of the real system with confidence. The proposed methodology was checked using the dataset of two large water utilities and was found that it can potentially help water utilities gain confidence in their analyses results by statistically signifying the results.

METHODOLOGY TO ENHANCE THE RELIABILITY OF DRINKING WATER PIPELINE PERFORMANCE ANALYSIS

PRUTHVI S. PATEL

GENRAL AUDIENCE ABSTRACT

Water utilities are facing monetary crises to maintain and expand services to meet the current as well as the future demands. Standard practice in pipeline infrastructure asset management is to collect data and predict the condition of pipelines using models and tools. There is a number of modeling techniques available for assessing the condition of the pipelines, but there is a massive shortage of methods to check the reliability of the results obtained using different modeling techniques. This study presents the methodology that can enhance the reliability of model results for water pipeline performance analysis which can potentially help water utilities to be proactive in fixing or replacing the pipelines with confidence. Different types of analyses on the data received from the two large water utilities (name confidential) were performed to understand and check the application of the proposed methodology in the real world and was found that it can potentially help water utilities gain confidence in their analyses results by statistically signifying the results.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. Sunil K. Sinha, for his assistance and guidance throughout this research.

I am also grateful to Dr. Roberto Leon and Dr. Lee Sears for serving on my committee and their valuable suggestions during this process.

I would like to acknowledge the United States Bureau of Reclamation who provided funding for this research. Along with this, I would like to recognize Virginia Tech, Sustainable Water Infrastructure Management (SWIM) center, and ICTAS II for providing infrastructure for conducting this research.

I am deeply grateful to my parents Shailesh Patel and Priya Patel and my sister Megha Patel who always believed in me and encouraged me throughout this process.

Thank you to all my colleagues at Virginia Tech especially Aprajita Lavania, Pururaj Singh Shekhawat and Anmol Vishwakarma who worked closely with me in providing data and insightful comments.

Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Introduction

The water infrastructure system is reported to be the most invaluable part of water supply system. (Giustolisi et al. 2006). With 1 million miles of drinking water pipes across the country and 240,000 breaks occurring each year, new solutions or techniques are required to serve the growing population demand and to maintain the required level of service.

A Grade "D" has been assigned to the drinking water infrastructure by the American Society of Civil Engineers (ASCE) in its 2017 infrastructure report card. "$1 trillion investment is needed to maintain and expand service to meet demands over the next 25 years" (ASCE, 2017). Today, water utilities are facing a huge monetary crisis as the funding of these needs is limited, and a run-to-fail ideology prevails in many regions.

Models and tools are essential for the decision support for infrastructure asset management. Overall, it has is observed that water utilities feel positive about their in-house models, but there is a widespread lack of methods to check the reliability of the model results. Adopting a new methodology can provide efficient, fast and reliable decision-making tool to handle six billion gallons of treated water through leaking pipes. (ASCE, 2017).

1.2 Objectives

This research aims at providing a deep insight regarding the technical approach to enhance the overall confidence of water utilities on their risk, failure and performance models for drinking water pipelines.

To achieve this aim, the thesis will meet the following objectives:

- To formulate a methodology consisting of relevant domains and levels of analysis for enhancing the reliability of water pipelines performance analysis results.
- To justify the proposed methodology by performing analysis using data from two large water utilities.
- To suggest efficient modeling techniques that can be successfully applied to the proposed methodology.
- To explain about advanced automatic algorithms which can be used to make predictions, real-world simulations, pattern recognition and classifications of the input data on large data such as Artificial Neural Networks (ANN)

1.3 Organization of the thesis

This research consists of seven chapters as shown in Figure 1.

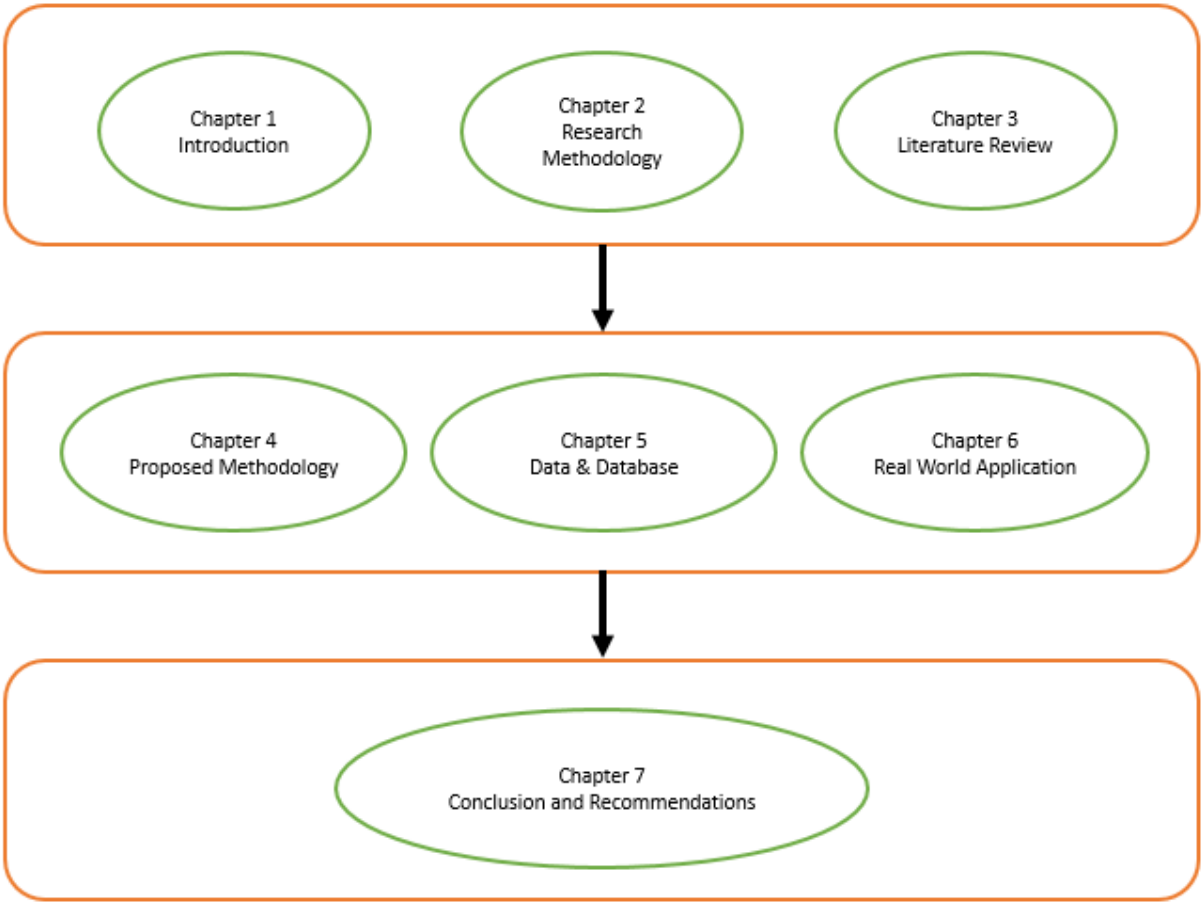

Figure 1 Organization of the thesis

- **Chapter 2** covers the research methodology.

- **Chapter 3** presents a thorough literature review on water pipe condition/failure models. Each model studied is summarized into three parts namely, model summary, model results and data requirements.

- **Chapter 4** elaborates on the proposed methodology that aims at enhancing the reliability of water pipeline performance analysis. Different types of domains and levels of analyses are discussed in detail in this chapter.

- **Chapter 5** discusses the typical problem that arises during data storage and collection and also talks about the steps involved in processing the data before it can be used for the analysis.

- **Chapter 6** presents the real world application of the proposed methodology by performing different types of analyses using the dataset of two large water utilities at different domains and levels as proposed.

- **Chapter 7** assess the overall research and provides future recommendations.

CHAPTER 2

RESEARCH METHODOLOGY

The methodology for this research consists of four steps starting with an extensive state-of-the-technology literature review, followed by classification of the analyses into three domains (Information, Engineering, and Management domain analysis) and three levels (Local, Regional and Global level analysis). After this, different types of analyses on the data received from the two large water utilities (name confidential) were performed to understand and check the application of the proposed methodology in the real world. Finally, recommendations are provided which can help utilities to become proactive in making effective decisions with a high level of precision and confidence in their results.

A detailed explanation of four steps are as follows:

Step 1: Initially a quantitative literature review was conducted by searching the keywords on the major online databases like ASCE database, Engineering Village, AWWA database to acquire academic publications from leading publishers on various failure/performance analysis models. The second part of the literature review consists of a qualitative stage where the most relevant information was synthesized to gain insight on the models which can be applied for the performance analysis of drinking water pipelines.

Step 2: This step consisted of classification of the analysis into three domains and three levels. The analysis is broadly divided into two categories (1) Quantitative analysis (data-driven mathematical models) and (2) Qualitative analysis (heuristic and empirical knowledge driven analysis).

Quantitative analysis is further classified into three domains and three levels (Figure 2). The three domains are information domain analysis, engineering domain analysis, and management domain analysis followed by three levels in each of the domain, namely, local, regional, and global level analysis. Adopting this approach towards analysis can help to complement the reliability of the results, as the model used for the analysis gets piloted at various water utilities.

Step 3: The third step comprised on performing the analysis on processed data obtained from two large water utilities to understand the credibility of the proposed methodology in the real world.

Step 4: In the final stage of this research, future recommendations are provided on different type of techniques that can be used for the proposed methodology to obtain highly accurate results.

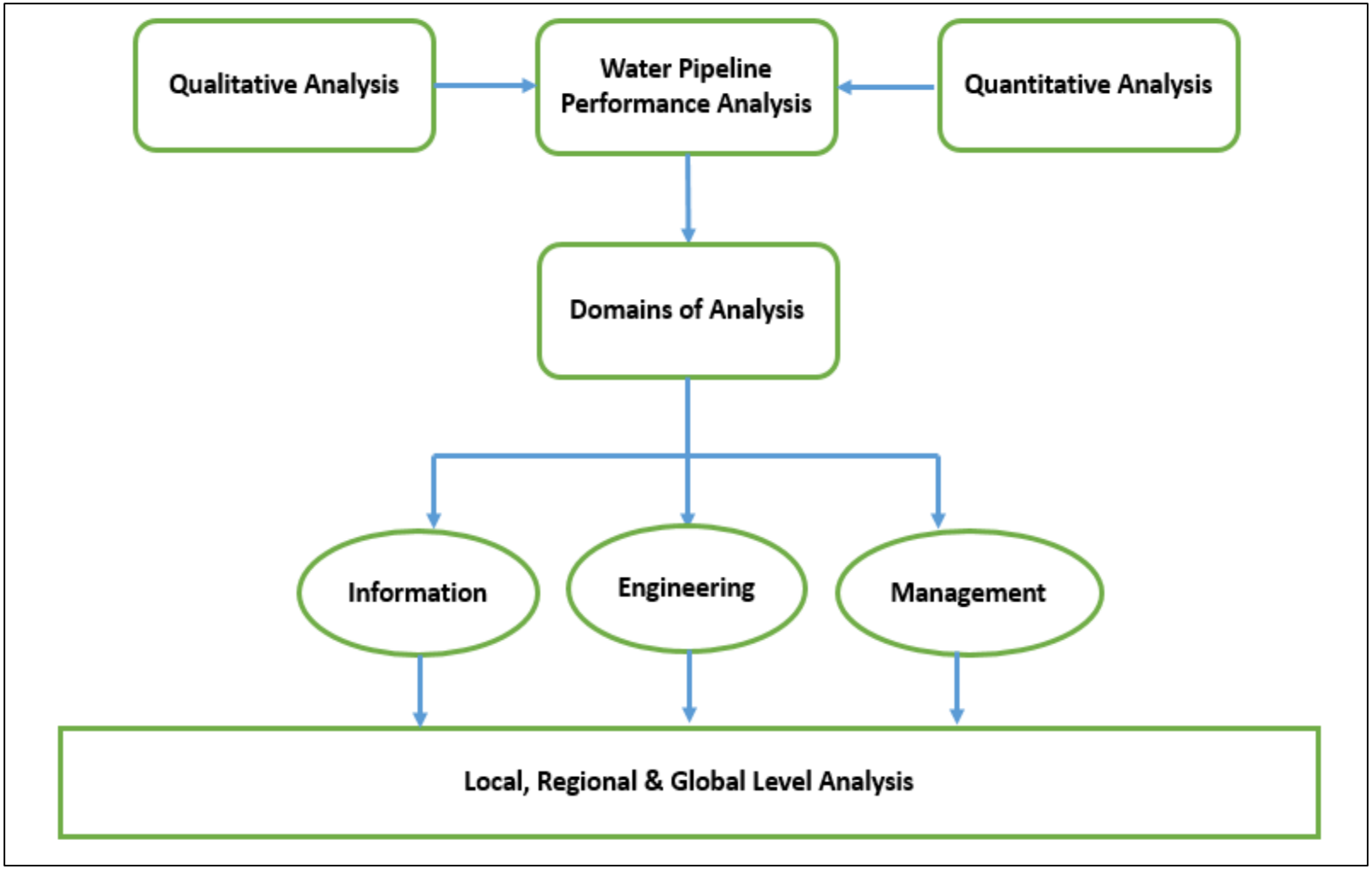

Figure 2 Classification of water pipeline performance analysis

CHAPTER 3

LITERATURE REVIEW

This chapter presents a state-of-the-technology literature review on condition curves and failure models found in the literature and utility practice.

Performance and condition prediction models for water pipes can be grouped into the following six categories:

1. Deterministic Models
2. Statistical Models
3. Probabilistic Models
4. Artificial Neural Networks
5. Fuzzy Logic Models
6. Heuristic Models

The following sections provide a comprehensive overview of different types of models developed in the above-mentioned categories.

3.1 Deterministic Models

Deterministic models are commonly used where the relationships between the components are certain. These kind of models are comparatively simple as compared to other categories of models.

Deterministic Model Approach

- Regression analysis, combined mechanistic-empirical analysis, and opinions from experienced engineers are the major procedures that have been used for developing existing deterministic models.
- Deterministic models mainly require the use of laboratory tests and specimen to obtain relationship information.
- Deterministic models are not suitable to account for nonlinearity in the behavior of soil and pipe materials.
- In infrastructure management, these models are generally used in defect propagation models such as corrosion propagation.

3.1.1 Review of Articles on Deterministic Models

Model Summary: (Rajani & Makar, 2000) describes a procedure to estimate the remaining service life of cast iron (CI) water mains by analyzing the effect of corrosion pit on structural

capacity of the pipe. Prediction on the factor of safety (FS) are made and then checked against the minimum threshold value by calculating residual resistance capacity, anticipated corrosion rates, and corrosion pit measurements. Correlation is then established between the factor of safety and remaining service life.

Model Results: A case study was presented to validate the proposed methodology. Results showed that large diameter pipes are more prone to breaks as compared to smaller pipes.

Data Requirements:

Pipe Diameter	Water Temperature
Thickness of Pipe	Frost Load Factor
Backfill Material	Traffic Reduction Factor
Soil Information	Bursting Tensile Strength
Load Factor	Ring Rupture Modulus
Pressure Surge Information	Fracture Toughness
Pressure Allowance	Manufactured Year
Age	Trench Depth

Model Summary: (Babovic et al. 2002) presented advanced data mining method to determine the risk of burst pipes using Bayesian networks. Purpose of developing risk assessment model is to study factors responsible for an increased risk of pipe failure. Bayesian networks also account for uncertainties.

Model Results: Method presented is applied using a Copenhagen Water case study. After performing sensitivity analysis, it is found that previous bursts are the main predictors among all the factors taken into consideration for the study. It is recommended to investigate the Bayesian network to check the reliability of the model.

Data Requirements:

Temperature Data	Pipe Material
Soil Type	Pipe Diameter
Previous Break Records	Thickness of Pipe
Age	Age when last repaired

3.2 Statistical Models

Statistical modeling is commonly used to predict the likelihood of occurrence of an event or the lifetime. The long-term historical database is required about the condition, and attributes of pipes are necessary to generate an expected outcome.

Statistical Model Approach

- Quality pipe infrastructure historical data are required to predict the future condition.
- The current condition of the asset is required to predict future condition.
- Processing through regression analysis is one of the common features of different types of statistical models used, but this approach is again limited if the richer data set is absent.

3.2.1 Review of Articles on Statistical Models

Model Summary: (Poulton et al. 2007) calculated the impact of pipe segment length on break predictions in water mains. Linearly Extended Yule Process (LEYP) was used to find break predictions for each segment. Intensity functions are used for the calculations that depend on age, the number of previous events and the covariates. Intensity function used includes the influence of previous events, wherein Weibull model represented the influence of age and Cox Proportional Hazard model represented the influence of the covariates.

Model Results: Models were verified based on the case study with data collected from Veolia Water in France. As per the results, the model is not sensitive to small pipe segments.

Data Requirements:

Pipe Diameter	Installation Year
Pipe Length	Type of Incident
Surface Type	Traffic Level
Water Pressure	Intervention Date
Joint Type	Soil Type

Model Summary: (Kleiner & Rajani, 2008) prioritized water mains for renewal using non-homogeneous Poisson model. Parameters were classified based on three classes: pipe dependent, time-dependent and pipe and time-dependent. Lipschitz Global Optimizer (LGO) algorithm was first used to train a model. Model validation was done by comparing the forecasted number of breaks with the observed failures.

Model Results: Case study was presented using the data from water utility in Canada. The model was found to be appropriate based on the goodness-of-fit test.

Data Requirements:

Pipe Material	Pipe Diameter
Installation Year	Pipe Length
X-Y coordinates of pipe nodes	Break Type

3.3 Probabilistic Models

Probabilistic modeling analyzes the probability of an event occurring. The likelihood of an event to happen is denoted by 1 and not to happen is denoted by 0.

Probabilistic Model Approach

- Probabilistic models require extensive data.
- Statistical models relying on a probabilistic relationship between the parameters are mostly used.
- Probabilistic Models handles inherited uncertainties very efficiently.
- Commonly used in pavement, bridge industry.

3.3.1 Review of Articles

Model Summary: (Davis et al. 2007) developed a failure rate prediction model for polyvinyl chloride (PVC) pipelines. The model uses internal defect data to determine the failure rates. Linear Elastic Fracture Mechanics (LEFM) theory is used to determine the time to brittle fracture. Monte Carlo simulation is used to approximate the lifetime probability distribution, and Weibull hazard function is used to estimate the failure rates.

Model Results: Model results were compared with the failure data form 17 UK water utilities. Predicted failure rates and observed failure rates resembled closely to each other.

Data Requirements:

Monte Carlo Simulation Parameters	Weibull scale parameters
No of Pipe Segments	Young's Modulus
Pipe Lengths	Visco-Elasticity
Simulation Time	Pipe Radius
Incremental Time Period	Internal Pressure
Fracture Toughness	Soil depth, Soil Unit Weight, and Modulus
Crack Growth Parameters	Residual Hoop Stress

3.4 Artificial Neural Networks

Artificial Neural Networks (ANN) is a method used to model pipe failure and condition rating of the pipeline system. Neural network comprises many elements called "neurons." Each neuron becomes very complicated when interconnected with each other.

ANN Model Approach

- High level of skills and training is required to develop these complex networks.

- "Quality labeled data are required for supervised training and predicting the future condition" (St. Clair & Sinha, 2013).

3.4.1 Review of Articles

Model Summary: (Zangenehmadar, 2016) used the artificial neural network (ANN) to estimate the remaining useful life of the pipelines. A different number of neurons, hidden layers and three training algorithms (Levenberg-Marquardt (LM), Bayesian Regularization (BR), and Scaled Conjugate Gradient (SCG)) were used for training the data. Generalized Regression Neural Network (GRNN) was also used and the results obtained were compared with the ANN results.

Model Results: It was observed that ANN predicted data more accurately than GRNN. GRNN started overestimating the reaming useful life (RUL) of pipelines after a certain age (age more than 70 years) and underestimated the (RUL) at an early age of pipelines. LM algorithm was recommended out of all the three algorithms for similar kind of study.

Data Requirements:

Pipe Diameter	Pipe Length
Breakage Rate	Pipe Material

3.5 Fuzzy Logic Models

Fuzzy logic is a mathematical method which is made up of membership functions that can account for uncertainty, ambiguities in an event.

Fuzzy Logic Model Approach

- Constructing fuzzy rule sets are challenging.
- Expert opinions are required in building membership functions.
- If-then rule statements are used to describe complex systems (Sivanandam, Sumathi, & Deepa, 2007).

3.5.1 Review of Articles

Model Summary: (St. Clair & Sinha, 2013) developed novel performance index for drinking water pipelines using Fuzzy interference methodology. Altogether, 27 parameters were used for this model. The fuzzy model can handle a number of input and output variables very easily as the model is made up of many membership functions which accounts for many uncertainties in an event.

Model Results: Model was tested using the data from water utilities in the USA, and it was found that fuzzy model accurately predicted the performance of the pipelines utilizing a number of if-then statements.

Data Requirements:

Pipeline Age	R/R type
Design Life	Pressure Class Exceeded
Vintage	Pressure Surges
Rehab	Adequate Fire Flow
C Factor	Pressure Complaints
Remaining Thickness	Discolored Water
Pipe Break and Leak records	Disturbances
Defect Type	Flooding
Live Load	Cathodic Protection
Material Type	Stray Currents
Dissimilar Metals	Soil Corrosivity

3.6 Heuristic Models

These models are mainly developed based on opinions from experts. Heuristic models are widely used for the problems that are not well understood.

Heuristic Model Approach

- Expert opinions are captured using the step-wise procedure.

3.6.1 Review of Articles

Model Summary: (Al-Barqawi & Zayed, 2006) proposed a condition assessment model using analytic hierarchy process. The process consists of eight steps. The first step in the process is to select the factors which affect the condition of water main. Priority vector and pair-wise comparison matrices are then developed for the mains. Weights are selected after verifying the consistency of the pair-wise comparison matrices using consistency analysis. Condition assessment value is then assigned based on the combination of different priority matrices.

Model Results: Condition assessment values ranged from 0 to 10. 10 is the excellent condition of the pipes whereas 0 being the most critical pipe. The case study was utilized to demonstrate the application of the model in the real world. The model provided the acceptable results concluding that the pipe age has the highest effect on the condition assessment ratings.

Data Requirements:

Type of Soil	Type of Road
Type of Service	Ground Water Level
Pipe Diameter	Pipe Material
Pipe Age	Breakage Rate
Hazen Williams C Factor	Cathodic Protection
Operational Pressure	

The following points can be concluded from the literature review:

1. There is a number of modeling techniques available for assessing the condition of the pipelines.

2. A large number of parameters are needed to estimate the pipe performance.

3. There is a shortage of methods to check the reliability of the results obtained using different modeling techniques.

CHAPTER 4

PROPOSED METHODOLOGY TO ENHANCE THE RELIABILITY OF WATER PIPELINE PERFORMANCE ANALYSIS

This chapter discusses each domain and each level of analysis proposed in the methodology in detail. As already discussed in chapter 2 of the thesis, the analysis is broadly classified into two categories.

(1) Quantitative Analysis
(2) Qualitative Analysis

4.1 Qualitative Analysis

This category of analysis mainly relies on suggestions provided by domain experts. It helps to develop hypotheses for quantitative study. Some conventional methods include group discussions, individual interviews, and surveys. Respondents are selected to fulfill a given questionnaire, and usually, the sample size is comparatively smaller.

4.2 Quantitative Analysis

Quantitative analysis refers to analysis that aims to understand and predict the behavior of the event through the use calculations, different modeling techniques, and real data.

Two types of analyses are further classified into three domains & levels as shown in Figure 2. All the three domains and levels of analysis are explained in detail in the current section, whereas suitable real-world examples corresponding to each domain and level is presented in chapter 6.

4.3 Domains of Analysis

The three domains of analysis are:

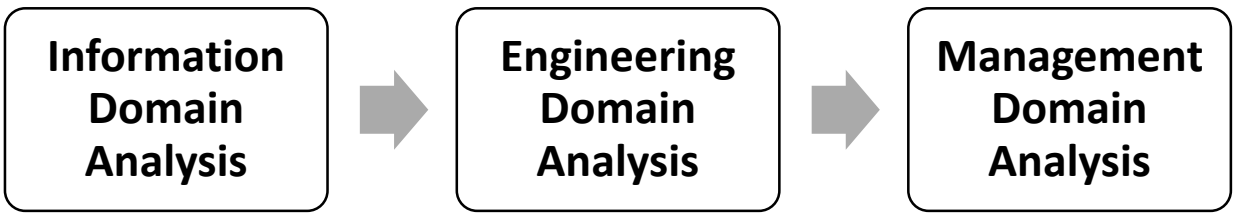

Figure 3 Domains of Analysis

4.3.1 Information Domain Analysis

The information level analysis provides support for high-level information regarding the trends in the system. This category is a high-level data mining method to get a general understanding of the current and past trends in the drinking water system.

For example, different types of pipe material can be plotted against the year in which they were installed to know about the installation trend of the network. Pipe breaks vs. corresponding months in which the breaks took place can show the effect of weather pattern on pipe performance. A number of breaks and leaks can be plotted against the diameter of pipes to get the idea about the trend of breakage distribution in the system (Figure 4). Also, many different types of hypotheses can be formulated using information analysis which can be tested using engineering analysis. The applications for the information domain analysis can be straightforward like calculating percentages of different pipe materials in a system to finding out one to one relation between the variables using regression analysis.

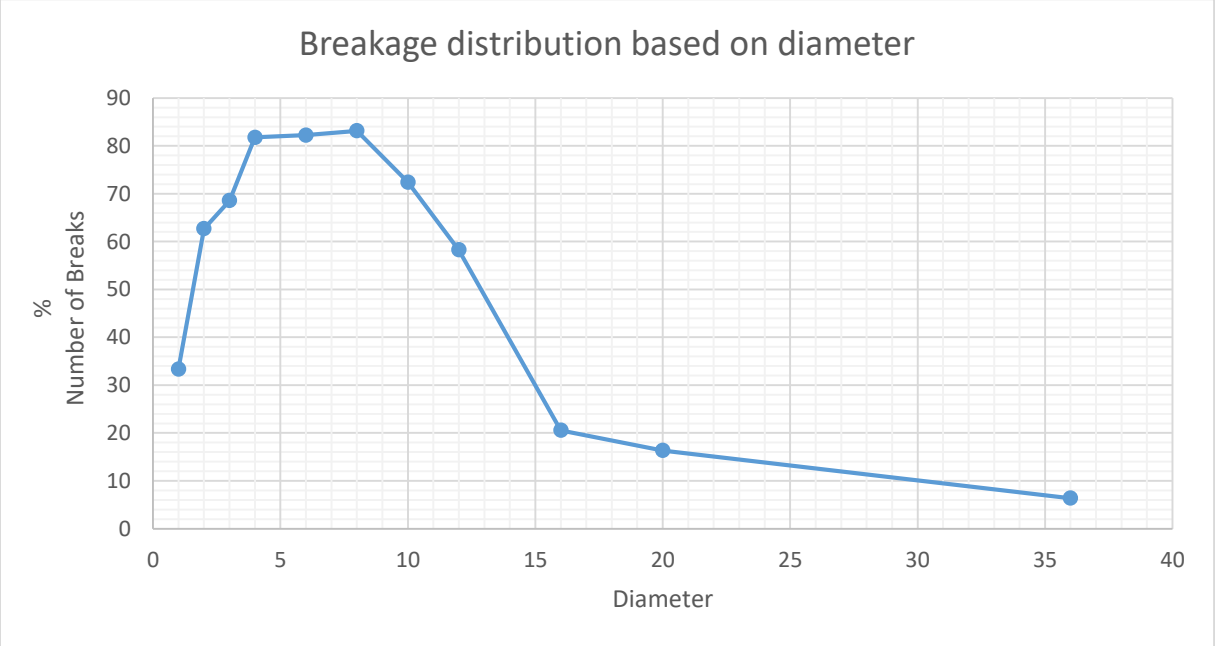

Figure 4 % Number of Breaks vs. Diameter (Breakage Distribution in the network of utility A)

4.3.2 Engineering Domain Analysis

Models in the area of performance assessment, risk assessment, performance prediction, condition prediction, failure prediction fall in the engineering categories. This analysis includes technical and engineering related issues. The applications for the engineering domain analysis can vary from weighted factor models to very advanced machine learning models such as Artificial Neural

Networks (ANN). An example of the analysis results obtained from weighted factor model and ANN model is compared and plotted as shown in Figure 5.

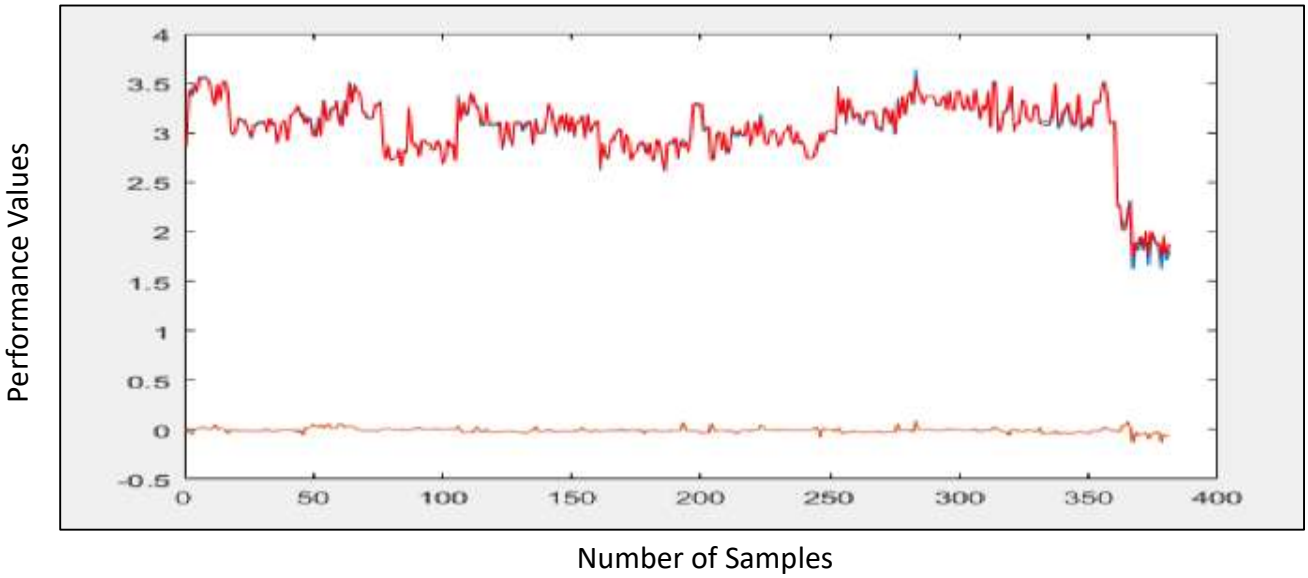

Figure 5 Observed Values vs. Predicted Performance Values (ANN results)

Figure 5 represents the graph of evaluated performance rating values (blue in color, ratings obtained from weighted factor performance model) vs predicted performance rating values (red in color, obtained performance rating results from ANN model) along with error values (Error Values = Predicted Values – Observed value, Yellow line - at the bottom of the graph). Weighted factor performance model and ANN model are discussed in detail in chapter 6 of the thesis. Table 1 illustrates the values obtained from both the models for a random sample from the entire dataset.

Table 1 Weighted factor performance model results vs. ANN model results

WEIGHTED FACTOR PERFORMANCE MODEL	ANN RESULTS	Error
3.12	3.1178	0.0020
3.01	3.0008	0.0086
3.14	3.1492	-0.0053
3.07	3.0857	-0.0122
3.09	3.1059	-0.0170
2.89	2.9084	-0.0195
3.01	3.0107	-0.0027
3.12	3.1178	0.0020
3.12	3.1178	0.0020
2.92	2.9201	-0.0003
3.17	3.1714	-0.0052
3.18	3.1905	-0.0088
3.22	3.2173	0.0003
3.24	3.272	-0.0353
3.17	3.1714	-0.0052
3.14	3.1997	-0.0644
3.10	3.0879	0.0087
3.16	3.1357	0.0228
3.14	3.1114	0.0316

4.3.3 Management Domain Analysis

Management level analysis aims at prioritization decisions such as where, when and how the budget should be spent. The applications for the management domain analysis can vary from predicting the future monetary demands based on the current condition of the pipelines in the network to estimating the optimal time to replace/rehabilitate the pipelines in the system. Figure 6 graphically illustrates the results of current and future needs regarding cost for the dataset from utility A concerning the diameter of pipelines installed in the system.

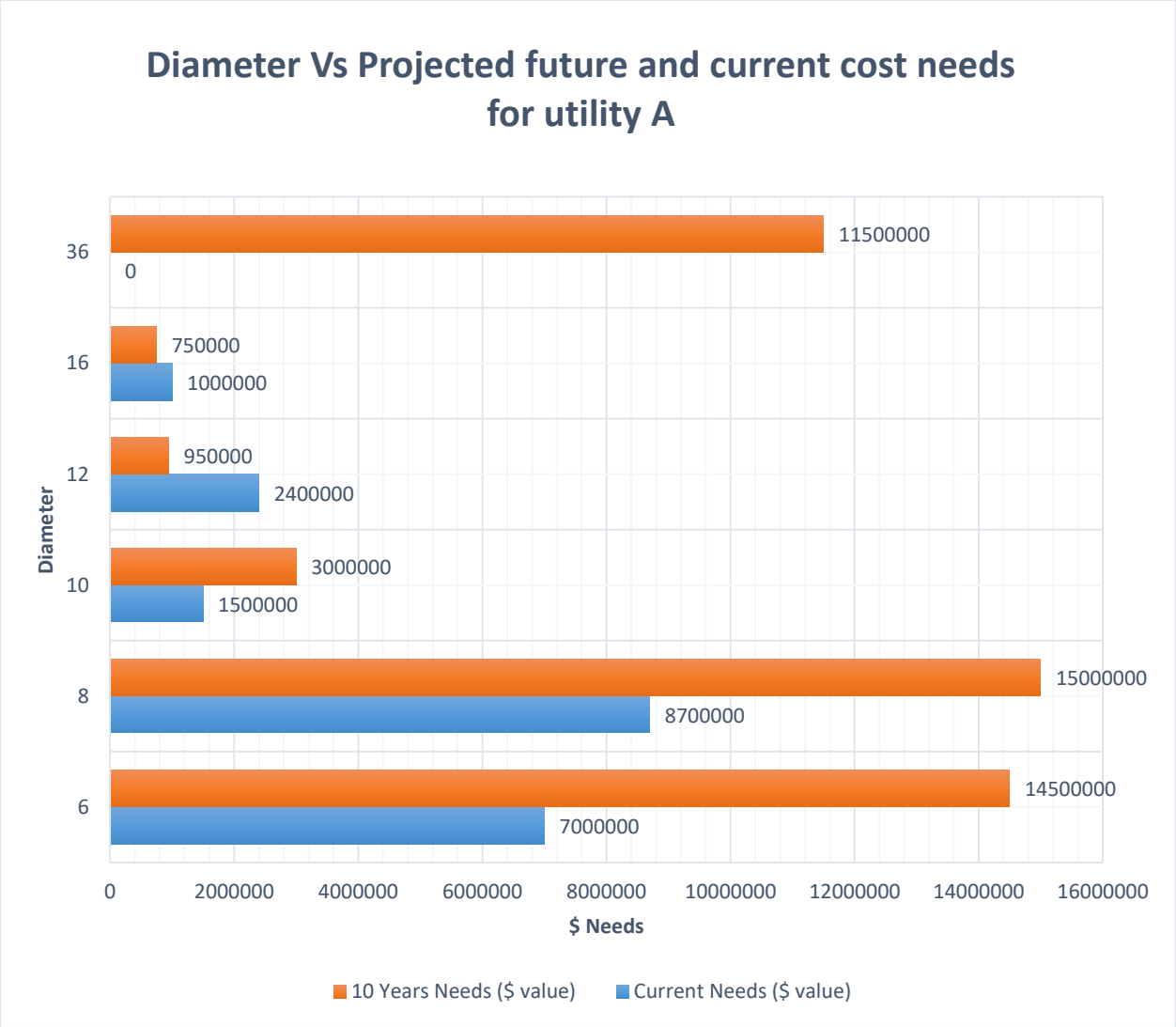

Figure 6 Predicted $ needs for utility A w.r.t. diameter of pipelines

4.4 Levels of Analysis

Each domain of analysis is further classified into three different levels which are explained in detail as follows:

4.4.1 Local Level Analysis

Analysis of one water utility is defined as local level analysis. All the domain analysis can be performed at the local level; however, the reliability of the results is low due to the lack of the time-dependent data, and also data from one utility is very limited. Hence, it becomes difficult to defend the formulated hypothesis statistically.

4.4.2 Regional Level Analysis

Regional analysis is the analysis of particular region and cohort. Cohorts can be made as per geography (Figure 7), temperature (Figure 8), soil, different types of pipe performance characteristics, etc.

Benefits of Cohort Analysis:

Cohort analysis breaks the data into related small groups rather than looking at the entire data as one unit. These cohorts, usually share common characteristics or experiences. Cohort analysis allows seeing patterns that particular dataset undergoes. This method can be very useful for this study in order to describe an aggregate of drinking water pipeline performance analysis results having in common a significant event in their lifetime, such as effect of soil corrosion on lined/unlined metallic pipelines, study on whether specific joints are falling apart in certain conditions or not, impact of loading on pipelines, etc.

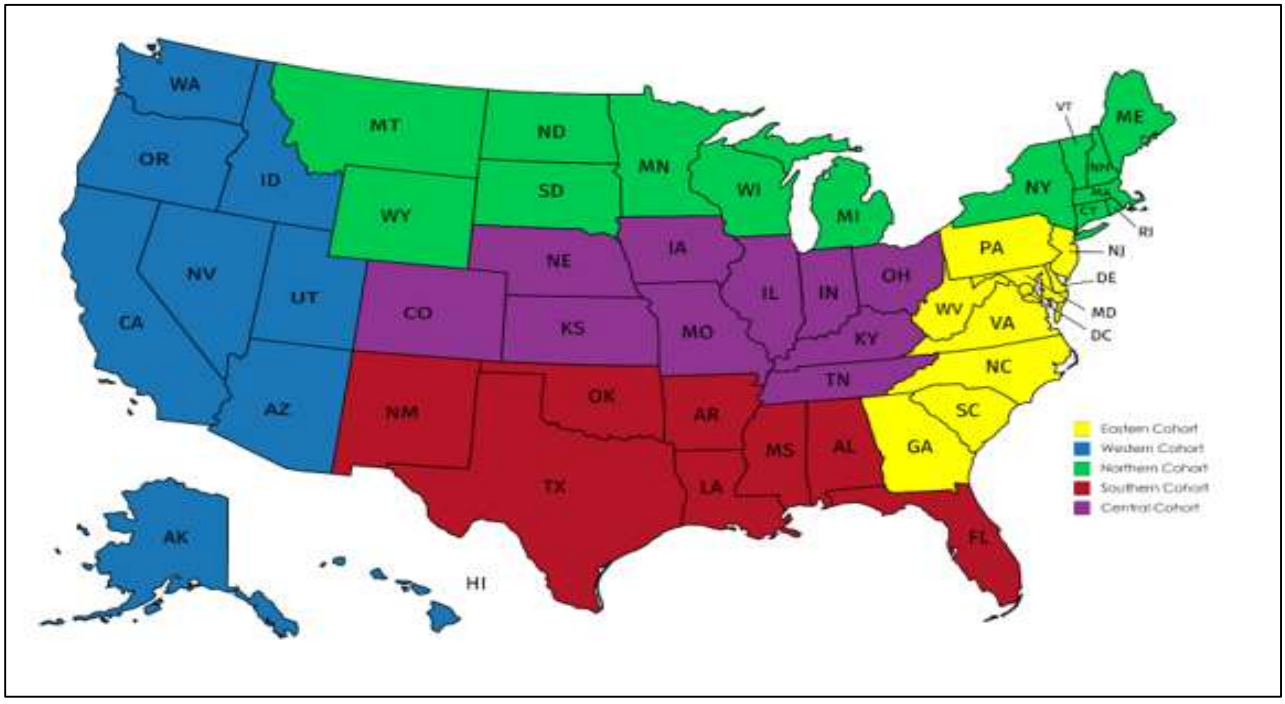

Figure 7 Cohorts formed as per geography

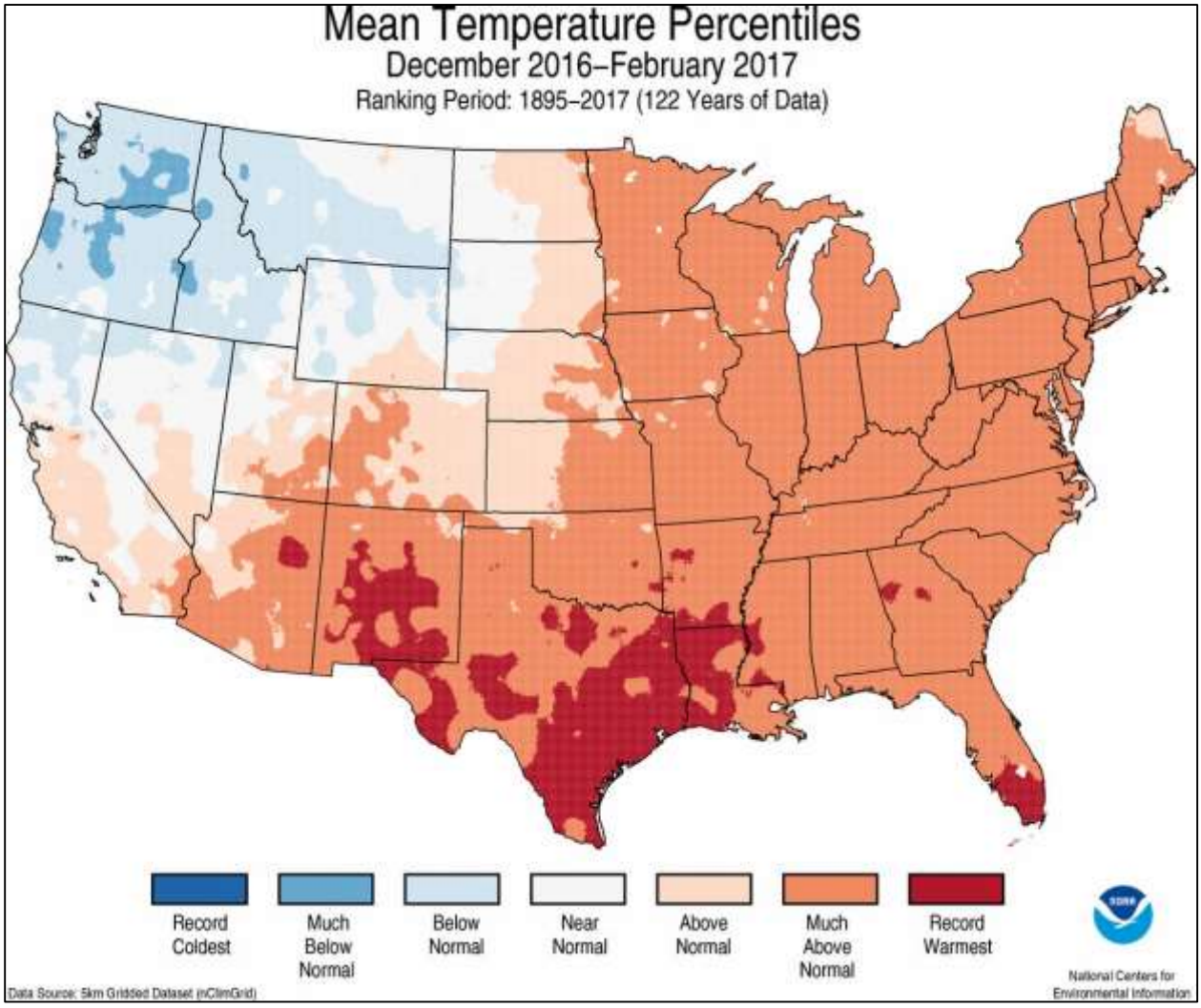

Figure 8 Cohorts composed as per mean temperature data (National Centers for Environmental Information 2017)

It is crucial to make sure that data and sample is an accurate representation of different pipe materials, failure type, temperatures, loading conditions, environmental conditions and other factors. As shown in Figure 7, cohorts are formulated as per geography, but it is not necessarily the actual representation of the proposed methodology. Ideally, cohorts should be based on similar type of characteristics observed within the entire dataset. For example, two places located at different geographical locations can form one cohort by soil characteristics. Cohort analysis will help utilities to enhance their model/tools by learning from the other utilities which are identified in the same cohort or dealing with similar problems in their network.

4.4.3 Global Level Analysis

The primary purpose of the global level analysis is to identify the high-level trends across the country and around the world. For example, in the leadite joint, the difference in the coefficient of thermal expansion between the joint and the pipe material appears to produce pipe failures. These type of trends can be observed all over the country irrespective of any place or cohort and can be categorized in the global level analysis.

CHAPTER 5

DATA AND DATABASE

Data quality and its representation is the first step before running any analysis. Analyzing the data that has not been carefully processed or cleaned can produce misleading results. Machine learning models largely depend on the data pre-processing. Data pre-processing requires in-depth knowledge of the data. A large number of variables are required to conduct the performance analysis of the pipelines. Each parameter or variable also contains uniqueness in it. Such kind of variables such as vintage, diameter, pipe type, breaks, etc. can be termed as a function of analysis as these variables can largely influence the results of the analysis.

5.1 Functions of Analysis

- Breaks: Recent pipe breaks are often considered more hazardous than past pipe breaks.
- Vintage: Pipe vintage determines metallurgy, the thickness of the pipe, available diameters, and various information of the different pipe materials. For example, pit cast method was predominantly used casting method before 1930 which got changed to centrifugally spun cast.
- Diameter: Small diameter pipelines fail in the entirely different mode as compared to large diameter pipelines. For example, the dominant mode of failure in small diameter metallic pipelines is circumferential cracks and bell splitting whereas, the major mode of failure in large diameter pipelines is longitudinal cracks and bell shearing.
- Pipe Type: The manufacturing process of the pipeline determines the resilience of the strength. Also, different materials have a varied design life, thickness, etc.

It is essential to understand the subdivisions in each parameter and should be appropriately incorporated as the input in the models. Weights and scores should be adequately adjusted to account for the complexity of each parameter.

5.2 Challenges in Data Collection

Each and every utility collects the data based on their knowledge and needs. There is no standardized specification of what data to collect during event. Ideally, the initial step should be collection of data on the frequency and nature of failures, preferably by material type, age and size. However, the definition of failure is not explicitly defined in this industry. As a result, it is difficult to distinguish between leaks and major ruptures. (Oxenford et al. 2012). Minor failures are often treated as maintenance events whereas major failures are handled by distinct investigations. Leaking service connections, small cracks, corrosion holes are often categorized as minor failures. These do not normally require meticulous investigations and are classified as burst events by the

most of the utilities. Utilities typically collect only basic data at sites and investigations carried for such small events are performed at very basic level. Also, there is lack of integration of failure data with enterprise databases.

5.3 Storage of Failure Information in Utility Records

Data in the utility is assembled from many departments such as:

1. Maintenance Department
2. Planning Department
3. Finance Department
4. Engineering Department

The role of engineering department is to identify shortcomings, analyze the collected data, develop reliable and robust models for serving growing population demands and maintain the required level of service. Each department collects separate sets of data. It is important to serve engineering department with correct set of failure data even though the failure dataset is not collected by them to enable it to make more informed, repair, renewal, and rehabilitation decisions.

Management of failure data in enterprise systems can be challenging because of the massive amounts of the data they include such as:

1. Pipe Identification Number
2. Failure Identification Number
3. Location
4. Failure Date
5. Failure Location
6. Failure Orientation
7. Failure Type
8. Breakage History
9. Ambient Temperatures
10. Water Temperature
11. Pipe Depth
12. Water pH
13. Nominal Pipe Diameter
14. Pipe Length
15. Soil Type
16. Moisture Content
17. Soil Redox Potential
18. Soil Resistivity
19. External Coating Status
20. Minimum and Maximum Wall Thickness
21. Other

Failure data is usually recorded in Geographic Information Systems (GIS) where they are a subset of pipelines data. All the attributes of the failure parameters must be combined with the pipeline dataset. This pool of data should be then provided to the engineering department for the analysis.

Most of the utilities have started collecting the failure attributes like diameter of pipeline, type of pipe, type of joint, depth of pipeline, installation year of the pipe, geographic location of the pipeline, date of the failure, information whether there is a minor leak or major break, failure type, cause of failure. From the current practice, it is observed that many of the water utilities have taken a step forward in collecting the failure data and are understanding its importance.

One of the problems that are observed in data collection practice of utility is that when another pipe segment replaces the pipeline then all the data already tagged with the old pipeline is lost from the current layer of the GIS and the information of new pipeline with new identification number gets recorded. It is necessary for the engineering department to have the information of the old pipeline for high-level, detailed analysis. It is highly recommended to record information of both, i.e., new and old pipe as the subset of the pipeline data.

Also, recording cause of failure in failure data records is highly useful for the analysis. It is important to record information accurately as many hypotheses and assumptions rely on that particular data. However, it is observed that many times the data collection person lacks the technical knowledge and fails to understand and document the difference between pipe failure and joint failure. It is highly suggested for the utilities to have two different set of attributes, i.e., joint failure and pipe failure cause so that it becomes easy for the technician to record the data and also for the engineer to perform the good analysis.

5.4 Data Preprocessing

Even before the data is used for the analysis, various steps are required to be performed on the dataset to convert into a meaningful format. Real-world data is often incomplete, inconsistent, and is probable to contain errors. Also as already discussed there is no common data standardization in the industry which leads to many problems even for running a same model for the different utility data. Some of the most common steps involved in pre-processing the data are presented in Figure 9.

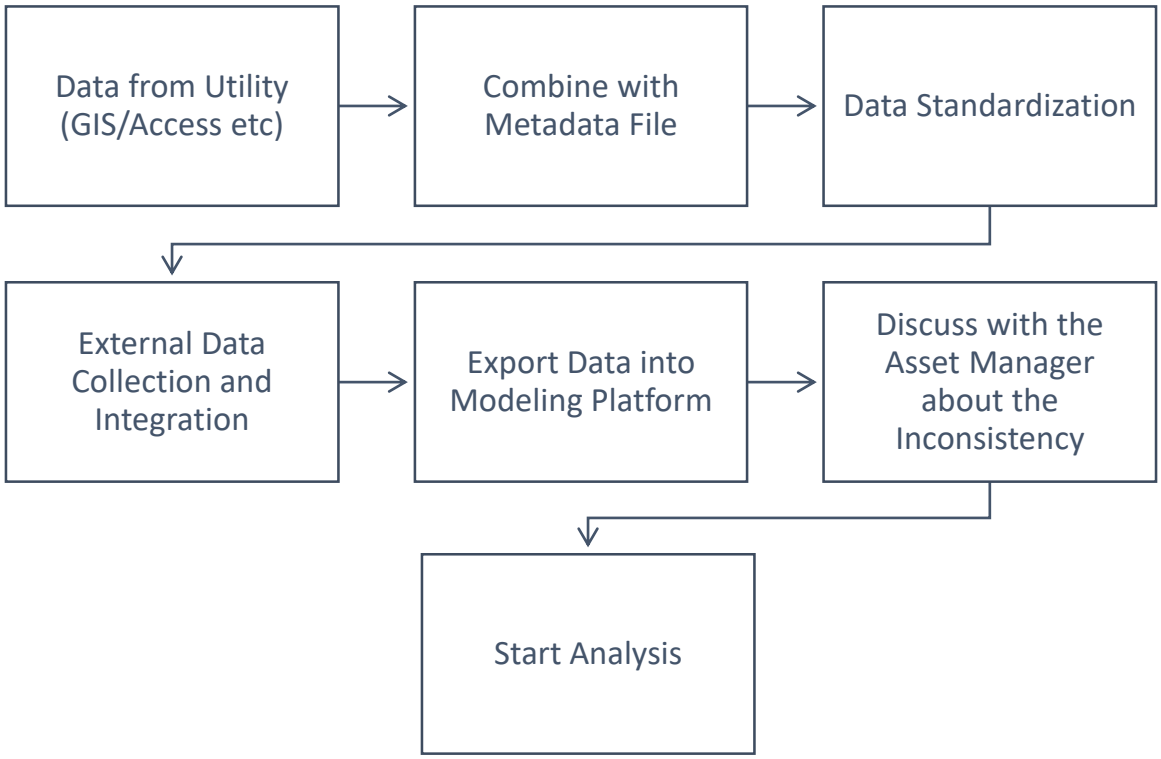

Figure 9 Steps involved in Data Preprocessing

Step 1) Data from utility: Utility data can be in many formats. The most common ones include Microsoft Excel, PDF documents, Microsoft Access, and GIS. Virginia Tech enterprise geodatabase was used to store and host GIS data pertinent to this research.

Step 2) Combining data file with metadata file: Data received from the water utilities contains many fields. However, not every field will be useful for the analysis. Also, some of the data columns will contain default values and default titles. This information is generally stored in a file called metadata file by most of the utility. It is vital to append all the information from the metadata file with the data file from the utility.

Step 3) Data Standardization or Data Mapping/Translation: Utilities store data in heterogeneous formats based on their organizational requirements. Pipeline materials like material, age, diameter, installation year, etc. show a great level of dissimilarity when compared with data collection practice of multiple utilities. Moreover, a number of parameters available across different utilities vary significantly. These demands for a standardized protocol to collect and code data values in common format for quality analysis and visualization. For this research, a personalized data model for drinking water pipeline infrastructure created by SWIM research team at Virginia Tech has been used.

Step 4) External data collection and integration*:* It is quite evident that parameters like soil type, traffic loading, land cover, climatic conditions, and stray currents have a strong influence on the performance of the pipelines. There are many open source platforms for data collection such as USGS, SSURGO, Tiger Files, and NOAA to name a few. Files obtained from these sources are spatially overlaid with the utility database in GIS. This rich pool of database can support good analysis.

Step 5) Export data into modeling platform: After having coupled the utility database with external parameters, a new set of data is then exported to the modeling platform. Modeling platform is generally selected based on the type of the analysis to be performed. Some conventional data analysis and modeling platform used are MATLAB, Microsoft Excel, KNIME, and ESRI's Web App Builder.

Step 6) Discussing the dataset with utility's asset manager: The last step before running analysis is to discuss the data with the asset manager. Many times, it is possible that data can be incomplete or inconsistent. Removing those set of data from the datasets reduces the sample size which leads to less statistically significant results. However, utility asset manager can always provide a good estimation about the data to be assumed and also about the fields to avoid from the analysis.

Step 7) Analysis: First step in the analysis should start with information domain analysis or trend analysis. The applications for the information domain analysis can be elementary like calculating percentages of different pipe materials in a system to finding out one to one relation between the variables using regression analysis. A considerable amount of time is utilized even for plotting simple graphs for large dataset. MATLAB script was written to deal with such large dataset (attached in Appendix, A1 – Preliminary analysis code). This script can plot simple column and pie charts for parameters such as age, material, number of breaks, and diameter. Also, the script can run correlation and simple linear regression (one-one relation). Engineering domain analysis is the next step after the data preprocessing and preliminary analysis. It is discussed in detail in the next chapter. Three step process for analysis is recommended for obtaining highly accurate results.

CHAPTER 6

REAL-WORLD APPLICATION OF PROPOSED METHODOLOGY USING DATASET FROM WATER UTILITIES

This chapter discusses the benefits of proposed methodology with suitable examples. Next step after data preprocessing is the analysis of the dataset at various domains and levels. All the analysis performed is done using a dataset of two water utilities. Only the cast iron pipelines from both the datasets are taken into consideration.

6.1 Information Domain Analysis

The information domain analysis provides support for high-level information regarding the trends and hotspots in the system. Many different types of hypotheses can be formulated using information domain analysis which can be later tested using engineering analysis to discover the scientific reasons behind the process.

The influence of weather pattern on the number of breaks in metallic pipelines is one of the hypotheses that were formulated. It is observed that number of breaks increases rapidly as the temperature decreases. This hypothesis is tested using the proposed methodology at both the levels, i.e., local and regional level.

6.1.1 Local Level Analysis Results

A simple column chart is plotted between the number of breaks and corresponding months in which the breaks were recorded. All the pipelines considered in this analysis are the cast-iron pipes. The result of the analysis performed on the utility A dataset is shown in Figure 10.

As observed from the graph, cold winter months (December-March) have a strong influence on the number of breaks recorded.

6.1.2 Regional Level Analysis Results

A similar trend was observed (as shown in Figure 11) from the dataset of Utility B when a number of breaks were plotted by months.

One of the reasons can be attributed to the majority of the pipelines in both the utilities are installed at a shallower depth (3ft to 5ft). Also, (Welter 2001) has provided evidence about the stresses developed in the pipes and drop in water temperature as the major cause of breaks in winter in his study.

One utility data is not enough to defend any hypothesis, but when the similar trends are observed at different levels with multiple datasets from different utilities lying in one cohort, helps to gain confidence in the analyses results. Hence, information domain analysis is an essential tool for formulating such hypotheses.

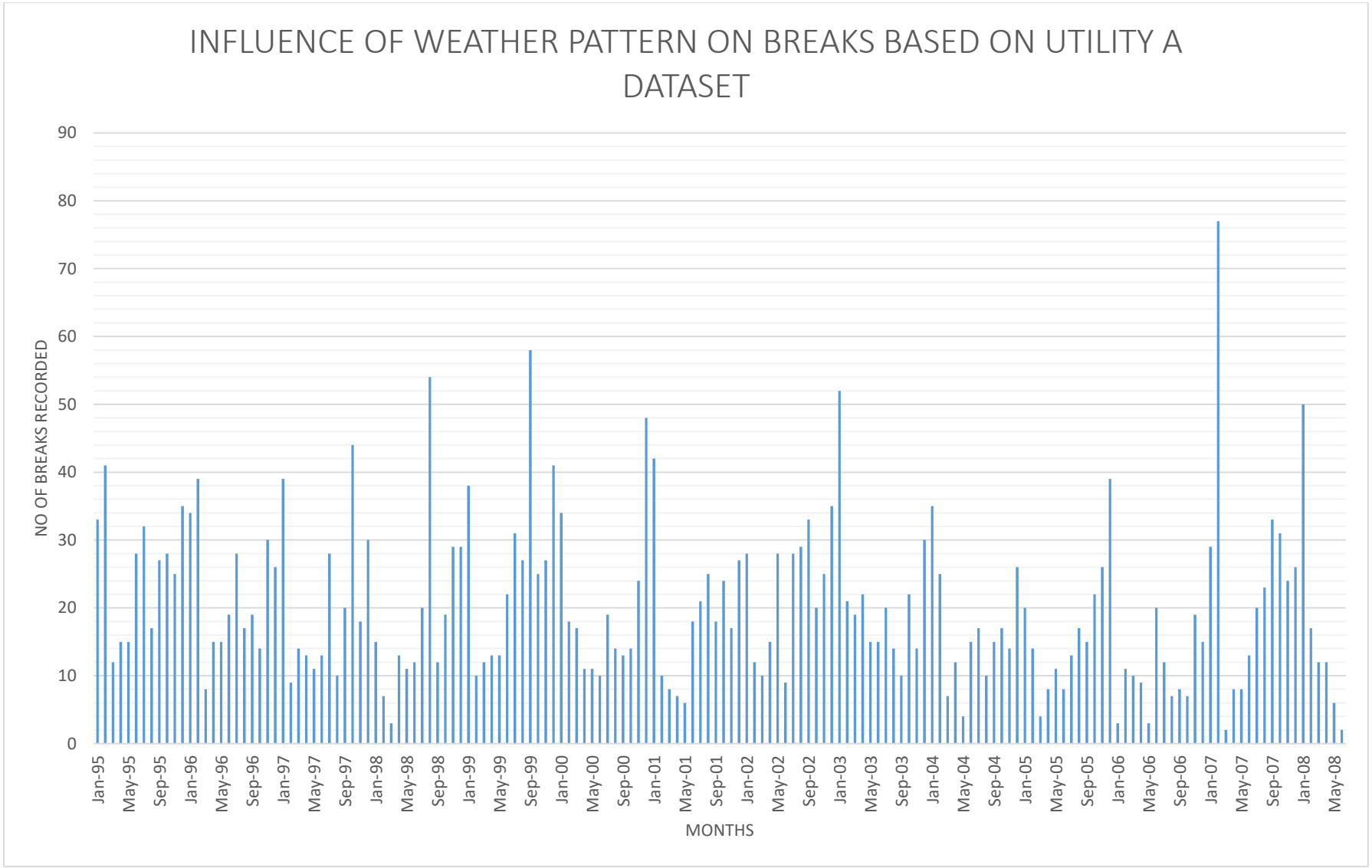

Figure 10 Influence of weather on breaks (utility A)

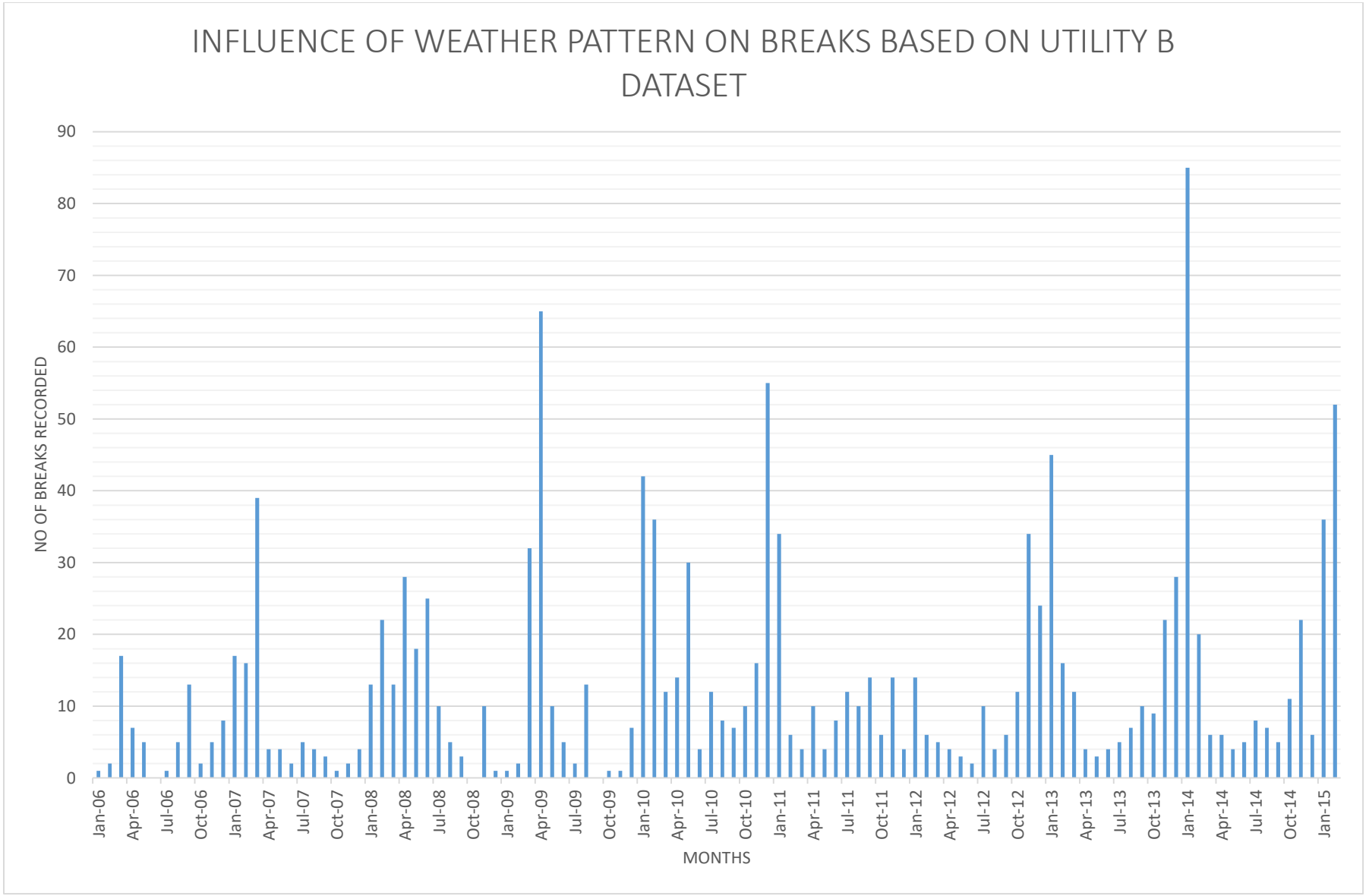

Figure 11 Influence of weather on breaks (utility B)

6.2 Engineering Domain Analysis

Models in the area of performance assessment, risk assessment, performance prediction, condition prediction, failure prediction fall in the engineering categories. For this category, weighted factor performance model (St. Clair & Sinha, 2013) is used to demonstrate the benefits of the proposed methodology. The weighted factor performance model is the network level model and was developed after working closely with the water utilities.

6.2.1 Weighted Factor Performance Model

The purpose of the weighted factor model is to evaluate the performance of the current condition of the pipeline network system. Altogether, 27 parameters were chosen to assess the pipeline performance. All the parameters are listed in Table 2.

The performance model is divided into four modules:

1. **Current Pipe Integrity** includes parameters such as age, design life, vintage, rehab, Hazen Williams C Factor, remaining thickness, tuberculation, leak, pipe break, breaks<5 years ago, defect type, and rehab type.
2. **Internal Condition** includes pressure class information, pressure surges, adequate fire flow, pressure complaints, and discolored water.
3. **External Stress** includes disturbances, flooding, live load, and material type.
4. **External Corrosion** includes dissimilar metals, cathodic protection, stray currents, soil corrosivity, and coating.

At the end of the analysis, each pipe will be tagged with the performance rating. The scale ranges from one to five with 1.0-1.5 representing performance is excellent, and 4.5-5.0 means the pipe performance is very poor. The color-coded performance index is shown in Figure 12.

Figure 12 Performance Index

Reliability index was developed to check the confidence of the performance index. Many utilities don't collect information regarding all the parameters used in the model. However, many of the parameters can be easily derived or downloaded. Sometimes an educated guess is also made for some parameters which are not easily obtainable.

The reliability percentage of the overall drinking water pipe performance score is calculated using a parameter reliability scale as shown in Table 3. For this analysis, reliability percentage for both the utility data is 76%.

Reliability levels are defined on a value ranging from 0 to 5. 0 is the least reliable data while 5 being the most reliable one.

Table 2 List of parameters used in weighted factor performance model

Dissimilar Metals	Pressure Complaints	Adequate Fire Flow
Disturbances	Pressure Surges	Age
Hazen Williams C Factor	Rehab	Breaks<5 Years Ago
Flooding	Rehab Type	Cathodic Protection
Leak	Remaining Thickness	Coating
Live Load	Soil Corrosivity	Defect Type
Material Type	Stray Currents	Design Life
Pipe Break	Tuberculation	Diameter
Pressure Class Exceeded	Vintage	Discolored Water

Table 3 Reliability Scale

Parameter Reliability	Value
Direct Record	5
Derived Indirectly	4
Educated Guess (High confidence)	3
Educated Guess (High confidence)	2
Educated Guess (High confidence)	1
No Data	0

6.2.2 Local Level Analysis Results

As per the analysis result, all the cast iron pipes installed before 1930 is rated "good" by the model along with 82% of the pipes installed after 1930. Rating distribution in percentage is shown in Figure 13.

Snapshot of the result (random sample from the entire dataset) from the weighted factor performance model is illustrated in Figure 14.

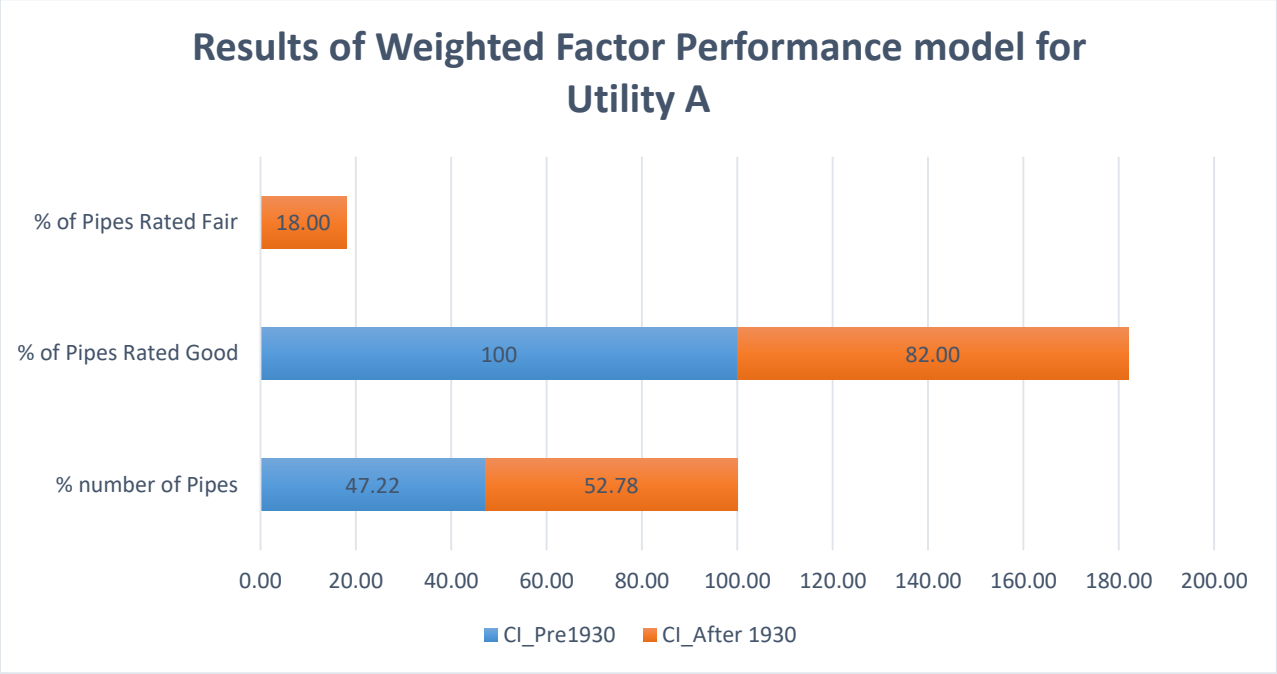

Figure 13 Results of weighted factor performance model for utility A

Installation Date	CURRENT INTEGRITY	INTERNAL CONDITION	EXTERNAL STRESS	EXTERNAL CORROSION	PERFORMANCE INDEX	RATING
CI_1970	3.26	2.96	2.60	2.34	2.79	FAIR
CI_1965	3.32	1.00	2.60	2.34	2.31	GOOD
CI_1967	3.26	1.92	2.60	2.34	2.53	FAIR
CI_1925	3.28	1.00	2.20	2.34	2.20	GOOD
CI_1923	3.28	1.00	2.20	2.34	2.20	GOOD
CI_1925	3.28	1.92	2.20	2.34	2.43	GOOD
CI_1925	3.28	1.92	2.20	2.34	2.43	GOOD
CI_1927	3.34	1.92	2.20	2.34	2.45	GOOD

Figure 14 Results from weighted factor performance model for utility A

6.2.3 Regional Level Analysis Results

The region defined for this analysis is the combined region of both the utilities (dataset of utility A + utility B). The purpose of this analysis is to perform the same analysis as it was performed for utility A and check whether the trends observed are still valid for the entire region or not. Results of the regional level analysis are illustrated in Figure 15.

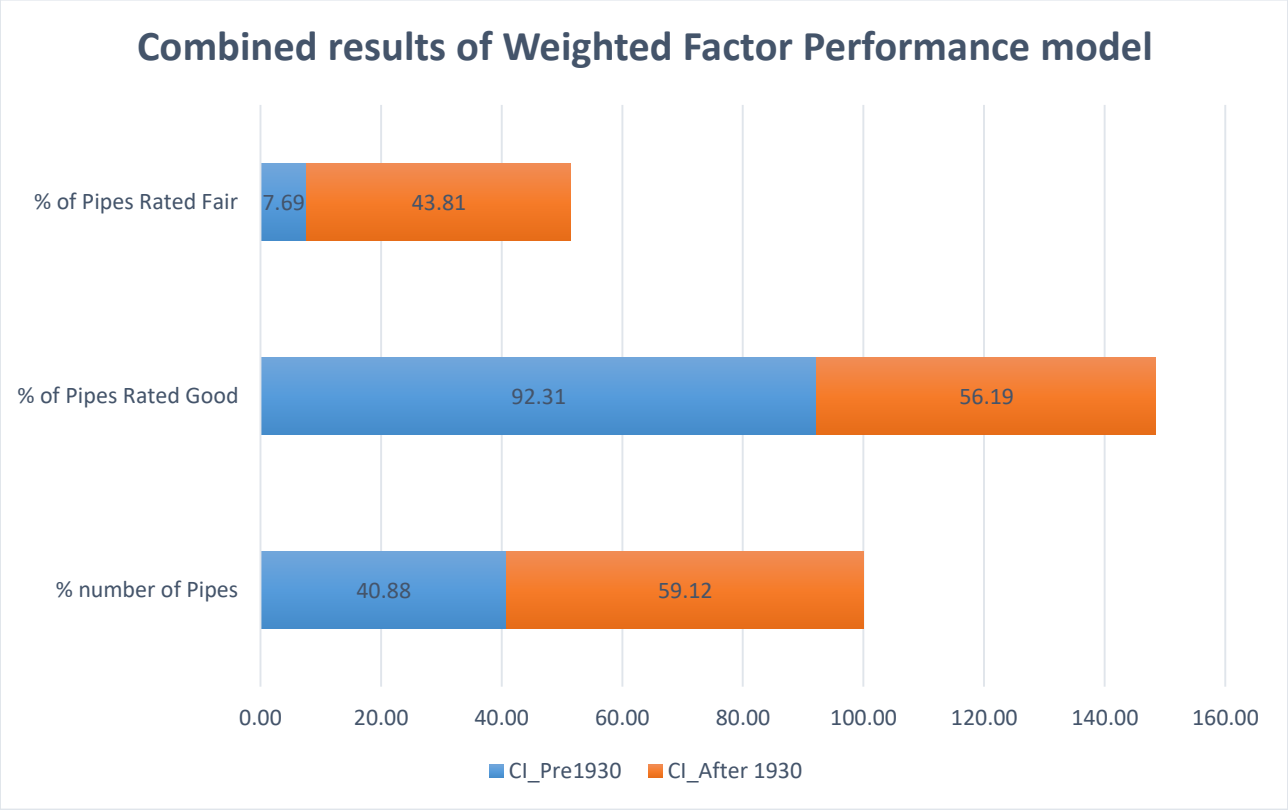

Figure 15 Combined Results of Weighted Factor Performance Model for both the utilities

Hence, according to the results, it can be inferred that majority of the cast iron pipes installed before 1930 is rated "good" by the performance model.

One of the possible reason could be the difference between the manufacturing process of the pipes. Gray cast iron was the dominantly used material from the mid-1800s to 1950s. It was estimated that till 1992, 48% of all the existing pipes were grey cast iron in the United States (Krimeyer et al. 1994). Also, before 1930, pipes used to be thicker as compared to the pipelines installed later on. Corrosion pits in a buried pipe undermine its resistance capacity. Hence thicker the pipe material, higher the resistance as the pipe will have more material to offer against the corrosion pits.

Hence, it can be concluded that performing engineering domain analysis at regional level enhances the reliability and confidence of the results and strongly supports that results are truly indicative of the trend in the population and are not occurring by any chance.

6.3 Management Domain Analysis

Management level analysis aims at prioritization decisions such as where, when and how the budget should be spent. The applications for the management domain analysis can vary from predicting the future monetary demands based on the performance of the pipe to estimate the optimal time to replace/rehabilitate the pipelines in the system.

6.3.1 Local Level Analysis Results

Different kinds of graphs are plotted to find out the economic needs. Figure 16 projects current as well as future cost needs of 10 years from now based on the performance rating from the weighted factor performance model. These kinds of plots can be conducive to the sustainability of drinking water supply.

The clustered bar is plotted between the diameter and estimated dollar needs based on utility A dataset.

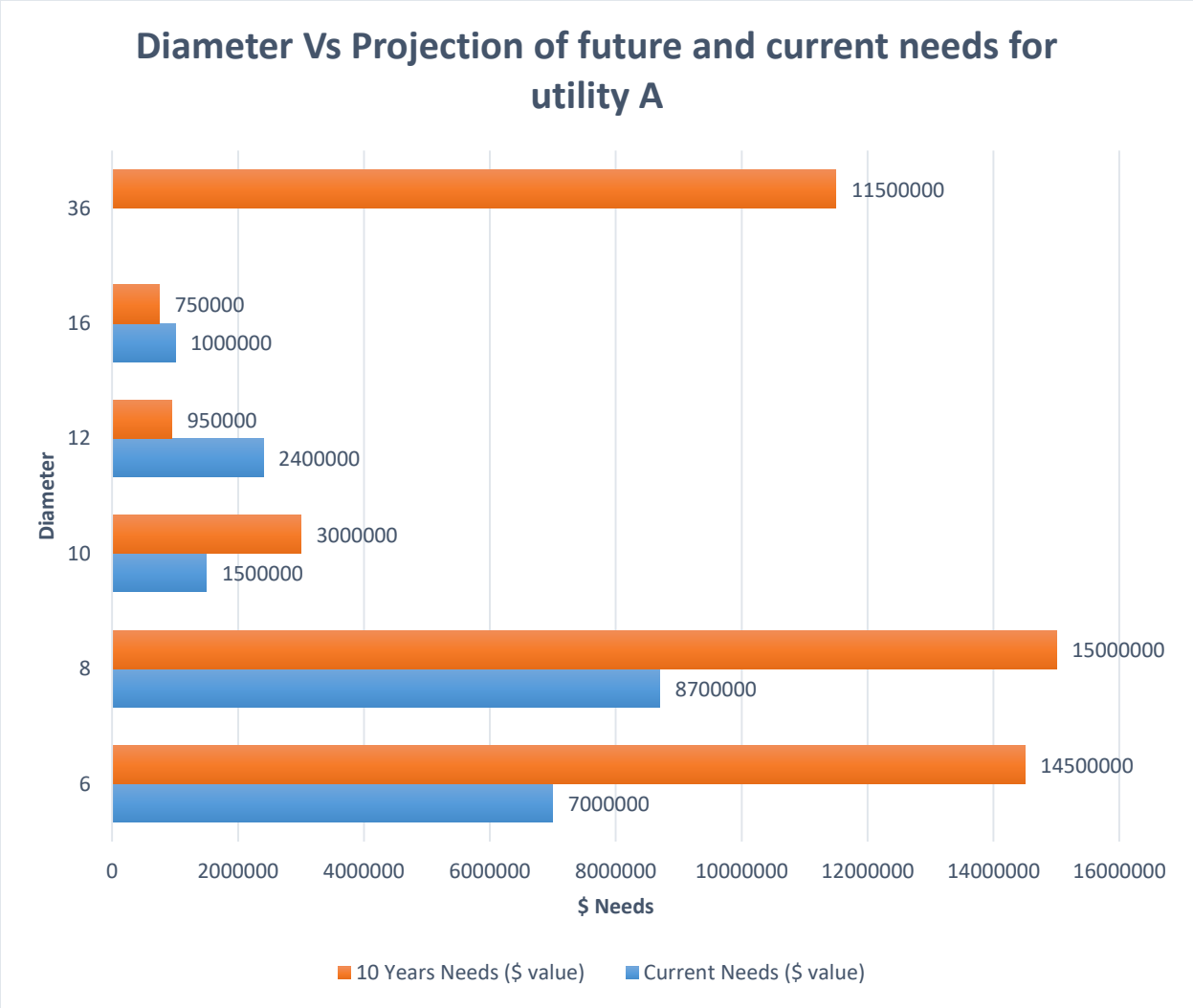

Figure 16 Current vs. Future needs based on utility A dataset

Figure 17 presents the comparison between installation cost and replacement cost per break based on different types of projects. Fixing and replacing a single pipe is costlier than replacing or repairing the group of pipes installed within a particular job site. A project is defined as the collection of pipes installed within one specific area.

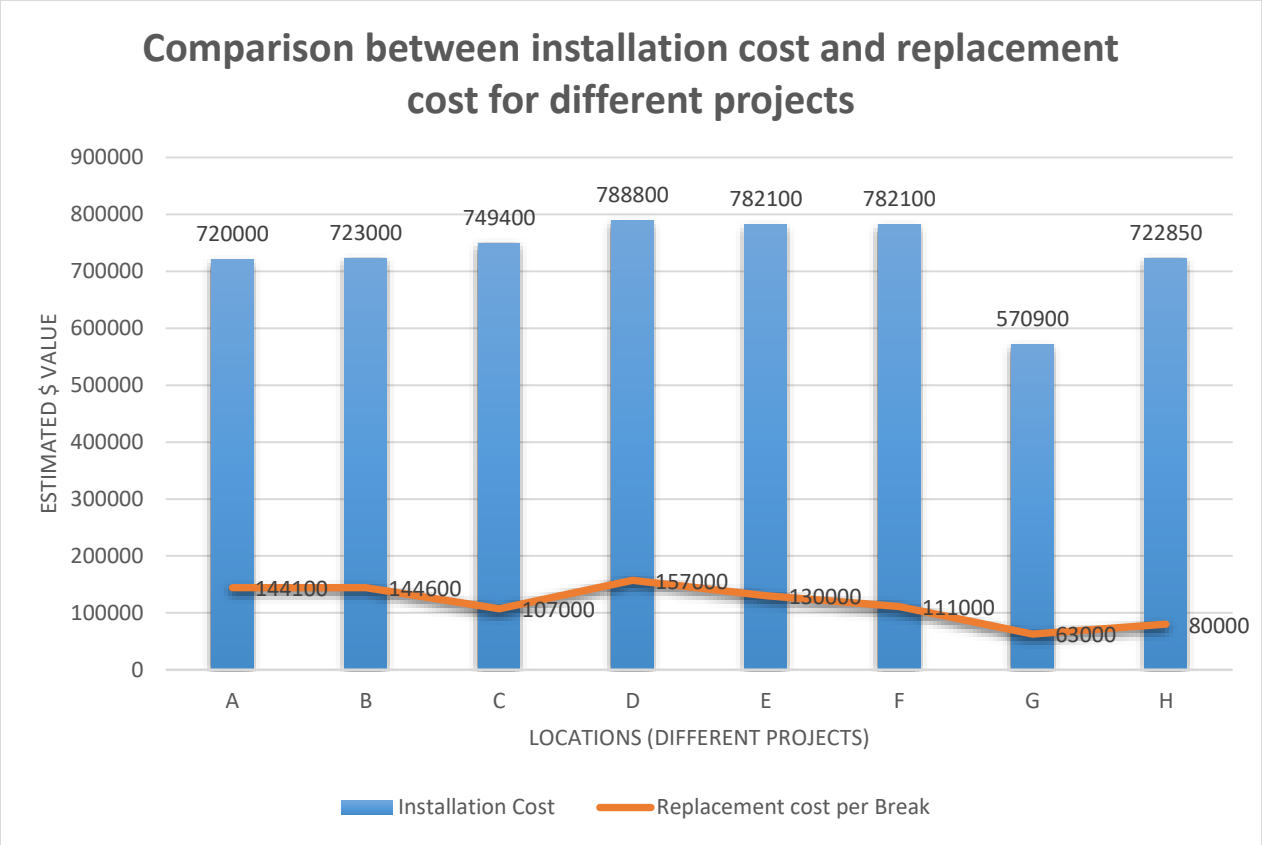

Figure 17 Estimated installation cost and replacement cost values by project

6.3.2 Regional Level Analysis Results

The region defined for this analysis is the region of utility A + utility B. Results of the regional level analysis are illustrated in Figure 18.

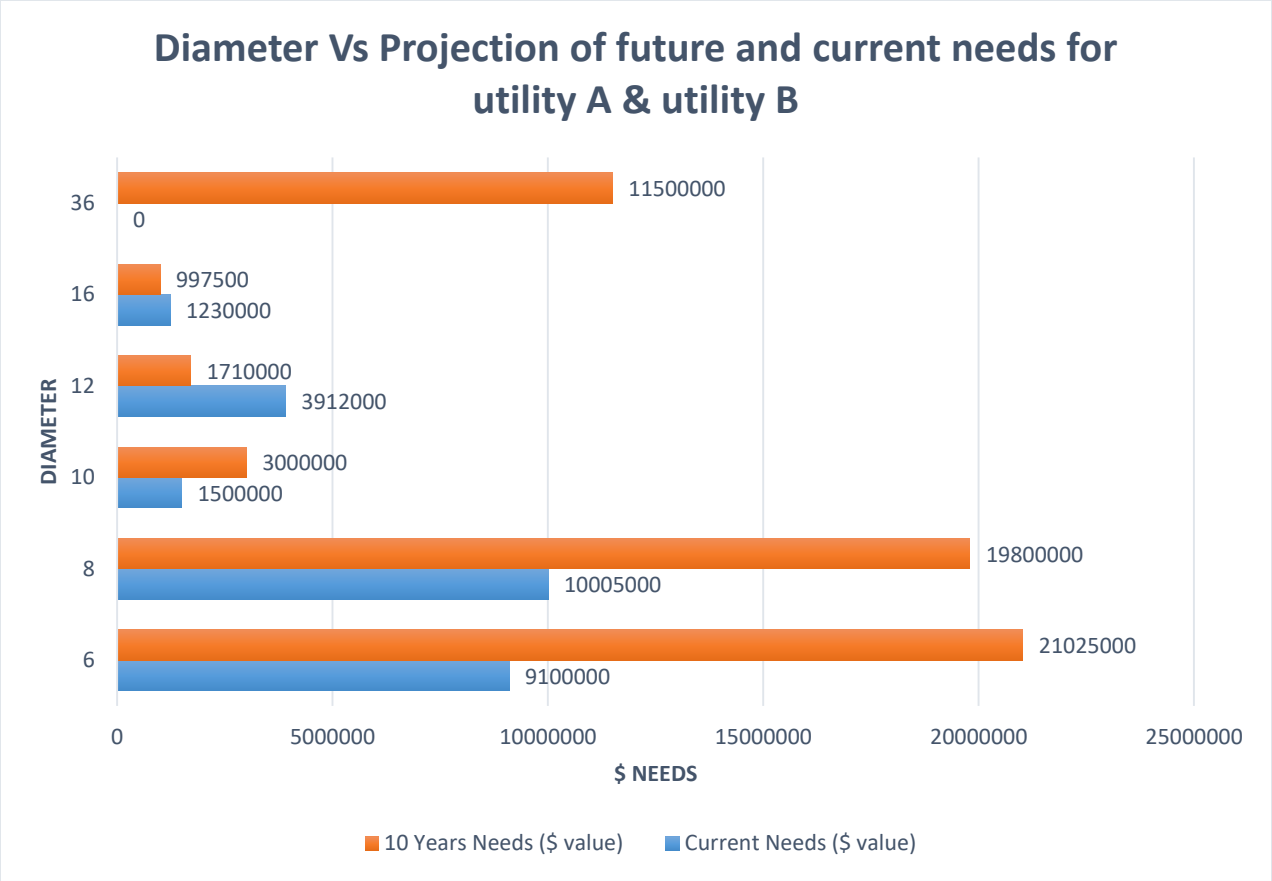

Figure 18 Current vs. Future needs based on combine dataset of both the utilities

It has been realized from the local and regional analysis that performing analysis on a larger dataset increases the confidence level. This happens because larger dataset leads to statistically significant results. In other words, statistically significant results are those results that are not likely to have occurred purely by chance (random error) and thereby have underlying causes for the occurrence.

Performing analysis at the regional level can provide several benefits like:

- Increases the likelihood that results are truly indicative of a trend in the population.
- Leads to greater accuracy.
- With a large sample, conclusions can be confidently drawn about the cohorts of a sample.
- The odds of capturing the outliers increases with increase in the size of the sample.

Mostly, management domain analysis is supported by the engineering domain analysis. Management analysis can only predict the future needs as good as the reliability of the input data going into the model (i.e., performance ratings from the weighted factor performance model).

Overall, piloting models at various levels can help utilities in enhancing the reliability of the results.

6.4 Three Step Analysis Process

The following section discusses some of the additional modeling techniques that can be easily and successfully applied to the proposed methodology by the water utilities. All the modeling techniques described below are analyzed at the local level, but it is highly suggested to pilot at the regional level.

Three step analysis process is suggested for the pipeline performance analysis as shown in Figure 19.

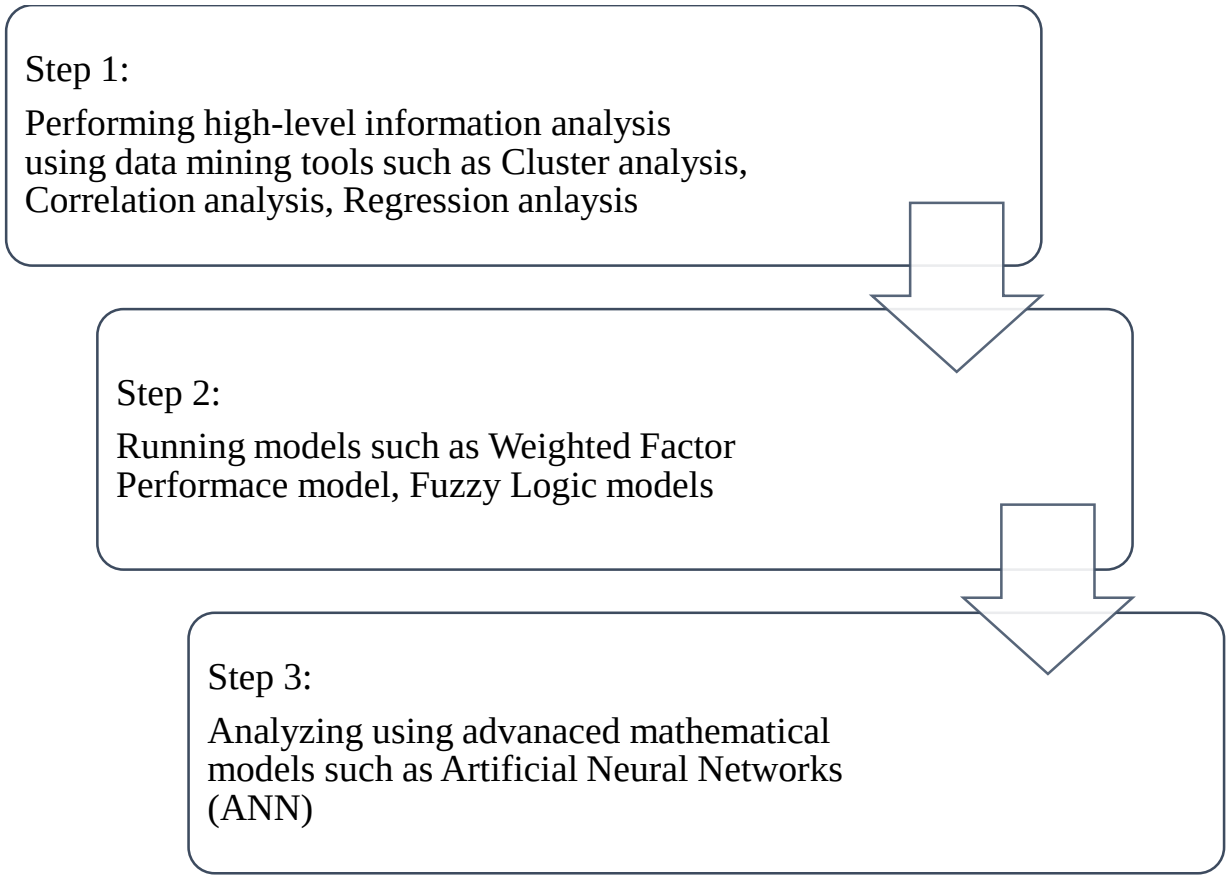

Figure 19 Three steps analysis process

Step 1: In information domain analysis, apart from developing 2D & 3D graphs, many data mining techniques such as cluster analysis, correlation analysis, and regression analysis can be used obtain a great deal of information at local, regional as well as at global level. Application of such data mining tools can vary from as few as two parameters to a large number of parameters.

6.4.1 Correlation Analysis

Correlation analysis is often used to check how close two variables are to having a linear relationship. Correlation analysis can be used to study the strength of a relationship between two or more variables at the network level. If the correlation is found between two variables, it means

that when there is a systematic change in one variable, there is also a systematic change in the other.

For example, a number of breaks in metallic pipelines can be correlated with soil corrosion (soil corrosion data can be easily obtained using USGS) to check the influence of the external corrosion on pipelines. This implies that one should expect a value closer to one if a number of breaks are found to increase with the increase in soil corrosivity level and a value closer to negative one if the case is opposite. +1 indicates the strongest positive correlation possible, and -1 indicates the strongest negative correlation possible.

Similarly, the correlation matrix can be used to check the relationships between a set of variables. In the correlation matrix, each random variable in the table is correlated with each of the other values in the table. This allows checking the pairs having the highest correlation in the table. Diagonal of the table is always one because the correlation between a variable and itself is always one. One such correlation matrix is illustrated in Figure 20. Correlation matrix shown in the figure consists of three parameters: tuberculation, Hazen Williams C factor, and remaining thickness of the pipelines. A summary of the three parameters used is described below.

Hazen Williams C Factor: It is the roughness coefficient of the inside of a pipe. The low value of C factor indicates the poor internal condition of the pipes.

Tuberculation: It is the formation of small mounds of corrosion inside a pipe. Increase in tuberculation build-up increases the roughness of the pipe which in turn decreases C factor of the pipe.

Remaining Thickness: It gives information regarding the remaining wall thickness.

	Tuberculation	*Hazen Williams C Factor*	*Remaining Thickness*
Tuberculation	1		
Hazen Williams C Factor	-0.54675459	1	
Remaining Thickness	-0.70913526	0.69723551	1

Figure 20 Correlation Matrix

It can be inferred from the correlation matrix results that tuberculation and C factors are negatively correlated with each other. Also as the tuberculation build-up increases, the wall thickness of the pipe starts decreasing. This is also reflected in the results of the correlation matrix. Hence, correlation analysis is a measure of the extent to which two or more variables are related.

6.4.2 Cluster Analysis (K-means clustering)

K-means cluster analysis is one of the means of massive dataset mining, which can be helpful to divide the entire dataset into k clusters in which each dataset belongs to the cluster with the nearest

mean. It is assumed that each group consisting of a group of pipes have similar parameters and are expected to have same breakage pattern.

Cluster analysis can quickly help to find the bursting pipe zones of similar failure mode, diameter, age, etc. Since the region divided in the cluster has certain relations already, any analysis performed after it will have a higher level of accuracy in the results. For example, (Farmani et al. 2017) in his study used K-means clustering analysis as a tool to improve the performance of pipe failure prediction models. Evolutionary polynomial regression was then employed for each cluster to predict the number of failures. The K-means clustering is applied here using the KNIME software as shown in Figure 21 to partition the dataset into a number of specified clusters based on diameter and age of pipelines (using a dataset of utility A).

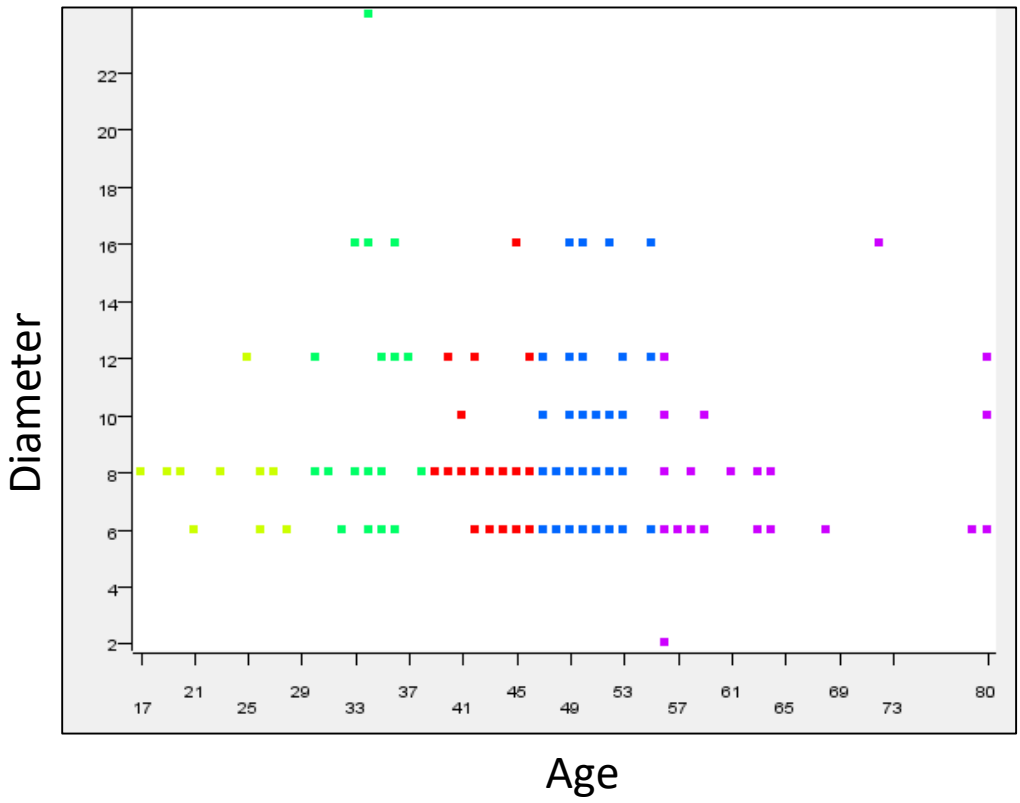

Figure 21 Results of K means clustering between diameter and age of the pipeline

6.4.3 Regression Analysis

Regression analysis is an essential tool for estimating the relationships between the variables. Regression analysis varies in complexity from simple linear regression to evolutionary polynomial regression models.

Figure 22 presents analysis results of linear, cubic and polynomial regression model analyzed on one homogeneous group of the dataset obtained after performing K-means cluster analysis on the dataset of utility A. For the regression model, independent variable considered is the diameter of the pipelines whereas dependent variable is the number of breaks for predicting pipe breaks in water distribution systems.

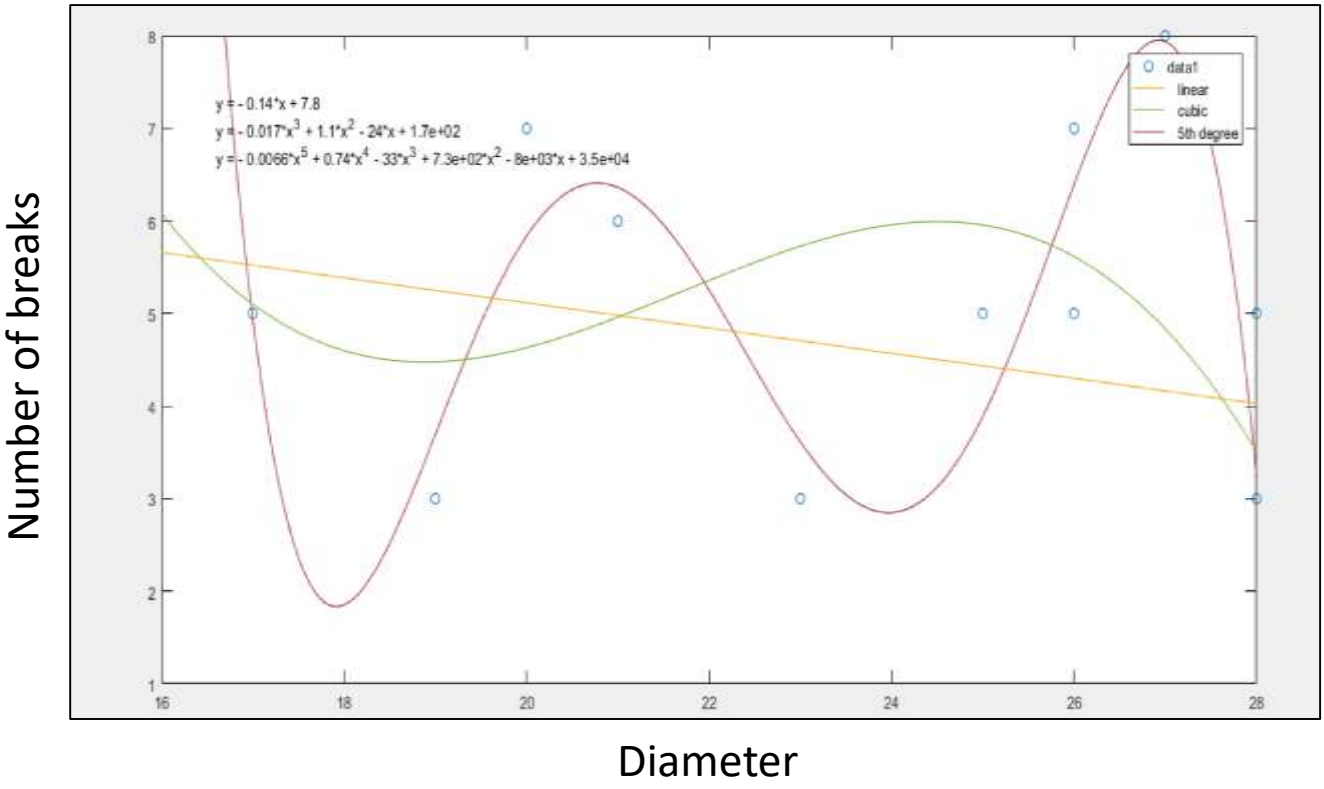

Figure 22 Results of Regression Analysis

The result of the tests suggests that 5th-degree polynomial line gave the best results for the specific case that was analyzed. The performance of the selected model is evaluated using the R^2 value. More specifically, the R^2 value for this particular analysis is 0.83. This approach can be used to obtain accurate predictions of failure for each of these homogeneous groups.

Hence, data mining tools can be an important tool to check whether there is a relationship between the variables or not, how are the variables related to each other and also to divide the dataset into homogeneous clusters.

Step 2: The next step in the process is to predict the overall performance of individual water pipelines. Models like weighted factor performance model (St. Clair & Sinha, 2013) and Fuzzy Logic models falls under this category. These models are generally the network level models.

Models in this category require a number of parameters than compared to parameters needed for the models and tools described in step 1 of the process.

Step 3: There is also need of some advance automatic algorithms which can be used to make predictions, real-world simulations, pattern recognition and classifications of the input data on large dataset such as Artificial Neural Networks (ANN), Genetic Algorithm (GA), Support Vector Machine (SVM), etc.

6.4.4 Artificial Neural Networks (ANN)

The artificial neural network is a mathematical model interconnected with processing units named neurons. It can deal with large dataset very efficiently. The architecture of ANN includes a number of neurons (nodes) in the input, output as well as a hidden layer. ANN learn through an iterative process and adjusts weights and biases to minimize the gap between the predicted and estimated values. Errors at each level are back propagated to reduce the overall error till the target answers coincide with a given value of tolerance. A number of hidden layers and hidden nodes decides the performance of ANN model. ANN structure is highly dependent on a number of hidden layers. "There is no unified theory for the determination of an optimal ANN structure" (Ahn et al. 2005).

An example of predicting the performance of pipelines using ANN is presented below. Levenberg–Marquardt algorithm is used for training the data. Neural Network code was generated using the neural fitting app in MATLAB. The code is attached in the appendix section (A2 – Neural Network Code). Levenberg–Marquardt is well-known for prediction, estimation and solving non-linear least squares fitting problems (Zangenehmadar, 2016). Dataset is randomly divided into groups of 70%, 15% and 15% for training, validation and testing the results respectively. The entire dataset contains 381 samples. Hence the distribution is 267,57 and 57 samples for training testing and validation respectively as shown in Figure 25. Estimated performance value is evaluated using the weighted factor performance model (St. Clair & Sinha, 2013). The single hidden layer is selected for the model and model is trained using 25 neurons.

The performance of the models is assessed based on R^2 value since it is easy to calculate and understand. The R^2 value oscillates between 0 and 1 and evaluates the percentage of total differences between the evaluated and predicted value concerning average. The coefficient of determination values of training, testing, and validation phases are displayed in Figure 24.
The results of the original model and the predicted values from ANN model are compared to validate the results of the model as shown in Figure 25.

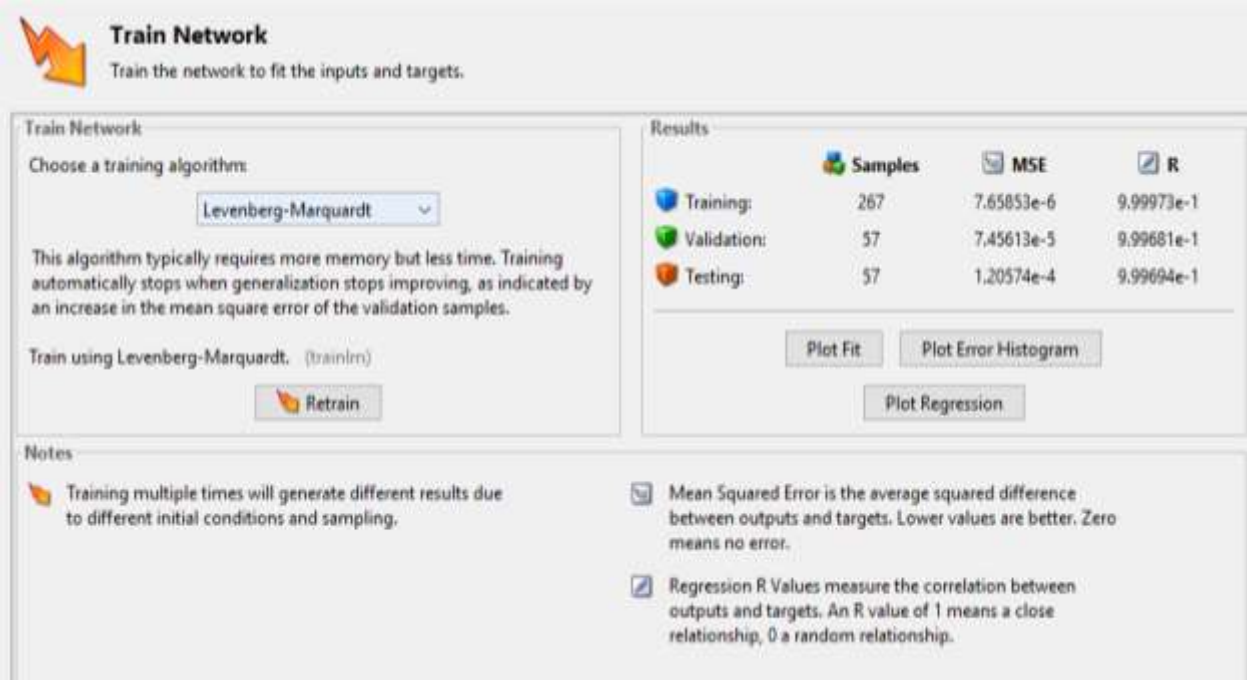

Figure 23 Training network details

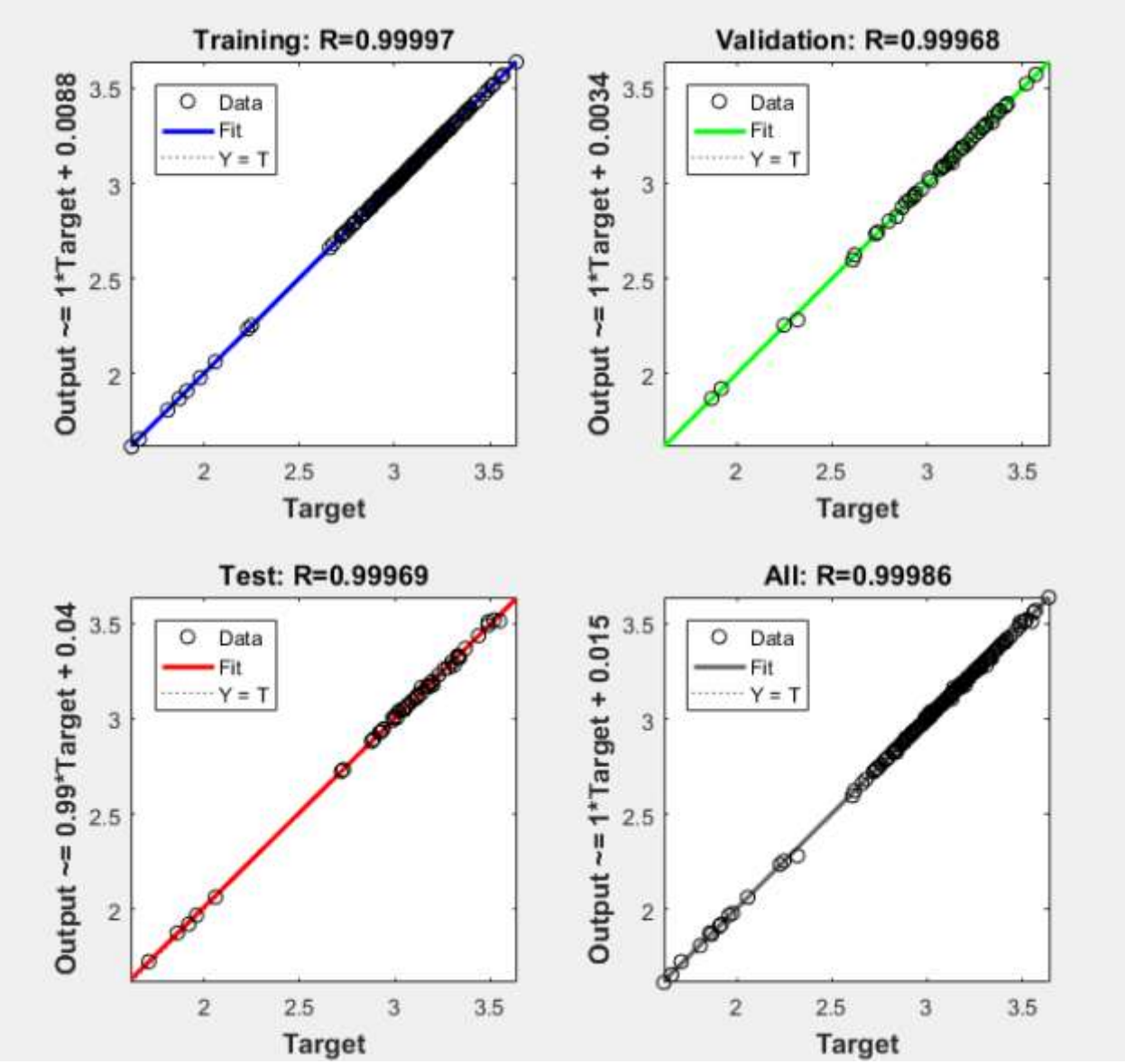

Figure 24 Coefficient of determination values for training, testing, and validation models

The horizontal axis and vertical axis show target values versus output values that are performance rating values for individual pipelines. The R^2 value of all data is displayed as well. The fitted line for all data is output = 0.99 × target + 0.04, and R^2 value is 99% which shows that the predicted values are very close to evaluated values.

Figure 25 represents the graph of evaluated performance rating values (blue in color, ratings obtained from weighted factor performance model) vs predicted performance rating values (red in color, obtained performance rating results from ANN model) along with error values (Error Values = Predicted Values – Observed value, Yellow line - at the bottom of the graph).

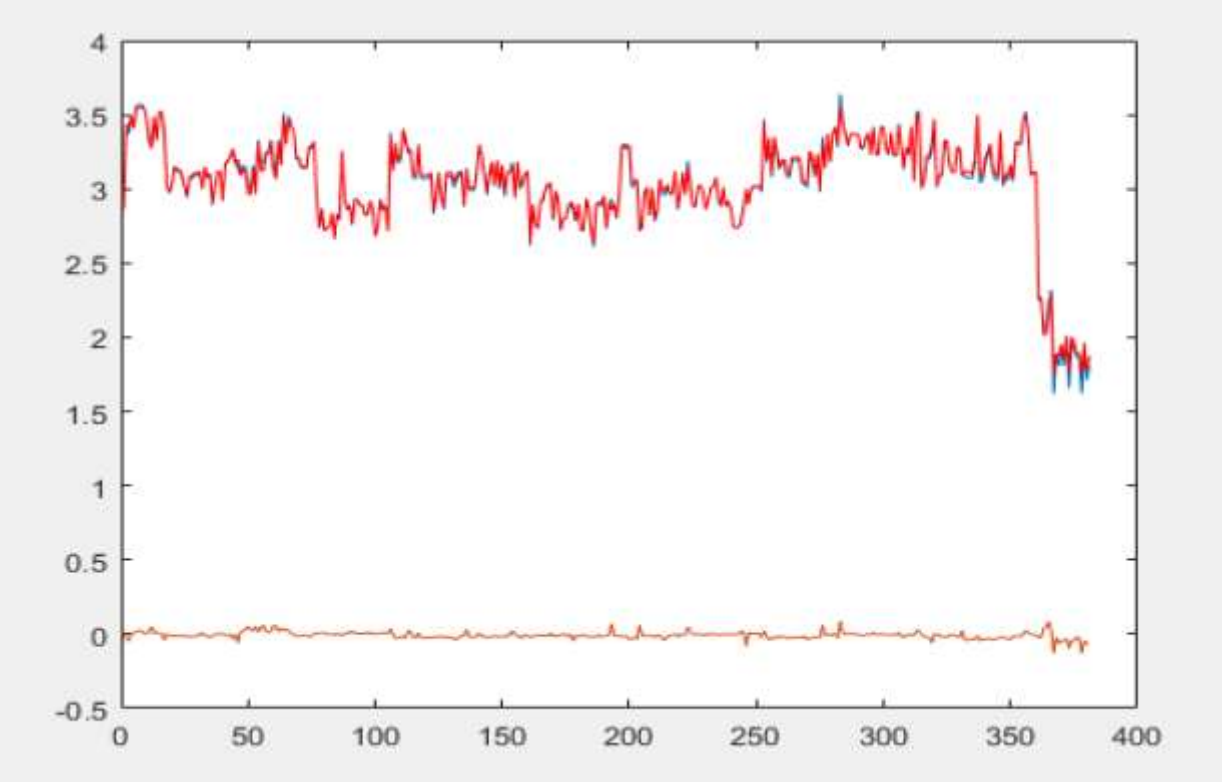

Figure 25 Observed Values vs. Predicted Performance Values (ANN results)

ANN model has a large number of benefits in big data analysis. However, it will completely be a "black-box" approach if the rich quality data is not used for the analysis. Because of the limited set of quality data, ANN model for this thesis is not piloted at various levels of analysis, but it is highly recommended to test the ANN at local, regional and global level. Based on the results, it seems that ANN can be used as a robust performance assessment tool for big data analysis.

CHAPTER 7

CONCLUSION AND FUTURE RECOMMENDATIONS

7.1 Conclusion

As water utilities attempt to be proactive and willing to adopt more robust models, piloting their new and existing models may provide them reliable results which will help them to be more efficient in their decision-making process.

The following points can be concluded from the methodology proposed:

- General understanding of current and past trends (information domain analysis) provides insight into the data and can help water utilities to observe unseen patterns which can be tested using engineering analysis or in the laboratory.
- It is essential for water utilities to pilot existing and even new models for practical use.
- A reliable water pipeline performance analysis model will assist water utilities in supporting decisions.
- Piloting models at various utilities will help utilities in enhancing the reliability of model outputs.
- Establishing a centralized modeling platform can assist water utilities for storing, updating, retrieving, modeling, and analyzing, the data at various domains and levels of analysis.
- Cohort analysis will help utilities to enhance their model/tools by learning from the other utilities which are identified in the same cohort or dealing with similar problems in their network.

The following points can be concluded from the analyses conducted:

- Correlation analysis is a good technique that can be used to study the strength of a relationship between two or more variables at network level using as few as 2-3 parameters to a large number of parameters by developing correlation matrix.
- K-means clustering analysis as an important tool which can be easily applied to improve the performance of pipe failure prediction models.

- ANN model has a large number of benefits in big data analysis.
- A number of hidden layers and hidden nodes decides the performance of ANN model.
- Three step process should be adopted to get increased accuracy in the analyses results:

    I. Step 1 includes performing high-level information analysis using data mining tools such as correlation analysis, cluster analysis, and regression analysis. Such kind of analyses can be easily implemented using two to three parameters.

    II. Step 2 includes running models such as weighted factor performance model or fuzzy logic models. Weights that are used in the model can be estimated through the step 1 of the process. These models can be used at the network level and can be run using easily available parameters from the utility data such as pipe diameter, pipe age, installation year, type of failure, number of breaks, depth of pipes, soil type, traffic loading, etc.

    III. Step 3 includes advanced mathematical models for the analysis such as neural networks. This kind of models is highly recommended for the big data analysis. Also, a large amount of quality data is required for the networks to make predictions, real-world simulations, and recognizing patterns.

7.2 Future Recommendations

I. There is dire need to establish nationwide standardized database platform. The database should include data aggregation from multiple data sources/departments such as maintenance department, planning department, finance department, and engineering department. This national database can be a valuable asset for further improving the condition of buried water infrastructure.

II. In regression analysis, only one variable is accounted as an independent variable. But it is highly recommended to use multiple variable regression models or evolutionary polynomial regression model for break prediction. The reason is numerous variables are responsible for breaks in pipelines, and it is not possible to capture and predict the number of breaks only by using a single parameter.

III. Neural networks can be coded to the extent that it can have almost zero error values. However, overfitting neural networks stop generalizing the network and start memorizing the values. As a result, there will still be a high accuracy in the training set data, but the accuracy of test data will crash after a certain point. Neural network

should be stopped when there is a peak in the test set data. Early stopping algorithm can be used for finding the peak in the test set data and stop neural network to train further for avoiding the overfitting problems.

BIBLIOGRAPHY

Al-Barqawi, H., & Zayed, T. (2006). Assessment Model of Water Main Conditions. Pipeline Division Speciality Conference, Illinois, 2006. p1-8, 8p: American Society of Civil Engineers.

ASCE. (2017). "2017 Infrastructure Report Card"_ Retrieved March 15, 2018, from https://www.infrastructurereportcard.org.

Babovic, V., Drecourt, J., Keijzer, M. & Hansen, P. (2002). "A Data Mining Approach to Modeling of Water Supply Assets." Urban Water 4(4):401-414.

Davis, P., Burn, S., Moglia, M., & Gould, S. (2007). "A Physical Probabilistic Model to Predict Failure Rates in Buried PVC Pipelines." Reliability Engineering and System Safety 92(9): 1258-1266.

Farmani, R., Kakoudakis, K., Behzadian, K., & Butler, D. (2017). Pipe Failure Prediction In Water Distribution Systems Considering Static and Dynamic Forces. Procedia Engineering 186(2017): 117-126.

Giustolisi, O., Laucelli, D., & Savic, D. A. (2006). “Development of rehabilitation plans for water mains replacement considering risk and cost-benefit assessment.” Civil Engineering and Environmental Systems, 175-190.

Kleiner, Y., & Rajani, B. (2008). Prioritising Individual Water Mains for Renewal.Proc., ASCE/EWRI World Environmental and Water Resources, Honolulu, Hawaii, National Research Council Canada-CNRC-NRC, 1-10.

Krimeyer, Richards, W., Gregory, J., & Smith, C. (1994). An assessment of Water Distribution Systems and Associated Research Needs. Denver,Co: AWWARF.

Oxenford, J., et al. (2012). "Key asset data for drinking water and wastewater utilities." Water Research Foundation Rep., Water Research Foundation, Denver.

Poulton, M., Le Gat, Y., & Bemond, B. (2007). The Impact of Pipe Segment Length on Break Predictions in Water Distribution Systems.Proc., LESAM-2nd Leading Edge Conference on Strategic Asset Management, Lisbon, Portugal.

Rajani, B., & Makar, J. (2000). "A Methodology to Estimate Remaining Service Life of Grey Cast Iron Water Mains." Canadian Journal of Civil Engineering 27: 1259-1272.

Sivanandam, S. N., Sumathi, S., & Deepa, S. N. (2007). Introduction to Fuzzy Logic using MATLAB. Heidelberg, Springer-Verlag Berlin Heidelberg.

St. Clair, A. M., & Sinha, S. (2013). Development of a Novel Performance Index and a Performance Prediction Model for Metallic Drinking Water Pipelines (Doctoral dissertation).

Welter, G. (2001). Predicting water main breaks in winter. AWWA Annual Conference, (pp. 17-21). Washington, DC.

WIN. (2000). Clean and Safe Water for the 21st Century. Washington, DC, Water Infrastructure Network (WIN).

Zangenehmadar, Z. (2016). Asset Management Tools for Sustainable Water Distribution Networks (Doctoral dissertation).

APPENDIX A

A1 Preliminary Analysis Code

```
char X1;
char X2;
char X3;
char X4;
char X5;
fig_num=0;
age = menu('Does the utility data contains the information on AGE of the
pipes:','yes','no');
diameter = menu('Does the utility data contains the information on DIAMETER
of the pipes:','yes','no');
pipetype = menu('Does the utility data contains the information on MATERIAL
TYPE of the pipes:','yes','no');
failuretype = menu('Does the utility data contains the information on FAILURE
TYPE of the pipes:','yes','no');
breakdata = menu('Does the utility data contains the information on NO OF
BREAKS of the pipes:','yes','no');
if age == 1
   X1 = input ('write down the name of the column corresponding to age of the
pipelines: \n', 's');
end
if diameter == 1
    X2 = input ('write down the name of the column corresponding to diameter
of the pipelines: \n', 's');
end
if pipetype == 1
   X3 = input ('write down the name of the column corresponding to the
material of the pipelines: \n', 's');
end
if failuretype == 1
   X4 = input ('write down the name of the column corresponding to the
failuretype of the pipelines: \n', 's');
end
if breakdata == 1
   X5 = input ('write down the name of the column corresponding to the no of
breaks of the pipelines: \n', 's');
end
[num,txt,~] = xlsread('BookA.xlsx');
index1=0;
 for i=1:length(txt(1,:))
     if strcmp(X1,txt(1,i))
        index1=i;
        break;
     end
 end
```

```
X1_val = num(:,index1);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
index2=0;
 for i=1:length(txt(1,:))
     if strcmp(X2,txt(1,i))
 index2=i;
break;
end
 end
X2_val = num(:,index2);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
index3=0;
 for i=1:length(txt(1,:))
     if strcmp(X3,txt(1,i))
 index3=i;
break;
     end
 end
X3_val = txt(2:end,index3);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
index4=0;
 for i=1:length(txt(1,:))
     if strcmp(X4,txt(1,i))
        index4=i;
        break;
     end
 end
X4_val = txt(2:end,index4);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
index5=0;
 for i=1:length(txt(1,:))
     if strcmp(X5,txt(1,i))
 index5=i;
break;
end
 end
X5_val = num(:,index5);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% filename = 'testdata1.xlsx';
% xlswrite(filename,C,'AMERICAN WATER.xlsx','A2');
A=[X1_val X2_val];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
AGE=A(:,1);
NO_OF_PIPES_AGE_LESS_THAN_OR_EQUAL_TO_25YEARS=sum(AGE<=25);
NO_OF_PIPES_AGE_GREATER_THAN_25_AND_LESS_OR_EQUAL_TO_50_YEARS=sum(AGE>25 &
AGE<= 50);
NO_OF_PIPES_AGE_GREATER_THAN_50_AND_LESS_OR_EQUAL_TO_75_YEARS=sum(AGE>50 &
AGE<= 75);
```

```
NO_OF_PIPES_AGE_GREATER_THAN_75_AND_LESS_OR_EQUAL_TO_100_YEARS=sum(AGE>75 &
AGE<= 100);
NO_OF_PIPES_AGE_GREATER_THAN_100_AND_LESS_OR_EQUAL_TO_150_YEARS=sum(AGE>100 &
AGE<= 150);
NO_OF_PIPES_AGE_GREATER_THAN_150YEARS=sum(AGE>150);
PERCENTAGE_AGE_DISTRIBUTION_PIPES=[NO_OF_PIPES_AGE_LESS_THAN_OR_EQUAL_TO_25YE
ARS NO_OF_PIPES_AGE_GREATER_THAN_25_AND_LESS_OR_EQUAL_TO_50_YEARS
NO_OF_PIPES_AGE_GREATER_THAN_50_AND_LESS_OR_EQUAL_TO_75_YEARS
NO_OF_PIPES_AGE_GREATER_THAN_75_AND_LESS_OR_EQUAL_TO_100_YEARS
NO_OF_PIPES_AGE_GREATER_THAN_100_AND_LESS_OR_EQUAL_TO_150_YEARS
NO_OF_PIPES_AGE_GREATER_THAN_150YEARS]
pie3(PERCENTAGE_AGE_DISTRIBUTION_PIPES)
legend('<25','25-50','50-75','75-100','100-15','>150')
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
DIAMETER=A(:,2);
NO_OF_PIPES_DIAMETER_LESS_THAN_OR_EQUAL_TO_4IN=sum(DIAMETER<=4);
NO_OF_PIPES_DIAMETER_GREATER_THAN_4_AND_LESS_OR_EQUAL_TO_16_IN=sum(DIAMETER>4
& DIAMETER<= 16);
NO_OF_PIPES_DIAMETER_GREATER_THAN_16_AND_LESS_OR_EQUAL_TO_36_IN=sum(DIAMETER>
16 & DIAMETER<= 36);
NO_OF_PIPES_DIAMETER_GREATER_THAN_36IN=sum(DIAMETER>36);
PERCENTAGE_DIAMETER_DISTRIBUTION_PIPES=[NO_OF_PIPES_DIAMETER_LESS_THAN_OR_EQU
AL_TO_4IN NO_OF_PIPES_DIAMETER_GREATER_THAN_4_AND_LESS_OR_EQUAL_TO_16_IN
NO_OF_PIPES_DIAMETER_GREATER_THAN_16_AND_LESS_OR_EQUAL_TO_36_IN
NO_OF_PIPES_DIAMETER_GREATER_THAN_36IN]
figure(2)
pie3(PERCENTAGE_DIAMETER_DISTRIBUTION_PIPES)
legend('<4','4-16','16-36','>36')
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[pipe_type, Ind_a, Ind_c]=unique(X3_val);
pipe_count = zeros(1,numel(pipe_type));
for i = 1:length(pipe_count)
    pipe_count(i) = sum (Ind_c == i);
end
pipe_type;
pipe_count;
Total_data_set=length(X1_val);
figure(3)
bar(pipe_count)
set(gca, 'XTickLabel', pipe_type)
ylabel('No Of Pipes');
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[failure_type, Ind_y, Ind_z]=unique(X4_val);
pipe_count1 = zeros(1,numel(failure_type));
for i = 1:length(pipe_count1)
    pipe_count1(i) = sum (Ind_z == i);
end
figure(4)
bar(pipe_count1)
```

```
set(gca, 'XTickLabel', failure_type)
ylabel('No Of Pipes');
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
figure(5)
scatter(X1_val,X5_val,'*','r')
xlabel('AGE');
ylabel('BREAKS');
b=mean(X1_val);
c=mean(X5_val);
b1=X1_val-b;
c1=X5_val-c;
bcproduct=b1.*c1;
bsquare=b1.^2;
csquare=c1.^2;
sumproduct=sum(bcproduct);
sumbsquare=sum(bsquare);
sumcsquare=sum(csquare);
product=sumbsquare*sumcsquare;
denom=sqrt(product);
r=sumproduct/denom
md1=fitlm(X5_val,X1_val)
```

A2 Neural Network Code

```
% Solve an Input-Output Fitting problem with a Neural Network
% Script generated by Neural Fitting app
% Created 04-Apr-2018 15:27:53
%
% This script assumes these variables are defined:
%
%   i - input data.
%   t - target data.

x = i;
t = t;

% Choose a Training Function
% For a list of all training functions type: help nntrain
% 'trainlm' is usually fastest.
% 'trainbr' takes longer but may be better for challenging problems.
% 'trainscg' uses less memory. Suitable for low memory situations.
trainFcn = 'trainlm';  % Levenberg-Marquardt backpropagation.

% Create a Fitting Network
hiddenLayerSize = 15;
```

```
net = fitnet(hiddenLayerSize,trainFcn);

% Setup Division of Data for Training, Validation, Testing
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;

% Train the Network
[net,tr] = train(net,x,t);

% Test the Network
y = net(x);
e = gsubtract(t,y);
performance = perform(net,t,y)

% View the Network
view(net)

% Plots
% Uncomment these lines to enable various plots.
%figure, plotperform(tr)
%figure, plottrainstate(tr)
%figure, ploterrhist(e)
%figure, plotregression(t,y)
%figure, plotfit(net,x,t)
```