Failure Rate Prediction Models of Water Distribution Networks

Seyed Farzad Karimian

A Thesis

In The Department of

Building, Civil and Environmental Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Applied Science (Civil Engineering) at

Concordia University

Montreal, Quebec, Canada

December 2015

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: **Seyed Farzad Karimian**

Entitled: **Failure Rate Prediction Models of Water Distribution Networks**
and submitted in partial fulfillment of the requirements for the degree of
MASTER OF APPLIED (Civil Engineering)

Complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. S. Li _____ Chair

Dr. G. Gopakumar _____ External to Program

Dr. Z. Zhu _____ Examiner

Dr. O. Moselhi _____ Co-Supervisor

Dr. T. Zayed _____ Co-Supervisor

Approved by: _____

December 2, 2015 _____

Abstract

Failure Rate Prediction Models of Water Distribution Networks

Seyed Farzad Karimian

The economic, social and environmental impacts of water main failures impose a great pressure on utility managers and municipalities to develop reliable rehabilitation/replacement plans. The Canadian Infrastructure Report Card 2012 stated that 15.4% of Canadian water distribution systems was in a "fair" to "very poor" condition with a replacement cost of CAD 25.9 billion. The "fair", "poor" and "very poor" conditions represent the beginning of deterioration, nearing the end of useful life and no residual life expectancy, respectively. The majority of municipalities in Canada do not possess complete dataset of water distribution networks. The annual number of breaks or breakage rate of each pipe segment is known as one of the most important criteria in condition assessment of water pipelines. The main objective of this research is to develop a research framework that circumvent the limitations of existing studies by: 1) identifying the most critical factors affecting water pipe failure rates, 2) determining the best mathematical expression for predicting water pipeline failure rate 3) developing deterioration curves, and 4) deploying sensitivity analysis to recognize the effect of each input change on the breakage rate.

The proposed research framework utilizes Best Subset regression to recognize the most effective factors on water pipelines. Best-Subset Algorithm is a procedure to find the best combination of variables to predict the water pipe failure rate among all possible candidates. Once the process of critical factor selection is performed, selected variables are

employed to predict the number of breaks of water pipes using Evolutionary Polynomial Regression (EPR). The EPR is an intuitive data mining technique performed in two stages: 1) the search for the best model using Multi-Objective Genetic Algorithm (MOGA), and 2) the parameter estimation for the model using Least Square Method. The predicted number of breaks, computed by EPR, is utilized to develop deterioration curves by applying Weibull distribution function. Finally, sensitivity analysis is performed to: 1) recognize the effect of changing each input on the failure rate, and 2) study the relationship between the selected inputs and the output.

The developed research framework is applied into two case studies to test its effectiveness. The case considers the water distribution networks in the City of Montréal, Canada and the City of Doha, Qatar. Physical factors, such as age, length, diameter and pipe material were identified as the most critical factors to affect the failure rate of pipes. The results indicate that the developed models successfully estimated the number of breaks for the City of Montreal and City of Doha with a maximum R-Squared of 89.35% and 96.27%, respectively. Also, it is tested by using 20% of each dataset and promising results were generated with a maximum R-Squared of 84.86% and 74.39% for dataset of Montreal and Doha respectively. This demonstrates the accuracy and robustness of the developed models in assessing and analyzing water distribution networks. The developed model is useful for municipalities and decision makers to prioritize the maintenance, repair, rehabilitation, and budget allocation of water distribution networks.

Acknowledgement

First of all, I would like to express my sincere gratitude to my supervisors, Dr. Osama Moselhi and Dr. Tarek Zayed for their invaluable guidance and continuous support during my Master program. Their valuable advice and positive attitudes along with constructive criticism made me capable of walking through this challenging but interesting journey.

I would also like to acknowledge the Qatar National Research Fund (a member of The Qatar Foundation) for their support and funding provided through the NPRP grant # (NPRP-5-165-2-055).

I wish to express my special gratitude to Dr. Ahmed Atef Youssef for his endless support in offering his time and continues assistance. I also acknowledge Dr. Laya Parvizsedghy for her valuable feedback and advice. In addition, I would like to thank my other colleagues in the Construction Automation Laboratory (BCEE). I had the opportunity to work in a professional and friendly environment. I would like to acknowledge the support of Ms. Nathalie Oum by sharing City of Montreal water distribution network data.

I feel indebted to my beloved parents for my entire life. They have taught me to believe in myself and encouraged me unconditionally. Sincere appreciation goes to my brothers for their continuous help and support.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Water pipelines are intensive capital assets, preserved through operation and maintenance, to meet customers' expectations and avoid catastrophic failures (Giustolisi et al. 2006). The 2013 American Society of Civil Engineers Report Card (ASCE 2013) rated the US drinking water networks with a score of D, which is interpreted as "Poor" condition. According to the American Water Works Association (AWWA), there are 240,000 water main breaks per year in the United States, imposing a total cost of $1 trillion on municipalities over the coming decades. Also, as the Canadian Infrastructure Report Card 2012 (CIRC 2012) shows, municipal drinking-water networks are ranked "Good: Adequate for now". Despite this overall rating, 15.4% of water distribution systems in Canada were ranked "fair" to "very poor" with a replacement cost of CAD 25.9 billion. The "fair", "poor" and "very poor" conditions would be interpreted as deterioration beginning to be reflected, nearing the end of useful life and no residual life expectancy respectively (CIRC 2012). Water main deterioration leads to a breakage rate increase and a hydraulic capacity decrease. According to CIRC 2012, 86 Canadian municipalities own a total of 719,630 km of water pipelines containing distribution pipes (≤350 mm diameter) and transmission pipes (>350 mm diameter).

According to the CIRC 2012, the majority of municipalities in Canada do not have complete data for buried infrastructure networks, including water and sewer networks. Besides, it is clear that testing, inspection and evaluation of the pipe physical specifications require a large amount of financial reserves, and in some cases, it is difficult to implement. For operators and managers, it is vital to develop models that can estimate the breakage rate of water pipes by using their available and limited historical data instead of relying on

models that either require extensive data collection practices or physical testing of pipes. If these models can detect the factors that are critical for estimating the breakage rate and utilize them to predict it, this will have a profound impact on decreasing their required operational budget. Recently, a data-mining technique titled Evolutionary Polynomial Regression (EPR) was developed by Giustolisi and Savic (2006). This type of regression generates several symbolic expressions that are understandable by specialists and professionals, based on various independent variables.

1.1 Problem Statement

This research has been inspired by a lack of comprehensive analysis in the water pipe failure prediction models. In accordance with the importance of water distribution networks, the major limitations with respect to this research are briefly described in this section. There is a lack of computational models to predict water pipe failure rate, to be generic and not limited to certain physical characteristics (Berardi et al. 2008). The majority of developed models in literature were limited to pipes with certain material type or diameter.

Furthermore, current practices do not justify why certain factors were selected for predicting the breakage rate (Berardi et al. 2008 and Xu et al. 2011). There is a need for a more comprehensive approach that starts with examining available datasets to extract factors statistically critical for predicting the breakage rate. As will be demonstrated later in this research, extracting and utilizing the most critical factors to estimate the breakage rate will improve the obtained statistical results.

In addition, current researches effort focused on modeling the water pipe failure rate without considering the interrelationships between considered variables and subsequently on estimating the failure rate. There is a need to develop failure rate models that consider such interrelationships with the ability to test and to determine the best mathematical symbolic expression to recognize the correlations among dependent and independent variables.

1.2 Research Objectives

The main objective of this research is to develop a generic framework for predicting water pipe failure rate. This main objective can be achieved through the following sub-objectives:

1) Identify and study the critical factors of predicting the number of breaks of water mains.
2) Develop models to predict the number of breaks of water mains.
3) Develop deterioration curves to predict the future condition of water pipelines.
4) Perform sensitivity analysis to recognize the most sensitive factors to the number of breaks of water mains.

1.3 Research Framework Overview

The proposed research framework consists of 6 main parts as shown in Figure 1-1 and described below:

1) Literature Review: The literature review is performed to identify current studies' limitations, which need to be investigated in this research. It starts by outlining and discussing the components of water distribution networks. Then,

it focuses on revealing the current state of the art of: 1) factors utilized in predicting the water pipeline failure rate and 2) models for predicting those failure rate and their limitations.

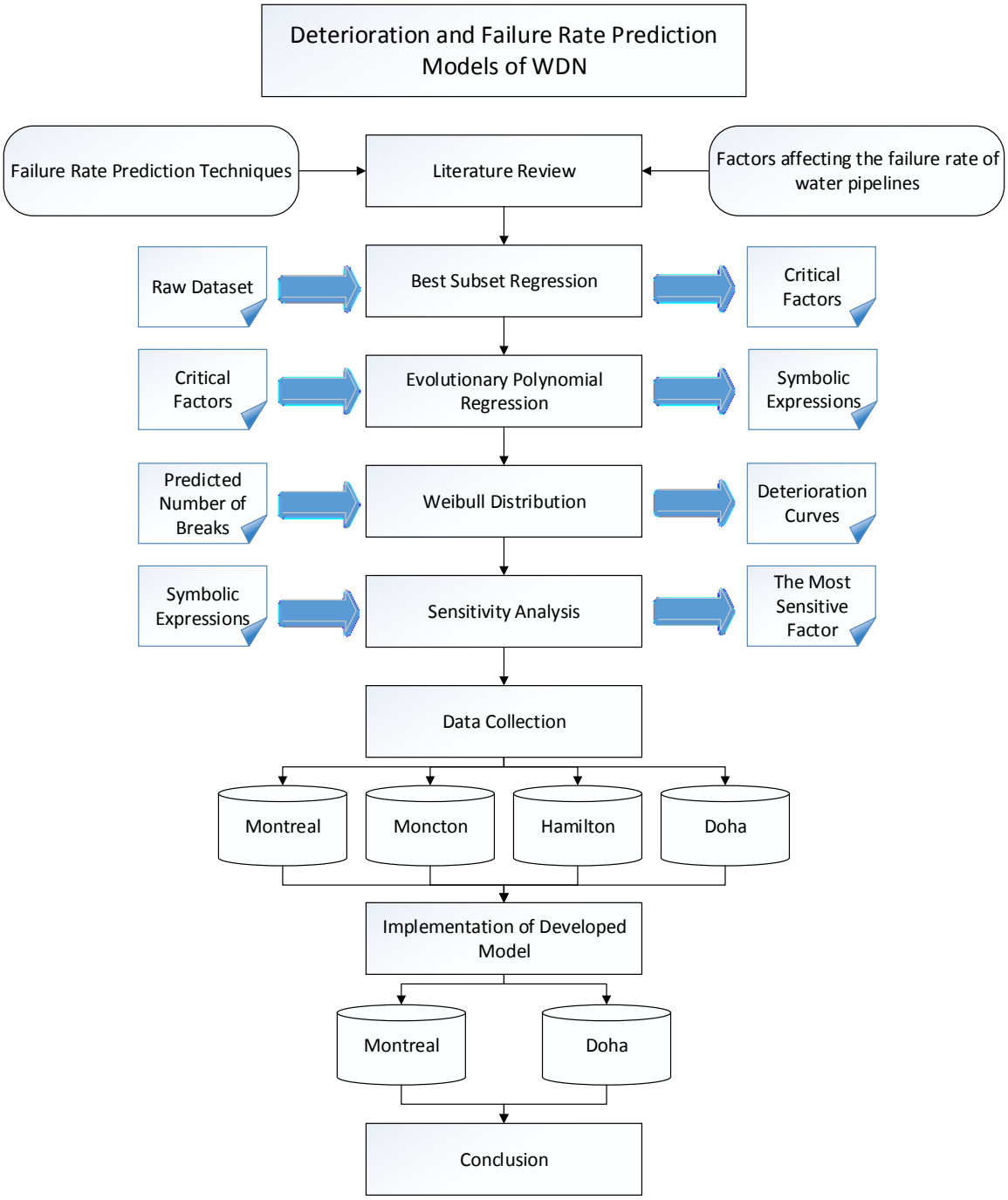


Figure 1-1 Research Framework Overview

2) Best Subsets regression: The best subset regression is an automated technique that recognizes the best-fitting regression models with factors specified by the user. In this study, this technique is used to find the best combination of independent variables to predict the number of breaks of water pipelines.
3) Evolutionary Polynomial Regression: This is a data-driven technique and is classified as a grey box method as it provides insight into the relationship between inputs and output (Giustolisi 2004). This method is performed in two stages: 1) a search for the best model using Multi-Objective Genetic Algorithm (MOGA), and 2) parameter estimation for the model using Least Square Method. The process of EPR should be coupled with engineering knowledge to verify if the generated equations and correlations between utilized inputs and output are reasonable. Two separated models are developed using datasets of Montréal and Doha, based on the most critical factors obtained from the best subset regression.
4) Weibull Distribution: Weibull reliability function is employed to generate deterioration curves as it poses three main advantages over the other methods to be described later in chapter 4. In general, Weibull-based models are widely used in different studies and applications to solve various problems (Jardine and Tsang, 2013). In this study, the value of number of breaks that is predicted using EPR is used to establish deterioration curves for several homogeneous clusters.
5) Sensitivity analysis: This technique is deployed to explore the effect of changing each input on the predicted output (i.e. number of breaks). Also, sensitivity analysis is utilized to verify if the existing relationships between the

selected inputs from the best subset algorithm and the predicted output from the EPR algorithm are reasonable in terms of engineering knowledge.

6) Data Collection: Four datasets of water distribution networks obtained from the City of Moncton, City of Hamilton and City of Montréal in Canada and the City of Doha in Qatar, were considered in this study. These four datasets were considered for understanding current practices of data collection. Their examination also serves us to obtain better understanding of the water pipe deterioration processes. In addition, these datasets are utilized to build up a comprehensive water pipeline assessment model for: 1) identifying the most critical factors, 2) determining the best mathematical form for predicting water pipe breakage rate and 3) providing deterioration curves and recognizing the most sensitive factors. Finally, a part of the same datasets was employed to check the proposed model's validity.

1.4 Thesis Organization

This thesis consists of 6 chapters and 3 appendices. The literature review is presented in chapter 2 and it starts with discussing the components of water distribution networks. Then, the factors and models to predict the failure rate in previous studies, along with their limitations, are presented. Evolutionary Polynomial Regression is described in details as well. At the end of this chapter, the limitations of previous studies are presented. Chapter 3 describes and analyzes four datasets: The City of Moncton, City of Hamilton and City of Montréal in Canada and City of Doha in Qatar. Chapter 4 contains the research framework and its developed models. Two case studies: City of Montréal and City of Doha, are used to test the developed model. Their analyses and results are presented in chapter 5.

Finally, chapter 6 highlights the contributions, limitations and recommendation of this study.

Chapter 2: Literature Review

2.1 Overview

This chapter starts by outlining and discussing the components of water distribution networks. The literature review focused on revealing the current state of the art of: 1) factors utilized in predicting the failure rate of water pipelines and 2) models for predicting the failure rate with their limitations. The factors utilized in predicting the failure rate of water pipelines were classified into two clusters based on; 1) whether these factors are static or dynamic through the lifecycle of water pipelines and 2) whether these factors are physical, environmental or operational. Failure rate models are reviewed with their drawbacks being highlighted. The failure rate models are clustered into four groups: deterministic, statistical, probabilistic, and artificial intelligence. Finally, this chapter concludes with a summary of the identified limitations in the previous studies. Figure 2-1 shows an overview of this chapter.

2.2 The components of Water Distribution Systems

Water distribution networks have three main parts: pipes, valves and flush hydrants. The pipes and valves are buried, thus the involved parties like municipalities and contractors need a detailed map to have a quick and precise access to the location of the pipes in case of emergency. Also, this map can be used in upgrading and improvement of the system.

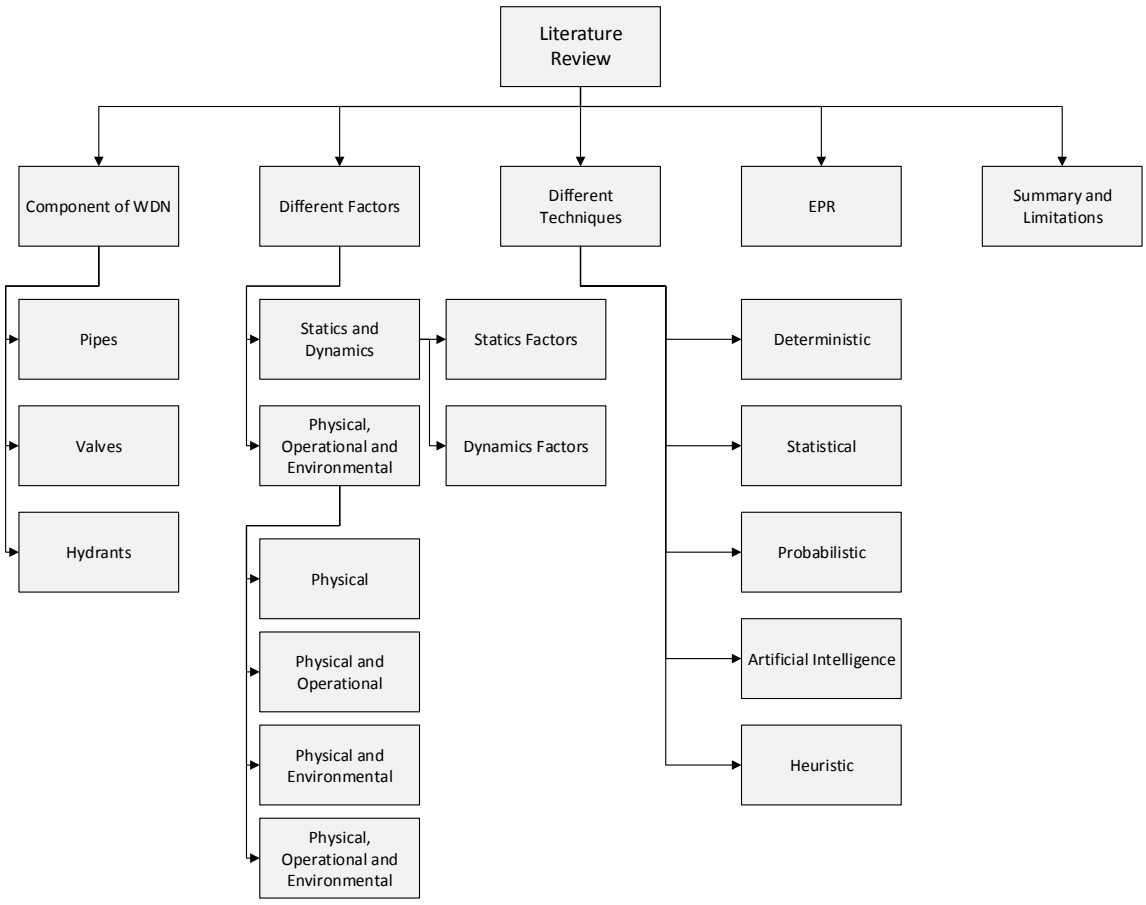


Figure 2-1 Overview of Literature Review

2.2.1 Pipes

As it can be seen in Figure 2-2, there are two main types of water pipes; transmission pipes and distribution pipes. Transmission pipes carry the water from the source to the treatment plant and storage tanks. These are the largest (>350 mm diameter) and thickest pipes in the system, therefore, the most expensive ones. For reducing the transmission cost, the location of the storage system should be as close as possible to the source of water.

Distribution pipes (≤350 mm diameter) carry out the water from storage tanks to the users. These pipes must be far at least 10 feet from sewers pipes and laid in separated

trench for water quality assurance purposes. The minimum diameter for distribution pipes is 2 inches while for serving the fire hydrant the 6 inches pipe is needed. To take into consideration the population growth, most of the decision makers try to use bigger pipes than the minimum size.

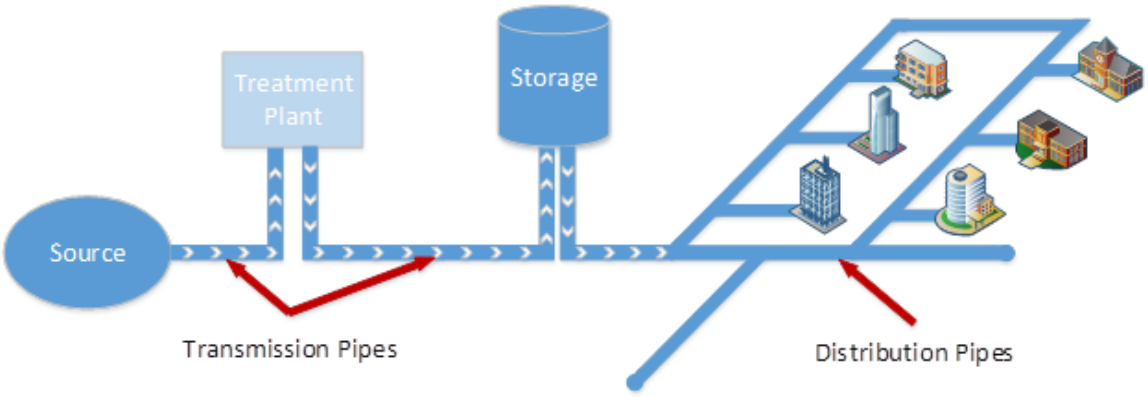


Figure 2-2 Water Supply Distribution System (Adopted from EPA, 2006)

Materials commonly used in water pipes can be divided into three main groups: metallic pipes, cement pipes and plastic pipes. Metallic pipes include gray cast iron pipe (GCIP), ductile cast iron pipe (DCIP), steel pipe and copper. Cement pipes such as asbestos cement (A.C.) pipe and in older systems concrete or fired clay. Plastic pipes include PVC (polyvinyl chloride) pipe and high-density polyethylene (HDPE) pipe.

2.2.2 Valves

Valves are one of the most important parts of water distribution networks. During the maintenance, valves can isolate the portion of the water that needs to be kept in the

system. Installing valves in suitable place minimizes the loss of service in water pipe rehabilitation and replacement. Valves that are not used for many years can be stuck or even broken if neglected. Thus, valve exercise program is an important part in water pipes maintenance.

2.2.3 Flush Hydrant

Flush hydrants are almost the only visible part of the water distribution networks. They must be located at the end of all lines to remove sediment, silt, rust, debris, or stagnant water from dead-ends. Flush hydrants should also be installed throughout the system to provide for periodic flushing to maintain high water clarity and quality. Fire hydrants are larger and more expensive than the flush hydrants and usually are connected to the larger pipes. But some of the municipalities use fire hydrants for flushing their lines.

2.3 Factors affecting the failure rate of water pipelines

In the last decade, the extensive research effort was made to develop models for predicting the failure rate of water pipelines. The factors utilized in these models were classified into two clusters based on; 1) whether these factors are static or dynamic through the lifecycle of water pipelines and 2) whether these factors are physical or environmental or operational. After reviewing previous studies, it was observed that the second classification is more common in recent research efforts.

2.3.1 Static and dynamic factors

Stone et al. (2002) categorized factors contributing to the failure of water pipelines into two groups: static factors and dynamic factors. The characteristics of static parameters do not depend on the time, but dynamic factors' specifications change over time. Static

parameters include the diameter, length, soil type, pipe material, etc. On the other hand, the age, cumulative number of breaks, soil corrosivity and water pressure are examples of dynamic factors influencing pipe failure rate. Osman and Bainbridge (2011) studied the effect of time-dependent variables like pipe age, temperature and soil moisture on the deterioration of water pipes. Static factors such as soil type, length, wall thickness and diameter of the pipe were not considered in their study because of the unavailability of reliable data.

2.3.2 Physical, environmental and operational factors

InfraGuide. (2003) classified the factors contributing to the deterioration of water pipes to three main categories; physical, environmental and operational as shown in Table 2-1. According to InfraGuide (2003), physical factors include pipe material, pipe wall thickness, pipe age, pipe vintage, pipe diameter, type of joints, thrust restraint, pipe lining and coating, dissimilar metals, pipe installation and pipe manufacture. In other researches, pipe length and buried depth are also known as physical factors.

InfaGuide (2003) considered pipe bedding, trench backfill, soil type, groundwater, climate, pipe location, disturbances, stray electrical currents, and seismic activity as the environmental factors. While, other researchers included rainfall, traffic and loading, and trench backfill as the environmental factors as well. Kabir et al. (2015b) studied the effect of soil type on the failure rate of water pipelines and highlighted that soil type can be classified further to major and minor factors. The five major soil's factors include soil electrical resistivity, soil pH, redox potential, soil sulfide contents and soil moisture. The five minor soil factors are; temperature of soil, oxygen contents, presence of acids, sulfates,

Table 2-1 Factors that contribute to water system deterioration (InfraGuide. 2003)

Factor		Explanation
Physical	Pipe material	Pipes made from different materials fail in different ways.
	Pipe wall thickness	Corrosion will penetrate thinner walled pipe more quickly.
	Pipe age	Effects of pipe degradation become more apparent over time.
	Pipe vintage	Pipes made at a particular time and place may be more vulnerable to failure.
	Pipe diameter	Small diameter pipes are more susceptible to beam failure.
	Type of joints	Some types of joints have experienced premature failure (e.g., leadite joints).
	Thrust restraint	Inadequate restraint can increase longitudinal stresses.
	Pipe lining and coating	Lined and coated pipes are less susceptible to corrosion.
	Dissimilar metals	Dissimilar metals are susceptible to galvanic corrosion.
	Pipe installation	Poor installation practices can damage pipes, making them vulnerable to failure.
	Pipe manufacture	Defects in pipe walls produced by manufacturing errors can make pipes vulnerable to failure. This problem is most common in older pit cast pipes.
Environmental	Pipe bedding	Improper bedding may result in premature pipe failure.
	Trench backfill	Some backfill materials are corrosive or frost susceptible.
	Soil type	Some soils are corrosive; some soils experience significant volume changes in response to moisture changes, resulting in changes to pipe loading. Presence of hydrocarbons and solvents in soil may result in some pipe deterioration.
	Groundwater	Some groundwater is aggressive toward certain pipe materials.
	Climate	Climate influences frost penetration and soil moisture. Permafrost must be considered in the north.
	Pipe location	Migration of road salt into soil can increase the rate of corrosion.
	Disturbances	Underground disturbances in the immediate vicinity of an existing pipe can lead to actual damage or changes in the support and loading structure on the pipe.
	Stray electrical currents	Stray currents cause electrolytic corrosion.
	Seismic activity	Seismic activity can increase stresses on pipe and cause pressure surges.
Operational	Internal water pressure, transient pressure	Changes to internal water pressure will change stresses acting on the pipe.
	Leakage	Leakage erodes pipe bedding and increases soil moisture in the pipe zone.
	Water quality	Some water is aggressive, promoting corrosion
	Flow velocity	Rate of internal corrosion is greater in unlined dead-ended mains.
	Backflow potential	Cross connections with systems that do not contain potable water can contaminate water distribution system.
	O&M practices	Poor practices can compromise structural integrity and water quality.

and sulfates reducing bacteria's.

The Internal water pressure, transient pressure, leakage, water quality, flow velocity, backflow potential, and O&M practices are examples of operational factors (InfraGuide 2003). Others considered the nature and date of last failure (e.g., type, cause, severity), nature of maintenance operations (e.g., TV inspections, pipe cleaning, cathodic protection), nature and date of last repair (e.g., type, length), water quality and construction method, as operational factors that affect water pipe's failure rate.

Researchers either used a single group of factors (i.e. physical only) or a combination of these groups to predict the failure rate of water pipelines (physical and operational, physical and environmental and physical, operational and environmental).

I. Physical factors

For the physical factors, the impact of these factors on predicting the failure rate of pipes was examined by several researchers (Berardi et al. (2008), Wang et al. (2009), Xu et al. (2011), Aydogdu and Firat (2014), Arsénio et al. (2014), Jenkins et al. (2014) and Kutyłowska (2015)). Berardi et al. (2008) utilized the six following factors for each pipe: 1) number of pipe's breaks recorded during the monitoring period; 2) pipe age; 3) number of properties supplied; 4) pipe length and 5) pipe nominal diameter (up to 250 mm). The whole dataset were clustered into several homogeneous groups (class) based on the age and diameter of the pipe. The authors considered age, length, diameter, number of properties supplied and number of pipes in each class as the inputs and number of pipe's breaks as the output. It should be noted that they did not take into consideration the material of the pipe as the input.

Wang et al. (2009) divided the dataset of Québec City into five groups based on the pipe material: gray cast iron, ductile iron with, ductile iron without lining, PVC, and concrete pipes. They considered three factors as the independent variables including pipe length, pipe age, and pipe diameter. In addition, higher orders and interactions of the first order terms of L, A, D such as: square of length, square of pipe age, square of pipe diameter, interaction of length and age (L*A), interaction of length and diameter (L*D), and interaction of age and diameter (A*D), are included in their inputs as well. These inputs are used to improve the accuracy of the model. They observed that pipe length had a great impact on the water pipe's failure. Xu et al. (2011) established a relationship between the number of pipe breaks and the following physical factors, the age, length and year of installation (age). The dataset of Beijing City was aggregated into several homogeneous groups based on the pipe diameter and pipe age. This database was divided into two parts based on the observation date, one of them was used for model development, and the other one was used for validation. They did not consider pipe material as input as well.

Aydogdu and Firat (2014) estimated the failure rate considering the age, diameter and length of water pipes as the independent variables. Historical records from the City of Malatya in Turkey during 2006–2012 were selected to develop and test their model. The authors divided the dataset to three groups based on the pipe material: PVC, cast iron and asbestos cement pipes. Then, they studied the relationship between the failure rate and the above-mentioned factors for each group separately. Aydogdu and Firat (2014) observed that the failure rate for the following three groups of pipes was the highest: pipes with lengths of 0–200 m, pipes with diameters of 110 cm, and pipes with ages in the interval of 15–20 years.

Arsénio et al. (2014) took into consideration the ground movement and pipe age as the inputs to estimate the breakage rate of the water mains. The water distribution network of an unknown Dutch drinking water company was selected as the study area. This dataset includes three types of pipes: PVC, asbestos cement and cast iron pipes. The authors demonstrated that the failure rate of pipes of all materials located in areas with high probability of ground movement was higher than the others. However, they did not consider other physical factors such as diameter, length and material of the pipes. While according to the previous finding physical factors are the most significant variables in water pipe failure occurrence.

Jenkins et al. (2014) addressed the problem of uncertain and limited data in Weibull hazard rate models for water distribution networks. They tried to fill the gap of data that were unknown material type and installation date. Whereas pipe length is used as the explanatory variable in many statistical models, the uncertainty associated with fitting the segment lengths, made it impossible to consider length in the model. Data had been provided by large utility that is located in the southeastern United States. Kutyłowska (2015) considered material, length, diameter, and installation date of the pipes to predict the failure rate of water mains. Historical data was collected from a Polish water distribution network during 2001-2006. They used 50%, 25%, and 25% of the database for training, testing and validation respectively.

II. Physical and Operational factors

Moliga et al. (2007) and Shirzad et al. (2014) added more parameters from various categories (operational and physical) as the independent variables to improve the reliability of their models. Moliga et al. (2007) identified a homogeneous group of cast iron water

mains by selecting pipes installed between 1953 and 1969 in Australia's database. This population was about 23% of the total network length. The pipes with the diameter less than 40mm were not included in this cohort. The explanatory variables in this study were age, length, diameter, wall thickness, corrosion rate, and operating water pressure.

Shirzad et al. (2014) took into the consideration an operational factor like hydraulic pressure in addition to physical factors to forecast the pipe burst rate. The age, length, diameter and buried depth were the physical parameters in their study. The authors collected their data from two cities in Iran: the City of Mashhad and the City of Mahabad. Asbestos pipes with diameter between 80 and 300 mm and polyethylene pipes with diameter between 32 and 160 mm were considered in Mashhad's database and Mahabad's database respectively.

III. Physical and Environmental factors

There has been an extensive effort in the previous studies to assess impact of physical and environmental factors on the failure rate prediction models of water mains (Asnaashari et al. (2013), Nishiyama and Filion (2014), Francis et al. (2014), Kabir et al. (2015a), Kimutai et al. (2015), and Kabir et al. (2015b)). Asnaashari et al. (2013) considered the soil type as an environmental factor, while the physical parameters were length, age, diameter and material of pipes. Moreover, the date of cement mortar lining (if implemented) and the date of cathodic protection (if implemented) were added to independent variables. They applied their model to predict pipe failure rate in the City of Etobicoke, Ontario, Canada. Based on the analysis of historical data, they found that failure rate is decreased following the initiation of the CP and CML programs.

Nishiyama and Filion (2014) developed a model to forecast pipe breaks in cast iron water mains considering the diameter, age and length of the pipes as the physical factors and the soil type as the environmental factor. The data was collected from the City of Kingston, Ontario. It contains cast iron, ductile iron, PVC, and concrete pressure pipes (CPP). The reduction in failure rate was observed in Kingston West, Kingston Central, and Kingston East because of the old pipe removal.

Francis et al. (2014) collected the pipe breaks and location data from a large city in Mid-Atlantic United States during 2010-2011 to construct a knowledge model for water pipe breaks. They were not able to collect pipe characteristics such as pipe age, pipe length and pipe material. Instead, they tried to gather publicly available proxies for some of this information. For example, they used the average house age at the census tract level to reach the approximate age of the water distribution network of that area. Also, population density was included in their study as a proxy for intensity of water use. They tried to find the possibility of correlations in population characteristics such as age, ethnic and racial composition with pipe age. Several soil types and some weather characteristics were considered as the environmental factors in their study. It should be mentioned that estimation method of pipe age and intensity of water use was novel but might be not accurate enough to model water pipe's breaks.

Kabir et al. (2015a) tried to develop a failure rate prediction model of water mains considering several physical factors (pipe diameter, pipe length, pipe age, and vintage) and environmental factors (freezing index, rain deficit, soil resistivity, soil corrosivity index, and land use). This model was implemented to predict the failure rate of cast iron and ductile iron pipes in the database of the City of Calgary, Alberta, Canada. The results

indicated that the behavior of CI and DI pipes is different from input's effect. Also, CI and DI pipes are more sensitive to soil resistivity and soil corrosivity index respectively.

Kimutai et al. (2015) studied effects of different covariates on the failure rate of water pipes. Pipe length, pipe diameter and pipe type were included in their study as the physical variables while they considered soil resistivity, freezing index (temperature), and rain deficit (precipitation) as environmental variables. Water distribution network of the City of Calgary was utilized as the case study. They concluded that the effect of physical factors on the failure rate of water mains were more significant than environmental factors.

Kabir et al. (2015b) considered pipe characteristics like age, diameter, length and vintage or manufacturing period to develop a failure rate prediction model for cast iron and ductile iron water mains. Also, soil resistivity and soil corrosivity index were taken into consideration to explore the dependence of the actual failure rate, soil resistivity and soil corrosivity index. Higher order and logarithmic factors (i.e. A2, log A) were included among independent variables in order to improve the accuracy of the model. This information was collected from water distribution network of the City of Calgary, Alberta, Canada. This database comprises different pipe types such as ductile iron (DI), cast iron (CI), asbestos cement concrete and concrete cylinder pipes, steel, copper, and plastic pipes.

IV. Physical, Operational and Environmental factors

Some others included physical, environmental and operational parameters at the same time to improve the effectiveness and robustness of the failure rate prediction models (Jafar et al. (2010), Wang et al. (2010), and Kabir et al. (2014)). Jafar et al. (2010) tried to model the failure rate and estimate the optimal replacement/rehabilitation time for an

individual pipe in water distribution networks. They employed five physical factors (pipe material, pipe diameter, pipe length, wall thickness, and pipe age), an operational factor (hydraulic pressure), and two environmental factors (soil type and pipe location) as the explanatory variables. The database was constructed by collecting 14 years historical data (during the observation period between 1991 and 2004) from a city in the north of France.

Wang et al. (2010) estimated the condition of the water pipes considering ten physical, environmental and operational parameters. At first, the factors were the diameter, age, coating (inner and outer), soil condition, bedding condition, trench depth, electrical recharge, operational pressure, material (steel, cast iron, and ductile iron), and the number of road lanes. Then after some numerical experiments of different factor combination, it was cleared that water pipe condition can be assessed without information of road lane, trench depth, and electric recharge. While, pipe age is the most important factor in assessing pipe condition. Kabir et al. (2014) studied the risk of failure of metallic water pipes (cast iron, ductile iron, galvanized, and steel) using a large variety of physical, environmental and operational factors. The considered factors were the diameter, age, length, wall thickness, water pressure and velocity, turbidity, free residual chlorine, color, season, water pH, freezing index, soil resistivity, soil pH, redox potential, sulphide content, moisture content, population, land use, and traffic and road type. All parameters were collected from water distribution network of the City of Kelowna, British Columbia, Canada.

The summary of all aforementioned studies is shown in Table 2-2. Figure 2-3 shows the frequency of parameters which were used in 19 different previous works including: industry and academia, for each category (physical, environmental and

operational). By examining closely these results, out of the 17 reviewed factors, there are nine physical factors, seven environmental factors, and just one operational factor. It shows the importance of physical factors in modeling failures of water pipes. Also, Kimutai et al. (2015) confirmed that physical factors are more critical in estimating the failure rate than environmental factors. In Figure 2-3, it is obvious that the most frequent factors utilized in previous studies to predict the failure rate of water pipelines are; age, diameter, length, soil type, and pipe material. Berardi et al. (2008) stated that pipe age, diameter and length are the most important variables in describing water pipe failure occurrence. Also, Wang et al. (2009) concluded that length has a great impact on water pipe's failure. Thus, in this study the major physical factors like age, diameter, length and pipe material are considered as the independent variables to predict the number of breaks and failure rate of water pipelines.

Table 2-2 Considered factors affecting water pipes failure rate by different researchers

	Physical Factors													Environmental Factors										Operational Factors						
	Pipe Material	Pipe Wall Thickness	Pipe Age	Pipe Length	Pipe Vintage	Pipe Diameter	Type of Joint	Thrust Restraint	Pipe Lining and Coating	Dissimilar Metals	Depth Laid	Pipe Installation	Pipe Manufacture	Pipe Bedding	Trench Backfill	Soil Type	Groundwater	Climate	Pipe Location	Disturbances	Stray Electrical Currents	Traffic and Loading	Seismic Activity	Internal Water Pressure, Transient Pressure	Leakage	Water Quality	Flow Velocity	Backflow Potential	O&M Practices	Other Factors
Moglia et al. (2007)		✓	✓	✓		✓																		✓						corrosion rate
Berardi et al. (2008)			✓	✓		✓																								Number of Properties Supplies
Wang et al. (2009)	✓		✓	✓		✓					✓																			
Jafar et al. (2010)	✓	✓	✓	✓		✓										✓			✓					✓						
Wang et al. (2010)	✓		✓			✓			✓					✓	✓	✓					✓	✓		✓						
Xu et al. (2011)			✓	✓		✓																								
Asnaashari et al. (2013)	✓		✓	✓		✓			✓							✓														
Arsénio et al. (2014)			✓																											Ground Movement
Shirzad et al. (2014)			✓	✓		✓					✓													✓						
Aydogdu and Firat (2014)			✓	✓		✓																								
Nishiyama and Filion (2014)			✓	✓		✓										✓														
Kabir et al. (2014)		✓	✓	✓		✓										✓						✓		✓		✓				
Jenkins et al. (2014)			✓			✓																								
Francis et al. (2014)			✓													✓		✓	✓											
Kutyłowska (2015)	✓		✓	✓		✓																								
Kabir et al. (2015a)			✓	✓	✓	✓										✓			✓											Number of Connection for Each Pipe
Kimutai et al. (2015)	✓			✓		✓										✓		✓												Soil Resistivity, Freezing Index, and Rain Deficit
Kabir et al. (2015b)			✓	✓	✓	✓										✓														Soil Resistivity and Soil Corrosivity Index

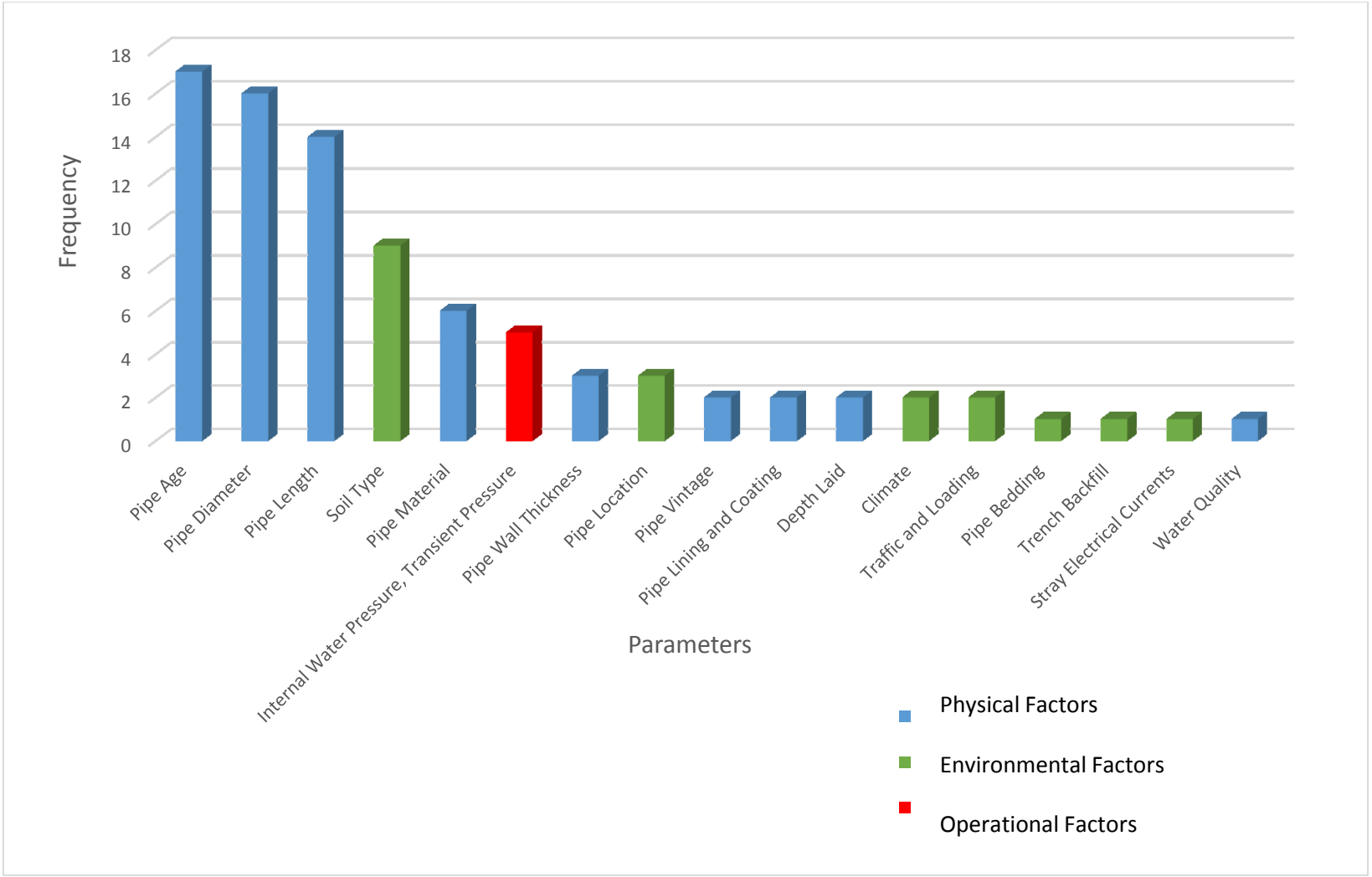


Figure 2-3 Frequency of effective parameters

2.4 Failure Rate Prediction Techniques

During the last three decades, researchers developed different models to predict the failure rate of water pipes for a reliable infrastructure management. These failure prediction models are classified into four categories; deterministic, statistical, probabilistic, artificial intelligence models such as artificial neural networks (ANN) and fuzzy logic. A summary of the reviewed models is shown in Table 2-3.

2.4.1 Deterministic Models

Deterministic models usually are used in cases where the relationship between inputs and output is clear. In two approaches the deterministic models can be applied: empirical and mechanistic. Empirical approach tries to find the relation between failure rates as the output and the features and attributes of a group of pipes as the inputs. While, the mechanistic approach can forecast the remaining useful life of an individual asset (just one pipe). The problem of these models is that a deterministic model can be applied just in specific location (Clair and Sinha 2012).

2.4.2 Statistical Models

This type of modeling is typically used to predict the useful life or time to failure of infrastructure assets (Lawless 1983). Statistical models are applied to homogeneous groups of pipes or other infrastructure assets and need recorded failures or data regarding asset's condition. In this approach, regression is utilized to build a model based on the historical data that can predict the failure or condition of water assets. In regression, the dependent variable is related to at least one of the independent variables.

Table 2-3 Prediction Models of Water Distribution Networks

Authors (Year)	Model Classification	Methodology	Output Type
Moglia et al. (2007)	Probabilistic	Monte-Carlo Simulation Framework	Probability of Failure for CI Pipes
Berardi et al. (2008)	Statistical	Evolutionary Polynomial Regression	Pipe Deterioration
Wang et al. (2009)	Statistical	Five Multiple Regression Models	Annual Break Rates
Li et al. (2009)	Probabilistic	Monte-Carlo Simulation	Remaining Useful Life
Jafar et al. (2010)	Artificial Intelligence	Six ANN Models	Failure Rate
Wang et al. (2010)	Statistical	Bayesian Inference	Deterioration Rate
Xu et al. (2011)	Statistical	Genetic Programming and Evolutionary Polynomial Regression	Deterioration Rate
Osman and Bainbridge (2011)	Statistical	Rate of Failure (ROF) and Transition State (TS)	Deterioration Rate
Asnaashari et al. (2013)	Artificial Intelligence	ANN and Multi Linear Regression	Failure Rate
Arsénio et al. (2014)	Statistical	Ground Movement Estimated by Radar Satellite Data	replacement-prioritization plan
Shirzad et al. (2014)	Artificial Intelligence	ANN and Support Vector Regression (SVR)	Pipe Burst

Authors (Year)	Model Classification	Methodology	Output Type
Aydogdu and Firat (2014)	Artificial Intelligence	Fuzzy Clustering and Least Squares Support Vector Machine (LS-SVM)	Failure Rate
Nishiyama and Filion (2014)	Artificial Intelligence	ANN	Pipe Breaks
Kabir et al. (2014)	probabilistic	Bayesian Belief Networks (BBN)	Risk of Failure
Jenkins et al. (2014)	probabilistic	Weibull Hazard	Failure Rate
Francis et al. (2014)	probabilistic	Bayesian Belief Networks (BBN)	Pipe Breaks
Kutyłowska (2014)	Artificial Intelligence	ANN	Failure Rate
Kabir et al. (2015a)	Statistical	Bayesian Weibull Proportional Hazard Model (BWPHM)	Failure Rate
Kimutai et al. (2015)	Statistical	Weibull proportional hazard model (WPHM), the Cox proportional hazard model (Cox-PHM), and the Poisson model (PM)	Pipe Failure
Kabir et al. (2015b)	probabilistic	Bayesian Belief Networks (BBN)	Failure Rate

It should be mentioned that this technique requires a large historical dataset that contains a number of data points collected over a period to develop a promising statistical model (Clair and Sinha 2012). Table 2-2 shows that in recent years many researchers have utilized statistical models (number of regression models) to forecast water pipes failure or pipes condition. There has been an extensive effort during the past decades to develop the failure rate prediction model by using statistical approach (Berardi et al. (2008), Wang et al. (2009), Wang et al. (2010), Xu et al. (2011), Osman and Bainbridge (2011), Arsénio et al. (2014), Kabir et al. (2015a), and Kimutai et al. (2015)). Berardi et al. (2008) developed a water pipe deterioration model using Evolutionary Polynomial Regression. As it is mentioned before, they used a dataset that was classified into homogeneous groups based on the age and diameter of the pipe. The developed model can predict the number of breaks in each group. Then, for predicting the failure rate for each pipe, a general structural deterioration model based on EPR aggregated model was developed.

Wang et al. (2009) utilized five multiple regression models for different pipe materials (gray cast iron, ductile iron without lining, ductile iron with lining, PVC, and hyprescon) to predict the annual break rate of individual water pipe rather than a homogeneous group. The overall model robustness was measured by F-test and the significant of each independent variable was measured by t-test. The model was validated using 20% of their collected dataset that was randomly selected. Wang et al. (2010) employed the Bayesian inference to assess the condition of water pipes. Ten factors from three pipe materials (cast iron, ductile cast iron, and steel) were used to generate factor weight. Based on the results of their model, the age of pipe is the most critical variable

while, the model was not sensitive to some factors like trench depth, electrical recharge, and some road lanes.

Xu et al. (2011) developed two prediction models for failure rate using Evolutionary Polynomial Regression and Genetic Programming, and then they compared the results of these two models. Results were measured based on; 1) error between predicted and actual data, 2) parsimony of generated equation, and 3) ability to justify the generated equations based on the engineering knowledge. The results showed that EPR has some advantages over GP in equation uniformity and parameters estimation, while GP was better to find the complex relations. Osman and Bainbridge (2011) employed two statistical deterioration models to predict future failures of water pipes: rate-of-failure models (ROF) and transition-state (TS) models. ROF model extrapolates the failure rate for a specific group of water pipes that were classified based on age and some environmental factors. This model does not differentiate the times between successive pipe breaks for an individual segment while, the transition-state model focuses on finding the time between successive failures for the water pipes. TS models are dependent on the availability of sufficient and accurate data, but ROF models can be applied to limited historical data.

The stresses in the buried pipes, which increase the probability of pipe failure, might be caused by the ground movement. This is a hypothesis that Arsénio et al. (2014) have worked on it. They estimated the ground movement using radar satellite data. Two different analyzes were done in their study: cell-based and pixel- based. The number of breaks of three types of water pipe was investigated: asbestos cement, PVC, and cast iron pipes. Kabir et al. (2015a) presented Bayesian Model Averaging method (BMA) to select the most critical explanatory variables. Then the Bayesian Weibull Proportional Hazard

Model (BWPHM) is applied to provide the survival curves and to forecast the failure rate of two pipe types: cast iron and ductile iron.

Kimutai et al. (2015) studied the effect of different independent variables on predicting the failure rate of water pipes using three statistical models: the Weibull proportional hazard model (WPHM), the Cox proportional hazard model (Cox-PHM), and the Poisson model (PM). Also, they used curve fitting techniques to estimate a baseline hazard function equation for the Cox-PHM and applied it on a dataset from the City of Calgary. The predicted breaks and actual breaks were compared using root mean square error (RMSE), mean absolute error (MAE), root relative squared error (RRSE) and relative absolute error (RAE).

2.4.3 Probabilistic Models

Probabilistic models analyze the probability of an event occurring (Creighton 1994). The probability of occurrence is one and the probability of the event that cannot happen is zero. The other probability of occurrence should be between 0 and 1 (Mitrani 1998). Information about asset conditions and attributes are required to develop a probabilistic model. The output or dependent variable would be a range of values instead of the specific number. These models need extensive data and typically used in infrastructure assets (Clair and Sinha 2012). It should be noted that the probabilistic approach commonly increases the computational complexity of the models (Moglia 2007). As shown in Table 2-2, many studies employed the probabilistic approach to develop water mains assessment models (Moglia et al. (2007), Li et al. (2009), Kabir et al. (2014), Jenkins et al. (2014), Francis et al. (2014), and Kabir et al. (2015b)). Moglia et al. (2007) developed a physical probabilistic failure prediction model based on the fracture mechanics of cast

iron water pipes. The random independent variables were added to the inputs, and then Monte-Carlo simulation technique was applied to deal with the computational complexity of the model. The developed model without failure data, degradation and load data, was not capable of estimating failure rates of water pipes. Whereas, with these data, it can predict failure rates more accurately.

Li et al. (2009) used the mechanically-based probabilistic model to predict remaining useful life and failure probability of buried pipes. They considered the effect of random inputs and used Monte-Carlo simulation framework to calculate cumulative distribution function (CDF) of remaining useful life of pipelines. But, they did not consider the correlation of defects for a pipeline having more than one corrosion defects. Also, they found CDF more suitable than probability density function (PDF) and reliability index in describing the probability of failure.

Kabir et al. (2014) assessed the risk of failure of metallic water pipes using a Bayesian Belief Network (BBN). Bayesian Belief Network can be interpreted as a probabilistic graphical model that can represent a collection of some covariates and their probabilistic relationships. This model recognizes the most vulnerable and sensitive pipe segments through the water pipe networks. The proposed model is good just for small to medium utilities with limited data. Jenkins et al. (2014) tried to address the problem of limited, incomplete, or uncertain data in water distribution networks. Two main modification were added to Weibull hazard rate models (WPHM) to improve the prediction performance of the models: the expert opinion and the spatial analysis. But these two modifications were not tested in the other utilities.

Francis et al. (2014) analyzed the water distribution systems to develop a pipe breaks prediction model using Bayesian Belief Networks (BBNs). They illustrated that assessing water pipe network is not only important for the failure prediction model but also is crucial for avoiding water loss and water quality degradation. Kabir et al. (2015b) stated that uncertainty regarding quality and quantity of databases became a major concern for failure prediction model development of infrastructure assets. Thus, they tried to reduce these uncertainties by developing failure prediction model for water mains using a new Bayesian belief network based data fusion model. The proposed model can identify the most vulnerable and sensitive pipe in the entire network, as well as the total number of pipes that require the immediate and appropriate action like maintenance, rehabilitation, and replacement.

2.4.4 Artificial Intelligence Models

In this literature review, Artificial intelligence models include Artificial Neural Networks and Fuzzy set theory models.

I. Artificial Neural Networks

Artificial Neural Network (ANN) is a method that can predict pipe failure and deterioration of infrastructure specially buried pipes. The ANN follows the pattern of the human brain using its generalization capabilities. Thus, this technique is able to process information even under large, complex, and uncertain environment. The high-quality database is needed for supervised training and forecasting the future condition of the pipes. Moreover, ANN needs several controlling factors including: number of hidden layers, the number of neurons in each hidden layer, activation functions, the number of training epochs, learning rate, and momentum term. However, ANN is considered as a "Black-

Box" technique. Therefore, it is not able to provide insight into the relationship between dependent and independents variables (Clair and Sinha 2012; Moselhi and Hegazy 1993, Atef et al. 2015, Shirzad et al. 2014).

Jafar et al. (2010) employed Artificial Neural Network (ANN) to analyze the urban water mains. Six ANN models that predict the failure rate of water pipes of a city in France were developed then, they tried to estimate the optimal rehabilitation/replacement time for the same network. These prediction models were tested and validated using cross-validation. In the first part of this article, data collection was explained then development and validation of ANN models were discussed. In the data collection part, correlation and chi2 method were applied to select the most critical inputs.

Asnaashari et al. (2013) studied two different methods to forecast the water pipe's failure rate. Multi Linear Regression (MLR) and Artificial Neural Networks (ANN) were utilized, and their results were compared. The value of R-Squared showed that the ANN model (R2=0.94) is more promising while the MLR technique (R2=0.75) is just good enough for preliminary assessment. Shirzad et al. (2014) compared the predictive performance of Artificial Neural Network (ANN) and Support Vector Regression (SVR) in forecasting the water pipe's breakage rate. In addition, they investigated the effect of hydraulic pressure (average and maximum hydraulic pressure values) on precision of predicting the pipe's failure rate. The results showed that the ANN model is more accurate, but it is not suitable for generalization purposes. Thus for management purposes, SVR might be more appropriate.

Nishiyama and Filion (2014) developed a model to predict breaks in the water supply system of the City of Kingston, Ontario using Artificial Neural Networks. A feed-

forward back propagation algorithm was utilized to improve the performance and minimize the errors. Moreover, they employed the mean square error, receiver operating characteristics curves, and a confusion matrix in order to measure the accuracy of their model. Kutyłowska (2014) predicted the failure rate of pipes in an urban water utility using ANN. They employed quasi-Newton approach to train the model. The house connections and distribution pipes are considered as two different sections in database, and the results for both were acceptable. According to the author, simplicity is the advantage of this model.

II. Fuzzy Logic

Fuzzy Logic is a mathematical method in the field of artificial intelligence that widely used by researchers to assign a value to a certain degree of membership instead of crisp values such as zero and one. This method is known to deal with systems that are subject to uncertainties and ambiguities. Fuzzy Logic is applicable in infrastructure assets like oil and gas, water, bridges and highways (Siler and Buckley 2005, Clair and Sinha 2012). Aydogdu and Firat (2014) incorporated two methods: fuzzy clustering and Least Squares Support Vector Machine (LS-SVM) in order to estimate the failure rate of water pipes. At first, they developed failure rate estimation model using LS-SVM, and then fuzzy clustering method is utilized to define nine sub-regions for predictive performance improvement of the model. Afterward, the results were compared to the results of Feed Forward Neural Network (FFNN) and Generalized Regression Neural Network (GRNN) methods. Finally, for model evaluation they employed some measurement indexes such as Correlation Coefficient (R), Efficieny (E) and Root Mean Square Error (RMSE).

2.5 Evolutionary Polynomial Regression

The Evolutionary Polynomial Regression (EPR) technique was first presented by Giustolisi and Savic (2006). The technique utilizes the huge potential of conventional numerical regression techniques and the strength of Genetic Algorithm in solving optimization problems (Xu et al. 2011).

Later, this approach was used by other researchers in several engineering fields. Savic et al. (2006) and Ugarelli et al. (2008) used EPR to model the sewer pipe failures. Berardi et al. (2008) and Xu et al. (2011) applied the EPR to develop deterioration models for water distribution networks. Rezania et al. (2008) utilized the EPR methodology to evaluate the uplift capacity of suction caissons and shear strength of reinforced concrete deep beams. Elshorbagy and El-Baroudy (2009) compared the EPR and Genetic Programming to develop the prediction model of soil moisture response. Guistolisi and Savic (2009) tested the EPR-MOGA (an improved EPR) to develop groundwater level prediction model based on monthly rainfall. El-Baroudy et al. (2010) utilized the EPR to develop the evapotranspiration process then compared the efficiency of Evolutionary Polynomial Regression to Artificial Neural Networks (ANNs) and Genetic Programming (GP). Markus et al. (2010) applied EPR, ANNs and the naive Bayes model to forecast weekly nitrate-N concentrations at a gauging station. Ahangar-Asr et al. (2011) applied EPR to predict mechanical properties of rubber concrete. Fiore et al. (2012) used EPR to provide the predicting torsional strength model of reinforced concrete beams.

2.6 Summary and Limitation of Previous Studies

In this chapter, the water distribution networks and their components were covered. Factors affecting the water pipe failure rate were discussed along with their classifications.

According to the literature review, the most significant independent variables for predicting the failure rate of water pipes are the physical factors especially the age, length and diameter of water pipes. Subsequently, the failure rate models were categorized to four groups: deterministic, statistical, probabilistic, artificial intelligence such as artificial, neural networks (ANN) and fuzzy logic. The required inputs, outputs and limitations of each model were discussed.

Water pipes are capital intensive assets preserved through operation and maintenance to meet customers' expectations and avoid failures and consequent catastrophes. The expected life time of water pipes ranges between 100-150 years (Infraguide, 2003). A robust and promising deterioration model for water pipes can assist municipalities in making rational decisions about the replacement/rehabilitation time of water pipes. As seen in Figure 2-3, a few studies considered the pipe material as one of the independent variables. In most cases, datasets were clustered into different groups, based on the pipe material, and then one model was developed for each group. Thus, there are several models just for one network that might be tough to implement in the real world.

Several techniques were utilized by the other authors. Particularly, Artificial Neural Networks (ANN) are commonly used in many studies. ANN is able to develop accurate prediction models in complex and uncertain environments.

However, EPR is selected because it does not require large datasets for training and unlike ANN, it enables the recognition of correlations among dependent and independent variables. Being as such, EPR is not a "Black-Box" technique, but it is classified as a "Grey-Box" technique that can provide insight into the relationship between inputs and the output. The process of development and selection of EPR contains the engineering

knowledge that allows the user to understand the generated equations and correlation between variables involved. In ANN, each attempt delivers particular output, which can be different in other attempts with the same inputs and features, while, in EPR or generally regressions, all similar attempts lead to the same equations as the output.

Chapter 3: Research Methodology

3.1 Introduction

Figure 3-1 shows the developed research methodology. Chapter 3 starts by presenting the Best Subset regression that identifies the most critical factors for predicting the failure rate of water mains. The datasets of Montréal and Doha are classified into homogeneous groups, based on age, material and diameter of the pipes. Afterwards, these groups along with the factors selected for predicating the number of breaks using best subset regression are forwarded to the EPR algorithm. The EPR algorithm is used because – based on the selected factors – it generates some mathematical expressions able to predict the number of breaks of water pipelines. In this study, both datasets are analyzed with EPR in order to generate equations which provide insight into relationships between inputs and the output. The user selects the best symbolic expression for predicting the failure rate based on two criteria: 1) fitness to the historical data, and 2) parsimony of the equation. The predicted number of breaks, as the output of the EPR algorithm, is used to develop deterioration curves, using Weibull distribution function. A description of Weibull distribution is presented in this chapter as well. Finally, a Sensitivity analysis is deployed to explore the effect of changing each input on the predicted output (i.e. number of breaks). Also, sensitivity analysis is used to verify if the existing relationships between the selected inputs from the best subset algorithm and the predicted output from the EPR algorithm are reasonable in terms of engineering knowledge.

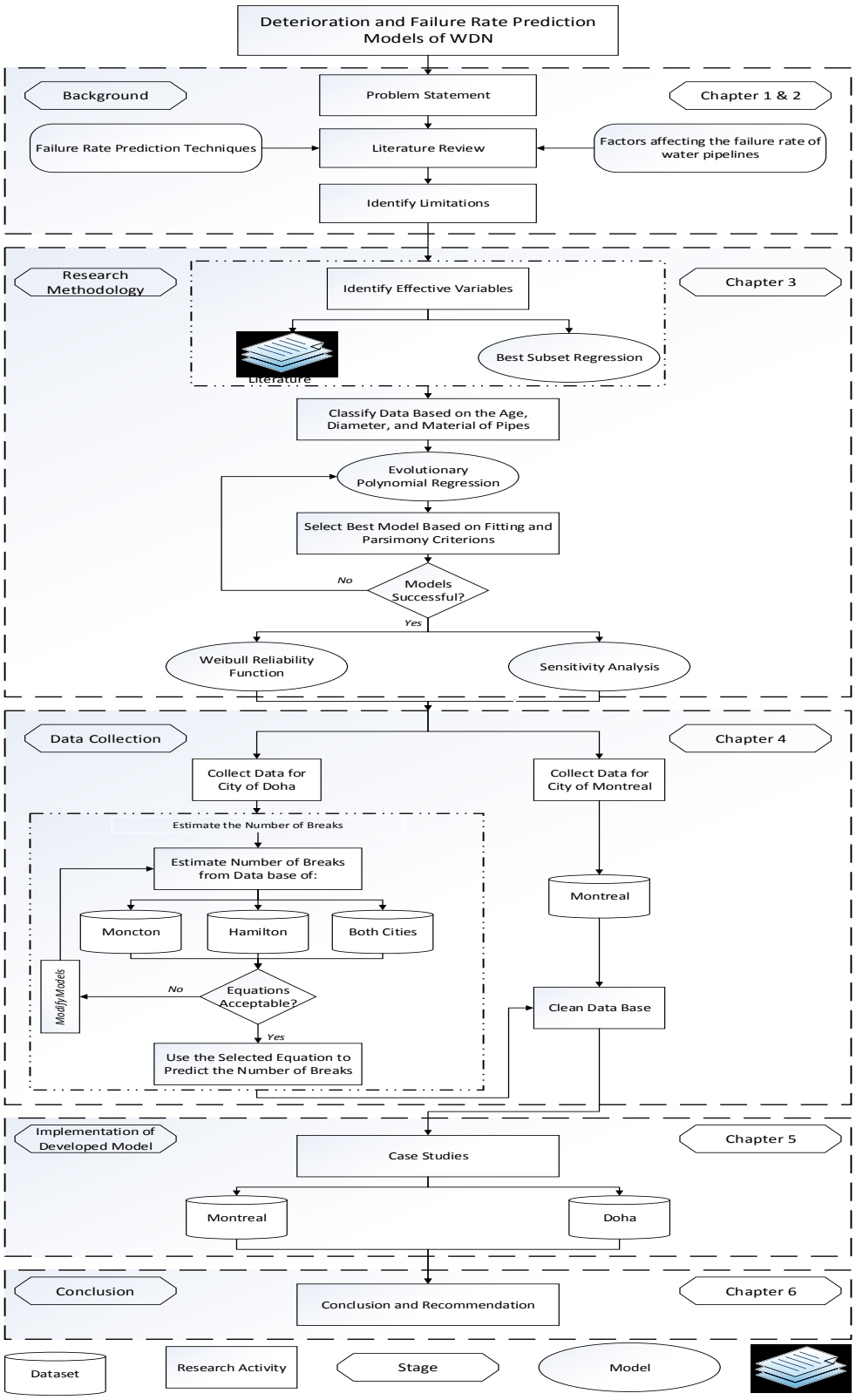


Figure 3-1 Research Methodology Flow Chart

3.2 Best Subset Regression

Best Subsets regression is an automated technique that recognizes the best-fitting regression models with factors specified by the user. In this study, Stepwise regression could be utilized as well. But, the Best Subset regression was selected because Stepwise regression does not assess all possible models. It rather constructs a model by adding and removing one variable at a time. Meanwhile, Best Subsets regression searches for all possible models and finally introduces the best candidates. Stepwise regression is simpler and Best Subset regression provides a model with more information (Minitab 17, 2015). Best Subset regression is not good for studies with a large number of independent variables. In such cases, finding the best combination of factors to predict water pipe failure rates and processing them take more time. But in this study, 4 to 5 independent variables are used to predict the failure rate in both datasets, which consequently makes using Best Subset regression suitable.

The independent variables for the water distribution network of the City of Montréal include 4 factors: Age, diameter, length and material of the pipes. The dataset of City of Doha includes 6 factors; Age, diameter, length, material, buried depth and elevation of pipes. The pipe material was almost constant for the entire dataset, thus it was excluded from this dataset. There were 1599 pipe segments, with only 3 steel pipes and the rest as ductile iron. Thus, these 3 segments were excluded from the dataset of Doha and only ductile iron pipes were considered in the Best Subset regression. The output in both cases is the number of breaks. The results of these two datasets are discussed in chapter 5.

The best subset regression is a procedure of finding the best combination of variables to predict the failure rate of water pipes on three main stages. First of all, all

possible combination of variables are identified. For example, if we have 7 independent variables, then there are 2^7 possible regression models. Secondly, out of all possible models, one or two models with the highest R-Squared among candidates with the same number of independent variables, are selected. The user can specify if a model as the best one is enough for same-size candidates or two or more models ought to be selected. Also, the minimum and maximum number of free predictors – i.e. independent variables – to add to the model can be specified by the user. Finally, further evaluation is required to select the best combination of independent variables, by using R-Squared, adjusted R-Squared, Mallows' Cp and square root of MSE (Iain Pardoe 2015, Minitab 17).

The selected combination of factors should have the highest R-Squared, the adjusted R-Squared and the smallest S (square root of MSE). The adjusted R-Squared penalizes the model when adding an extra independent variable does not improve the existing model's accuracy. In comparing models with the same size, the R-Squared is the most useful criterion. However, models with different number of independent variables are compared, based on the adjusted R-Squared and Mallows' Cp index (Wang 2006). The value of Mallows' Cp should be close to the number of predictors plus the number of constant terms, which is usually one (Minitab 17, 2015). For example, if there are 6 independent variables (predictors), the best model should have a Mallows' Cp close to 7. The R-Squared, adjusted R-Squared, and Mallows' Cp are calculated with equations number [1], [2], and [3] respectively:

[1] $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (SST = SSR + SSE)$$

[2] $$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SSE}{SST}\right) = 1 - \left(\frac{n-1}{SST}\right) MSE$$

[3] $$C_p = \frac{SSE_p}{MSE_{all}} - (n - 2p)$$

Where SSR is the sum of squares due to regression, SSE is the sum of squares due to error, n is the number of samples, p is the number of independent variables plus 1, MSE is the mean squared error, SSE_p is SSE for the best model with p predictors and MSE_{all} is the MSE for the model with all predictors (Iain Pardoe 2015). Furthermore, SSE, SSR and MSE are calculated with the following equations:

[4] $$SSR = \sum(\hat{Y}_i - \bar{Y})^2$$

[5] $$SSE = \sum(Y_i - \hat{Y}_i)^2$$

[6] $$MSE = \frac{SSE}{N-d}$$

Where $\bar{Y}$ is the average value of data, $\hat{Y}$ is the value predicted by the model, N is the number of samples and d is the number of independent variables. Figure 3-2 shows the average value line ($\bar{Y}$) and the best fit line ($\hat{Y}$) in a sample scatter plot.

Figure 3-3 illustrates a sample sheet of best subset regression in Minitab 17 statistical package. As it can be seen, there are two windows, the lower one containing the dataset sheet and the upper one showing the table of results. In this sample, there are 7 independent variables (V1 to V7) and one dependent variable (V8). The table of results in the upper window includes 13 columns and 14 rows. Column 1 shows the number of considered variables in each model and columns 2 to 6 show the value of R-Squared, adjusted R-Squared, predicted R-Squared, Mallows' Cp and S respectively. The last 7 columns specify which variables are in the model. Each row represents some information about a model. In this attempt, the user selects the best set of inputs from among 13 possible

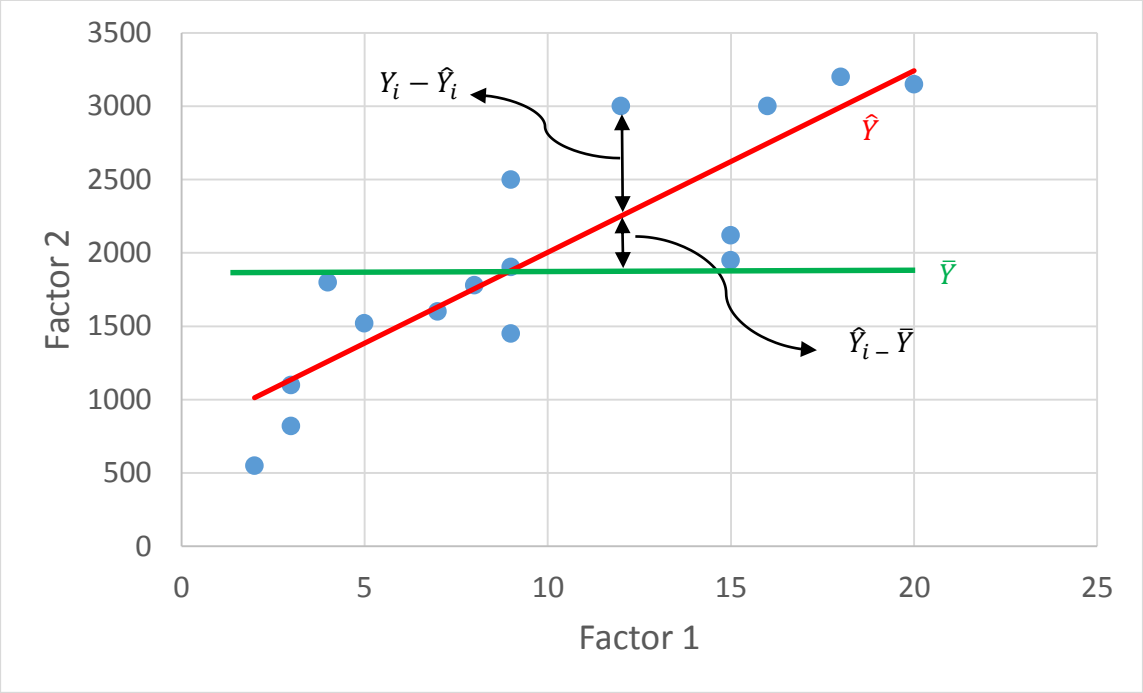


Figure 3-2 $\bar{Y}$ and $\hat{Y}$

combinations of variables. It can be concluded that the last model is the best one, because it has the largest R-Squared, the adjusted R-Squared and the lowest MSE. In addition, the value of Mallows' Cp is exactly 8, which is sum of number of variables (7) plus one.

Finally, the determined factors were considered as the independent variables to develop failure rate prediction model by using EPR.

3.3 Classification

Once, the process of factor selection is conducted, both datasets will be classified into several homogeneous groups, based on the age, diameter and pipe material. The objective of this classification is clustering pipe segments into classes with the same age, diameter and material. The following equations are used to achieve this objective:

[7] $$A_{class} = \frac{\Sigma_{class}(L_p \cdot A_p)}{L_{TA}}$$

[8] $$D_{class} = \frac{\Sigma_{class}(L_p \cdot D_p)}{L_{TD}}$$

Where, L_{TA} and L_{TD} are the total length of pipes with the same age and diameter respectively. Also L_p, A_p and D_p are length, age and diameter of each segment in the group

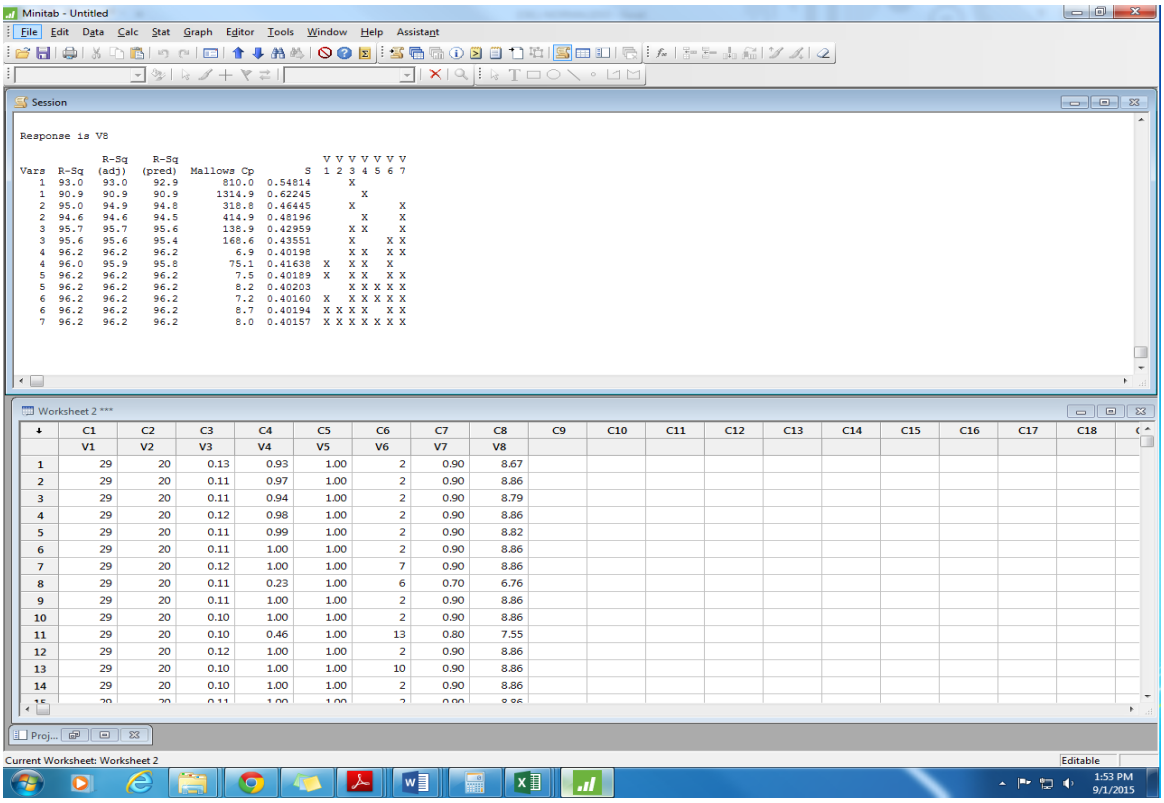


Figure 3-3 Sample Sheet of Best Subset Regression (Minitab 17)

(Berardi et al. 2008). There are several categories within the same class of age, diameter and material for each dataset. It should be mentioned that other physical factors of pipe, e.g. thickness, length, etc., can be utilized as the grouping criteria in different studies. But in this research, these three factors were selected for classification. Age was selected to take the indirect effect of time-varying solicitation on water mains into account, since from an engineering point of view, the higher the duration of solicitation, the higher the chemical and mechanical harmful effects on pipes. These effects can be caused by several factors

such as soil condition, traffic loads, and etc. (Berardi et al. 2008). A schematic view of classification features is shown in Figure 3-4.

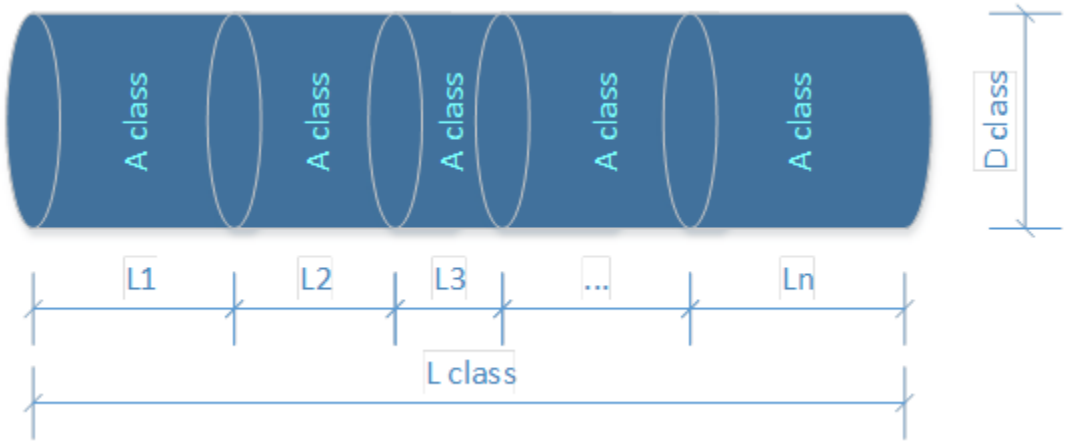


Figure 3-4 Classification's Features (Adopted from Berardi et al. (2008))

Other equivalent factors in each dataset can be calculated by different mathematical functions such as sum and average. In the dataset of Montréal, the length and number of breaks of each class were computed by summing the ones corresponding to each pipe segment. Likewise, in the Doha dataset, the same calculations were performed for the length and number of breaks while factors such as pipe elevation and burial depth were calculated by computing the average of related features of pipes in that group.

As an example, Table 3-1 shows a sample data of 10 different pipe segments. Age, length, diameter, material, buried depth and elevation are independent variables. Table 3-2 shows the classification of this sample data, based on age, diameter and material. As it can be seen in this table, the length of each class is calculated by summing up the length of all pipes with the same age, diameter and material. Also, the buried depth and elevation for each class are calculated by computing the average from pipes with the same features. For

example, class number 1 contains DI pipes with age of 30 years and diameter of 150 mm, which are 102 and 105 segments.

Table 3-1 Sample Data

Identification No.	Age (Year)	Length (m)	Diameter (mm)	Material	Buried Depth (m)	Pipe Elevation (m)
101	75	2	200	PVC	0.5	7
102	30	5	150	DI	1	8
103	75	6.5	200	PVC	1	8
104	100	4.5	300	DI	1.5	10
105	30	10.5	150	DI	2	14
106	75	3	200	PVC	0.5	12
107	100	15	300	DI	2.5	17
108	75	40	200	PVC	1	15
109	30	12	100	GI	1	18
110	100	6	300	DI	2	15

Table 3-2 Classified Sample Data

Pipes in each Class	Class	Age class	Length	Diameter class	Material class	Buried Depth	Pipe Elevation
102, 105	1	30	15.5	150	DI	1.5	11
109	2	30	12	100	GI	1	18
101, 103, 106, 108	3	75	51.5	200	PVC	0.75	10.5
104, 107, 110	4	100	25.5	300	DI	2	14

3.4 Evolutionary Polynomial Regression

Evolutionary Polynomial Regression is a data-driven technique and is classified as a grey box method, according to the color coding classification system. The color coding classification system categorizes mathematical models into three groups, based on available information; the white box models, black box models and grey box models. In the white box technique, the mathematical structure and parameters are already recognized. In the grey box technique, the mathematical structure is recognized by physical insight but

some data need to estimate parameters. And in the black box technique, the mathematical structure and parameters are not known and both should be recognized through the available data (Giustolisi 2004).

EPR is selected because it does not require large datasets for training and, unlike ANN, it provides insight into the relationship between the inputs and output. The process of EPR should be coupled with engineering knowledge to verify if the generated equations and correlations between utilized inputs and output are reasonable. In ANN, each attempt delivers particular output, able to vary in subsequent attempts when the same inputs are used. While in EPR or generally regressions, all similar attempts leads to generating the same equations.

The software of this method, EPR MOGA - XL tool version 1.0, was first developed by Giustolisi and Savic in 2006. The original code of this software has been developed in MATLAB environment (MATLAB®) and deployed as an Excel add-in function.

This algorithm attempts at generating a number of symbolic expressions that can predict the number of breaks of water mains, based on historical data. From among these generated symbolic expressions, the user will choose the best expression, based on the observed fitness and parsimony of the equation. The fitness to the observed data is measured by the value of R-Squared, while the number of terms and factors in each expression should be minimum to fulfill the requirement for parsimony. The process of creating the symbolic expressions contains two stages. In the first stage, the EPR finds the best model structure by using Multi-Objective Genetic Algorithm (MOGA). Then, the appropriate values for constant are estimated by Least-Squares optimization (LS) (Berardi et al. 2008).

Figure 3-5 shows the interface of EPR software. In this software, results can be shown as a Scatter plot or Cartesian plot. There are seven structures of the symbolic expression used to represent the relationship between the inputs and the output. The user select the best symbolic expressions according to the prior knowledge of the nature of the expected relation between the inputs and the output.

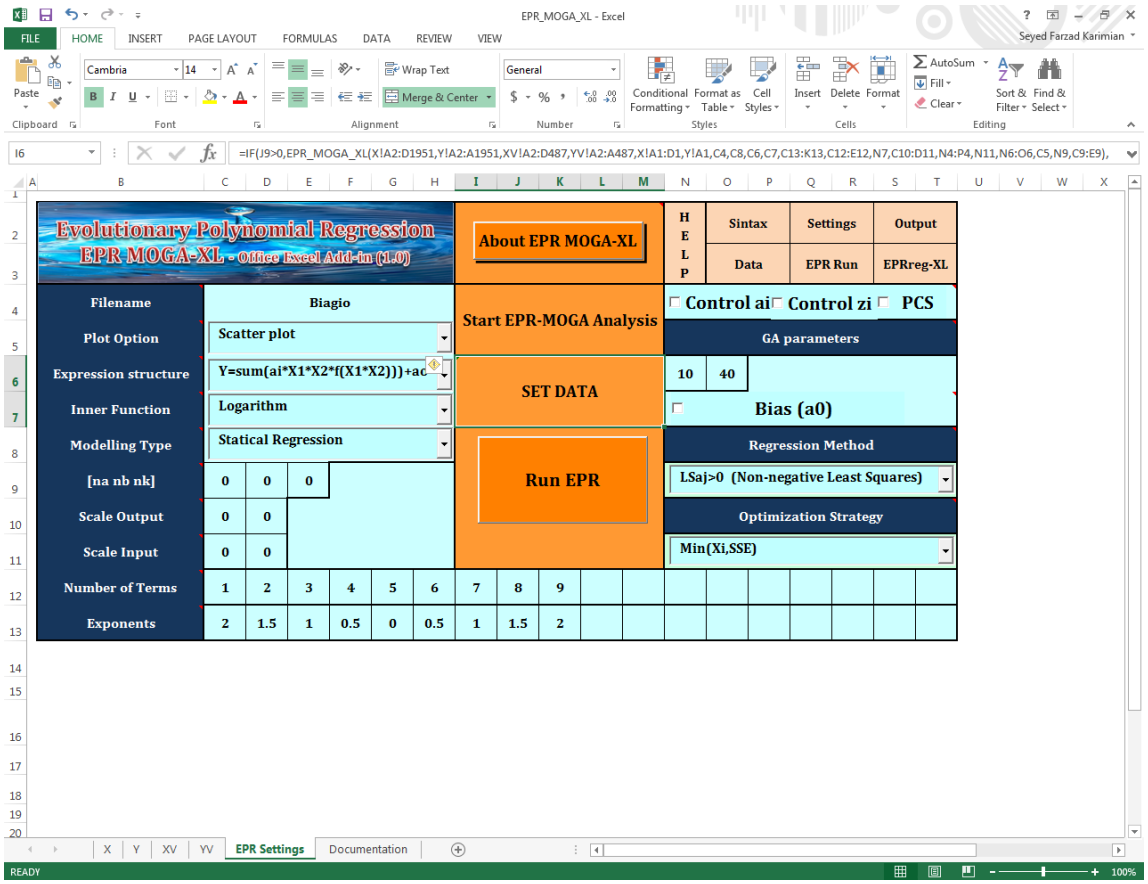


Figure 3-5 Interface of EPR Software

These seven structures are as follows:

$$[9]\ Y = a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)} \cdot f\left((X_1)^{ES(j,k+1)} \dots (X_k)^{ES(j,2k)}\right)$$

$$[10]\ Y = a_0 + \sum_{j=1}^{m} a_j \cdot f\left((X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)}\right)$$

$$[11]\ Y = a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)} \cdot f\left((X_1)^{ES(j,k+1)}\right) \dots f\left((X_k)^{ES(j,2k)}\right)$$

$$[12]\ Y = \log\left(a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)}\right)$$

$$[13]\ Y = \exp\left(a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)}\right)$$

$$[14]\ Y = \sin\left(a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)}\right)$$

$$[15]\ Y = \tan\left(a_0 + \sum_{j=1}^{m} a_j \cdot (X_1)^{ES(j,1)} \dots (X_k)^{ES(j,k)}\right)$$

Where, X_k is the kth explanatory variable, ES is the matrix of unknown exponents to be defined by the user, f is inner function selected by the user (can be no function, logarithm, exponential, tangent hyperbolic, or secant hyperbolic), a_j are unknown polynomial coefficients, m is the number of polynomial terms and a_0 is the bias term. During the generating symbolic expressions, if the EPR cannot find appropriate combination of terms containing f(x), it deselects this function (Giustolisi et al. 2011).

EPR rounds the output to the nearest integer number if the classification is selected as the Modelling Type. Thus, in scenarios where the real number was considered as a dependent variable, Statical Regression should be chosen. The Dynamical Regression can be selected as the modeling type in time series models. The normalization (if required) can be accomplished by EPR. The user, therefore, needs to specify the range wherein the inputs or output should be scaled (i.e. between 0 and 1). The maximum number of terms in every equation in each run can be specified by the user. The nomination of exponents should be limited to specific values – i.e. [-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2] – wherein the positive and negative values represent the direct and inverse relationship between dependent and independent variables and their amounts show how significant the inputs are. It must be

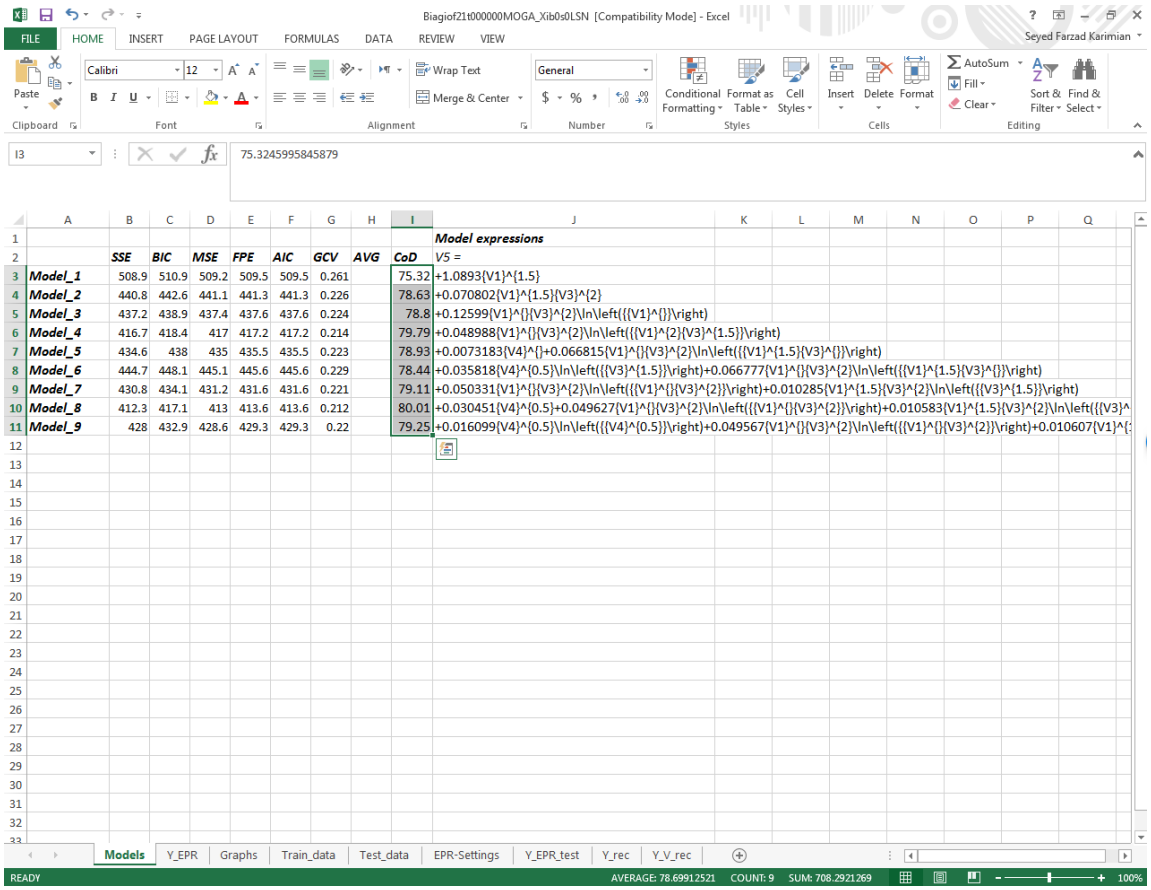
remarked that, the zero value should be considered in the matrix of exponents to make the EPR capable of removing variables not powerful enough in predicting the output (Giustolisi et al. 2011).

The "GA" is the number of generation and depends on several factors such as the number of independent and dependent variables, the number of terms and that of exponents. Furthermore, the user can force EPR to generate the expression with only positive value of constant coefficients (a_j>0). During the the EPR modeling phase, it returns several expressions based on the models' accuracy and parsimony. The model parsimony is implemented by optimizing the number of terms Min(a_j,SSE), the number of independent variables Min(X_i,SSE) or both strategies Min(a_j,X_i,SSE). These options are the user's input, defined in the optimization strategy scroll down box of the EPR model (Giustolisi et al. 2011). Finally, training and testing datasets are defined as follows: 1) X tab is for defining the training input, 2) Y tab is for defining the training output, 3) XV tab is for the testing input and 4) YV tab is for the testing output.

EPR produces five different types of result files including: Excel file, EPR fitting criteria, pareto, symbolic expressions, and scatter plot for each model. The Excel result file, contains 9 separated sheets are: Models, Y_EPR, Graphs, Train_data, Test_data, EPR-Setting, and Y_EPR_test. Figure 3-6 shows Models sheet and the figures of other sheets are shown in Appendix A.

The Models sheet contains all generated models from EPR with their coefficients, factors and exponents. The following parameters are generated for measuring accuracy of the EPR algorithm: SSE (Sum of Squared Error), BIC (Best Information Criterion), MSE (Mean Squared Error), FPE (Final Prediction Error of Akaike), AIC (Akaike's Information

Theoretic), GCV (Generalised Cross-Validation), AVG (Average Error) and CoD (Coefficient of Determination or R-Squared). MSE and CoD are calculated by equations number [6] and [1] respectively.



	A	B	C	D	E	F	G	H	I	J
1										***Model expressions***
2		***SSE***	***BIC***	***MSE***	***FPE***	***AIC***	***GCV***	***AVG***	***CoD***	*V5 =*
3	***Model_1***	508.9	510.9	509.2	509.5	509.5	0.261		75.32	+1.0893{V1}^{1.5}
4	***Model_2***	440.8	442.6	441.1	441.3	441.3	0.226		78.63	+0.070802{V1}^{1.5}{V3}^{2}
5	***Model_3***	437.2	438.9	437.4	437.6	437.6	0.224		78.8	+0.12599{V1}^{}{V3}^{2}\ln\left({{V1}^{}}\right)
6	***Model_4***	416.7	418.4	417	417.2	417.2	0.214		79.79	+0.048988{V1}^{}{V3}^{2}\ln\left({{V1}^{2}{V3}^{1.5}}\right)
7	***Model_5***	434.6	438	435	435.5	435.5	0.223		78.93	+0.0073183{V4}^{}+0.066815{V1}^{}{V3}^{2}\ln\left({{V1}^{1.5}{V3}^{}}\right)
8	***Model_6***	444.7	448.1	445.1	445.6	445.6	0.229		78.44	+0.035818{V4}^{0.5}\ln\left({{V3}^{1.5}}\right)+0.066777{V1}^{}{V3}^{2}\ln\left({{V1}^{1.5}{V3}^{}}\right)
9	***Model_7***	430.8	434.1	431.2	431.6	431.6	0.221		79.11	+0.050331{V1}^{}{V3}^{2}\ln\left({{V1}^{}{V3}^{2}}\right)+0.010285{V1}^{1.5}{V3}^{2}\ln\left({{V3}^{1.5}}\right)
10	***Model_8***	412.3	417.1	413	413.6	413.6	0.212		80.01	+0.030451{V4}^{0.5}+0.049627{V1}^{}{V3}^{2}\ln\left({{V1}^{}{V3}^{2}}\right)+0.010583{V1}^{1.5}{V3}^{2}\ln\left({{V3}^
11	***Model_9***	428	432.9	428.6	429.3	429.3	0.22		79.25	+0.016099{V4}^{0.5}\ln\left({{V4}^{0.5}}\right)+0.049567{V1}^{}{V3}^{2}\ln\left({{V1}^{}{V3}^{2}}\right)+0.010607{V1}^{

Figure 3-6 Excel Sheets File of EPR's Result

However, the other above mentioned indexes are computed using the following equations:

$$[16] \quad BIC = \left(1 + d\frac{\log N}{N}\right).SSE$$

[17] $$FPE = \left(\frac{1+ {}^{d}/_{N}}{1- {}^{d}/_{N}} \right) . SSE$$

[18] $$AIC = \left(1 + 2\frac{d}{N} \right) . SSE$$

[19] $$GCV = \frac{SSE}{(N-d)^2}$$

[20] $$AVG = 100 . \frac{1}{N} \sum_{i=1}^{N} \frac{SSE}{Y_i}$$

Where N is the number of samples and d is the number of independent variables.

The Y_EPR and Y_EPR_test sheets show the output for each generated model, based on the training and testing sets respectively. The graph sheet facilitates the process of generating figures for comparing predicted outputs with actual observations. Also, the expression and the value of CoD and SSE are shown in the graph sheet. There are two graphs in one sheet to visually identify the difference between them. The train and test data are both in the next two sheets. The content of these two sheets are exactly the same as X, Y, XV and YV sheets in the main EPR file. Also, the EPR-Setting shows the user interface in the current run of that file. In addition, Y_rec and Y_V_rec sheets contain the data that are reconstructed by EPR for train set and test set respectively (Giustolisi et al. 2011).

Figure 3-7 shows a sample of the EPR-fitting criteria graph. In this graph, horizontal axis shows the number of terms in each generated expression, while vertical axis shows the normalized value of different criteria – i.e. SSE, BIC, MSE, FPE, AIC, GCV, CoD and AVG.

In each run, EPR produces several scatter plots for each model. In these graphs, the predicted values of the output are compared with the actual data. As Figure 3-8 shows, the

horizontal axis demonstrates the value of the predicted output, while the vertical one shows actual – i.e. the experimental – data. These graphs are provided for separately training and testing each model. At the top of the graph, the symbolic expression of model number 8 and related CoD are shown as well.

Figure 3-8 shows the Scatter Plot with the sample data. As it can be seen, horizontal axis demonstrates the value of 1-CoD, while the vertical one shows the number of considered factors in each model ($^{d}/_{N}$). At the top of the graph, the function structure is shown as well. The Scatter plot shows generated models as the points in a graph. Based on the selecting criteria, already explained, the best model should be chosen from the lower

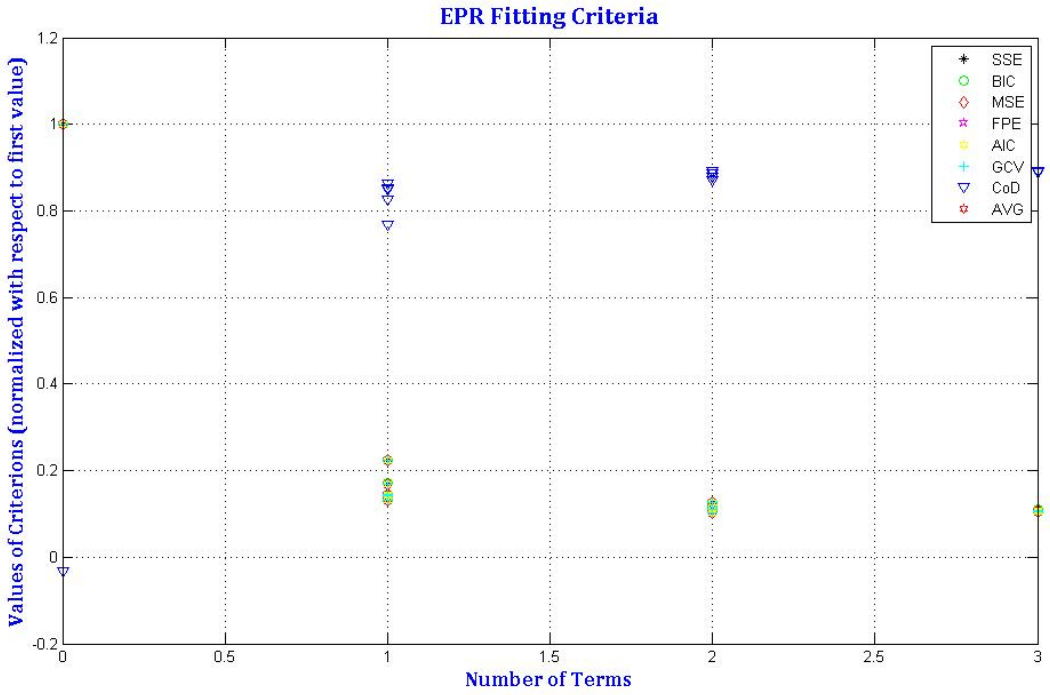


Figure 3-7 EPR Fitting Criteria

$$V5 = +0.030451V4^{0.5} + 0.049627V1V3^2 \ln\left(V1V3^2\right) + 0.010583V1^{1.5}V3^2 \ln\left(V3^{1.5}\right)$$

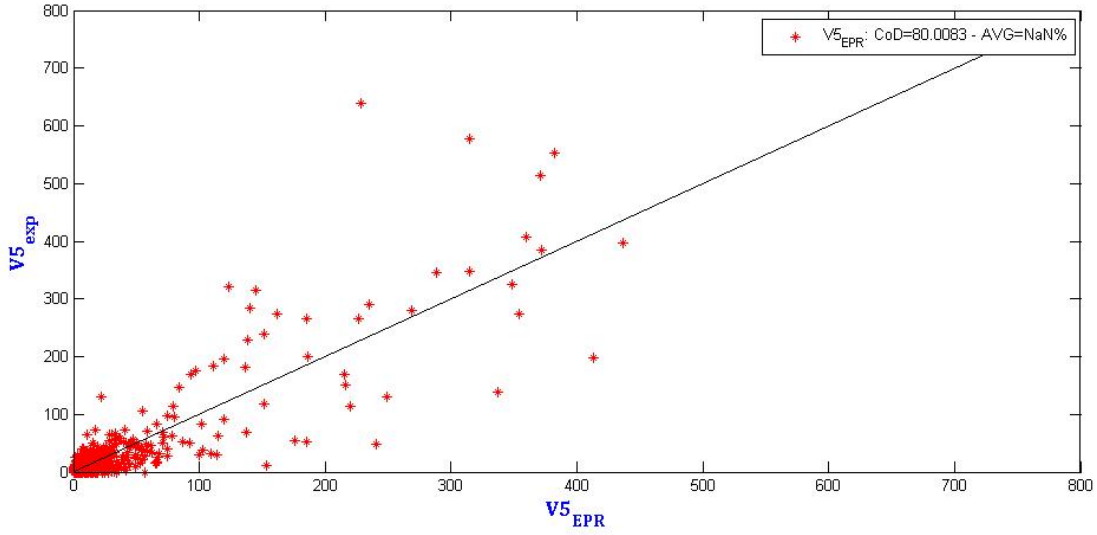


Figure 3-8 Scatter Plot (Predicted Value vs. Actual Data)

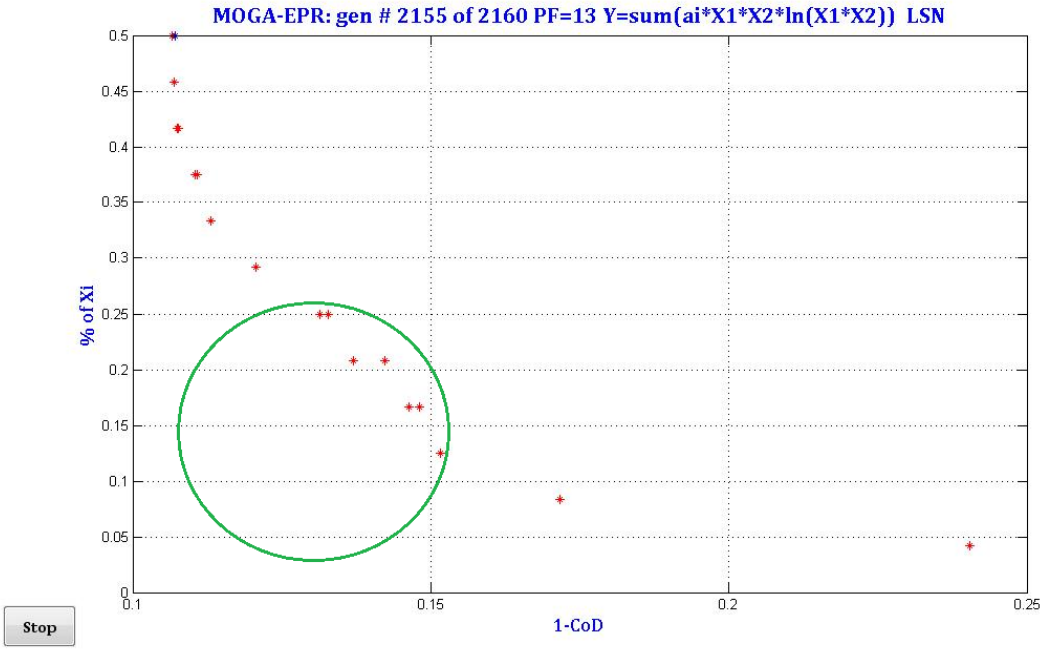


Figure 3-9 Pareto Graph (Trade of between Accuracy and Simplicity)

left corner of the graph, specified by a green circle in Figure 3-9, since in this area, the number of variables is minimum and the value of R-Squared (CoD) is maximum. These criteria fulfill requirements respectively for model parsimony and model fitness.

Based on the dataset size and level of complexity, several numbers of symbolic expressions are generated at the end of each run. Figure 3-10 shows 8 expressions that predict the output (V5), considering 4 independent variables including V1, V2, V3, and V4.

EPR-MOGA Symbolic expressions

(1) $V5 = +1.0893V1^{1.5}$

(2) $V5 = +0.070802V1^{1.5}V3^{2}$

(3) $V5 = +0.12599V1V3^{2}\ln(V1)$

(4) $V5 = +0.048988V1V3^{2}\ln(V1^{2}V3^{1.5})$

(5) $V5 = +0.0073183V4 + 0.066815V1V3^{2}\ln(V1^{1.5}V3)$

(6) $V5 = +0.035818V4^{0.5}\ln(V3^{1.5}) + 0.066777V1V3^{2}\ln(V1^{1.5}V3)$

(7) $V5 = +0.050331V1V3^{2}\ln(V1V3^{2}) + 0.010285V1^{1.5}V3^{2}\ln(V3^{1.5})$

(8) $V5 = +0.030451V4^{0.5} + 0.049627V1V3^{2}\ln(V1V3^{2}) + 0.010583V1^{1.5}V3^{2}\ln(V3^{1.5})$

Figure 3-10 Generated Symbolic Expressions

As mentioned before, the number of breaks is considered as the output of the model developed by EPR in this study. However, the value of the breakage rate is required in order to provide the deterioration curve using Weibull reliability function. Thus, the

following equation is used to transform the number of breaks as the output to the breakage rate:

[21] Breakage Rate = Number of Breaks / Length (km) / Age (yr)

The EPR can generate models to forecast the output, based on either one or several inputs. In other words, it can construct Multi Input Single Output (MISO) and/or Single Input Single Output (SISO) models. It should be noted that the limited missing data points can be recreated by using the linear interpolation by EPR. Thus, the model can be developed with an incomplete historical dataset although linear interpolation is not very accurate to reconstruct it.

3.5 Weibull Distribution

Finally, Weibull distribution is employed to generate deterioration curves. In general, Weibull-based models are widely used in different studies and applications to solve various problems (Jardine and Tsang, 2013). It has been used in the past for various building components, structural performance and infrastructure performance of subway networks by Grussing et al. (2006), Semaan (2011), and Gkountis (2014) respectively.

This technique has three advantages. As the most important one, this approach needs a few number of historical data while the other methods, such as the Markovian models, require the input of a significantly larger amount of data (Grussing 2006). The Weibull approach requires just two types of inputs to predict the future condition of the water pipelines: The age of the pipe and breakage rate (no. of breaks / km/ yr). Contrary to other methods, this one can be used to model either an individual pipe or the whole

network. Furthermore, different parameters of Weibull reliability function can be calculated easily as it is discussed later in this chapter.

The Weibull probability distribution function is calculated by the following equation:

[22] $$f(t) = \frac{b}{a}.\left(\frac{t-t_0}{a}\right)^{b-1}.e^{-\left(\frac{t-t_0}{a}\right)^b}$$

Where t is the time, t_0 is the location parameter, a is the scale parameter and b is the shape/slope parameter. In addition, the cumulative Weibull distribution function (cdf) is calculated as follows:

[23] $$F(t) = 1 - e^{-\left(\frac{t-t_0}{a}\right)^b}$$

Thus, the Weibull reliability function of a distribution is one minus the cumulative Weibull distribution function. Then, the Weibull reliability function is calculated by equation number [24], transformed to equation number [25] for the purpose of this study:

[24] $$R(t) = 1 - F(t) = e^{-\left(\frac{t-t_0}{a}\right)^b}$$

[25] $$R(t) = c\,.\,e^{-\left(\frac{t}{a}\right)^b}$$

Where, R(t) is the condition of pipe and c is the initial condition factor. The value of c is one in this study because the value of the R(t) is one at t = 0:

$$1 = c\,.\,e^{-\left(\frac{0}{a}\right)^b} = c\,.\,e^{0}$$

So: $c = 1$

Therefore, following equation is used to calculate the pipe's condition, based on the failure rate:

$$[26] \qquad R(t) = e^{-\left(\frac{t}{a}\right)^b}$$

Where, R(t) is the pipe's condition, t is the pipe's age, b is the shape parameter and 1/a is the failure rate. The value of b should be odd and more than one. In this study, this value is equal to 3 because it provides the smoothest inclination (Semaan 2011).

In some previous studies, especially in oil and gas pipelines and subway networks (Seaman, 2011), the values of performance threshold and minimum performance were assumed. But in water pipelines, there is no need for that because the failure in water pipelines is less costly and critical than the one in the oil and gas pipelines and subway networks.

3.6 Sensitivity Analysis

A possible definition of Sensitivity Analysis is the study of how uncertainty of the output of a model can be caused by different sources of uncertainty in the model inputs (Saltelli et al. 2004). In this study, the sensitivity analysis was performed for both cases to identify the effect of each independent variable on the pipe failure when water pipes age. The rationality of inputs-output relationship in the selected symbolic expression was studied as well. Generally, this technique depends on one or more independent variables. But in this study, the effect of changing only one parameter over a specific time period was investigated. The sensitivity analysis is discussed in details in the next chapter.

3.7 Summary and Conclusion

In this chapter, different parts of the developed research framework were described in details. In the first part, the most critical factors, affecting failure rates of water pipes, are identified by using best Subset regression. The best combination of independent variables are selected out of all possible candidates. Once, the process of factor selection is performed, each dataset will be classified into homogeneous groups based on the age, diameter and material of pipes. Then, homogeneous groups are forwarded to EPR in order to generate some mathematical expressions that predict number of breaks of water pipelines. EPR algorithm is performed in two stages: 1) Search for the best model using Multi-Objective Genetic Algorithm (MOGA) and 2) a parameter estimation for the model by using Least Square Method. Among all generated expressions, the user selects the best one based on two criteria: 1) Fitness to the historical data and 2) the parsimony of the equation. The predicted number of breaks obtained from the best symbolic expression is employed to generate deterioration curves by using Weibull distribution. Finally, the sensitivity analysis was conducted to: 1) recognize the effect of changing each input on the breakage rate and 2) study the rationality of relationship between the selected inputs and the output.

Chapter 4: Data Collection

In this study, four sets of data from four municipalities were considered for developing failure rate prediction models; City of Moncton, City of Hamilton and City of Montréal in Canada and City of Doha in Qatar. As the physical characteristics of water pipes in different datasets are generic and the results obtained using the Hamilton and Moncton datasets were very close, these two datasets were used to estimate the number of breaks in the City of Doha. Then, datasets of Montréal and Doha were employed to develop EPR models for predicting failure rates of water mains. A description of data collection is presented in this chapter.

4.1 City of Montréal

The city of Montréal has a population of 1.8 million, and its land area is around 365.1 square kilometers. Figure 4-1 shows the GIS map of the City and its water distribution network. In this city, there are six water treatment plants and 14 reservoirs. The City of Montréal owns 5045 kilometers of water distribution networks containing 4305 km distribution pipes and 740 km transmission pipes (Paul 2014). The original excel file of the dataset of Montréal comprises of 125,828 pipe segments that include various information such as: pipe ID, installation date, diameter, length, material, manager and owner, rehabilitation date, and rehabilitation type. It comprises of 56.55% Cast Iron (CI), 26.61% Ductile Iron (DI), 10.47% Cementitious (Asbestos and Concrete Cylinder), 5.54% Plastic pipes (PVC and Polyethylene), 0.77% Steel, 0.05% Copper, and 0.01% Galvanized Iron (GI). The CI pipes are installed during 1862–2015 and DI pipes are mostly installed during 1951–2015. Figure 4-2 shows the number of breaks in the water distribution networks between 1861 and 2015. By closely examining Figure 4-2 below, the trend of using DI and

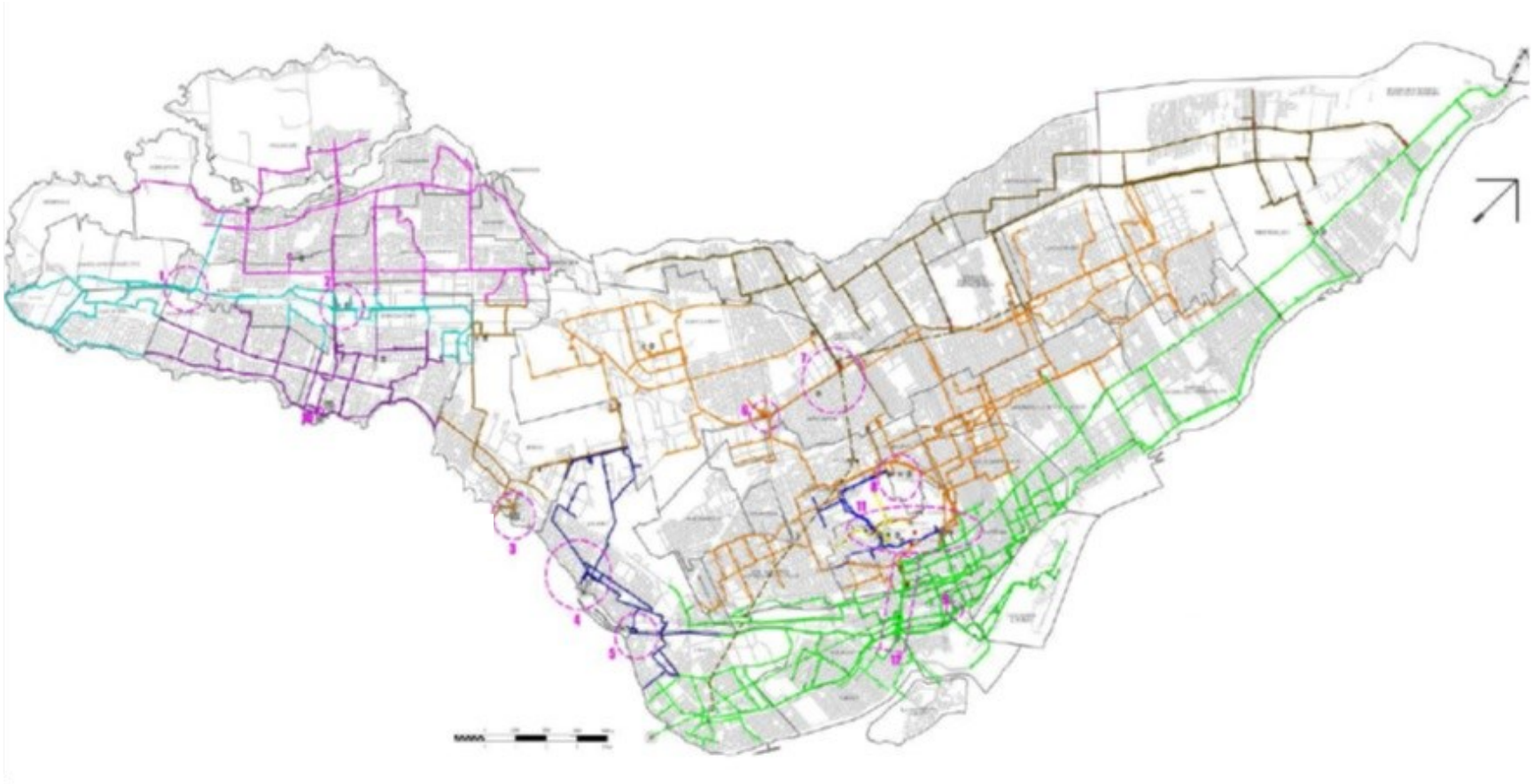


Figure 4-1 Water Distribution Networks of City of Montreal (Paul 2014)

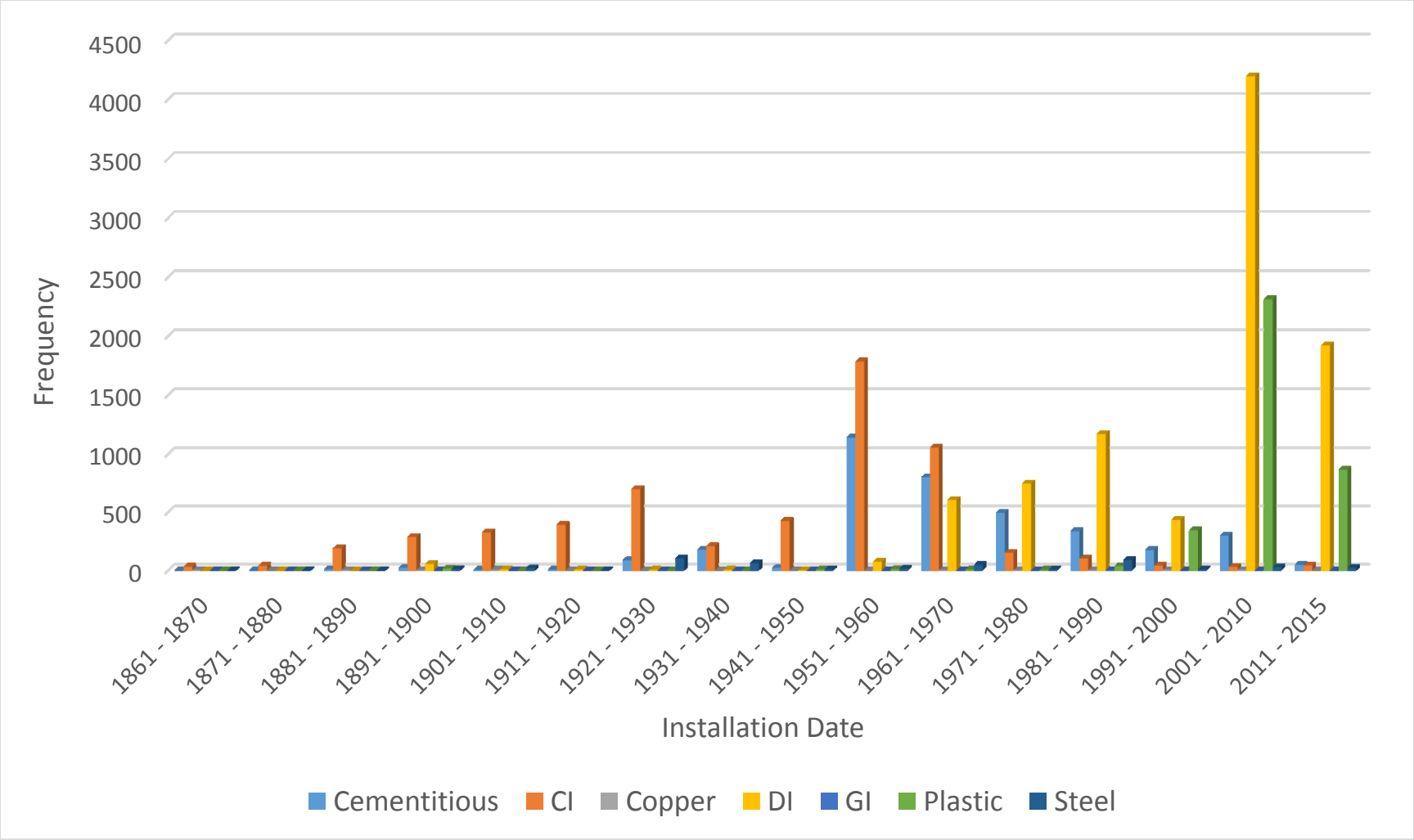


Figure 4-2 Number of Breaks per segment for Pipes with Different Material Installed Between 1861 and 2015 for City of Montreal

plastic pipes to replace other pipe types is increasing during 25 years ago. It also shows that the number of breaks for DI and plastic pipes, installed between 2001 and 2015, are increased significantly in the same period. It might be because of either poor installation techniques or low-quality material.

Figure 4-3 shows the number of breaks and their year of occurrence for different pipe material. The municipality of Montréal started to perform systematic recording of pipe failure since 1972, and the dataset contains a total of 22,735 pipe breaks so far. Figure 4-3 also shows that the number of breaks for CI pipes has steadily increased since 1986 and reached the peak in 2001-2005 interval, before falling slightly during the recent 15 years.

The dataset of the City of Montréal contains information about pipe's (age, length, diameter, material) and the related pipe failures. The units of age, length, and diameter are the year, Km, mm respectively in collected data. Also, the date and the type of rehabilitation was recorded for each pipe as well. As shown in Chapter 2, age, length, diameter, and pipe material are the most frequent independent variables utilized for predicting failure rates of water pipes. The original file of the dataset of Montréal, which was provided by the municipality, contains two separated excel spreadsheets: water pipes' attributes and related water pipes' breaks. Thus, it was required to incorporate these two files into the single file in which pipe's attributes and pipe's breaks are available for each segment. Table 4-1 shows a summary of some statistical measurements of quantitative factors for the City of Montréal.

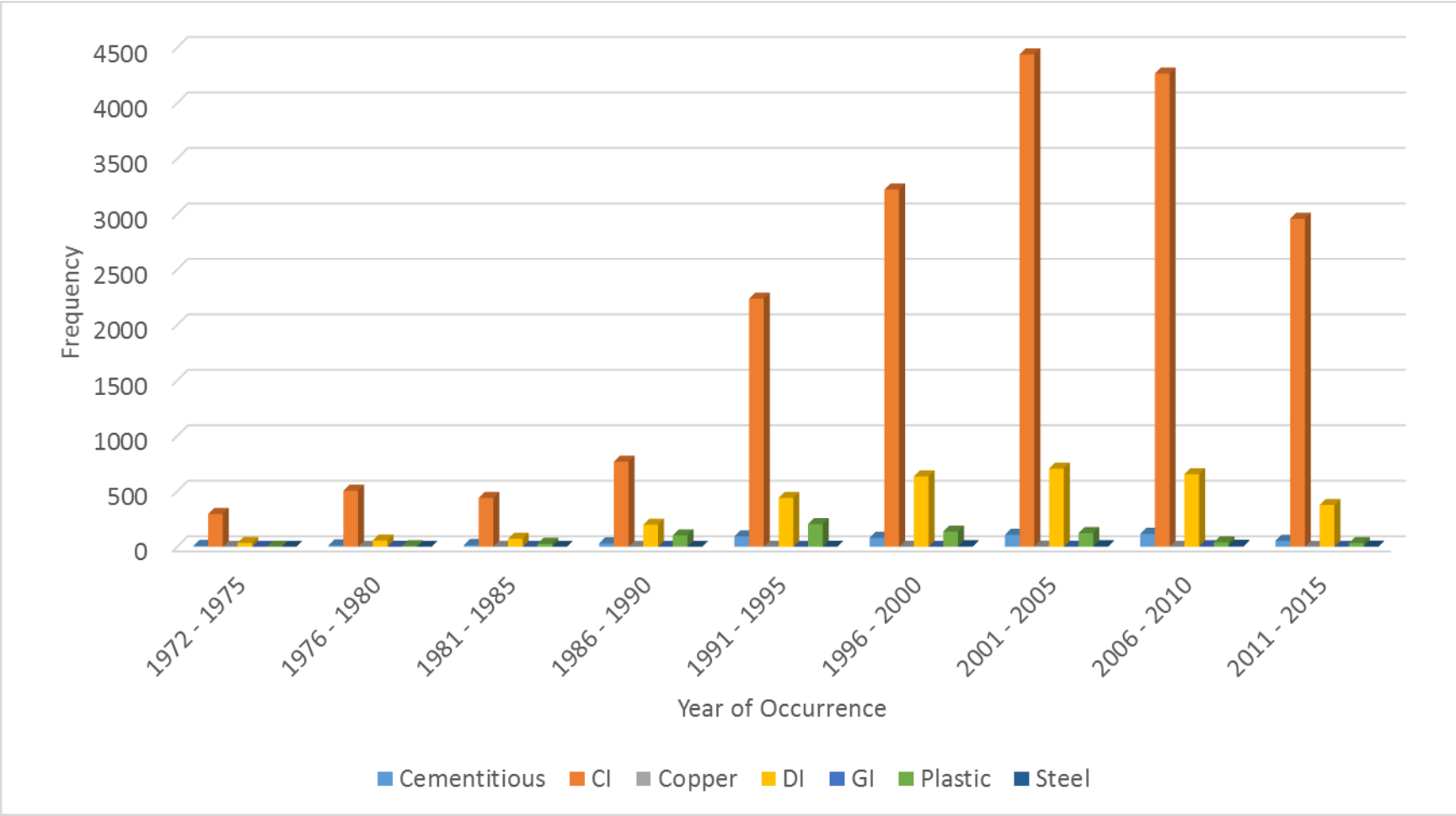


Figure 4-3 Number of Breaks per segment for Pipes with Different Material for City of Montreal

Table 4-1 Quantitative data attributes of city of Montréal

Attribute	Mean	Min	Max	Sdv
Age (year)	55.34	1	199	30.52
Diameter (mm)	245.98	20	3900	153.73
Length (m)	43.04	0.15	543.62	65.11
No. of Breaks	1.76	0	25	1.40

4.2 City of Doha

Qatar is one of the highest water consumers in the world. The amount of water consumption per capita is 500 liters a day that is quadruple the normal range in Europe (HSBC, 2014). The city of Doha has a population of 796,947 while its land area is around 132.1 square kilometers. The city of Doha owns 1,926 kilometers of water distribution networks (Kahramaa, 2009). It comprises of 99.99% Ductile Iron and 0.01% Steel pipes (just three segments out of 1599 segments). Thus, only ductile iron pipes were considered for this dataset.

The dataset of the city of Doha includes: age (year), length (km), diameter (mm), wall thickness (mm), pipe material, buried depth (m), and pipe elevation (m). Always, there is a strong relation between diameter and wall thickness in the water pipes. The pipes with higher diameter are thicker than the pipes with the smaller diameter. Thus, the wall thickness was recognized as a redundant variable and removed from the set of inputs. A summary of some statistical measurements for the dataset of City of Doha is shown in Table 4-2.

Table 4-2 Quantitative data attributes of City of Doha

Attribute	Mean	Min	Max	Sdv
Age (year)	11.94	2	3	4.40
Diameter (mm)	227.44	80	1400	280.60
Wall Thickness (mm)	11.13	10	26.10	3.08
Length (m)	34.45	0.04	546.85	52.02
Buried Depth (m)	0.56	0.50	6.31	0.43
Pipe Elevation (m)	12.59	7	18	3.46

The number of breaks was not available in dataset of Doha. Lack of such data prevents working with EPR because this technique takes into account the number of breaks or breakage rate as a dependent variable in order to develop a pipe failure prediction model. Therefore, it was necessary to estimate the number of breaks for the city of Doha from similar infrastructure datasets. The physical characteristics of water pipes in different datasets are generic (Karimian et al. 2015). In fact the results obtained using the Hamilton and Moncton datasets were very close. In view of this finding and the insufficient data collected from Doha, it was required to estimate the number of breaks in Doha based on datasets of Hamilton and Moncton. Several attempts were carried out using different regression models to estimate the number of breaks of water mains based on the pipe's age. The developed equations for each dataset and their features will be provided later in case study chapter. The result of City of Doha's analysis will be presented in Chapter 5 as well.

4.3 City of Moncton

The City of Moncton located in New Brunswick, Canada and has a population of 64,128, while its land area is around 142 square kilometers. The City of Moncton owns 500 kilometers of water distribution networks. This dataset contains 540 pipe segments which comprise of Cast Iron, Ductile Iron, and Asbestos. It includes: age (year), breakage rate (breaks/yr/km), C-factor, Diameter (mm), RUL (year), and wall thickness (mm). Table 4-3 shows a summary of some statistical measurements for the dataset of City of Moncton.

Table 4-3 Quantitative data attributes of City of Moncton (Atef et al. 2015)

Attribute	Mean	Min	Max	Sdv
Age (years)	46.02	10	106	19.93
Breakage rate (breaks/year/km)	0.67	0	5	0.68
C-factor	70.01	10	120	20
Pipe Diameter (mm)	795.35	100	2400	3.78
RUL (years)	103.97	44	140	19.93
Wall Thickness (mm)	6.03	3.5	8	0.45

4.4 City of Hamilton

The City of Hamilton located in Ontario, Canada and has a population of 519,949 while its land area is around 1,138 km². The City of Hamilton owns 1,891 km of water mains, in which estimated value for replacement of these pipes is around $1.8 billion (SOI Report, 2005). This dataset includes five quantitative variables and two qualitative

variables that are: age (year), buried depth (m), flow pressure, length (m), diameter (mm), material, and soil type. Hamilton dataset comprises of Cast Iron, Ductile Iron, PVC, and HDPE. Table 4-4 shows a summary of some statistical measurements for the dataset of City of Hamilton.

Table 4-4 Quantitative data attributes of City of Hamilton (Atef et al. 2015)

Attribute	Mean	Min	Max	Sdv
Age (years)	59.73	8	113	21.08
Buried Depth (m)	1.56	0	2.1	0.17
Flow Pressure	31.61	0	95	24.36
Length (m)	62.15	0.3	472	75.13

4.5 Data Filtering

For reducing errors and uncertainty of these datasets, several steps were performed. First of all, datasets of Montreal and Doha were cleaned and filtered. All segments with missing or incomplete information were removed from the datasets. Some historical records were irrational and inconsistent, so these records were removed as well. In some cases, there was a chance for the missing or irrational data to be reconstructed based on the other attributes' value or experts' opinion, but they were ignored for preventing the inaccurate result.

Both datasets, contain pipe material as a qualitative attribute that was converted to a quantitative attribute to apply with EPR. Thus, the qualitative variable, which is pipe

material, was transformed to the numerical ones. The pipes were sorted based on their rigidity and for each type one value was assigned. For example, if there are four different types of material, each number from 1 to 4 was assigned to a specific pipe material. The maximum number was assigned to the hardest pipe material; in other words, the harder the material, the larger the allocated number, and the vice versa.

Finally, two datasets were classified into homogeneous groups based on age, diameter, and material of the pipe. A detailed discussion about classification is presented in the research methodology chapter.

Chapter 5: Implementation of Developed Models

5.1 Introduction

In this chapter, two case studies: City of Montréal and City of Doha are analyzed and used to test and validate the developed model. As it can be seen in Figure 5-1, these two datasets include 5 and 6 subsections respectively. The chapter starts by discussing the effort made in identifying the most critical factors using Best Subset regression. Then, classifying each dataset into clusters of homogenous pipe segments with the same age, diameter and material are discussed. The EPR model is then applied to these clustered sets and results of testing and validating the model are reported and discussed. Afterwards, deterioration curves, which are developed using Weibull distribution function, are presented. Sensitivity analysis is utilized to study how the output can be apportioned to different sources of uncertainty in its inputs.

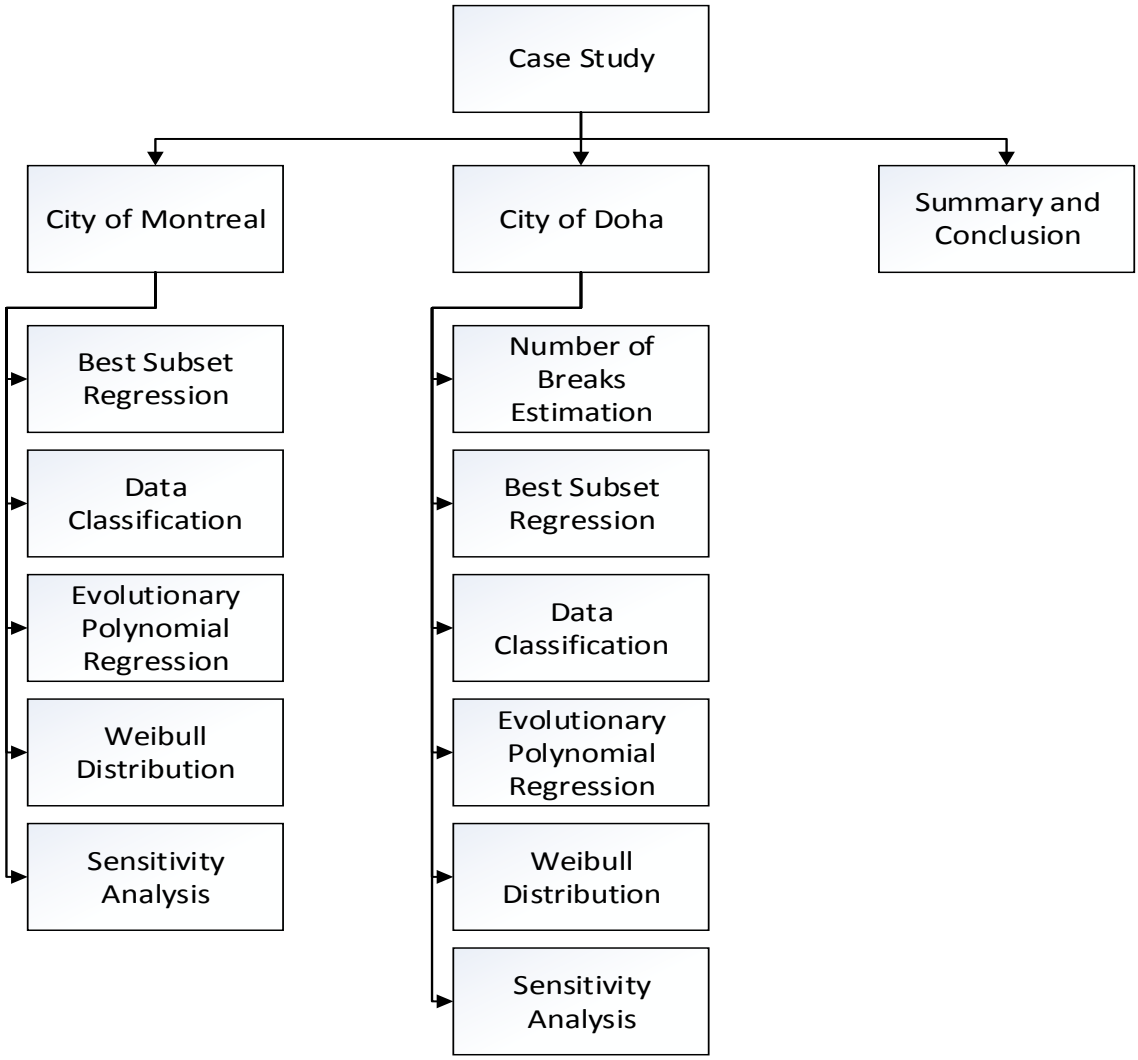


Figure 5-1 Chapter Overview

5.2 City of Montréal

The dataset is used in this section belongs to City of Montréal, Quebec, Canada. As it was discussed in data collection chapter, this city has a population of 1.8 million and its land area is around 365.1 square kilometers. The City of Montréal owns 5045 kilometers of water distribution networks containing 4305 km distribution pipes and 740 km transmission pipes (Paul, 2014).

5.2.1 Best Subset Regression

Best Subset regression is implemented to recognize the most critical factors for predicting number of breaks for water pipelines. Best Subset regression was employed by using Minitab 17 statistical package. Dataset of Montréal contains four independent variables including: length, diameter, age, and material of pipes.

Figure 5-2 shows the result for the dataset of Montréal. As it can be seen in the upper window, models number 5 and 7 have the highest value of R-Squared, adjusted R-Squared, and predicted R-Squared (68.9%, 68.9%, and 68.1%). The value of S (i.e. square

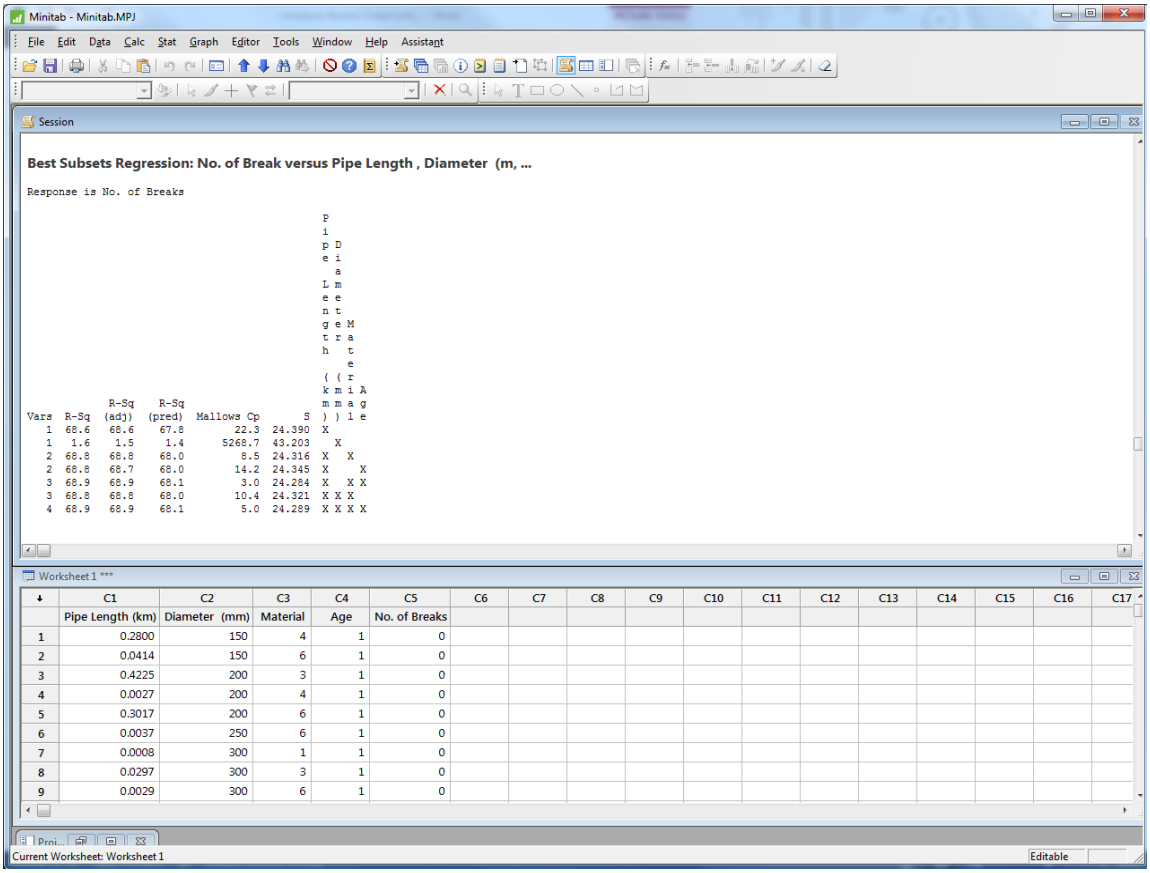


Figure 5-2 Best Subset Regression for City of Montréal

root of MSE) for model number 5 and 7 are 24.286 and 24.289, respectively. While, the value of Mallows' Cp for these two models are three and five respectively. The values of S are almost equal in both models, however the value for Mallows' Cp is 5 for model number 7. As discussed in chapter 4, the value for Mallows' Cp should be close to five in this study (number of independent variables plus one). Thus, it is concluded that model number 7 includes the best combination of factors for predicting the number of breaks.

5.2.2 Data Classification

The objective of the classification is clustering pipe segments into classes that have the same age, diameter and material. The original excel file of the dataset of Montréal comprises of 125,828 pipe segments. After data filtering, the dataset was classified into 2,436 homogeneous groups based on the age, diameter and pipe material. The length and the number of breaks of each class were computed by summing corresponding ones of each pipe segment. Samples of the original data and the classified data of dataset of Montréal are provided in Appendix B.

5.2.3 Evolutionary Polynomial Regression

In this study, Evolutionary Polynomial Regression generated twelve symbolic expressions, which are used to predict the number of breaks for water pipes in the City of Montréal. Table 5-1 shows these expressions and their related R-Squared scores. At the right side of expressions, L, D, A, and M represent the length, diameter, age, and material of the water pipelines, and the left side shows the output that is the number of breaks. As discussed in chapter 4, among all generated symbolic expressions, the best model should be chosen based on the fitness to the historical data and parsimony of the equation. In this study model number 10 was selected as it fulfills the requirement for these two criterions

Table 5-1 Symbolic Expressions for Montréal dataset and related R-Squared

Model #	***Expressions***	R^2 ***(%)***
1	$No.of\ Breaks = 3.4446 \times 10^{-5} \times L^{1.5}$	76.90
2	$No.of\ Breaks = 1.3197\ \frac{L^{1.5}}{D^2}$	82.51
3	$No.of\ Breaks = 0.08546\ \frac{L^{1.5}\ M^2}{D^2}$	85.04
4	$No.of\ Breaks = 1.8835\ \frac{L^{1.5}}{D^2}\ \ln\left(\frac{M^2}{A^{0.5}}\right)$	85.37
5	$No.of\ Breaks = 0.24999\ \frac{L^{1.5}\ A^{0.5}}{D^2}\ ln\left(\frac{M^2}{A^{0.5}}\right)$	86.40
6	$No.of\ Breaks = 0.092319 \times L^{0.5} + 0.23417\frac{L^{1.5}\ A^{0.5}}{D^2}\ \ln\left(\frac{M^2}{A^{0.5}}\right)$	87.04
7	$No.of\ Breaks = 0.12036\ \frac{L^{1.5}\ M^2}{D^2} + 4.8297 \times 10^{-7}\ \frac{L^2\ A^{1.5}}{D^2}\ \ln\left(\frac{1}{L}\right)$	88.03
8	$No.of\ Breaks = 0.008929\ \frac{L^{1.5}\ A}{D^2}\ \ln\left(\frac{1}{L^{0.5}}\right) + 0.069455\ \frac{L^{1.5}\ M^{1.5}\ A^{0.5}}{D^2}$	88.72
9	$No.of\ Breaks = 0.086502\ L^{0.5} + 0.00051089\ \frac{L^{1.5}\ A^{1.5}}{D^2}\ \ln\left(\frac{1}{L^{0.5}}\right) + 0.021313\ \frac{L^{1.5}\ M^2\ A^{0.5}}{D^2}$	88.86
10	$No.of\ Breaks = 0.017785\frac{L^{1.5}\ M^2\ A^{0.5}}{D^2} + 6.1833 \times 10^{-6}\ \frac{L^{1.5}\ A^2}{D\ M^2}\ \ln\left(\frac{D^{1.5}}{L}\right)$	89.35
11	$No.of\ Breaks = 0.00044077\ L + 0.017413\frac{L^{1.5}\ M^2\ A^{0.5}}{D^2} + 8.8604 \times 10^{-5}\ \frac{L^{1.5}\ A^2}{D^{1.5}\ M^2}\ \ln\left(\frac{D^{1.5}}{L}\right)$	89.21
12	$No.of\ Breaks = 0.00057323L + 0.049651\ \frac{L^{1.5}\ M^{1.5}\ A^{0.5}}{D^2}\ \ln(M^{0.5}) + 8.8156 \times 10^{-5}\ \frac{L^{1.5}\ A^2}{D^{1.5}\ M^2}\ \ln\left(\frac{D^{1.5}}{L}\right)$	89.30

which are having the highest R-Squared (89.35%) and including just two terms. According to the other models, it is observed that introducing a third polynomial term decreases the model fitness. The other accuracy indexes such as SSE, BIC, MSE, FPE, AIC, and GCV of all models are shown in Table 5-2. As it can be seen in this table, the minimum values

of all indexes are for model number 10. It confirms that this model is the best one in predicting the output.

Table 5-2 Accuracy Indexes for Montreal Dataset

	SSE	***BIC***	***MSE***	***FPE***	***AIC***	***GCV***
Model #1	476.4	478.2	476.6	476.9	476.9	0.245
Model #2	360.6	362	360.8	361	361	0.185
Model #3	308.5	309.7	308.7	308.8	308.8	0.158
Model #4	301.8	303	302	302.1	302.1	0.155
Model #5	280.5	281.5	280.6	280.7	280.7	0.144
Model #6	267.4	269.4	267.6	267.9	267.9	0.137
Model #7	246.9	248.8	247.1	247.4	247.4	0.127
Model #8	232.7	234.5	232.9	233.2	233.2	0.12
Model #9	229.7	232.4	230.1	230.4	230.4	0.118
Model #10	219.7	221.4	219.9	220.2	220.2	0.113
Model #11	222.6	225.2	223	223.3	223.3	0.115
Model #12	220.8	223.4	221.1	221.5	221.5	0.114

Figure 5-3 shows Pareto graph of expressions that were generated based on the Montreal dataset. As it was mentioned before, each point represents a generated symbolic expression. The selected model (model #10) is specified by the black arrow. The horizontal axis shows the value of one minus R-Squared (1-CoD) while the vertical axis shows the number of considered factors in each model.

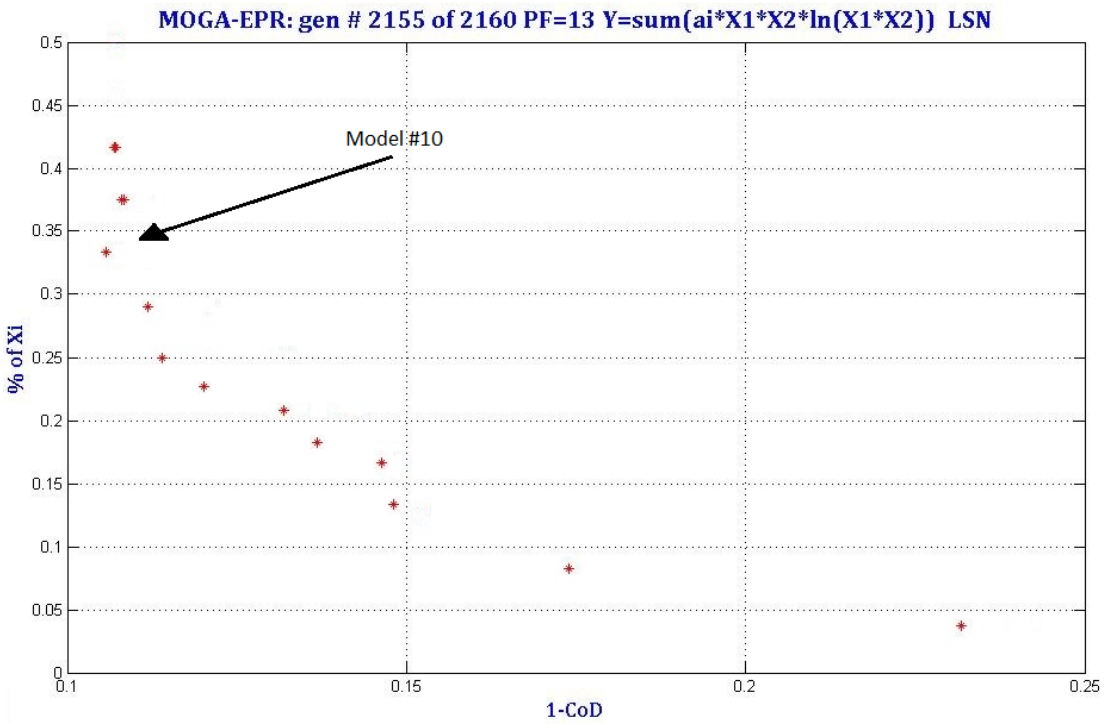


Figure 5-3 Pareto of Montreal Dataset

The dataset of Montreal randomly divided into two subsets (Training and Testing). As it can be seen in Table 5-3, 1950 (80%) samples were used for training and 486 (20%) samples were used for testing. It should be mentioned that testing samples were not exposed to the model during its development. Figures 5-4 and 5-5 show scatter plots that depict the relationship between the predicted and the actual number of breaks for training and testing datasets respectively. In these graphs, the vertical axis shows the actual number of breaks (experimental), while the horizontal axis shows the predicted value of the number of breaks. The values of R-Squared (CoD) are shown in the top right corner of each plot (i.e. 89.35% and 84.86% for training and testing respectively). At the top of the Figure 5-

4, the symbolic expression of model number 10 is shown. Scatter plots for the training and testing results of other symbolic models are shown in Appendix C.

Table 5-3 Montreal Dataset Size

City	Training Size	Testing Size	Total Size
Montreal	1950 (80%)	486 (20%)	2436 (100%)

$$No.of\,Breaks = +0.017785\frac{L^{1.5}M^{2}A^{0.5}}{D^{2}} + 6.1833e-006\frac{L^{1.5}A^{2}}{DM^{2}}\ln\left(\frac{D^{1.5}}{L}\right)$$

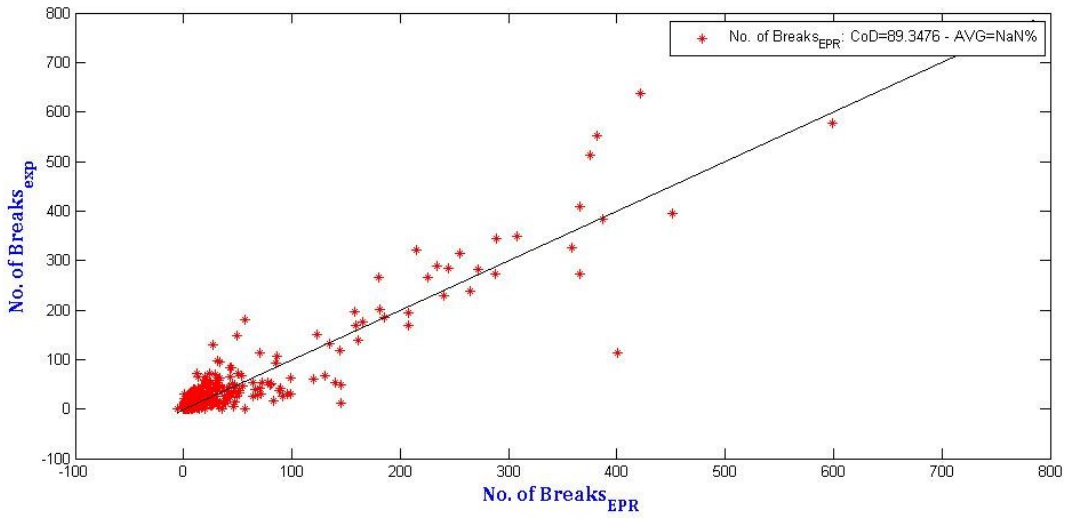


Figure 5-4 Scatter Plot of Model #10 for Training for Montreal Dataset

5.2.4 Weibull Distribution

The value of the number of breaks that was predicted in the previous section is used to establish deterioration curves using Weibull reliability function. It should be mentioned that the value of the number of breaks was transformed to a breakage rate by diving it by the pipe age (year) and length (km). Weibull reliability function can be used to model either

an individual pipe or the entire network. Thus, providing a curve for each pipe segment is

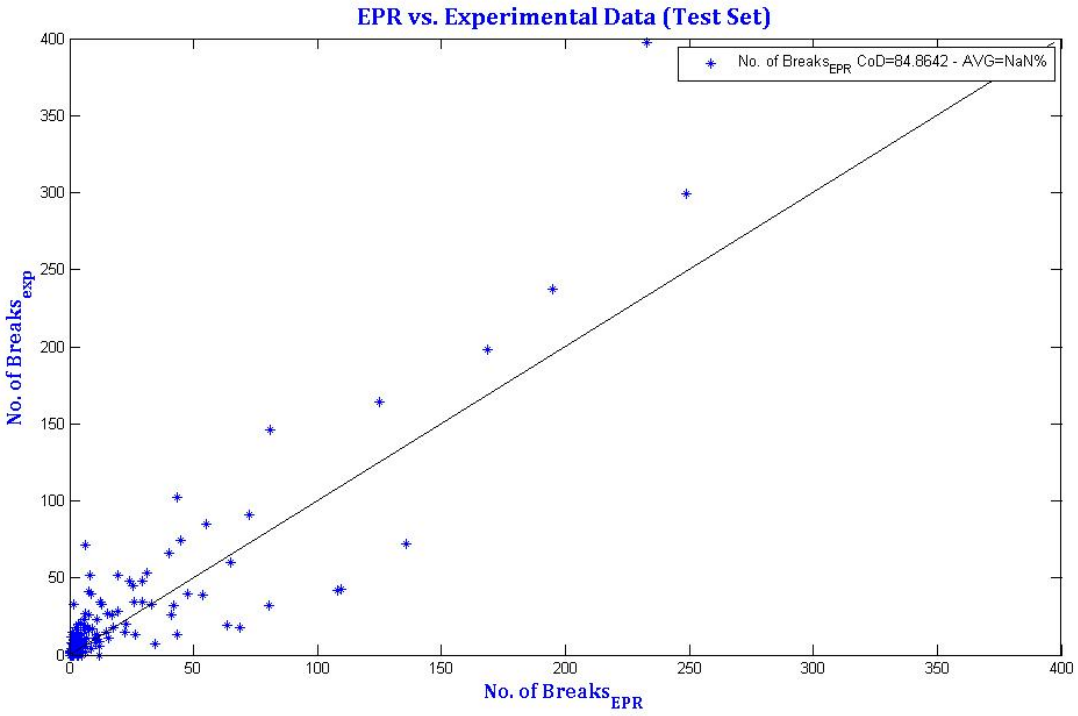


Figure 5-5 Scatter Plot of Model #10 for Testing for Montreal Dataset

possible. However, as it can be seen in Table 5-4, the dataset of Montreal was clustered into 18 clusters with a deterioration curve for each of them. The dataset was clustered based on length (short, medium, and large), diameter (small and large), and material (M1, M2, and M3) of pipes. For the pipe length, three subcategories were defined: short ($l \leq 300m$), medium ($300m < l \leq 2000m$), and long ($l > 2000m$). According to the literature (CIRC 2012), for the pipe diameter, two subcategories were defined: small ($D \leq 350mm$) and large ($D > 350mm$).

Table 5-4 Different Clusters and Related Features for Montreal Dataset

Cluster	***Features***
1	Length: Short, Diameter: Small, Material : M1
2	Length: Short, Diameter: Small, Material : M2
3	Length: Short, Diameter: Small, Material : M3
4	Length: Short, Diameter: Large, Material : M1
5	Length: Short, Diameter: Large, Material : M2
6	Length: Short, Diameter: Large, Material : M3
7	Length: Medium, Diameter: Small, Material : M1
8	Length: Medium, Diameter: Small, Material : M2
9	Length: Medium, Diameter: Small, Material : M3
10	Length: Medium, Diameter: Large, Material : M1
11	Length: Medium, Diameter: Large, Material : M2
12	Length: Medium, Diameter: Large, Material : M3
13	Length: Long, Diameter: Small, Material : M1
14	Length: Long, Diameter: Small, Material : M2
15	Length: Long, Diameter: Small, Material : M3
16	Length: Long, Diameter: Large, Material : M1
17	Length: Long, Diameter: Large, Material : M2
18	Length: Long, Diameter: Large, Material : M3

In addition, based on historical records of pipes failure, three sub categories were defined for the pipe material: M1, M2, and M3 which belong to group of same material pipes with low, moderate, and high rate of failure respectively. Thus, 18 different deterioration curves were generated for 18 Clusters.

Figures 5-6 and 5-7 show deterioration curves for Cluster number 7 and 16 respectively. In each graph, vertical axis shows the condition of the pipe while horizontal axis represents the age of the pipe. By observing closely Figure 5-6, pipe condition starts from 1 (the best condition) and then decreases slightly to the zero (the worse condition). Also, condition of Cluster number 7 starts to decrease sooner than number 16. This observation, confirms that the probability of failure in pipes with large diameter is lower than pipes with small diameter. Typically, any kind of rehabilitation increases pipe reliability and decreases probability of failure. These graphs were developed without considering the effect of rehabilitation on decreasing the failure rate of water pipes. Therefore, the failure rate of water pipes should be updated when rehabilitation action is being considered or applied.

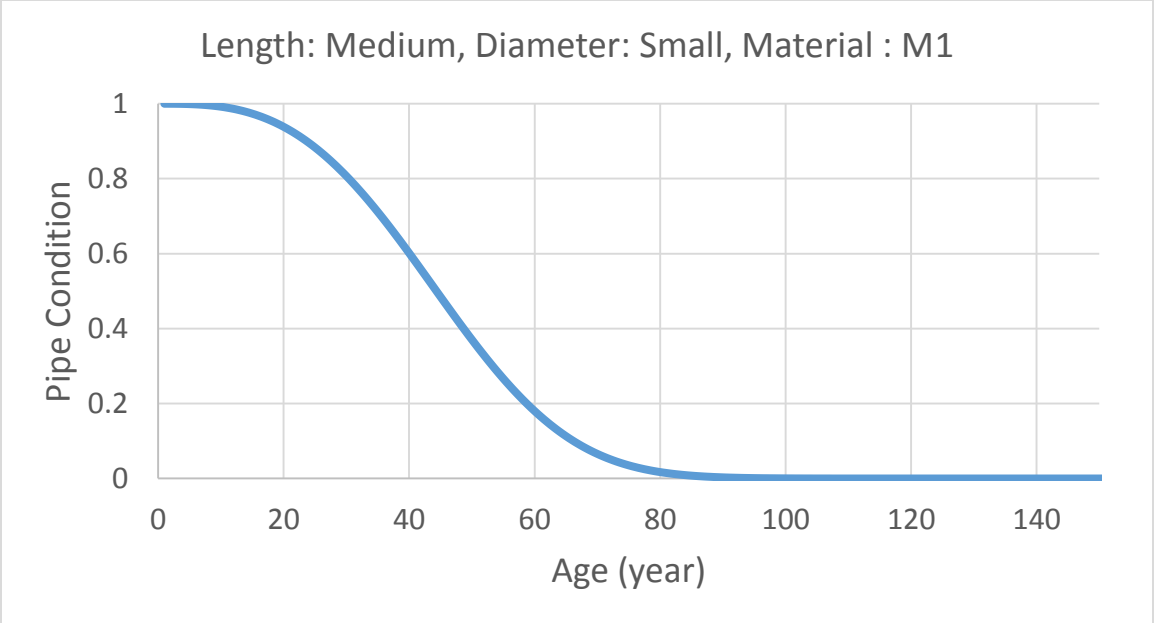


Figure 5-6 Deterioration Curve for Cluster number 7

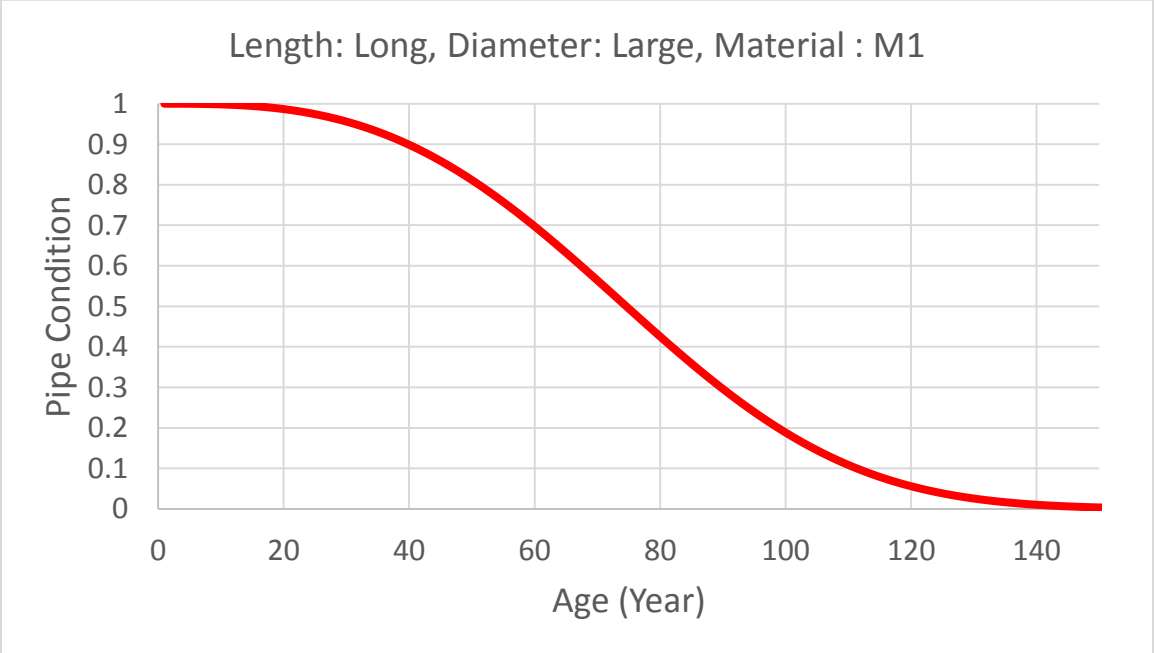


Figure 5-7 Deterioration Curve for Cluster number 16

5.2.5 Sensitivity Analysis

Sensitivity analysis was performed to identify the effect of changing each variable on the predicted number of breaks of any pipe approaches the end of its useful life. Figures 5-8, 5-9, and 5-10 show the effect of diameter, length, and pipe material on the number of breaks respectively as the pipe ages. In each one, a factor, which is the aim of the study, was changed while the rest ones were constant. In these graphs, the vertical axis shows the number of breaks while the horizontal axis represents the age of the pipe. It is clear from these figures that the number of breaks is increased when pipes approach the end of their useful life. As it can be seen in Figure 5-8, the number of breaks for pipes with the small diameter is higher than large diameter pipes. In the other words, the smaller the diameter of the pipe, the higher its value of the number of breaks will be. This can be justified because the wall thickness of smaller pipes is thinner than the larger ones, which allows the pipe to be corroded faster (El-Abbasy et al. 2014). Figure 5-9 shows that the number of breaks for longer length pipes is higher than pipes with the shorter length. These observations confirm previous findings in the literature about the relation between pipe's failure rate and its length and diameter (Berardi et al. 2008). Figure 5-10 shows the sensitivity analysis for pipe with different materials. As it was discussed in chapter 3, the qualitative variables should be converted to quantitative variables to apply with EPR. In this study, pipe materials were divided into six groups based on their historical pipe's failure. Number 1 was assigned to pipes with the lowest historical failure rate while number 6 was assigned to the pipes with highest historical failure rate.

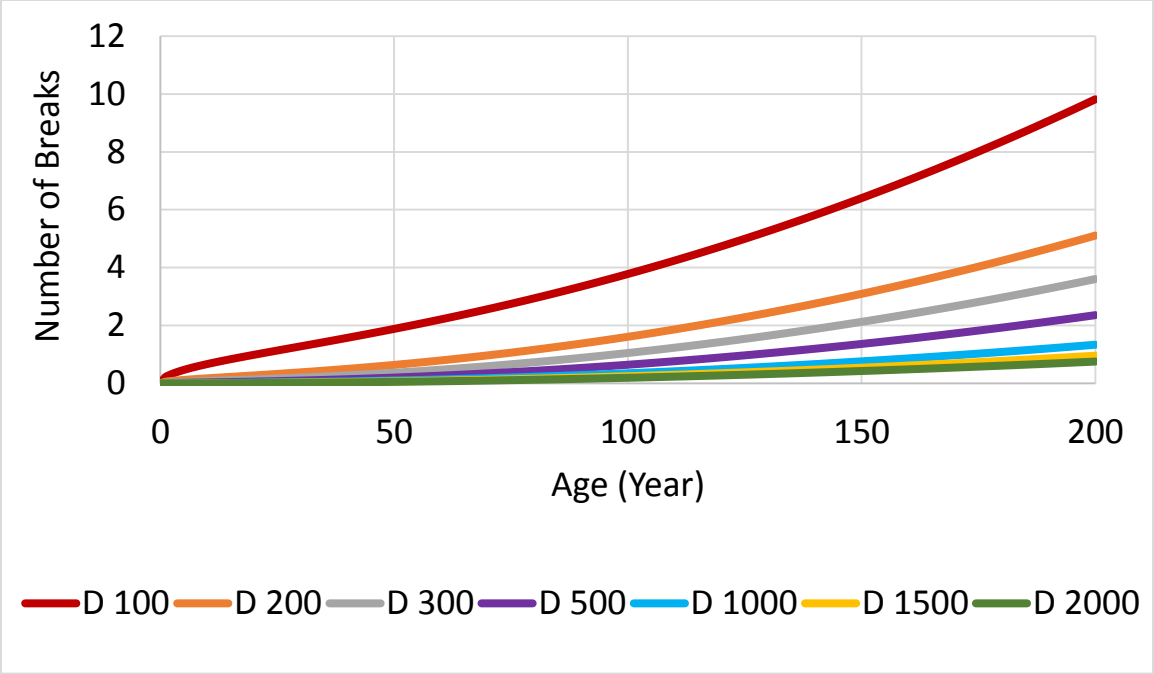


Figure 5-8 Number of Breaks for Different Pipe Diameter

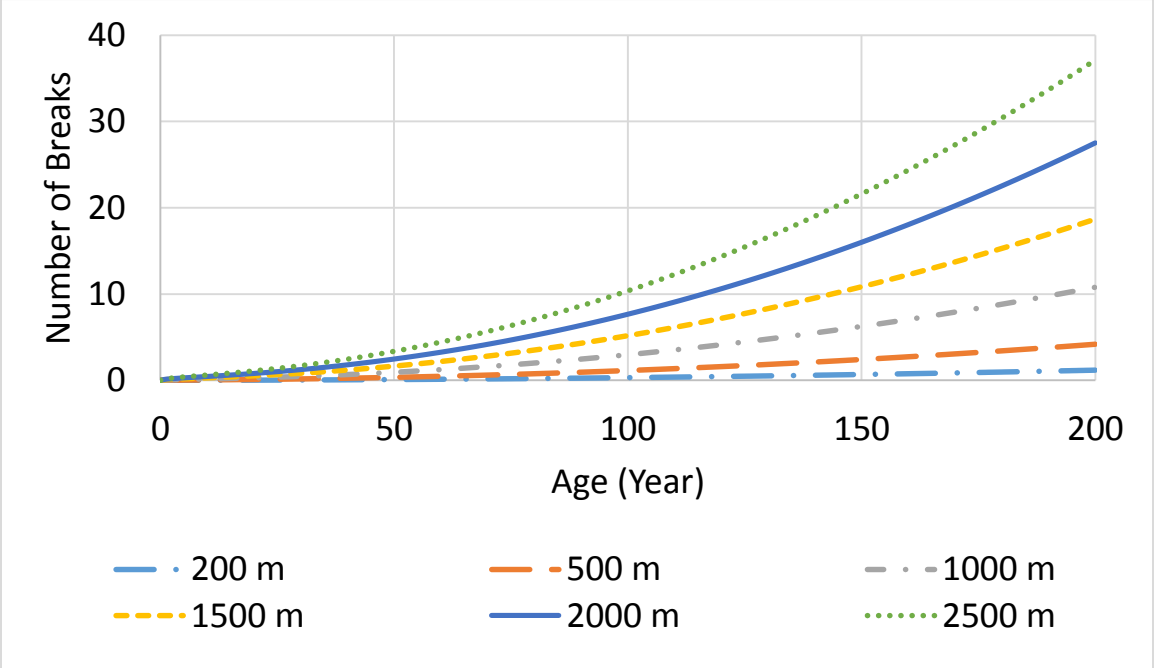


Figure 5-9 Number of Breaks for Different Pipe Length

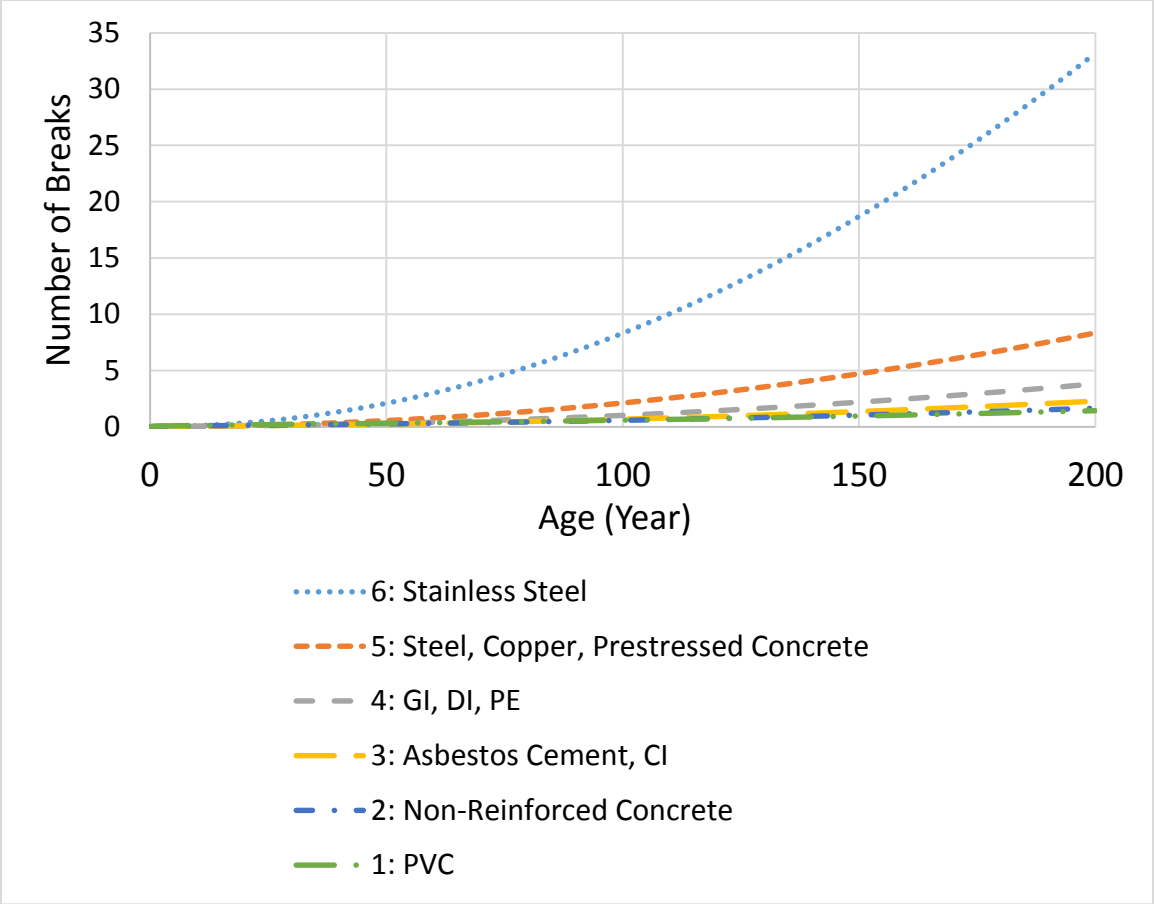


Figure 5-10 Number of Breaks for Different Pipe Material

Figure 5-11 shows the effect of changing of all input factors on the number of breaks of water pipelines. This graph was developed for two purposes. The first one is to understand the interrelationship between the number of breaks and its input factors. The second purpose is to determine what the most sensitive independent variables are. The vertical axis represents the number of breaks (Logarithmic Scale) whereas the horizontal axes represent the value of each factor. Since, each factor has its own unit, the horizontal axis was plotted using the normalized value from 0.01 to 1. However, for a better visualization the actual values of each factor are listed in a separated table below the normalized values. The corresponding values of the number of breaks are listed in another

table as well. The results confirm the direct relation between age and length as inputs and number of breaks as the output. It means that the number of breaks increases when the age and length of the pipe increase. Also, there is an inverse relationship between pipe's diameter and the number of breaks of the pipes. In other words, the number of breaks increases when the pipe diameter decreases. Among these four curves, changing the value of number of breaks in gray curve (pipe diameter) is more than the others which shows that the most sensitive factor in this model is pipe diameter.

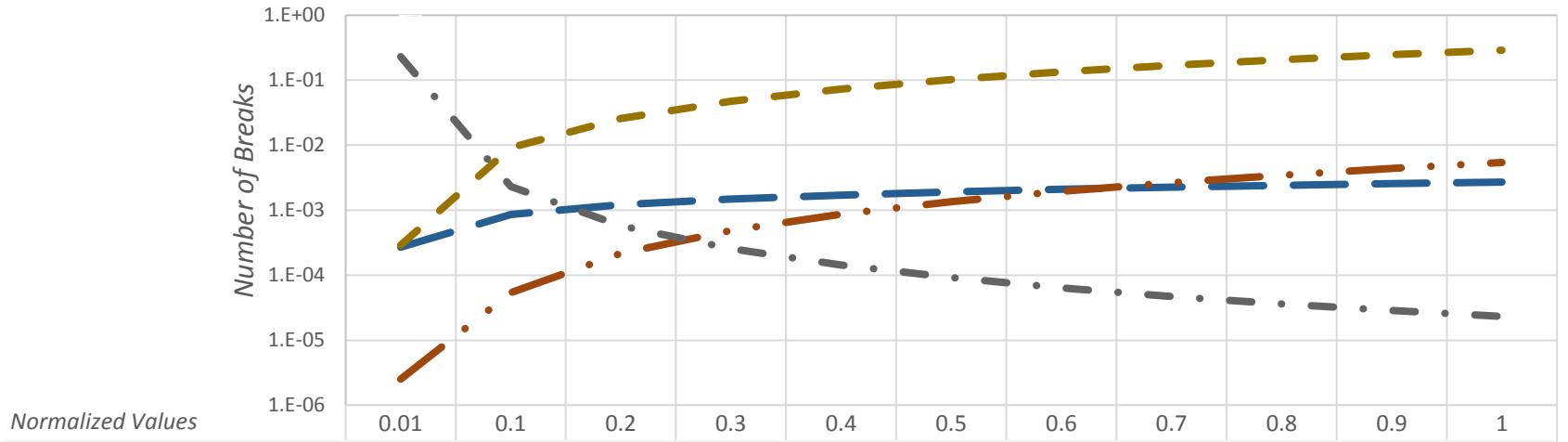
5.3 City of Doha

The dataset is used in this section is for City of Doha, Qatar. As it was discussed in data collection chapter, this city has a population of 796,947, while its land area is around 132.1 square kilometers. The city of Doha owns 1,926 kilometers of water distribution networks (Kahramaa, 2009).

5.3.1 Number of Breaks Estimation

As it mentioned in data collection, the number of breaks was not available in the dataset of Doha. Lack of such data prevents developing prediction models with EPR, because it considers the number of breaks as the output. Thus, it was required to estimate the number of breaks based on the other available datasets.

The physical characteristics of water pipes in different datasets are generic. In fact, the results obtained using the Hamilton and Moncton data were very close. In view of this finding and the insufficient data collected from Doha, it was required to use the developed model based on historical records from Hamilton and Moncton to estimate the number of



Normalized Values	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Factors' Real Value											
Age (yr)	1.00	19.00	39.00	59.00	79.00	99.00	119.00	139.00	159.00	179.00	199.00
Material	1.00	1.45	1.96	2.46	2.97	3.47	3.98	4.48	4.99	5.49	6.00
Diameter (mm)	20.00	372.73	764.65	1156.57	1548.48	1940.40	2332.32	2724.24	3116.16	3508.08	3900.00
Length (m)	0.15	7769.37	16401.84	25034.31	33666.78	42299.25	50931.72	59564.18	68196.65	76829.12	85461.59
Corresponding Number of Breaks											
Age	2.70E-04	8.55E-04	1.21E-03	1.48E-03	1.71E-03	1.91E-03	2.09E-03	2.26E-03	2.42E-03	2.56E-03	2.70E-03
Material	2.52E-06	5.39E-05	2.16E-04	4.85E-04	8.63E-04	1.35E-03	1.94E-03	2.64E-03	3.45E-03	4.37E-03	5.39E-03
Diameter	2.30E-01	2.30E-03	5.75E-04	2.56E-04	1.44E-04	9.22E-05	6.41E-05	4.71E-05	3.61E-05	2.85E-05	2.31E-05
Length	2.89E-04	9.13E-03	2.58E-02	4.74E-02	7.30E-02	1.02E-01	1.34E-01	1.69E-01	2.06E-01	2.46E-01	2.89E-01

Figure 5-11 Sensitivity Analysis for Montreal Dataset

breaks in Doha. It should be mentioned that population, city size and pipe characteristics of the City of Doha are similar to the City of Moncton and Hamilton as well. Estimation of the number of breaks was implemented by considering age as an input, which can be found in all datasets. Three different models were developed by applying regression analysis of Excel using datasets of Moncton and Hamilton. In the first two models, the data of each city was used separately, while, in the third one the combined data for both cities was utilized. In each model, the data was clustered into different groups based on the pipe age. It means that pipes with same age were put in one group. The breaks per length (m) was calculated for each age-class by computing the average of the number of breaks for the same group. Several attempts were conducted to reach the best model using different datasets. Since, in the dataset of Doha, there are no pipes older than 33 years, it was not necessary to keep pipes with the age of 34 and more, therefore they could be deleted in the new inventories. Finally, the model that utilized the large number of data points and gave the best performance based on the R-Squared (R2) was chosen to estimate the number of breaks for the city of Doha.

Figures 5-12, 5-13, and 5-14 show the result of regression (based on the No. of Breaks per Length (m)) of Moncton, Hamilton and mixing of both cities, respectively. The equation of each inventory and R-Square (R2) are shown in Table 5-5. It can be seen that the developed models of Moncton and both Cities are acceptable; while, the one that belongs to the City of Hamilton is not promising enough to be used on Doha. Finally, number of breaks per length that was obtained from these equations should be multiplied

by length of related pipe segments to calculate the estimated number of breaks of Doha's dataset.

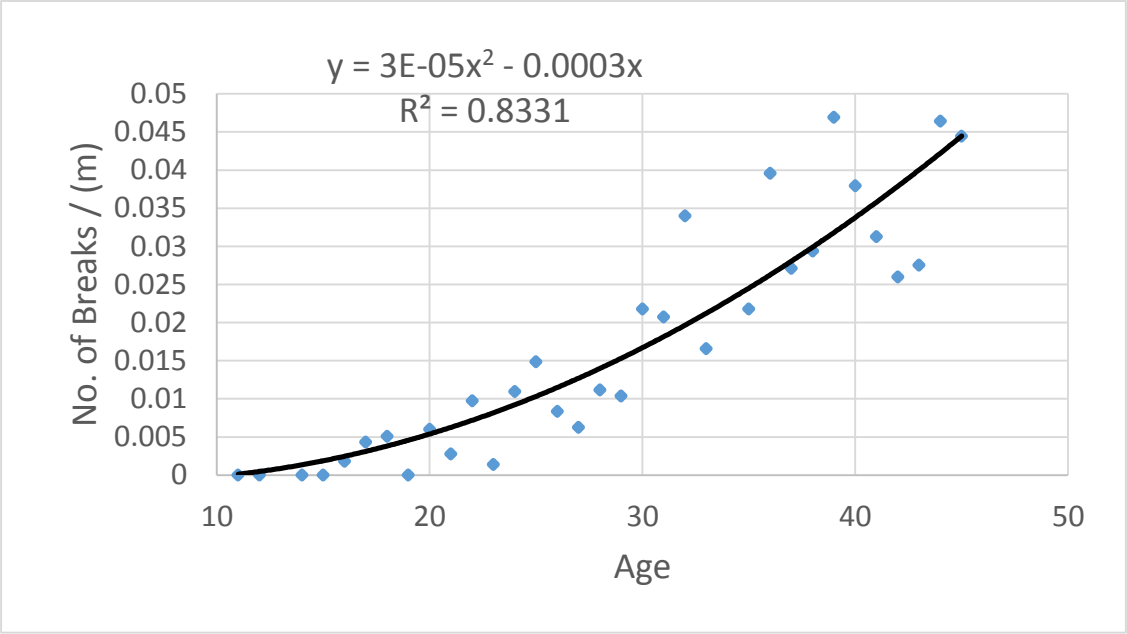


Figure 5-12 Scatter Plot of No. of Breaks per Length (m) and Age of Moncton Dataset

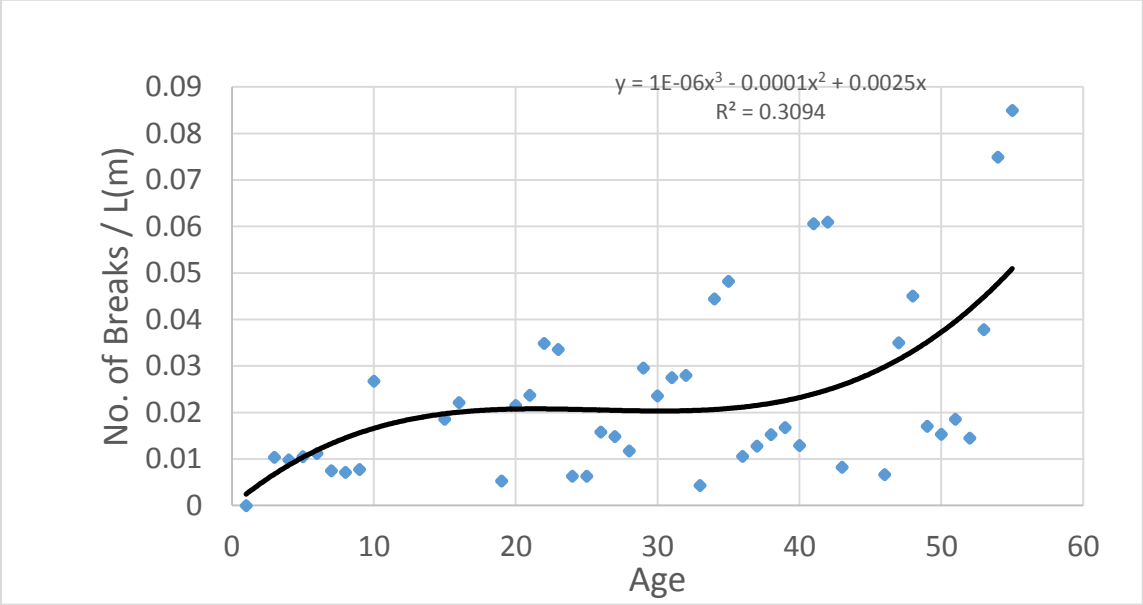


Figure 5-13 Scatter Plot of No. of Breaks per Length (m) and Age of Hamilton Dataset

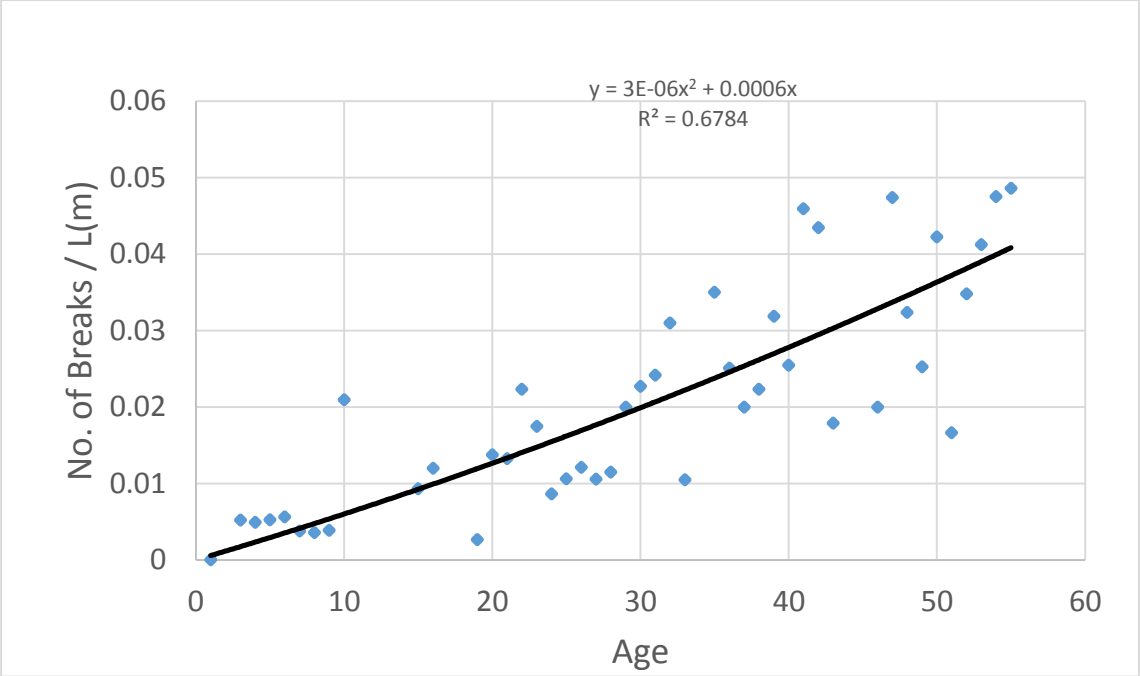


Figure 5-14 Scatter Plot of No. of Breaks per Length (m) and Age of Both Datasets

Table 5-5 Equations and related R-Squares

Different Datasets	Equations	R-Squared (%)
Moncton	$y = 3E\text{-}05\ x^2 - 0.0003\ x$	83.31
Hamilton	$y = 1E{-}06x^3 - 0.0001x^2 + 0.0025x$	30.94
Mixing of Both Cities	$y = 3E\text{-}06\ x^2 + 0.0006\ x$	67.84

Once, the number of breaks for the City of Doha was estimated, the analysis for this dataset is conducted. Figure 5-15 shows the number of breaks for ductile iron and steel pipes installed between 1981 and 2013. The highest range of pipes failure belongs to the period of 1996-2000. While from 1991 to 1995 and from 2001 to 2010, the frequencies of

pipe failures are almost equal. Having a higher pipe failure in a specific period of time can be caused by poor installation methods or low-quality materials.

Figure 5-16 demonstrates the number of breaks for pipes with different diameter and installation date. As it can be seen, the number of breaks for pipes with the smaller diameter is higher than the pipes with the larger diameter. This confirms the previous findings regarding the inverse relationship between failure rate and pipe diameter. The highest number of breaks belongs to the pipes with 100mm diameter that were installed between 1996 and 2000.

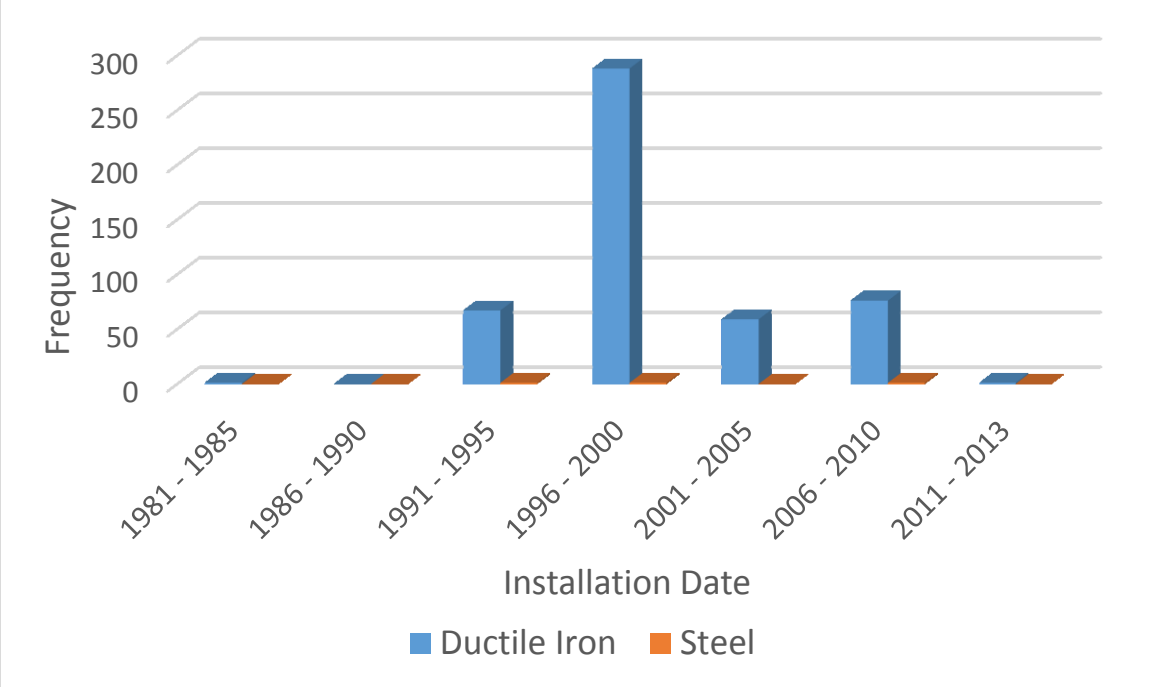


Figure 5-15 Number of Breaks per segment for pipes Installed between 1981 and 2013 for City of Doha

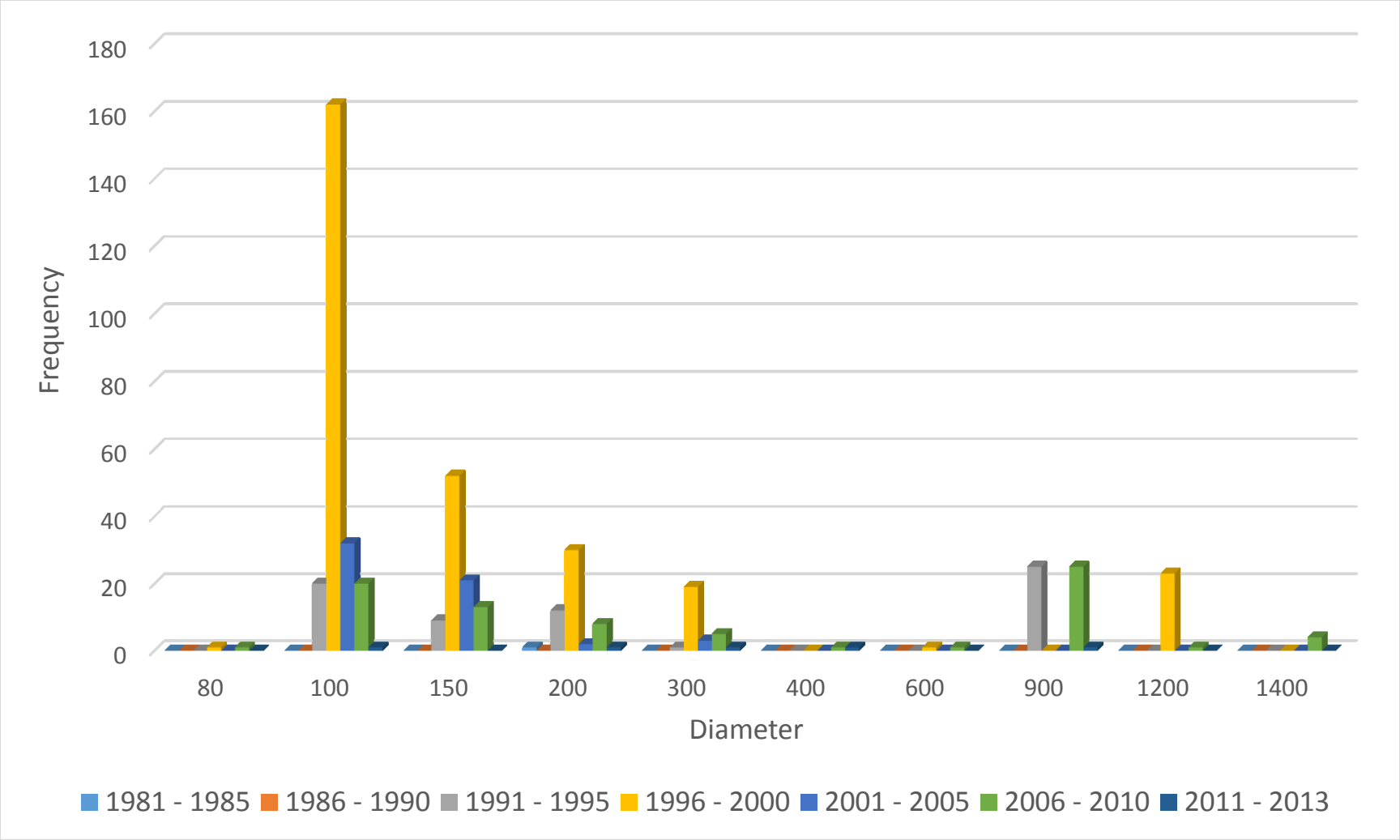
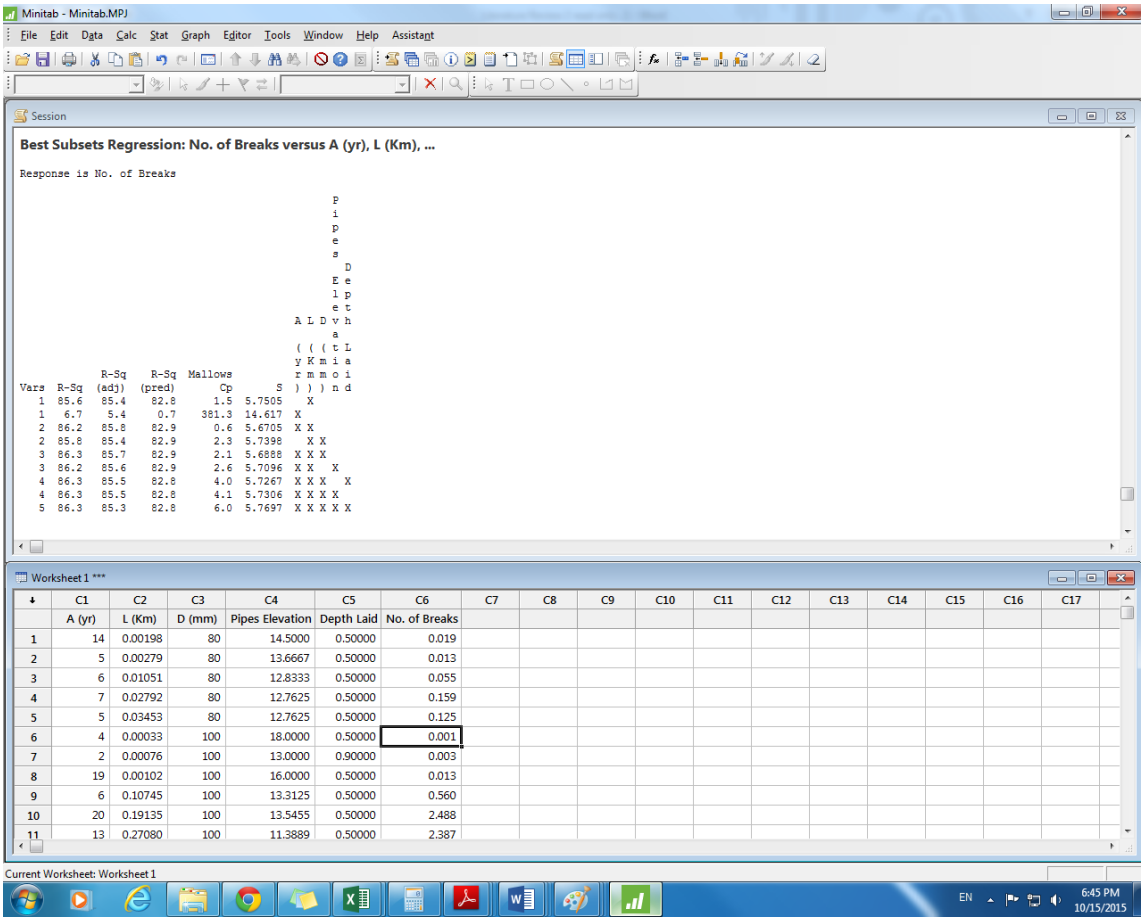


Figure 5-16 Number of Breaks per segment for Pipes with Different Diameter for City of Doha

5.3.2 Best Subset Regression

As it is mentioned in chapter 3, Dataset of Doha contains six independent variables including: the length, diameter, age, material, buried depth, and elevation of the pipe. However, this dataset comprises of 99.99% Ductile Iron and 0.01% Steel pipes thus, only ductile iron pipes were considered in this study. Figure 5-17 shows results of the best subset analysis for the City of Doha. As it can be seen in the upper window, there are nine possible sets of inputs in this dataset. All of them except model number 2 have the high value of R-Squared, adjusted R-Squared, and predicted R-Squared. However, the value for



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17
	A (yr)	L (Km)	D (mm)	Pipes Elevation	Depth Laid	No. of Breaks											
1	14	0.00198	80	14.5000	0.50000	0.019											
2	5	0.00279	80	13.6667	0.50000	0.013											
3	6	0.01051	80	12.8333	0.50000	0.055											
4	7	0.02792	80	12.7625	0.50000	0.159											
5	5	0.03453	80	12.7625	0.50000	0.125											
6	4	0.00033	100	18.0000	0.50000	0.001											
7	2	0.00076	100	13.0000	0.90000	0.003											
8	19	0.00102	100	16.0000	0.50000	0.013											
9	6	0.10745	100	13.3125	0.50000	0.560											
10	20	0.19135	100	13.5455	0.50000	2.488											
11	13	0.27080	100	11.3889	0.50000	2.387											

Figure 5-17 Best Subset Regression for City of Doha

Mallows' Cp should be equal to the number of independent variables plus one. In this dataset, there are five independent variables, thus only model number 9 has an acceptable value for Mallows' Cp, which is 6. Therefore, the of factors model 9 are selected as the most critical factors.

5.3.3 Data Classification

In this dataset, the aim of data classification is clustering the City of Doha dataset into groups that have the same age and diameter. The original excel file of the dataset of Doha comprises of 1,599 pipe segments. After data filtering, this dataset was classified into 72 homogeneous groups which is very smaller when compared with the dataset of Montreal. The length and the number of breaks of each class were computed by summing corresponding ones of each pipe segment. The original dataset and its classifications are provided in Appendix B.

5.3.4 Evolutionary Polynomial Regression

In the dataset of Doha, twelves symbolic expressions were generated to predict the number of breaks of water pipelines. Table 5-6 shows these expressions along with their related R-Squared scores. On the right side of symbolic expressions A, L, D, PE, and BD represent Age, Length, Diameter, Pipe Elevation, and Buried Depth respectively. It can be seen that age, length and diameter of the pipes are the most commonly used variables for estimating the number of breaks while buried depth and pipe elevation has been introduced in only the last five expressions. As discussed earlier, the best model should be chosen among all expressions based on the model fitness and parsimony. Model number 9 was selected as the best one, even though all models have acceptable R-Squared scores. The selected model has the highest value of R-Squared and is less complicated than models

number 10, 11, and 12. Accuracy indexes such as SSE, BIC, MSE, FPE, AIC, and GCV are shown in Table 5-7. Also by observing the results shown in this table, model number 9 has the minimum values in all indexes. This observation confirms that this model is the most promising one in predicting the output.

Table 5-6 Symbolic Expressions for Doha dataset and related R-Squared

	Expressions	R^2 ***(%)***
1	$No.\ of\ Breaks = 1.1493\ A^{0.5}\ L$	90.66
2	$No.\ of\ Breaks = 1.8169\times10^{-6}\frac{1}{L^2}+1.1471\ A^{0.5}\ L$	94.33
3	$No.\ of\ Breaks = 3.3795\times10^{-6}\frac{A^{0.5}}{L^2}+1.1466\ A^{0.5}\ L$	94.99
4	$No.\ of\ Breaks = 1.146\ A^{0.5}\ L+1.7317\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^{1.5}}\right)$	95.21
5	$No.\ of\ Breaks = 0.0066601\ln(D^{0.5})+1.1721\ A^{0.5}\ L+1.6884\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^2}\right)$	95.87
6	$No.\ of\ Breaks = 1.4624\times10^{-5}\frac{1}{L^{1.5}}\ln(D^{0.5})+\ 1.1467A^{0.5}L+2.0317\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^2}\right)$	95.58
7	$No.\ of\ Breaks = 3.4001\times10^{-5}\frac{A^{0.5}}{L^{1.5}}\ln(D^{0.5})+1.1473A^{0.5}L+2.4526\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^2}\right)$	95.7
8	$No.\ of\ Breaks = 3.2006\times10^{-5}\frac{A^{0.5}}{L^{1.5}}\ln(D^{0.5}PE^{0.5})+1.1477A^{0.5}L+2.8169\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^2}\right)$	95.91
9	$No.\ of\ Breaks = 3.4191\times10^{-7}\frac{A^{0.5}}{L^2D^{0.5}}\ln(PE^{1.5}BD^{0.5})+1.1463A^{0.5}L+3.5493\times10^{-5}\frac{A^2}{L^2}\ln\left(\frac{1}{A^{1.5}}\right)$	96.27
10	$No.\ of\ Breaks = 3.4559\times10^{-7}\frac{A^{0.5}}{L^2D^{0.5}}\ln(PE^{1.5}BD^{0.5})+1.1464A^{0.5}L+3.5792\times10^{-6}\frac{A^2}{L^2BD^{0.5}}\ln\left(\frac{1}{A^{1.5}}\right)$	96.1
11	$No.\ of\ Breaks = 3.817\times10^{-7}\frac{A^{0.5}}{L^2D^{0.5}}\ln\left(\frac{PE^{1.5}BD^{0.5}}{A^{0.5}}\right)+1.1464A^{0.5}L+3.5748\times10^{-6}\frac{A^2}{L^2BD^{0.5}}\ln\left(\frac{1}{A^{1.5}}\right)$	96.02
12	$No.\ of\ Breaks = 3.822\times10^{-7}\frac{A^{0.5}BD^{0.5}}{L^2D^{0.5}}\ln\left(\frac{PE^{1.5}BD^{0.5}}{A^{0.5}}\right)+1.1465A^{0.5}L+3.5775\times10^{-6}\frac{A^2}{L^2BD^{0.5}}\ln\left(\frac{1}{A^{1.5}}\right)$	96.05

Table 5-7 Accuracy Indexes for Doha Dataset

	SSE	***BIC***	***MSE***	***FPE***	***AIC***	***GCV***	***AVG***
Model #1	23.37	25.01	23.78	24.19	24.18	0.417	754.9
Model #2	14.19	15.18	14.44	14.69	14.68	0.253	5244
Model #3	12.53	14.29	12.98	13.43	13.4	0.232	8430
Model #4	11.99	13.67	12.42	12.85	12.82	0.222	6873
Model #5	10.35	11.8	10.72	11.09	11.06	0.191	8179
Model #6	11.07	13.4	11.68	12.28	12.22	0.212	11276
Model #7	10.77	13.03	11.35	11.94	11.88	0.206	16335
Model #8	10.24	12.39	10.8	11.36	11.3	0.196	11262
Model #9	9.327	11.29	9.836	10.34	10.29	0.179	8675
Model #10	9.774	11.83	10.31	10.84	10.79	0.187	10329
Model #11	9.958	12.05	10.5	11.04	10.99	0.191	11784
Model #12	9.897	11.98	10.44	10.98	10.92	0.19	7839

Figure 5-18 shows the Pareto graph of Doha dataset. Model number 9 is marked with a black arrow while other models are shown as red dots. In this graph, the vertical axis shows the number of independent variables, which were considered in each model. While, the horizontal axis represents the value of one minus R-Squared (1-CoD) for each model. Same as the dataset of Montreal, this dataset was divided randomly to two parts for training and testing. As shown in Table 5-8, 80% of dataset were used for training and 20% were used for testing. Scatter Plots for training and resting of model number 9 are shown in Figures 5-19 and 5-20 respectively. Scatter plots of other models are provided in Appendix C as well. These graphs compare the predicted and actual values of the number of breaks. The vertical axis shows the actual number of breaks (experimental) while the horizontal

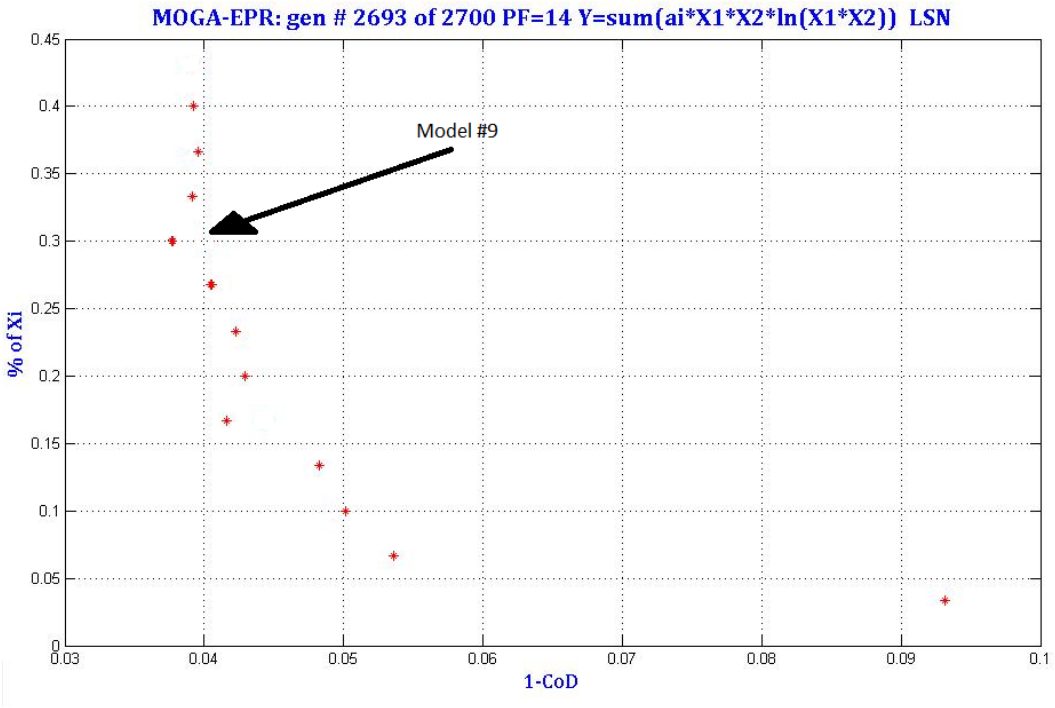


Figure 5-18 Pareto of Doha Dataset

axis shows the predicted value of the number of breaks. The values of R-Squared (CoD) is shown in top corners of graphs, which are 96.09% and 74.39% for training and testing respectively.

Table 5-8 Doha Dataset Size

City	Training Size	Testing Size	Dataset Size
Doha	58 (80%)	14 (20%)	72 (100%)

$$Breaks = +3.4191e-007\frac{A^{0.5}}{L^2 D^{0.5}}\ln\left(PE^{1.5}\,DeL^{0.5}\right) + 1.1463A^{0.5}L + 3.5493e-005\frac{A^2}{L^2}\ln\left(\frac{1}{A^{1.5}}\right)$$

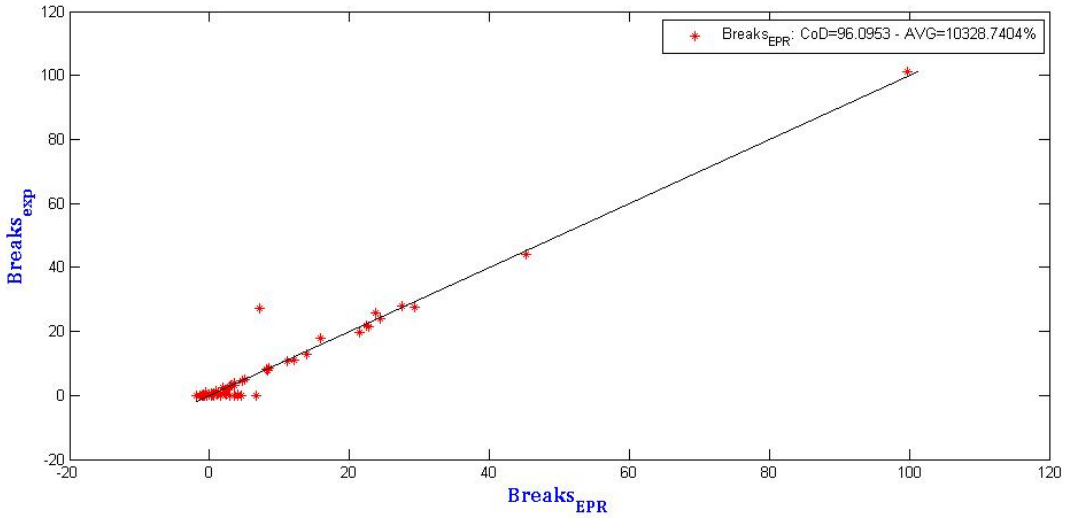


Figure 5-19 Scatter Plot of Model #9 for Training for Doha Dataset

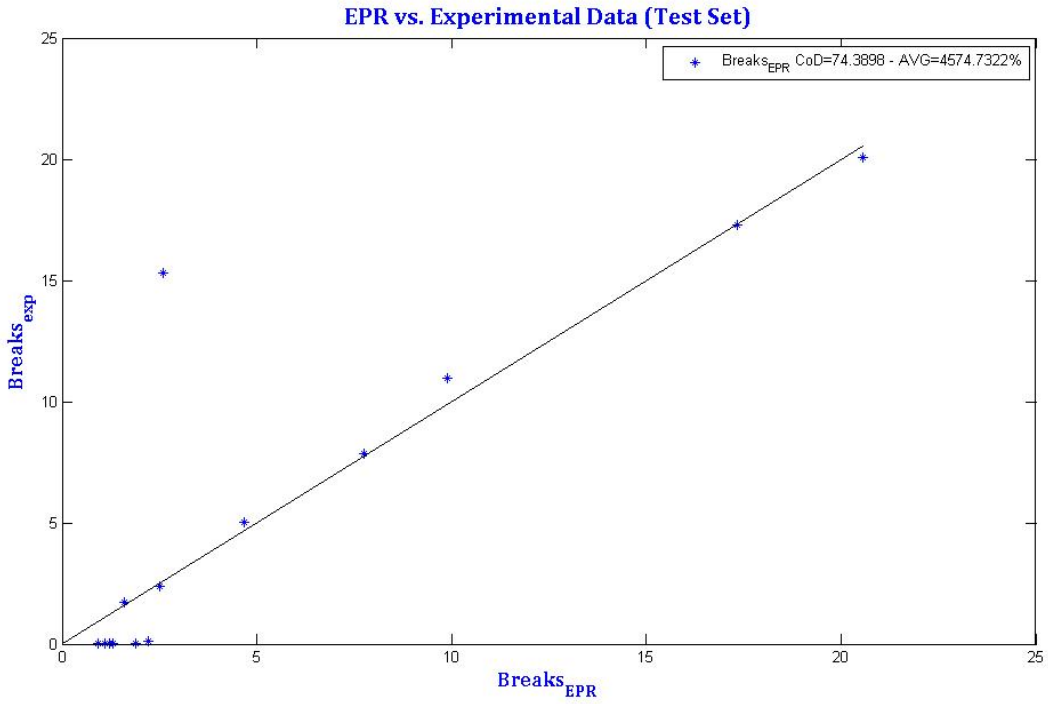


Figure 5-20 Scatter Plot of Model #9 for Testing for Doha Dataset

5.3.5 Weibull Distribution

The dataset of Doha comprises of 99.99% Ductile Iron thus, it clustered into 6 clusters just based on length and diameter of the pipes. Table 5-9 shows these 6 clusters and related features of each of them. For the pipe length, three subcategories were defined: short ($l \leq 300m$), medium ($300m < l \leq 2000m$), and long ($l > 2000m$). For the pipe diameter, two subcategories were defined based on the literature: small ($D \leq 350mm$) and large ($D > 350mm$). The number of breaks for each pipe segment, which was predicted in the previous section, is transformed to a breakage rate by dividing it by age and length. The result is used in this section to provide deterioration curves using Weibull reliability function. Figures 5-21 and 5-22 show deterioration curves for clusters number 3 and six respectively.

Table 5-9 Different Clusters and Related Features for Doha Dataset

Cluster	***Features***
1	Length: Short, Diameter: Small
2	Length: Medium, Diameter: Small
3	Length: Long, Diameter: Small
4	Length: Short, Diameter: Large
5	Length: Medium, Diameter: Large
6	Length: Large, Diameter: Large

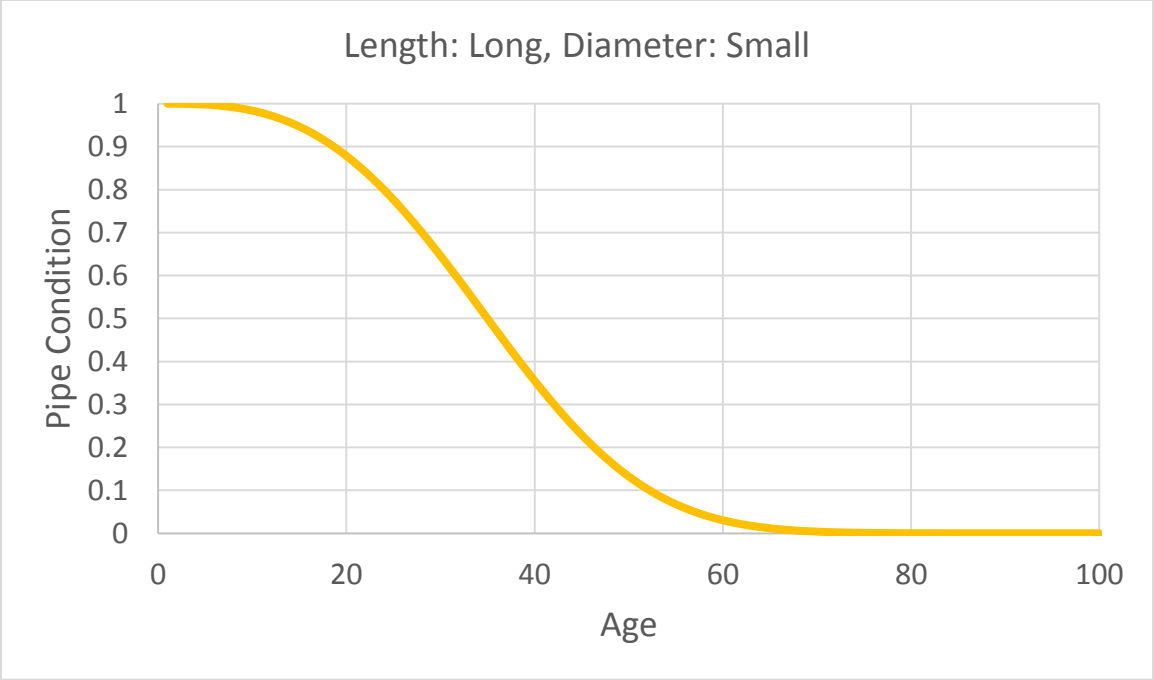


Figure 5-21 Deterioration Curve for Cluster Number 3

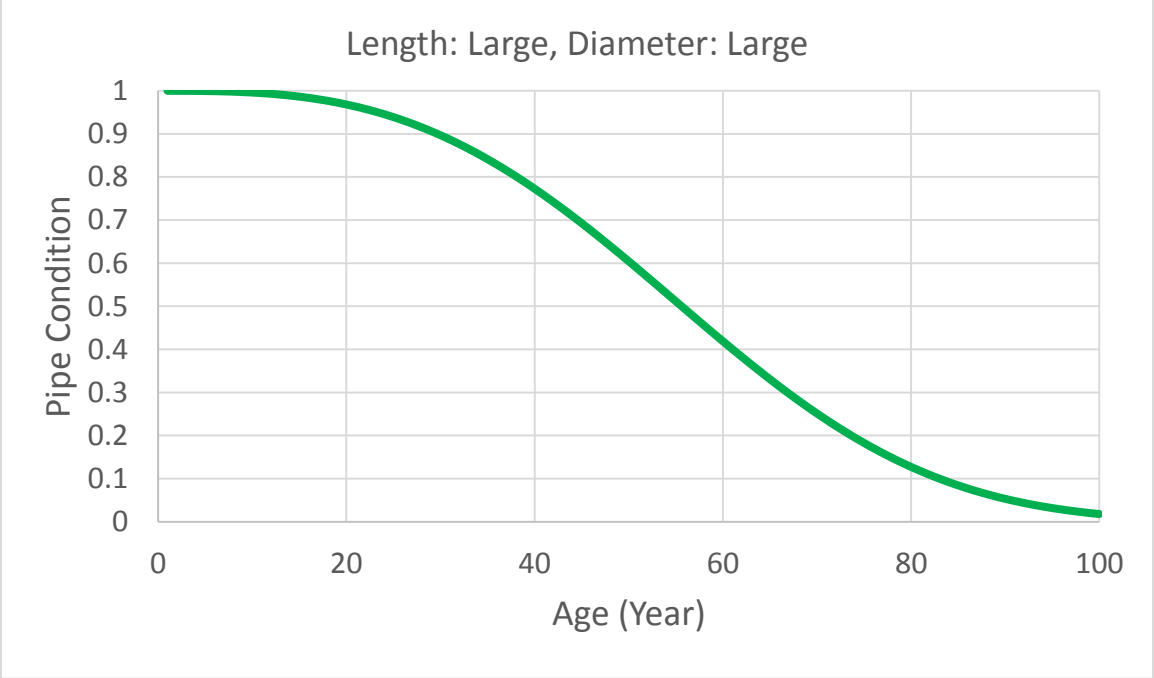


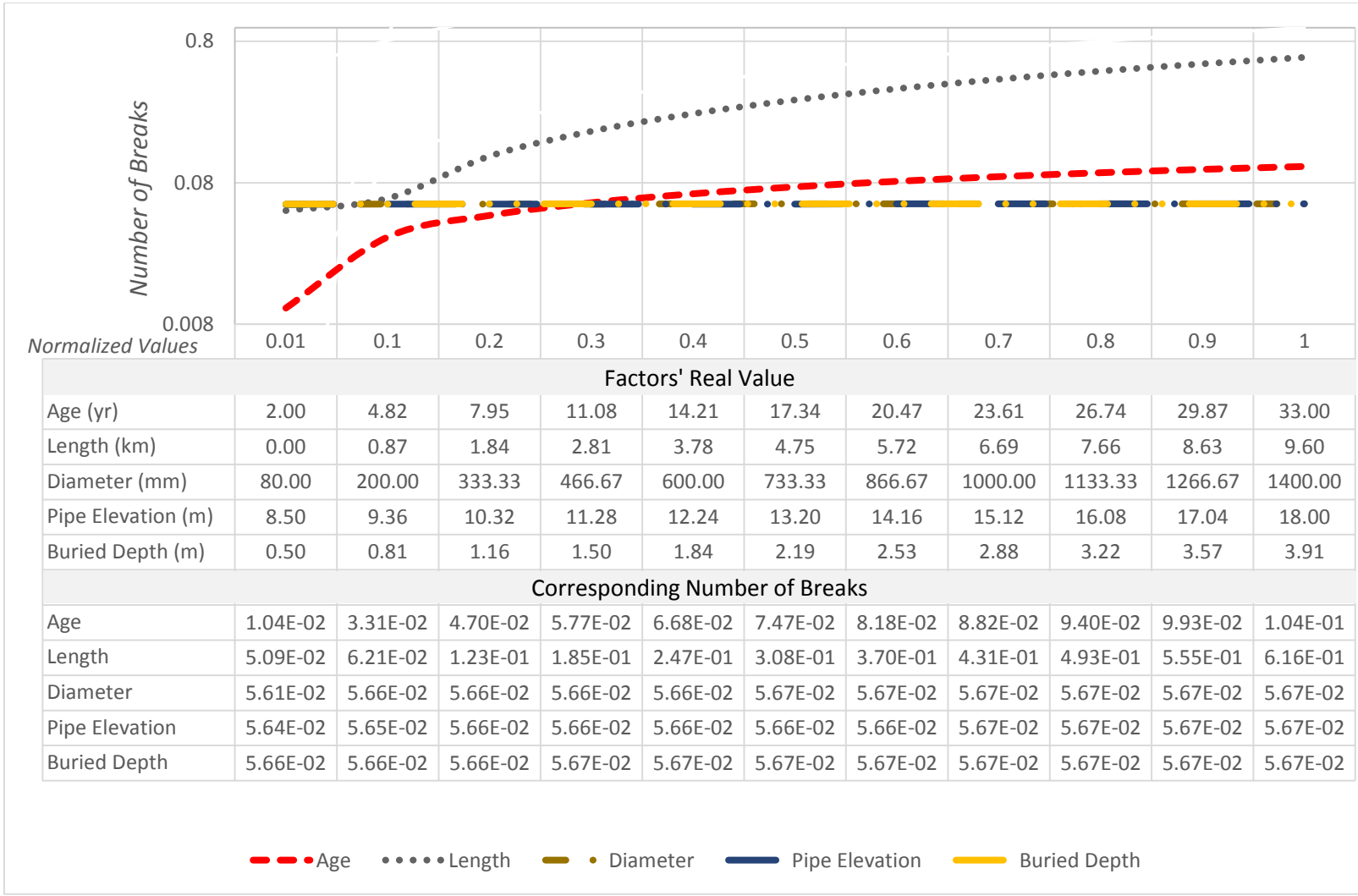
Figure 5-22 Deterioration Curve for Cluster Number 6

In each graph, the vertical axis shows the condition of the pipe while the horizontal axis represents the age of the pipe. As it can be seen, in both Figures pipe condition starts from 1 (the best condition) and then decreases slightly to the zero (the worse condition). It can be concluded that pipes with smaller diameter are more prone to failure than pipes with larger diameter.

5.3.6 Sensitivity Analysis

Figure 5-23 shows the sensitivity analysis for the dataset of Doha. The vertical axis represents the number of breaks (Logarithmic Scale) whereas the horizontal axis represents the value of each factor. Since, each factor has its own units, the horizontal axis was plotted using the normalized value from 0.01 to 1. However, for a better visualization the actual value of each factor and the corresponding value of the number of breaks are listed in two separated tables below the normalized values. The result confirms the previous finding from Montreal dataset that there is a direct relation between age and length as inputs and number of breaks as the output. In other words, the number of breaks increases when the age and length of the pipe increase. Also, it is concluded that pipe elevation and buried depth do not affect the water pipe failure significantly. By examining the above figure, the number of breaks is almost constant while the values of pipe elevation and buried depth are increasing. Also, it is found that in this study the first and second most sensitive independent variables are age and length of the pipe respectively. Thus, further analysis was done on these two factors.

Figures 5-24 shows the effect of different pipe length on the number of breaks while the water pipeline is aging. It can be seen that the number of breaks is increasing as the pipe is approaching the end of its service. The slope of curves shows the rate of increasing



Factors' Real Value											
Age (yr)	2.00	4.82	7.95	11.08	14.21	17.34	20.47	23.61	26.74	29.87	33.00
Length (km)	0.00	0.87	1.84	2.81	3.78	4.75	5.72	6.69	7.66	8.63	9.60
Diameter (mm)	80.00	200.00	333.33	466.67	600.00	733.33	866.67	1000.00	1133.33	1266.67	1400.00
Pipe Elevation (m)	8.50	9.36	10.32	11.28	12.24	13.20	14.16	15.12	16.08	17.04	18.00
Buried Depth (m)	0.50	0.81	1.16	1.50	1.84	2.19	2.53	2.88	3.22	3.57	3.91
Corresponding Number of Breaks											
Age	1.04E-02	3.31E-02	4.70E-02	5.77E-02	6.68E-02	7.47E-02	8.18E-02	8.82E-02	9.40E-02	9.93E-02	1.04E-01
Length	5.09E-02	6.21E-02	1.23E-01	1.85E-01	2.47E-01	3.08E-01	3.70E-01	4.31E-01	4.93E-01	5.55E-01	6.16E-01
Diameter	5.61E-02	5.66E-02	5.66E-02	5.66E-02	5.66E-02	5.67E-02	5.67E-02	5.67E-02	5.67E-02	5.67E-02	5.67E-02
Pipe Elevation	5.64E-02	5.65E-02	5.66E-02	5.66E-02	5.66E-02	5.66E-02	5.66E-02	5.67E-02	5.67E-02	5.67E-02	5.67E-02
Buried Depth	5.66E-02	5.66E-02	5.66E-02	5.67E-02	5.67E-02	5.67E-02	5.67E-02	5.67E-02	5.67E-02	5.67E-02	5.67E-02

Figure 5-23 Sensitivity Analysis for Doha Dataset

of the number of breaks. Thus, long length pipes have higher breakage rate than short length pipes.

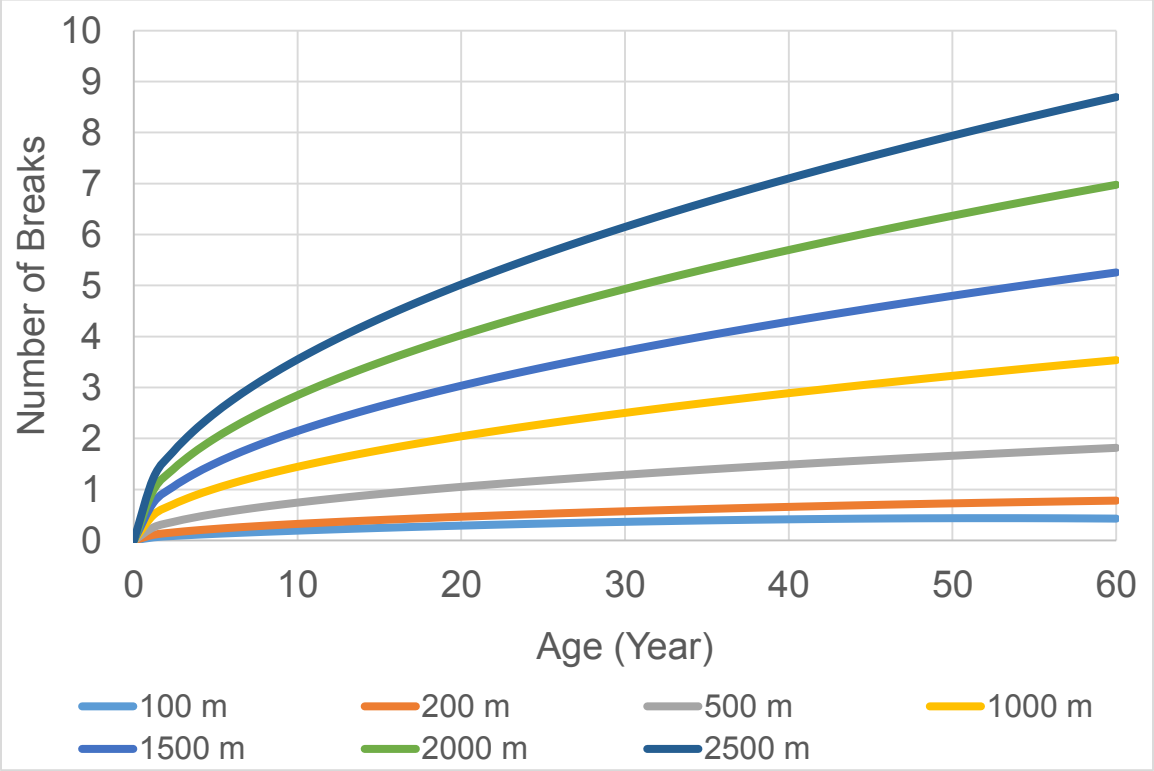


Figure 5-24 Number of Breaks for different Pipe Length

5.4 Summary and Conclusion

This chapter presented the analysis of the developed model on two case studies: The City of Montreal in Canada and the City of Doha in Qatar. The developed model encompasses three main computational techniques: Best Subset regression, Evolutionary Polynomial Regression, and Weibull reliability analysis. Best Subset regression was utilized to determine the most critical factors for predicting the number of breaks in water pipelines. Then, the selected critical factors were used to generate 12 symbolic expressions

by EPR. Subsequently, the predicted number of breaks by EPR is utilized as an input to generate deterioration curves by using Weibull distribution function.

The only difference between these two datasets is lack of information about the number of breaks in the dataset of Doha. Lack of such data prevents developing the prediction models by using EPR and, therefore, there was a need to estimate the number of breaks in the dataset of Doha. When examining the results obtained from the two datasets of Hamilton and Moncton, it was found out that these datasets were very close. Hence, the model developed based on them was used to estimate the number of breaks in Doha.

Data collection was performed for both cases to cluster pipe segments into classes that have the same specifications. The dataset of Montreal was classified based on age, diameter and material of pipes, while dataset of Doha was classified based on age and diameter of the pipes because it mostly comprises of Ductile Iron.

Sensitivity analysis was performed for both datasets to identify the effect of changing each independent variable on the water pipe failure rate when pipe gets older. The rationality of relationship between inputs and output in selected symbolic expression was studied as well.

Based on the Best Subset regression results, it was concluded that all available factors should be considered as inputs in EPR for predicting the number of breaks. Then, 12 symbolic expressions were generated by using EPR. Among them the best one was selected based on different criteria such as fitting to the actual data, the parsimony of generated equation and the possibility of justifying the equations in terms of reasonable relationship between inputs and output. In the end, two deterioration curves as samples

were presented for each dataset. As Weibull reliability function can be used for an individual pipe, providing a curve for each pipe segment is possible.

Chapter 6: Conclusion and Future Work

6.1 Summary and Conclusion

The increasing failure rates of water pipes are caused by the low maintenance and the aging of water distribution networks (Asnashari 2013). Failure prediction models can help utilities and municipalities prioritize the replacement/rehabilitation of water pipelines. The end result is more cost effective plans for the condition assessment and improved level of service. Recently, there has been considerable efforts in developing failure prediction models for water pipes as covered in the literature review of this thesis. This study presented a research framework that circumvent the limitations highlighted in Chapter 2 by: 1) identifying the most critical factors affecting failure rates of water pipes, 2) determining the best mathematical expression for relating the identified factors with the target output – i.e. breakage rates, 3) using the best mathematical formula to construct deterioration curves and 4) deploying the sensitivity analysis to recognize the effects of changing each input on the breakage rate.

Best Subset regression was utilized to find the best combination of variables for predicting breakage rates of water pipes. The technique was capable of extracting the most critical factors for predicting breakage rates using the numbers of statistical indices such as R2, Mallows' Cp and square root of MSE. However, this technique is not suitable for case studies with a large number of independent variables as the computational time needed to process and find the best combination of factors will significantly increase. But in this study, 4 and 5 independent variables were used to predict the number of breaks in the City of Montreal and Doha respectively. Therefore, Best Subset regression was capable of

finding the best factors in a timely fashion. In dataset of Montreal, the model, which includes all available independent variables, being age, diameter, length and material of pipes, was selected as the best one. Also, for the dataset of Doha, age, diameter, length, material, buried depth and elevation of the pipe were selected as the most critical factors.

Subsequently, EPR algorithm was deployed to generate a number of symbolic expressions able to predict the number of breaks of water mains. For each dataset, 12 symbolic expressions were generated and among them, the best one was chosen based on the observed fitness and parsimony of the equation. The process of creating the symbolic expressions contains two stages: 1) Finding the best model structure using Multi-Objective Genetic Algorithm and 2) estimating the appropriate values for constants using Least-Squares optimization (Berardi et al. 2008).

The predicted number of breaks, calculated by the best symbolic expression, was employed to construct deterioration curves by using Weibull reliability functions. Weibull distribution was utilised because it needs a few number of historical data and can also be used to model either an individual pipe or the whole network. Datasets of Montreal and Doha were grouped into 18 and 6 clusters respectively and a deterioration curve was developed for each group.

The sensitivity analysis was performed for both datasets to: 1) identify the effect of each independent variable on the breakage rate when water pipes are aging and 2) study the rationality of relationship between the selected inputs and the output. In dataset of Montreal, it was concluded that the pipe diameter is the most sensitive factor. In dataset of Doha, however, age and length of the pipe were identified as the most sensitive factors.

6.2 Research Contribution

This study provides a newly developed research framework for predicting the number of breaks for water pipes. As the most significant contributions of this research:

1) The most critical factors for predicting the failure rate of water mains were identified from the available literature and historical data.
2) The failure rate prediction models for water distribution networks were developed, considering the interrelationships among the most critical factors. Furthermore, different types of pipe material were considered as an independent variable. The result of this model was used to provide deterioration curves of water pipelines.
3) Two types of sensitivity analysis were conducted for each dataset, aiming to: 1) identify the effect of each independent variable on the breakage rate and 2) study the rationality of relationship between the selected inputs and the output.

6.3 Limitations

The developed methodology has some limitations, listed as follows:

1) Lack of available data prevented considering more inputs such as soil type, which was identified as one of the most important factors in predicting failure rate of water distribution networks.
2) The effect of third party, mechanical damages, construction defects and corrosion were not considered in this study.
3) The developed methodology does not take into consideration the effect of rehabilitation on water pipelines.

6.4 Future Works

The recommendation for future works can be divided into two areas: 1) Research enhancement and 2) research extensions. These two areas are summarized as follows:

6.4.1 Research Enhancement

1) Considering additional effective factors such as soil type in water distribution networks as inputs because the proposed methodology is flexible to include more contributing factors. According to the literature, the soil type was identified as one of the most important factors in predicting the failure rate of water mains.
2) Developing a user-friendly interface wherein the user inserts the pipe's specifications in order to obtain the most critical factors, the best mathematical form for predicting water pipe failures, deterioration curves and the most sensitive factor as outcomes. Also, this tool can be adapted to a web version to be accessible for interested parties across the world.
3) Implementing the developed research framework in more water distribution networks (other than North America and Middle East) in order to explore its capabilities and investigate the result validity with more datasets.
4) Considering the effect of third party, mechanical damages and construction defects in developing the prediction failure rate models for water distribution networks.
5) Investigating how the rehabilitation of water pipes can affect the deterioration curves. Considering this effect leads to more accurate and realistic deterioration curves to be generated.

6.4.2 Research Extensions

1) Maintenance, repair and rehabilitation plan can be prioritized based on the result of this study. Also, the budget allocation and life cycle cost optimization models can be integrated with this methodology to establish a more comprehensive framework for water pipes management.
2) Modifying the developed framework in order to be applicable in other infrastructure assets such as sewer pipelines, roads and bridges.

References

Achim, D., Ghotb, F., & McManus, K. (2007). Prediction of water pipe asset life using neural networks. *Journal of Infrastructure Systems, 13*(1), 26-30.

Alireza Ahangar-Asr, Faramarzi, A., Javadi, A. A., & Giustolisi, O. (2011). Modelling mechanical behaviour of rubber concrete using evolutionary polynomial regression. *Engineering Computations, 28*(4), 492-507. doi:10.1108/02644401111131902

Arsénio, A. M., Dheenathayalan, P., Hanssen, R., Vreeburg, J., & Rietveld, L. (2014). Pipe failure predictions in drinking water systems using satellite observations. *Structure and Infrastructure Engineering,* (ahead-of-print), 1-10.

ASCE (2013). Report Card for America's Infrastructure. Available Online at: http://www.infrastructurereportcard.org/

Asnaashari, A., McBean, E. A., Gharabaghi, B., & Tutt, D. (2013). Forecasting watermain failure using artificial neural network modelling. *Canadian Water Resources Journal, 38*(1), 24-33.

Atef, A. (2015). Asset management tools for municipal infrastructure considering interdependency and vulnerability (Degree of Doctor of Philosophy).

Aydogdu, M., & Firat, M. (2015). Estimation of failure rate in water distribution network using fuzzy clustering and LS-SVM methods. *Water Resources Management, 29*(5), 1575-1590.

AWWA. (1999). APPENDIX A-notes on procedures for soil survey tests and observation and their interpretations to determine wether polyethyline encasement should be used. (No. ANSI/AWWA C105/A21.5).American Water Works Association.

Berardi, L., Kapelan, Z., Giustolisi, O., & Savic, D. (2008). Development of pipe deterioration models for water distribution systems using EPR. *Journal of Hydroinformatics, 10*(2), 113-126.

Clair, A. M. S., & Sinha, S. (2014). Development of a standard data structure for predicting the remaining physical life and consequence of failure of water pipes. *Journal of Performance of Constructed Facilities,* DOI: 10.1061/(ASCE)CF.1943-5509.0000384.

Creighton, J. H. (2012). A first course in probability models and statistical inference Springer Science & Business Media, QA273.C847.

El-Abbasy, M. S., Senouci, A., Zayed, T., Mirahadi, F., & Parvizsedghy, L. (2014). Condition prediction models for oil and gas pipelines using regression analysis. *Journal of Construction Engineering and Management, 140*(6), 04014013.

El-Baroudy, I., Elshorbagy, A., Carey, S., Giustolisi, O., & Savic, D. (2010). Comparison of three data-driven techniques in modelling the evapotranspiration process. *Journal of Hydroinformatics, 12*(4), 365-379.

El Chanati, H. (2015). Performance assessment of water network infrastructure (Degree of Master of Applied Science).

Elshorbagy, A., & El-Baroudy, I. (2009). Investigating the capabilities of evolutionary data-driven techniques using the challenging estimation of soil moisture content. *Journal of Hydroinformatics, 11*(3-4), 237-251.

EPA (2009), State of Technology Review Report on Rehabilitation of Wastewater Collection and Water Distribution Systems. United States Environmental Protection Agency, Cincinnati, OH, USA. Available Online at: http://nepis.epa.gov/Adobe/PDF/P1008C45.pdf

Fiore, A., Berardi, L., & Marano, G. C. (2012). Predicting torsional strength of RC beams by using evolutionary polynomial regression. *Advances in Engineering Software, 47*(1), 178-187.

Francis, R. A., Guikema, S. D., & Henneman, L. (2014). Bayesian belief networks for predicting drinking water distribution system pipe breaks. *Reliability Engineering & System Safety, 130*, 1-11.

Gillenwater, J.Water distribution system: Pipes, valves, and flush hydrants. Retrieved from http://ohioline.osu.edu/b910/b910_13.html.

Giustolisi, O., & Savic, D. (2009). Advances in data-driven analyses and modelling using EPR-MOGA. *Journal of Hydroinformatics, 11*(3-4), 225-236.

Giustolisi, O. (2004). Using genetic programming to determine chezy resistance coefficient in corrugated channels. *Journal of Hydroinformatics, 6*, 157-173.

Grussing, M. N., Uzarski, D. R., & Marrano, L. R. (2006). Condition and reliability prediction models using the weibull probability distribution. Paper presented at the *Proc., 9th Int. Conf. on Applications of Advanced Technology in Transportation (AATT),* 19-24.

Iason, G. (2014). Infrastructure performance assessment of subway networks (Degree of Master of Applied Science).

Jafar, R., Shahrour, I., & Juran, I. (2010). Application of artificial neural networks (ANN) to model the failure of urban water mains. *Mathematical and Computer Modelling, 51*(9), 1170-1180.

Jardine, A. K., & Tsang, A. H. (2013). Maintenance, replacement, and reliability: Theory and applications CRC press, TS192.J37.

Jenkins, L., Gokhale, S., & McDonald, M. (2014). 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.

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

Kabir, G., Tesfamariam, S., Francisque, A., & Sadiq, R. (2015). Evaluating risk of water mains failure using a bayesian belief network model. *European Journal of Operational Research, 240*(1), 220-234.

Kabir, G., Tesfamariam, S., & Sadiq, R. (2015). Predicting water main failures using bayesian model averaging and survival modelling approach. *Reliability Engineering & System Safety, 142*, 498-514.

Karimian, F., Elsawah, H., Zayed, T., Moselhi, O., & Al Hawari, A. (2015). Forecasting breakage rate in water distribution networks using evolutionary polynomial regression. *The Canadian Society for Civil Engineering 5th International/11th Construction Specialty Conference, University of British Columbia, Vancouver,* 10.14288/1.0076483.

Kimutai, E., Betrie, G., Brander, R., Sadiq, R., & Tesfamariam, S. (2015). Comparison of statistical models for predicting pipe failures: Illustrative example with the city of calgary water main failure. *Journal of Pipeline Systems Engineering and Practice,* DOI: 10.1061/(ASCE)PS.1949-1204.0000196.

Kleiner, Y., & Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: Statistical models. *Urban Water, 3*(3), 131-150.

Kutyłowska, M. (2015). Neural network approach for failure rate prediction. *Engineering Failure Analysis, 47*, 41-48.

Lawless, J. (1983). Statistical methods in reliability. *Technometrics, 25*(4), 305-316.

Li, S., Yu, S., Zeng, H., Li, J., & Liang, R. (2009). Predicting corrosion remaining life of underground pipelines with a mechanically-based probabilistic model. *Journal of Petroleum Science and Engineering, 65*(3), 162-166.

Markus, M., Hejazi, M., Bajcsy, P., Giustolisi, O., & Savic, D. (2010). Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in illinois. *Journal of Hydroinformatics, 12*(3), 251-261.

Mena industry report: Doha invests heavily in infrastructure. (2014). Retrieved from https://globalconnections.hsbc.com/uae/en/articles/doha-invests-heavily-infrastructure.

Mitrani, I. (1998). Probabilistic modelling Cambridge University Press.

Moglia, M., Davis, P., & Burn, S. (2008). Strong exploration of a cast iron pipe failure model. *Reliability Engineering & System Safety, 93*(6), 885-896.

Moselhi, O., & Hegazy, T. (1993). Markup estimation using neural network methodology. *Computing Systems in Engineering, 4*(2), 135-145.

O'Connor, D. (2002). Report of the walkerton inquiry: Part two, a strategy for safe drinking water. Toronto, Ontario: Ontario Ministry of the Attorney General, Queen's Printer for Ontario.

Osman, H., & Bainbridge, K. (2010). Comparison of statistical deterioration models for water distribution networks. *Journal of Performance of Constructed Facilities, 25*(3), 259-266.

Pardoe, L. (2015). Stat 501. Unpublished manuscript.

Paul, S. M. (2014). Unpublished manuscript.

Rajani, B., & Kleiner, Y. (2001). Comprehensive review of structural deterioration of water mains: Physically based models. *Urban Water, 3*(3), 151-164.

Rezania, M., Javadi, A. A., & Giustolisi, O. (2008). An evolutionary-based data mining technique for assessment of civil engineering systems. *Engineering Computations, 25*(6), 500-517. doi:10.1108/02644400810891526

Savic, D., Giustolisi, O., Berardi, L., Shepherd, W., Djordjevic, S., & Saul, A. (2006). Modelling sewer failure by evolutionary computing. *Proceedings of the ICE-Water Management, 159*(2), 111-118.

Scheidegger, A., Leitão, J. P., & Scholten, L. (2015). Statistical failure models for water distribution pipes–A review from a unified perspective. *Water Research, 83*, 237-247.

Semaan, N. (2011). Structural performance model for subway networks (Degree of Doctor of Philosophy).

Shirzad, A., Tabesh, M., & Farmani, R. (2014). A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water distribution networks. *KSCE Journal of Civil Engineering, 18*(4), 941-948.

Siler, W., & Buckley, J. J. (2005). Fuzzy expert systems and fuzzy reasoning John Wiley & Sons, QA76.76.E95S557.

Sivanandam, S., Sumathi, S., & Deepa, S. (2007). *Introduction to fuzzy logic using MATLAB* Springer. SPIN 11764601 89/3100/SPi.

SOI Report (2005). State of the Infrastructure Report: City of Hamilton. Available Online at: http://www.myhamilton.ca/myhamilton/CityandGovernment/CityDepartments/PublicWorks/CapitalPlanning/Asset+Management/State+of+the+Infrastructure+Report.html

St. Clair, A. M., & Sinha, S. (2012). State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models! *Urban Water Journal, 9*(2), 85-112.

Stone, S. L., Dzuray, E. J., Meisegeier, D., Dahlborg, A., Erickson, M., & Tafuri, A. N. (2002). *Decision-support tools for predicting the performance of water distribution and wastewater collection systems* US Environmental Protection Agency, Office of Research and Development. EPA/600/R-02/029.

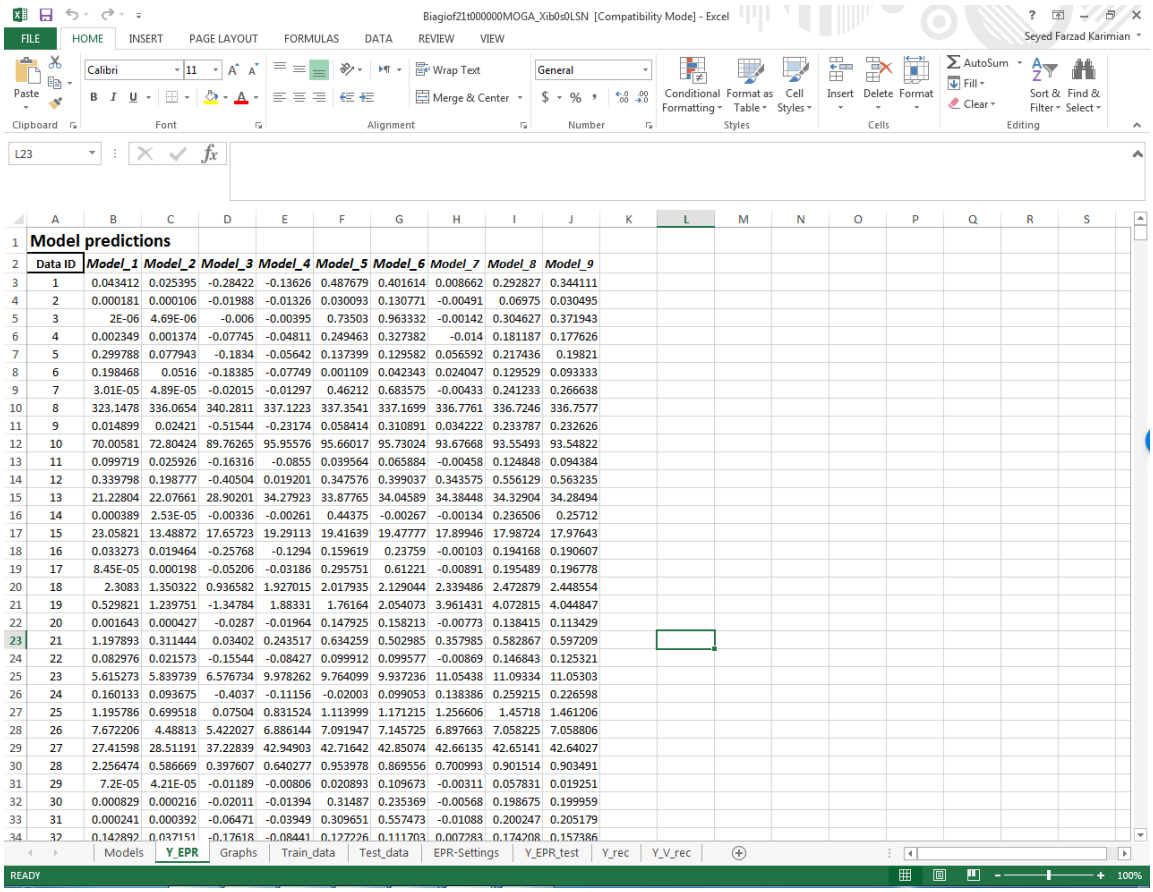
Ugarelli, R., Kristensen, S. M., Røstum, J., Sægrov, S., & Di Federico, V. (2009). Statistical analysis and definition of blockages-prediction formulae for the wastewater network of oslo by evolutionary computing. *Water Science and Technology, 59*(8), 1457-1470.

Wang, C., Niu, Z., Jia, H., & Zhang, H. (2010). An assessment model of water pipe condition using bayesian inference. *Journal of Zhejiang University SCIENCE A, 11*(7), 495-504.

Wang, Y., Zayed, T., & Moselhi, O. (2009). Prediction models for annual break rates of water mains. *Journal of Performance of Constructed Facilities, 23*(1), 47-54.

Xu, Q., Chen, Q., Li, W., & Ma, J. (2011). Pipe break prediction based on evolutionary data-driven methods with brief recorded data. *Reliability Engineering & System Safety, 96*(8), 942-948.

Appendix A



Biagiof21t000000MOGA_Xib0s0LSN [Compatibility Mode] - Excel

Seyed Farzad Karimian

Model predictions

Data ID	Model_1	Model_2	Model_3	Model_4	Model_5	Model_6	Model_7	Model_8	Model_9
1	0.043412	0.025395	-0.28422	-0.13626	0.487679	0.401614	0.008662	0.292827	0.344111
2	0.000181	0.000106	-0.01988	-0.01326	0.030093	0.130771	-0.00491	0.06975	0.030495
3	2E-06	4.69E-06	-0.006	-0.00395	0.73503	0.963332	-0.00142	0.304627	0.371943
4	0.002349	0.001374	-0.07745	-0.04811	0.249463	0.327382	-0.014	0.181187	0.177626
5	0.299788	0.077943	-0.1834	-0.05642	0.137399	0.129582	0.056592	0.217436	0.19821
6	0.198468	0.0516	-0.18385	-0.07749	0.001109	0.042343	0.024047	0.129529	0.093333
7	3.01E-05	4.89E-05	-0.02015	-0.01297	0.46212	0.683575	-0.00433	0.241233	0.266638
8	323.1478	336.0654	340.2811	337.1223	337.3541	337.1699	336.7761	336.7246	336.7577
9	0.014899	0.02421	-0.51544	-0.23174	0.058414	0.310891	0.034222	0.233787	0.232626
10	70.00581	72.80424	89.76265	95.95576	95.66017	95.73024	93.67668	93.55493	93.54822
11	0.099719	0.025926	-0.16316	-0.0855	0.039564	0.065884	-0.00458	0.124848	0.094384
12	0.339798	0.198777	-0.40504	0.019201	0.347576	0.399037	0.343575	0.556129	0.563235
13	21.22804	22.07661	28.90201	34.27923	33.87765	34.04589	34.38448	34.32904	34.28494
14	0.000389	2.53E-05	-0.00336	-0.00261	0.44375	-0.00267	-0.00134	0.236506	0.25712
15	23.05821	13.48872	17.65723	19.29113	19.41639	19.47777	17.89946	17.98724	17.97643
16	0.033273	0.019464	-0.25768	-0.1294	0.159619	0.23759	-0.00103	0.194168	0.190607
17	8.45E-05	0.000198	-0.05206	-0.03186	0.295751	0.61221	-0.00891	0.195489	0.196778
18	2.3083	1.350322	0.936582	1.927015	2.017935	2.129044	2.339486	2.472879	2.448554
19	0.529821	1.239751	-1.34784	1.88331	1.76164	2.054073	3.961431	4.072815	4.044847
20	0.001643	0.000427	-0.0287	-0.01964	0.147925	0.158213	-0.00773	0.138415	0.113429
21	1.197893	0.311444	0.03402	0.243517	0.634259	0.502985	0.357985	0.582867	0.597209
22	0.082976	0.021573	-0.15544	-0.08427	0.099912	0.099577	-0.00869	0.146843	0.125321
23	5.615273	5.839739	6.576734	9.978262	9.764099	9.937236	11.05438	11.09334	11.05303
24	0.160133	0.093675	-0.4037	-0.11156	-0.02003	0.099053	0.138386	0.259215	0.226598
25	1.195786	0.699518	0.07504	0.831524	1.113999	1.171215	1.256606	1.45718	1.461206
26	7.672206	4.48813	5.422027	6.886144	7.091947	7.145725	6.897663	7.058225	7.058806
27	27.41598	28.51191	37.22839	42.94903	42.71642	42.85074	42.66135	42.65141	42.64027
28	2.256474	0.586669	0.397607	0.640277	0.953978	0.869556	0.700993	0.901514	0.903491
29	7.2E-05	4.21E-05	-0.01189	-0.00806	0.020893	0.109673	-0.00311	0.057831	0.019251
30	0.000829	0.000216	-0.02011	-0.01394	0.31487	0.235369	-0.00568	0.198675	0.199959
31	0.000241	0.000392	-0.06471	-0.03949	0.309651	0.557473	-0.01088	0.200247	0.205179
32	0.142892	0.037151	-0.17618	-0.08441	0.127226	0.111703	0.007283	0.174208	0.157386

Models | Y_EPR | Graphs | Train_data | Test_data | EPR-Settings | Y_EPR_test | Y_rec | Y_V_rec

READY

Figure A - 1 Y_EPR Sheet of Excel Result File

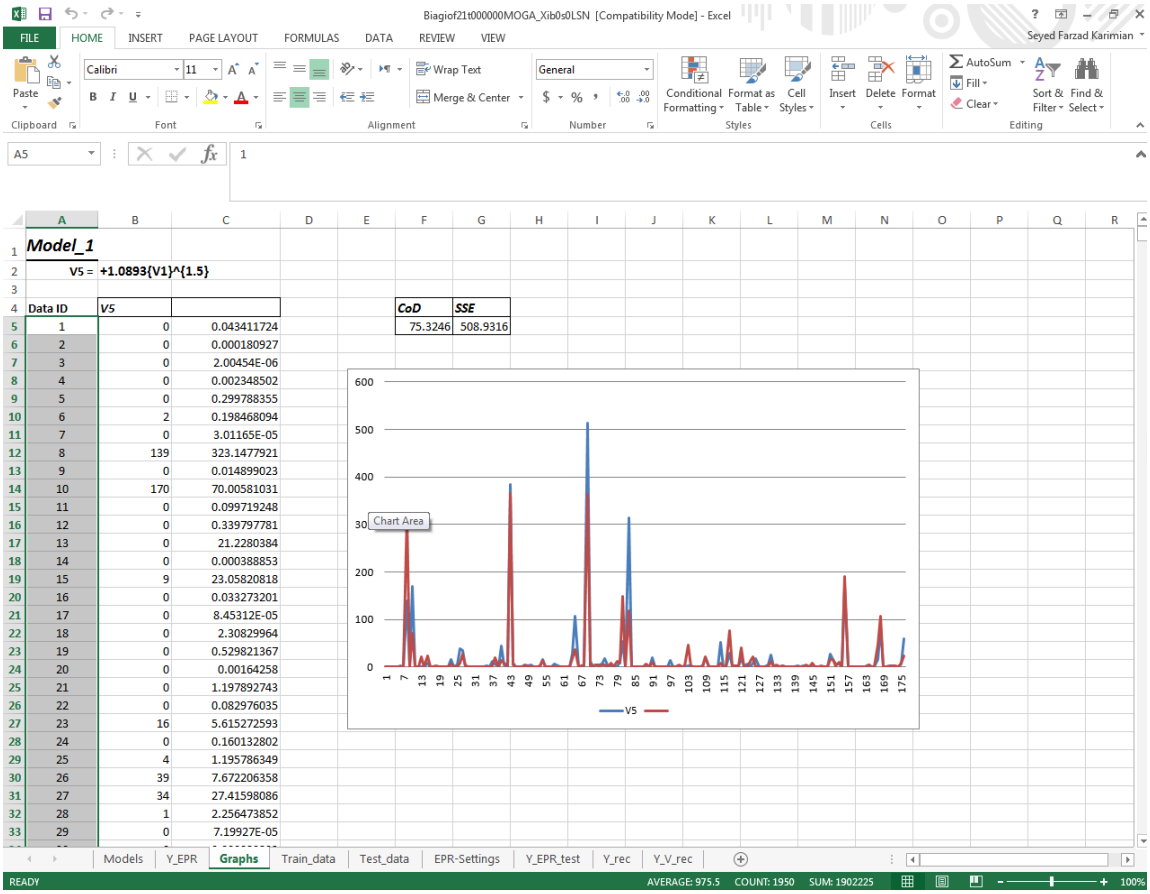
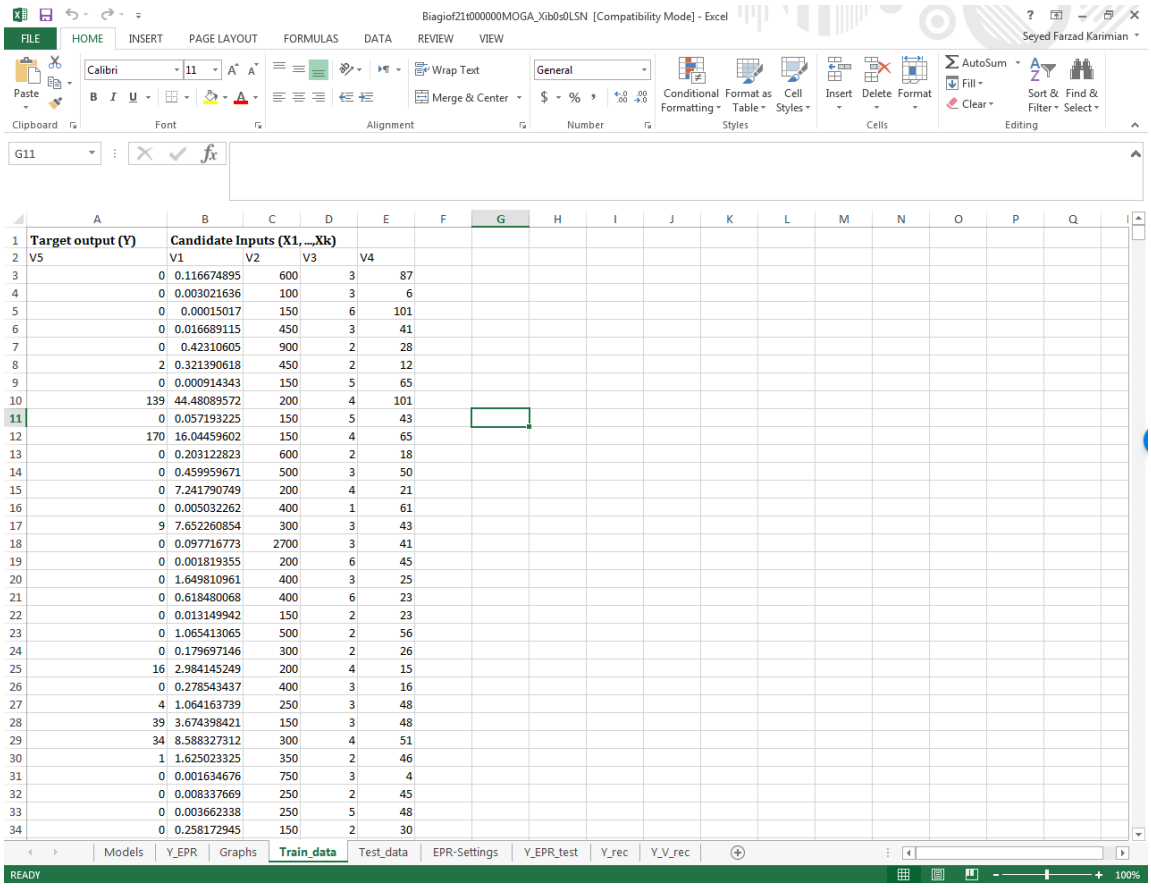


Figure A - 2 Graphs Sheet of Excel Result File



	A	B	C	D	E
1	**Target output (Y)**	**Candidate Inputs (X1, ...,Xk)**			
2	V5	V1	V2	V3	V4
3	0	0.116674895	600	3	87
4	0	0.003021636	100	3	6
5	0	0.00015017	150	6	101
6	0	0.016689115	450	3	41
7	0	0.42310605	900	2	28
8	2	0.321390618	450	2	12
9	0	0.000914343	150	5	65
10	139	44.48089572	200	4	101
11	0	0.057193225	150	5	43
12	170	16.04459602	150	4	65
13	0	0.203122823	600	2	18
14	0	0.459959671	500	3	50
15	0	7.241790749	200	4	21
16	0	0.005032262	400	1	61
17	9	7.652260854	300	3	43
18	0	0.097716773	2700	3	41
19	0	0.001819355	200	6	45
20	0	1.649810961	400	3	25
21	0	0.618480068	400	6	23
22	0	0.013149942	150	2	23
23	0	1.065413065	500	2	56
24	0	0.179697146	300	2	26
25	16	2.984145249	200	4	15
26	0	0.278543437	400	3	16
27	4	1.064163739	250	3	48
28	39	3.674398421	150	3	48
29	34	8.588327312	300	4	51
30	1	1.625023325	350	2	46
31	0	0.001634676	750	3	4
32	0	0.008337669	250	2	45
33	0	0.003662338	250	5	48
34	0	0.258172945	150	2	30

Figure A - 3 Train Data Sheet of Excel Result File

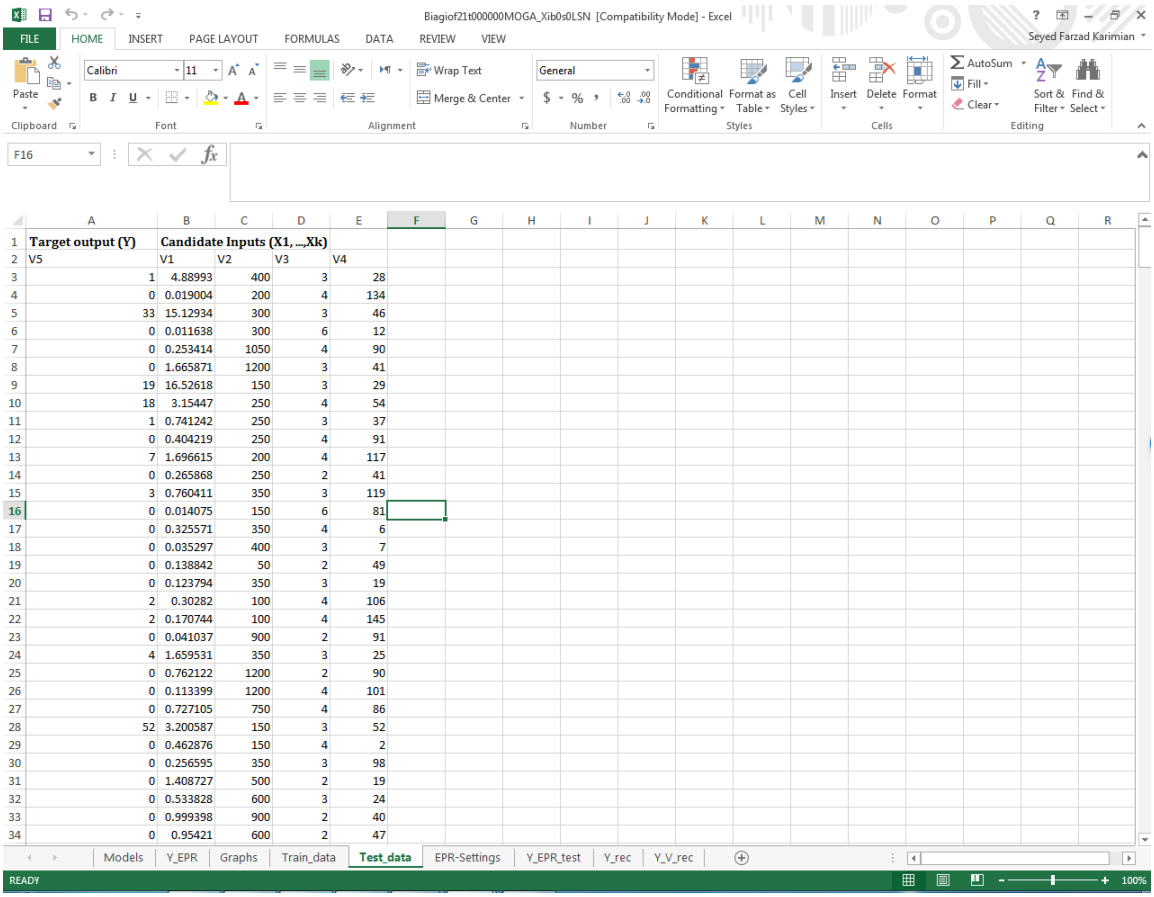


Figure A - 4 Test Data Sheet of Excel Result File

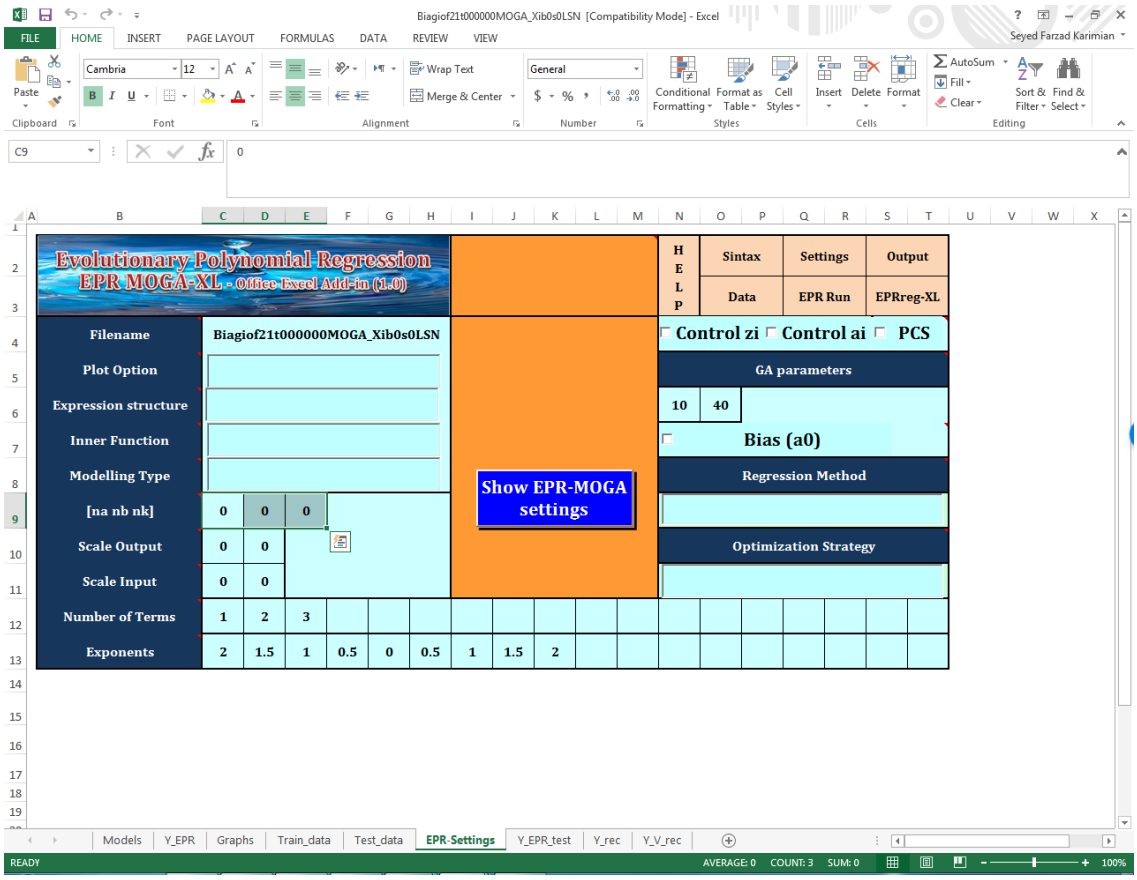
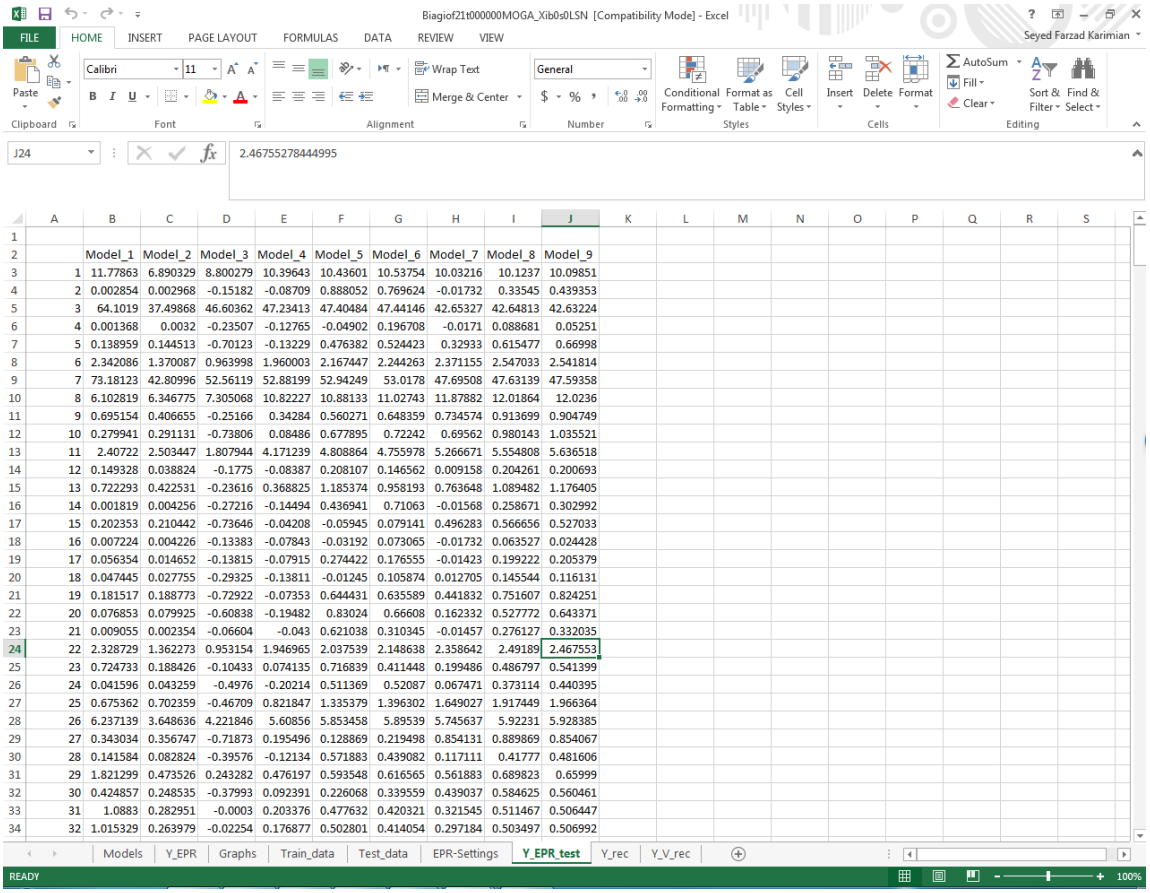


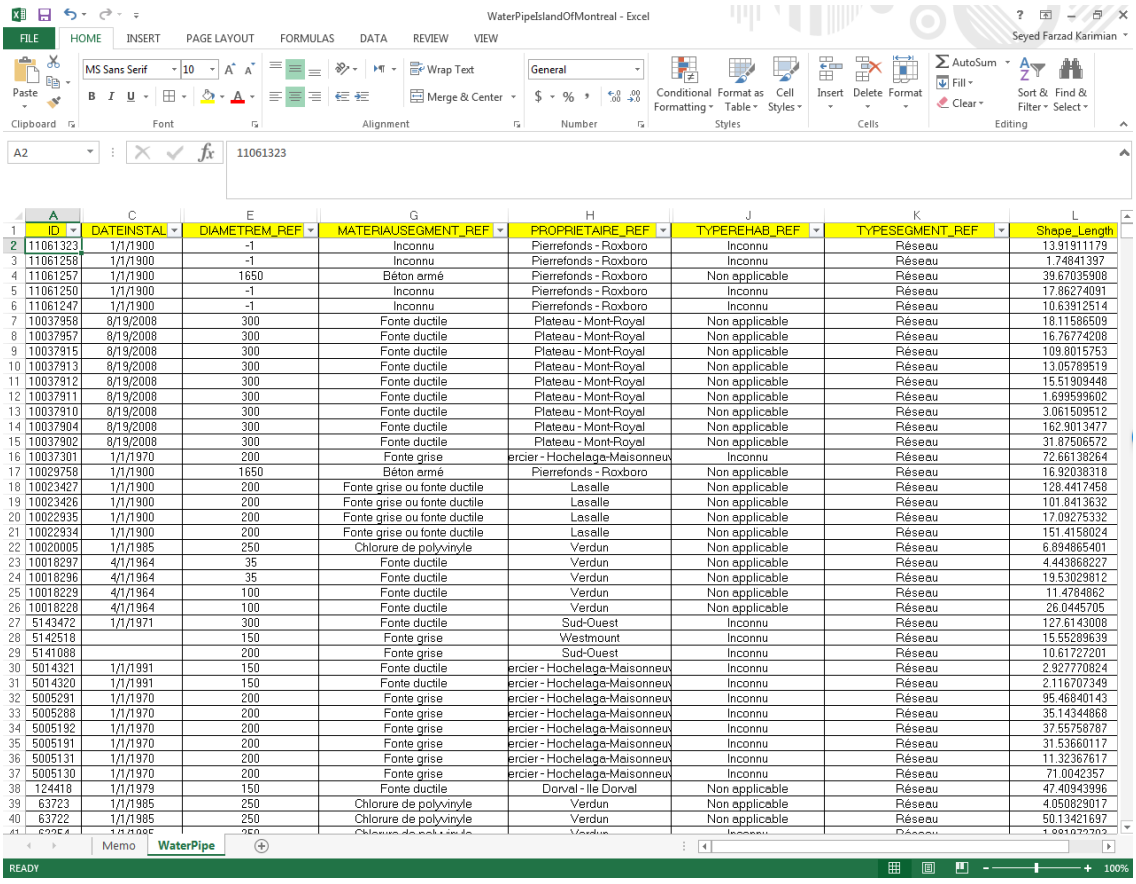
Figure A - 5 EPR Setting Sheet of Excel Result File



	A	B	C	D	E	F	G	H	I	J
1										
2		Model_1	Model_2	Model_3	Model_4	Model_5	Model_6	Model_7	Model_8	Model_9
3	1	11.77863	6.890329	8.800279	10.39643	10.43601	10.53754	10.03216	10.1237	10.09851
4	2	0.002854	0.002968	-0.15182	-0.08709	0.888052	0.769624	-0.01732	0.33545	0.439353
5	3	64.1019	37.49868	46.60362	47.23413	47.40484	47.44146	42.65327	42.64813	42.63224
6	4	0.001368	0.0032	-0.23507	-0.12765	-0.04902	0.196708	-0.0171	0.088681	0.05251
7	5	0.138959	0.144513	-0.70123	-0.13229	0.476382	0.524423	0.32933	0.615477	0.66998
8	6	2.342086	1.370087	0.963998	1.960003	2.167447	2.244263	2.371155	2.547033	2.541814
9	7	73.18123	42.80996	52.56119	52.88199	52.94249	53.0178	47.69508	47.63139	47.59358
10	8	6.102819	6.346775	7.305068	10.82227	10.88133	11.02743	11.87882	12.01864	12.0236
11	9	0.695154	0.406655	-0.25166	0.34284	0.560271	0.648359	0.734574	0.913699	0.904749
12	10	0.279941	0.291131	-0.73806	0.08486	0.677895	0.72242	0.69562	0.980143	1.035521
13	11	2.40722	2.503447	1.807944	4.171239	4.808864	4.755978	5.266671	5.554808	5.636518
14	12	0.149328	0.038824	-0.1775	-0.08387	0.208107	0.146562	0.009158	0.204261	0.200693
15	13	0.722293	0.422531	-0.23616	0.368825	1.185374	0.958193	0.763648	1.089482	1.176405
16	14	0.001819	0.004256	-0.27216	-0.14494	0.436941	0.71063	-0.01568	0.258671	0.302992
17	15	0.202353	0.210442	-0.73646	-0.04208	-0.05945	0.079141	0.496283	0.566656	0.527033
18	16	0.007224	0.004226	-0.13383	-0.07843	-0.03192	0.073065	-0.01732	0.063527	0.024428
19	17	0.056354	0.014652	-0.13815	-0.07915	0.274422	0.176555	-0.01423	0.199222	0.205379
20	18	0.047445	0.027755	-0.29325	-0.13811	-0.01245	0.105874	0.012705	0.145544	0.116131
21	19	0.181517	0.188773	-0.72922	-0.07353	0.644431	0.635589	0.441832	0.751607	0.824251
22	20	0.076853	0.079925	-0.60838	-0.19482	0.83024	0.66608	0.162332	0.527772	0.643371
23	21	0.009055	0.002354	-0.06604	-0.043	0.621038	0.310345	-0.01457	0.276127	0.332035
24	22	2.328729	1.362273	0.953154	1.946965	2.037539	2.148638	2.358642	2.49189	2.467553
25	23	0.724733	0.188426	-0.10433	0.074135	0.716839	0.411448	0.199486	0.486797	0.541399
26	24	0.041596	0.043259	-0.4976	-0.20214	0.511369	0.52087	0.067471	0.373114	0.440395
27	25	0.675362	0.702359	-0.46709	0.821847	1.335379	1.396302	1.649027	1.917449	1.966364
28	26	6.237139	3.648636	4.221846	5.60856	5.853458	5.89539	5.745637	5.92231	5.928385
29	27	0.343034	0.356747	-0.71873	0.195496	0.128869	0.219498	0.854131	0.889869	0.854067
30	28	0.141584	0.082824	-0.39576	-0.12134	0.571883	0.439082	0.117111	0.41777	0.481606
31	29	1.821299	0.473526	0.243282	0.476197	0.593548	0.616565	0.561883	0.689823	0.65999
32	30	0.424857	0.248535	-0.37993	0.092391	0.226068	0.339559	0.439037	0.584625	0.560461
33	31	1.0883	0.282951	-0.0003	0.203376	0.477632	0.420321	0.321545	0.511467	0.506447
34	32	1.015329	0.263979	-0.02254	0.176877	0.502801	0.414054	0.297184	0.503497	0.506992

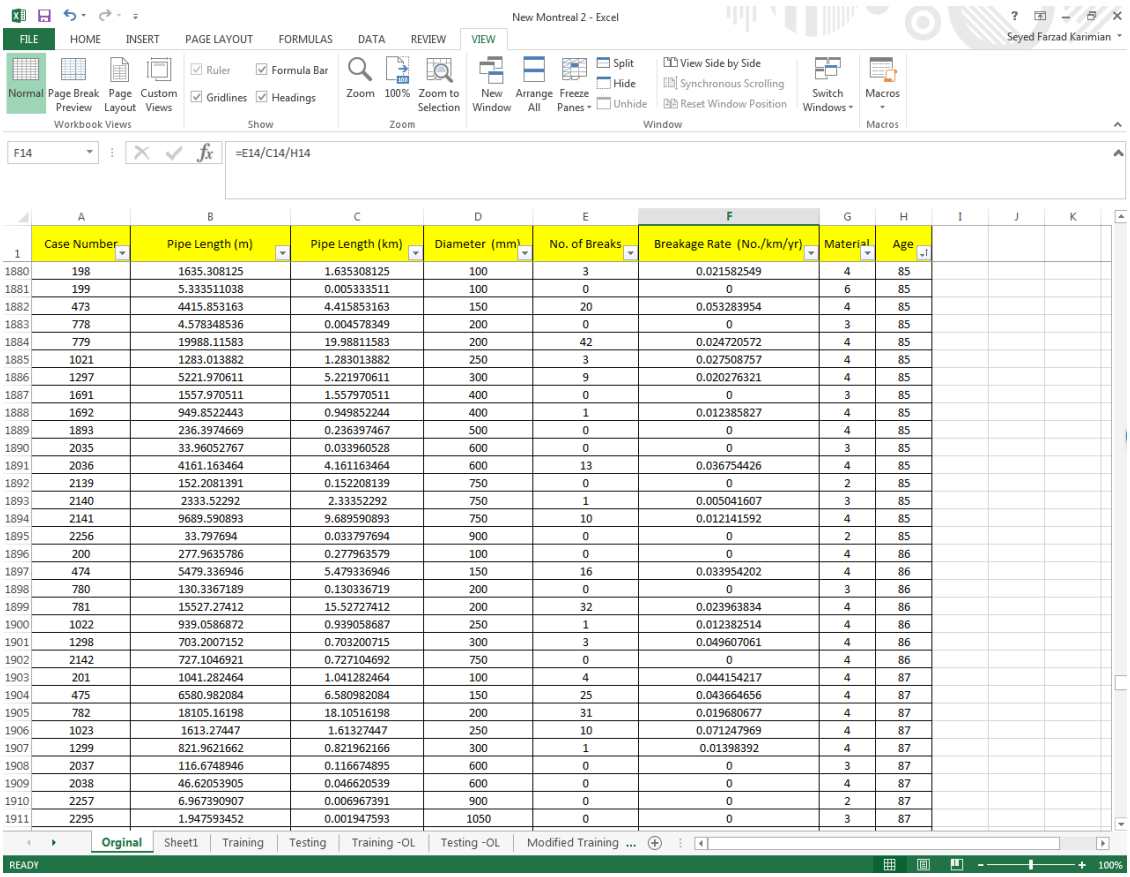
Figure A - 6 Y-EPR Test Sheet of Excel Result File

Appendix B



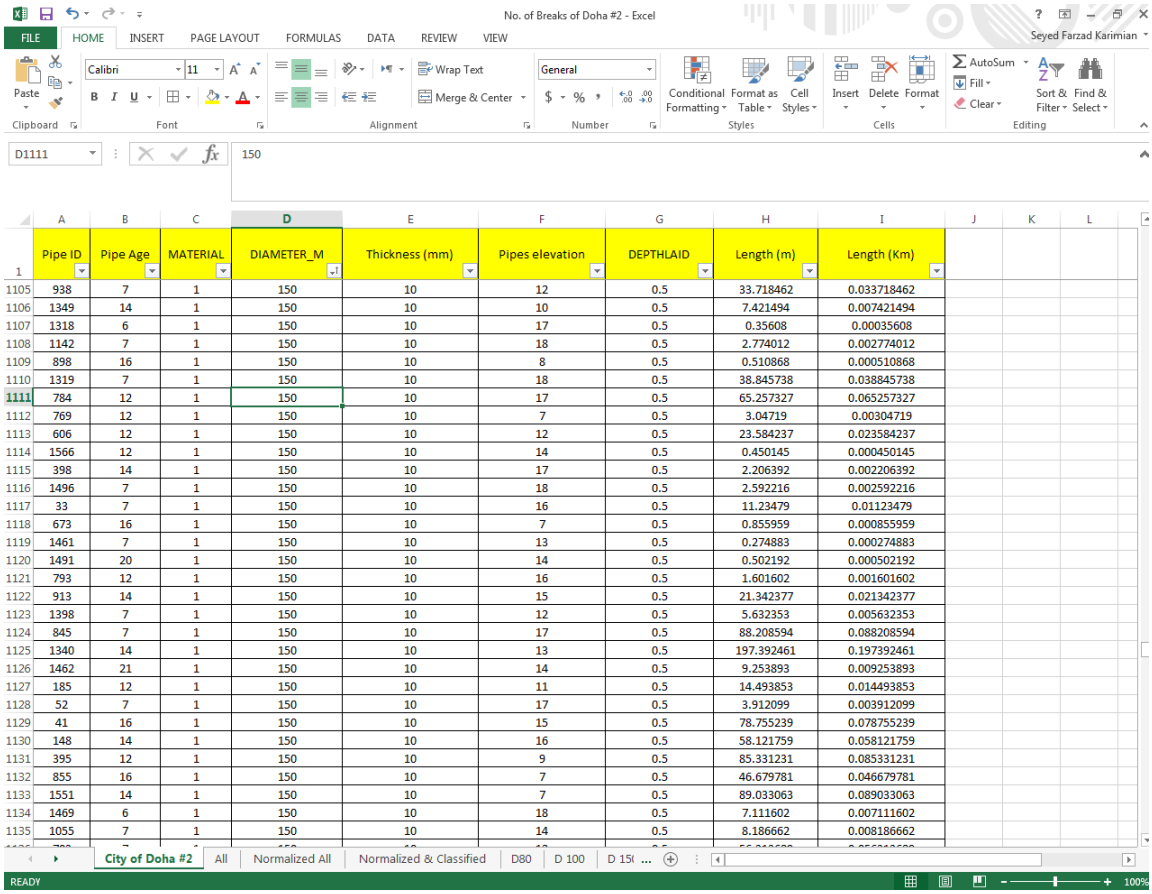
ID	DATEINSTAL	DIAMETREM_REF	MATERIAUSEGMENT_REF	PROPRIETAIRE_REF	TYPEREHAB_REF	TYPESEGMENT_REF	Shape_Length
11061323	1/1/1900	-1	Inconnu	Pierrefonds - Roxboro	Inconnu	Réseau	13.91911179
11061258	1/1/1900	-1	Inconnu	Pierrefonds - Roxboro	Inconnu	Réseau	1.74841397
11061257	1/1/1900	1650	Béton armé	Pierrefonds - Roxboro	Non applicable	Réseau	39.67035908
11061250	1/1/1900	-1	Inconnu	Pierrefonds - Roxboro	Inconnu	Réseau	17.86274091
11061247	1/1/1900	-1	Inconnu	Pierrefonds - Roxboro	Inconnu	Réseau	10.63912514
10037958	8/19/2008	300	Fonte ductile	Plateau - Mont-Royal	Non applicable	Réseau	18.11586509
10037957	8/19/2008	300	Fonte ductile	Plateau - Mont-Royal	Non applicable	Réseau	16.76774208
10037915	8/19/2008	300	Fonte ductile	Plateau - Mont-Royal	Non applicable	Réseau	109.8015753
10037913	8/19/2008	300	Fonte ductile	Plateau - Mont-Royal	Non applicable	Réseau	13.05789519
10037912	8/19/2008	300	Fonte ductile	Plateau - Mont-Royal	Non applicable	Réseau	15.51909448
10037911	8/19/2008	300	Fonte ductile	Plateau - Mont-Royal	Non applicable	Réseau	1.699599602
10037910	8/19/2008	300	Fonte ductile	Plateau - Mont-Royal	Non applicable	Réseau	3.061509512
10037904	8/19/2008	300	Fonte ductile	Plateau - Mont-Royal	Non applicable	Réseau	162.9013477
10037902	8/19/2008	300	Fonte ductile	Plateau - Mont-Royal	Non applicable	Réseau	31.87506572
10037301	1/1/1970	200	Fonte grise	ercier - Hochelaga-Maisonneuv	Inconnu	Réseau	72.66138264
10029758	1/1/1900	1650	Béton armé	Pierrefonds - Roxboro	Non applicable	Réseau	16.92038318
10023427	1/1/1900	200	Fonte grise ou fonte ductile	Lasalle	Non applicable	Réseau	128.4417458
10023426	1/1/1900	200	Fonte grise ou fonte ductile	Lasalle	Non applicable	Réseau	101.8413632
10022935	1/1/1900	200	Fonte grise ou fonte ductile	Lasalle	Non applicable	Réseau	17.09275332
10022934	1/1/1900	200	Fonte grise ou fonte ductile	Lasalle	Non applicable	Réseau	151.4158024
10020005	1/1/1985	250	Chlorure de polyvinyle	Verdun	Non applicable	Réseau	6.894865401
10018297	4/1/1964	35	Fonte ductile	Verdun	Non applicable	Réseau	4.443868227
10018296	4/1/1964	35	Fonte ductile	Verdun	Non applicable	Réseau	19.53029812
10018229	4/1/1964	100	Fonte ductile	Verdun	Non applicable	Réseau	11.4784862
10018228	4/1/1964	100	Fonte ductile	Verdun	Non applicable	Réseau	26.0445705
5143472	1/1/1971	300	Fonte ductile	Sud-Ouest	Inconnu	Réseau	127.6143008
5142518		150	Fonte grise	Westmount	Inconnu	Réseau	15.55289639
5141088		200	Fonte grise	Sud-Ouest	Inconnu	Réseau	10.61727201
5014321	1/1/1991	150	Fonte ductile	ercier - Hochelaga-Maisonneuv	Inconnu	Réseau	2.927770824
5014320	1/1/1991	150	Fonte ductile	ercier - Hochelaga-Maisonneuv	Inconnu	Réseau	2.116707349
5005291	1/1/1970	200	Fonte grise	ercier - Hochelaga-Maisonneuv	Inconnu	Réseau	95.46840143
5005288	1/1/1970	200	Fonte grise	ercier - Hochelaga-Maisonneuv	Inconnu	Réseau	35.14344868
5005192	1/1/1970	200	Fonte grise	ercier - Hochelaga-Maisonneuv	Inconnu	Réseau	37.55758787
5005191	1/1/1970	200	Fonte grise	ercier - Hochelaga-Maisonneuv	Inconnu	Réseau	31.53660117
5005131	1/1/1970	200	Fonte grise	ercier - Hochelaga-Maisonneuv	Inconnu	Réseau	11.32367617
5005130	1/1/1970	200	Fonte grise	ercier - Hochelaga-Maisonneuv	Inconnu	Réseau	71.0042357
124418	1/1/1979	150	Fonte ductile	Dorval - Ile Dorval	Non applicable	Réseau	47.40943996
63723	1/1/1985	250	Chlorure de polyvinyle	Verdun	Non applicable	Réseau	4.050829017
63722	1/1/1985	250	Chlorure de polyvinyle	Verdun	Non applicable	Réseau	50.13421697

Figure B - 1 Original dataset of Montreal



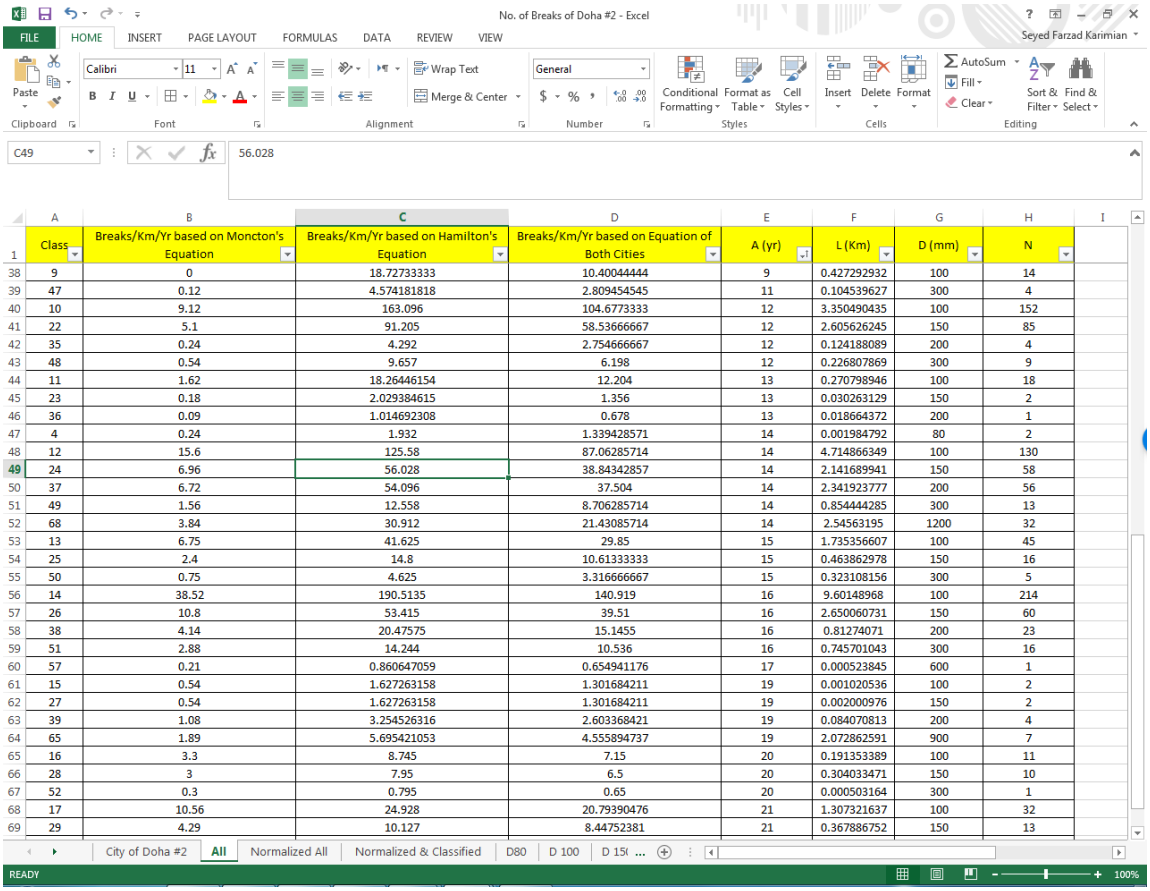
	Case Number	Pipe Length (m)	Pipe Length (km)	Diameter (mm)	No. of Breaks	Breakage Rate (No./km/yr)	Material	Age
1880	198	1635.308125	1.635308125	100	3	0.021582549	4	85
1881	199	5.333511038	0.005333511	100	0	0	6	85
1882	473	4415.853163	4.415853163	150	20	0.053283954	4	85
1883	778	4.578348536	0.004578349	200	0	0	3	85
1884	779	19988.11583	19.98811583	200	42	0.024720572	4	85
1885	1021	1283.013882	1.283013882	250	3	0.027508757	4	85
1886	1297	5221.970611	5.221970611	300	9	0.020276321	4	85
1887	1691	1557.970511	1.557970511	400	0	0	3	85
1888	1692	949.8522443	0.949852244	400	1	0.012385827	4	85
1889	1893	236.3974669	0.236397467	500	0	0	4	85
1890	2035	33.96052767	0.033960528	600	0	0	3	85
1891	2036	4161.163464	4.161163464	600	13	0.036754426	4	85
1892	2139	152.2081391	0.152208139	750	0	0	2	85
1893	2140	2333.52292	2.33352292	750	1	0.005041607	3	85
1894	2141	9689.590893	9.689590893	750	10	0.012141592	4	85
1895	2256	33.797694	0.033797694	900	0	0	2	85
1896	200	277.9635786	0.277963579	100	0	0	4	86
1897	474	5479.336946	5.479336946	150	16	0.033954202	4	86
1898	780	130.3367189	0.130336719	200	0	0	3	86
1899	781	15527.27412	15.52727412	200	32	0.023963834	4	86
1900	1022	939.0586872	0.939058687	250	1	0.012382514	4	86
1901	1298	703.2007152	0.703200715	300	3	0.049607061	4	86
1902	2142	727.1046921	0.727104692	750	0	0	4	86
1903	201	1041.282464	1.041282464	100	4	0.044154217	4	87
1904	475	6580.982084	6.580982084	150	25	0.043664656	4	87
1905	782	18105.16198	18.10516198	200	31	0.019680677	4	87
1906	1023	1613.27447	1.61327447	250	10	0.071247969	4	87
1907	1299	821.9621662	0.821962166	300	1	0.01398392	4	87
1908	2037	116.6748946	0.116674895	600	0	0	3	87
1909	2038	46.62053905	0.046620539	600	0	0	4	87
1910	2257	6.967390907	0.006967391	900	0	0	2	87
1911	2295	1.947593452	0.001947593	1050	0	0	3	87

Figure B - 2 Dataset of Montreal after Classification



	A	B	C	D	E	F	G	H	I
1	Pipe ID	Pipe Age	MATERIAL	DIAMETER_M	Thickness (mm)	Pipes elevation	DEPTHLAID	Length (m)	Length (Km)
1105	938	7	1	150	10	12	0.5	33.718462	0.033718462
1106	1349	14	1	150	10	10	0.5	7.421494	0.007421494
1107	1318	6	1	150	10	17	0.5	0.35608	0.00035608
1108	1142	7	1	150	10	18	0.5	2.774012	0.002774012
1109	898	16	1	150	10	8	0.5	0.510868	0.000510868
1110	1319	7	1	150	10	18	0.5	38.845738	0.038845738
1111	784	12	1	150	10	17	0.5	65.257327	0.065257327
1112	769	12	1	150	10	7	0.5	3.04719	0.00304719
1113	606	12	1	150	10	12	0.5	23.584237	0.023584237
1114	1566	12	1	150	10	14	0.5	0.450145	0.000450145
1115	398	14	1	150	10	17	0.5	2.206392	0.002206392
1116	1496	7	1	150	10	18	0.5	2.592216	0.002592216
1117	33	7	1	150	10	16	0.5	11.23479	0.01123479
1118	673	16	1	150	10	7	0.5	0.855959	0.000855959
1119	1461	7	1	150	10	13	0.5	0.274883	0.000274883
1120	1491	20	1	150	10	14	0.5	0.502192	0.000502192
1121	793	12	1	150	10	16	0.5	1.601602	0.001601602
1122	913	14	1	150	10	15	0.5	21.342377	0.021342377
1123	1398	7	1	150	10	12	0.5	5.632353	0.005632353
1124	845	7	1	150	10	17	0.5	88.208594	0.088208594
1125	1340	14	1	150	10	13	0.5	197.392461	0.197392461
1126	1462	21	1	150	10	14	0.5	9.253893	0.009253893
1127	185	12	1	150	10	11	0.5	14.493853	0.014493853
1128	52	7	1	150	10	17	0.5	3.912099	0.003912099
1129	41	16	1	150	10	15	0.5	78.755239	0.078755239
1130	148	14	1	150	10	16	0.5	58.121759	0.058121759
1131	395	12	1	150	10	9	0.5	85.331231	0.085331231
1132	855	16	1	150	10	7	0.5	46.679781	0.046679781
1133	1551	14	1	150	10	7	0.5	89.033063	0.089033063
1134	1469	6	1	150	10	18	0.5	7.111602	0.007111602
1135	1055	7	1	150	10	14	0.5	8.186662	0.008186662

Figure B - 3 Original Dataset of Doha



	A	B	C	D	E	F	G	H	I
1	Class	Breaks/Km/Yr based on Moncton's Equation	Breaks/Km/Yr based on Hamilton's Equation	Breaks/Km/Yr based on Equation of Both Cities	A (yr)	L (Km)	D (mm)	N	
38	9	0	18.72733333	10.40044444	9	0.427292932	100	14	
39	47	0.12	4.574181818	2.809454545	11	0.104539627	300	4	
40	10	9.12	163.096	104.6773333	12	3.350490435	100	152	
41	22	5.1	91.205	58.53666667	12	2.605626245	150	85	
42	35	0.24	4.292	2.754666667	12	0.124188089	200	4	
43	48	0.54	9.657	6.198	12	0.226807869	300	9	
44	11	1.62	18.26446154	12.204	13	0.270798946	100	18	
45	23	0.18	2.029384615	1.356	13	0.030263129	150	2	
46	36	0.09	1.014692308	0.678	13	0.018664372	200	1	
47	4	0.24	1.932	1.339428571	14	0.001984792	80	2	
48	12	15.6	125.58	87.06285714	14	4.714866349	100	130	
49	24	6.96	56.028	38.84342857	14	2.141689941	150	58	
50	37	6.72	54.096	37.504	14	2.341923777	200	56	
51	49	1.56	12.558	8.706285714	14	0.854444285	300	13	
52	68	3.84	30.912	21.43085714	14	2.54563195	1200	32	
53	13	6.75	41.625	29.85	15	1.735356607	100	45	
54	25	2.4	14.8	10.61333333	15	0.463862978	150	16	
55	50	0.75	4.625	3.316666667	15	0.323108156	300	5	
56	14	38.52	190.5135	140.919	16	9.60148968	100	214	
57	26	10.8	53.415	39.51	16	2.650060731	150	60	
58	38	4.14	20.47575	15.1455	16	0.81274071	200	23	
59	51	2.88	14.244	10.536	16	0.745701043	300	16	
60	57	0.21	0.860647059	0.654941176	17	0.000523845	600	1	
61	15	0.54	1.627263158	1.301684211	19	0.001020536	100	2	
62	27	0.54	1.627263158	1.301684211	19	0.002000976	150	2	
63	39	1.08	3.254526316	2.603368421	19	0.084070813	200	4	
64	65	1.89	5.695421053	4.555894737	19	2.072862591	900	7	
65	16	3.3	8.745	7.15	20	0.191353389	100	11	
66	28	3	7.95	6.5	20	0.304033471	150	10	
67	52	0.3	0.795	0.65	20	0.000503164	300	1	
68	17	10.56	24.928	20.79390476	21	1.307321637	100	32	
69	29	4.29	10.127	8.44752381	21	0.367886752	150	13	

Figure B - 4 Dataset of Doha after Classification

Appendix C

$$No.of\,Breaks = +3.4446e-005L^{1.5}$$

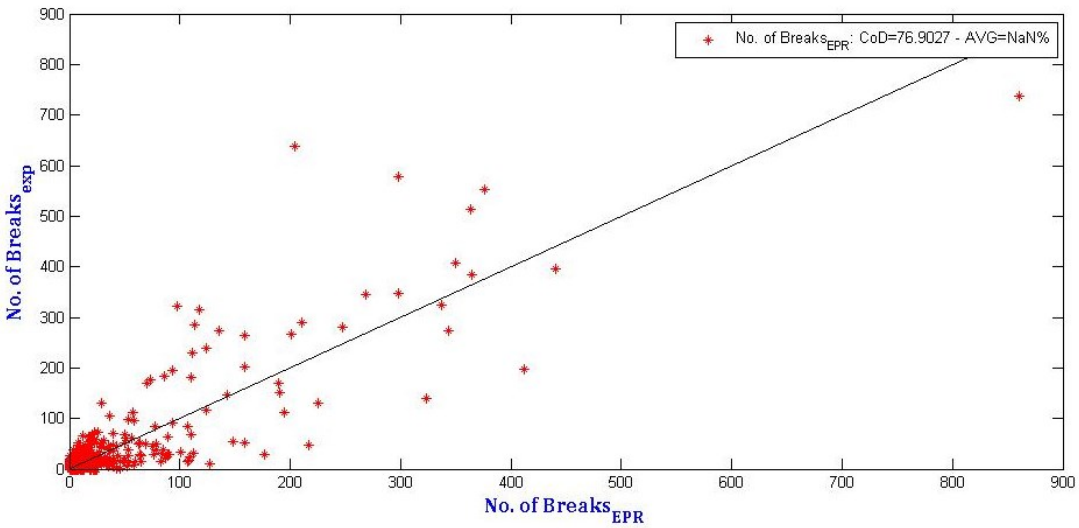


Figure C - 1 Scatter Plot of Model #1 for Training for Montreal Dataset

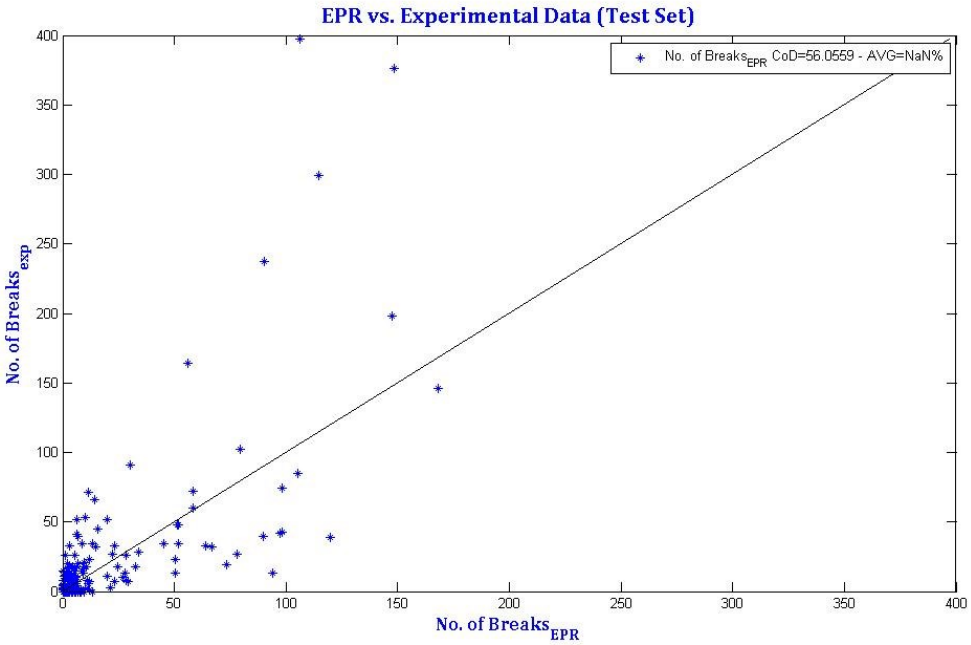


Figure C - 2 Scatter Plot of Model #1 for Testing for Montreal Dataset

$$No.of Breaks = +1.3197 \frac{L^{1.5}}{D^2}$$

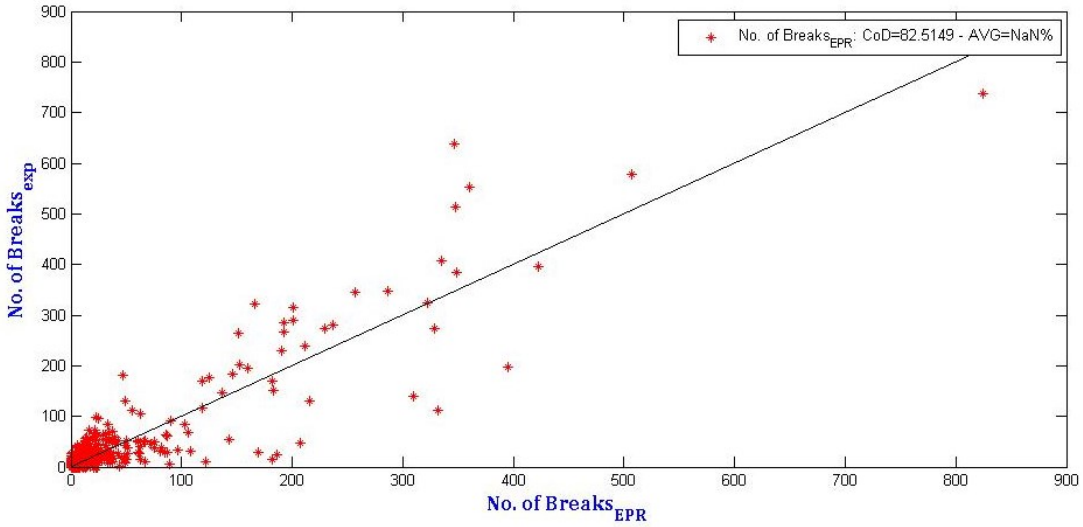


Figure C - 3 Scatter Plot of Model #2 for Training for Montreal Dataset

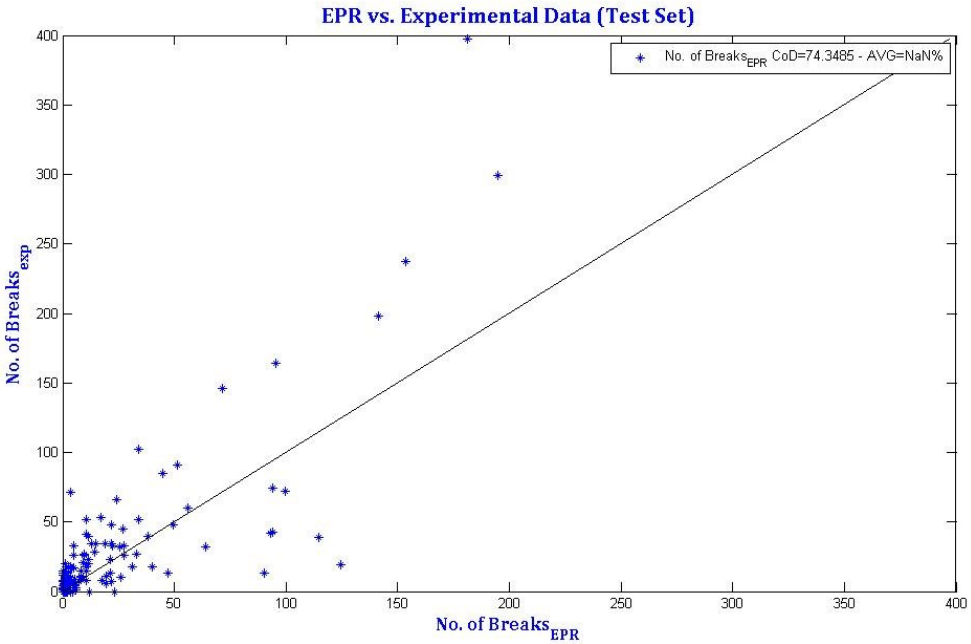


Figure C - 4 Scatter Plot of Model #2 for Testing for Montreal Dataset

$$No.of Breaks = +0.08546 \frac{L^{1.5} M^2}{D^2}$$

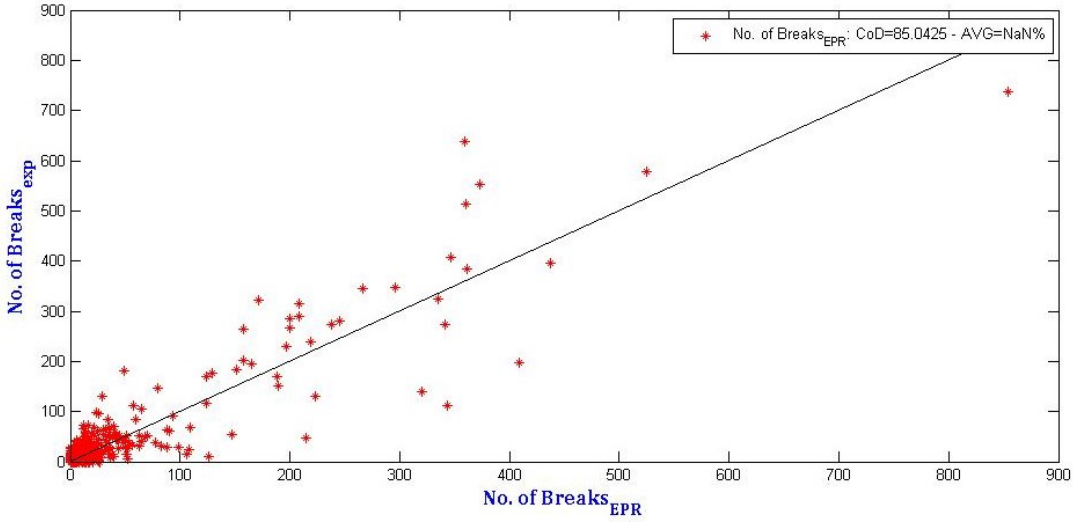


Figure C - 5 Scatter Plot of Model #3 for Training for Montreal Dataset

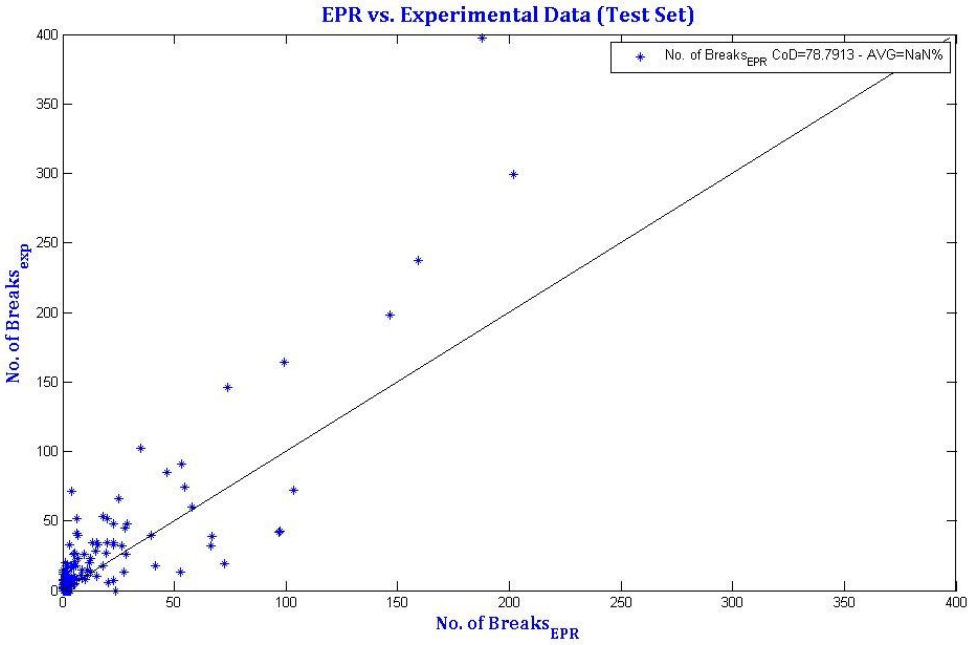


Figure C - 6 Scatter Plot of Model #3 for Testing for Montreal Dataset

$$No.of\,Breaks = +1.8835\frac{L^{1.5}}{D^{2}}\ln\left(\frac{M^{2}}{A^{0.5}}\right)$$

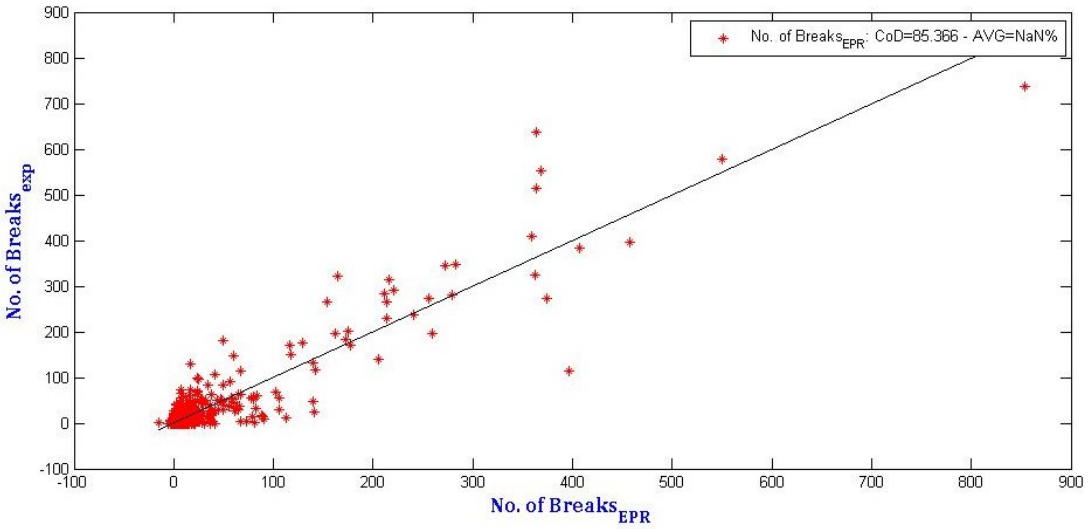


Figure C - 7 Scatter Plot of Model #4 for Training for Montreal Dataset

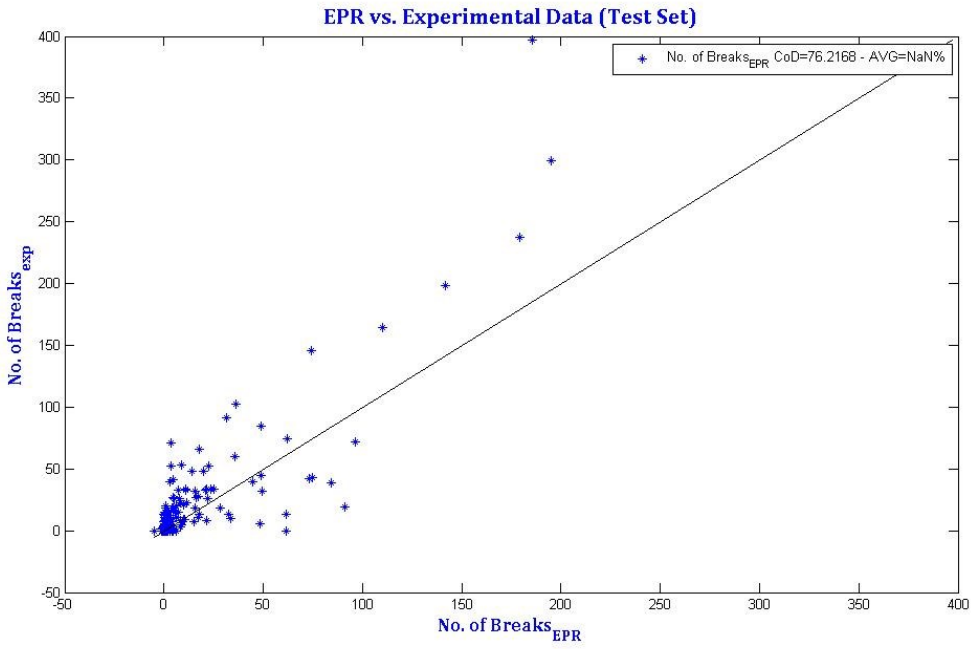


Figure C - 8 Scatter Plot of Model #4 for Testing for Montreal Dataset

$$No.of\,Breaks = +0.24999\frac{L^{1.5}A^{0.5}}{D^2}\ln\left(\frac{M^2}{A^{0.5}}\right)$$

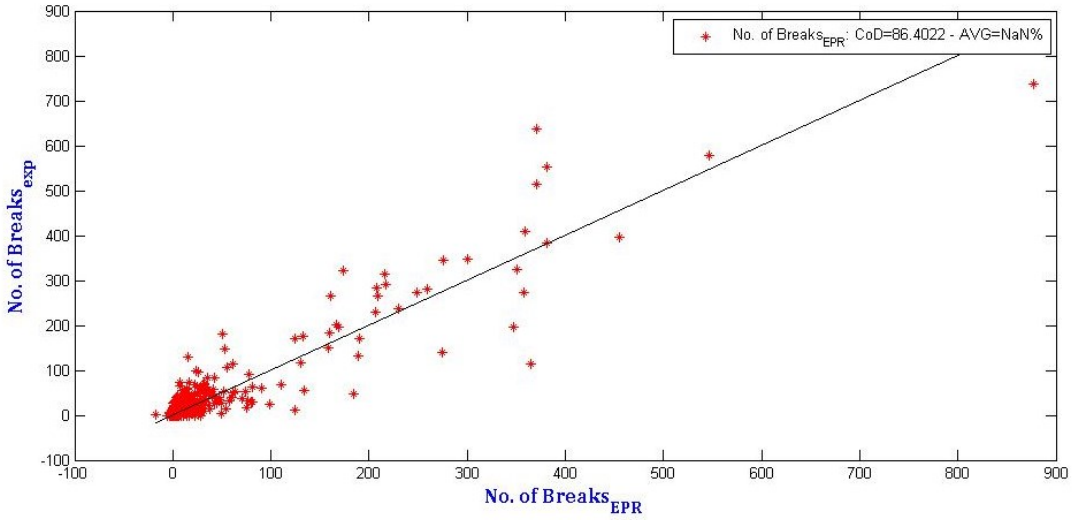


Figure C - 9 Scatter Plot of Model #5 for Training for Montreal Dataset

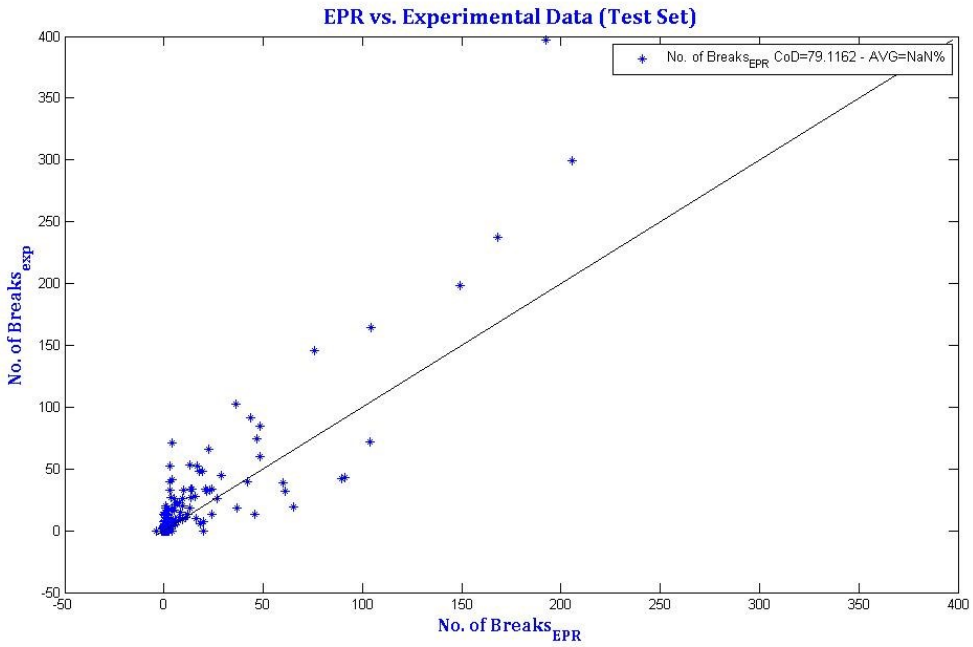


Figure C - 10 Scatter Plot of Model #5 for Testing for Montreal Dataset

$$No.of\,Breaks = +0.092319L^{0.5} + 0.23417\frac{L^{1.5}A^{0.5}}{D^2}\ln\left(\frac{M^2}{A^{0.5}}\right)$$

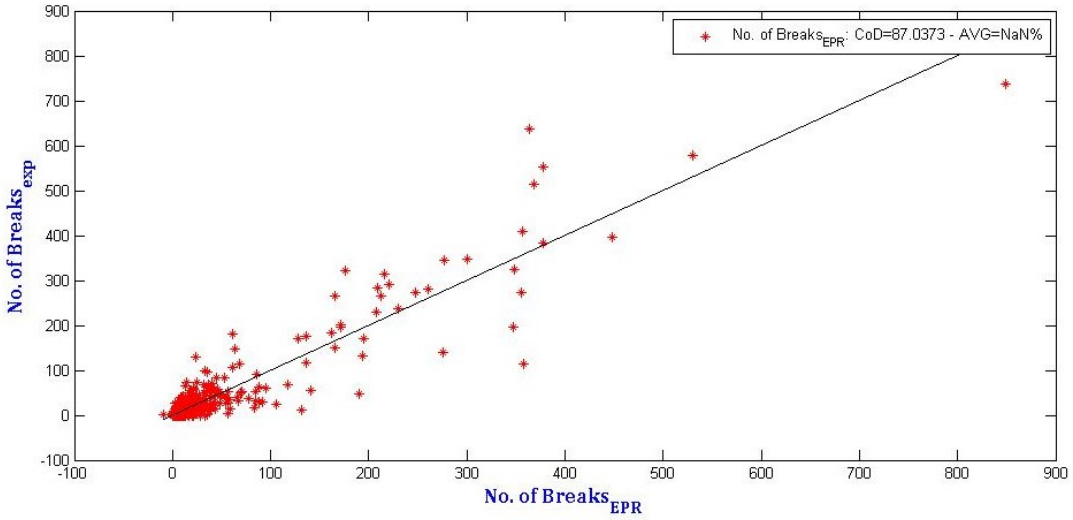


Figure C - 11Scatter Plot of Model #6 for Training for Montreal Dataset

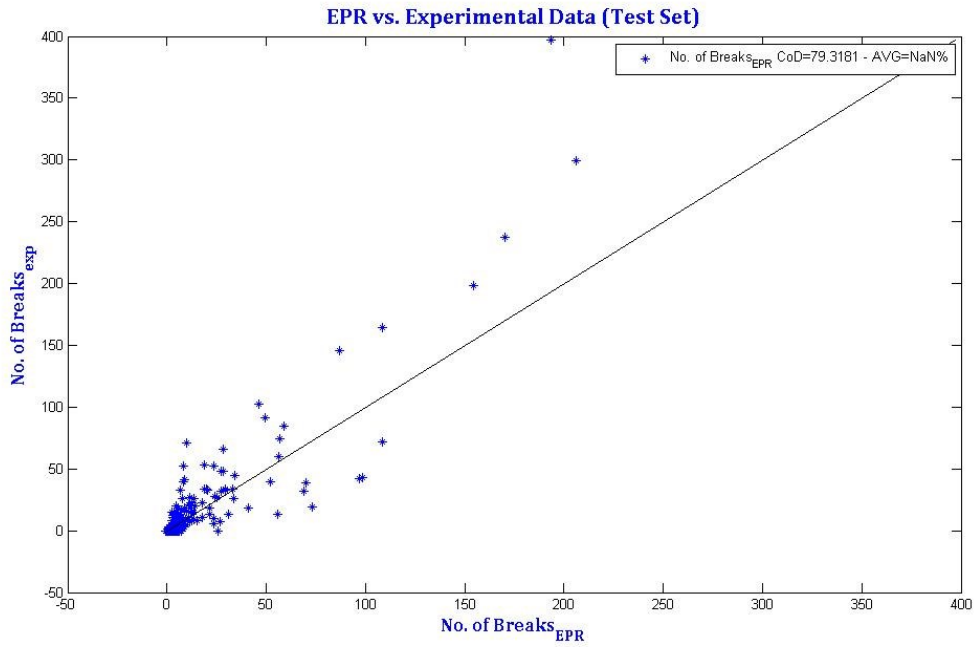


Figure C - 12 Scatter Plot of Model #6 for Testing for Montreal Dataset

$$No.of\,Breaks = +0.12036\frac{L^{1.5}M^2}{D^2} + 4.8297e-007\frac{L^2A^{1.5}}{D^2}\ln\left(\frac{1}{L}\right)$$

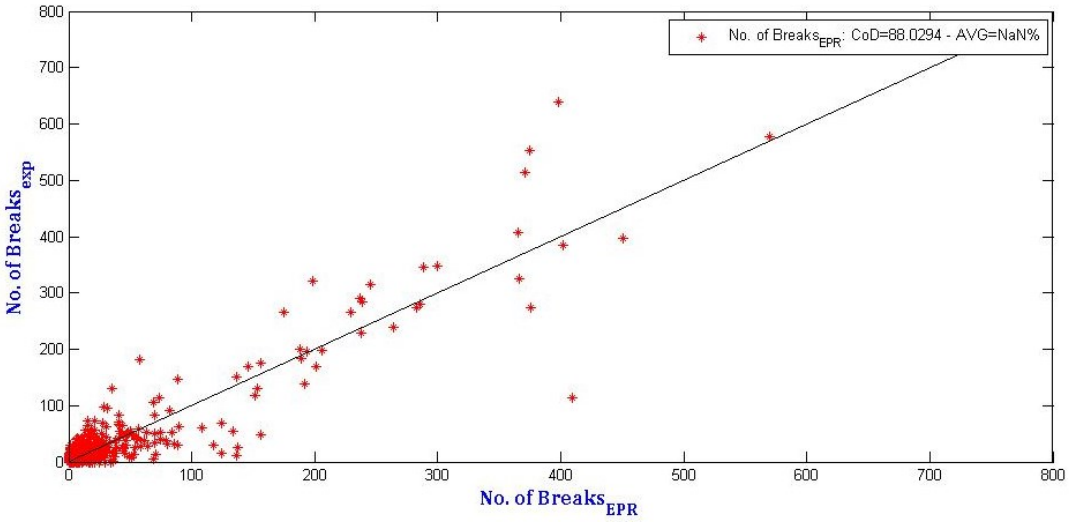


Figure C - 13 Scatter Plot of Model #7 for Training for Montreal Dataset

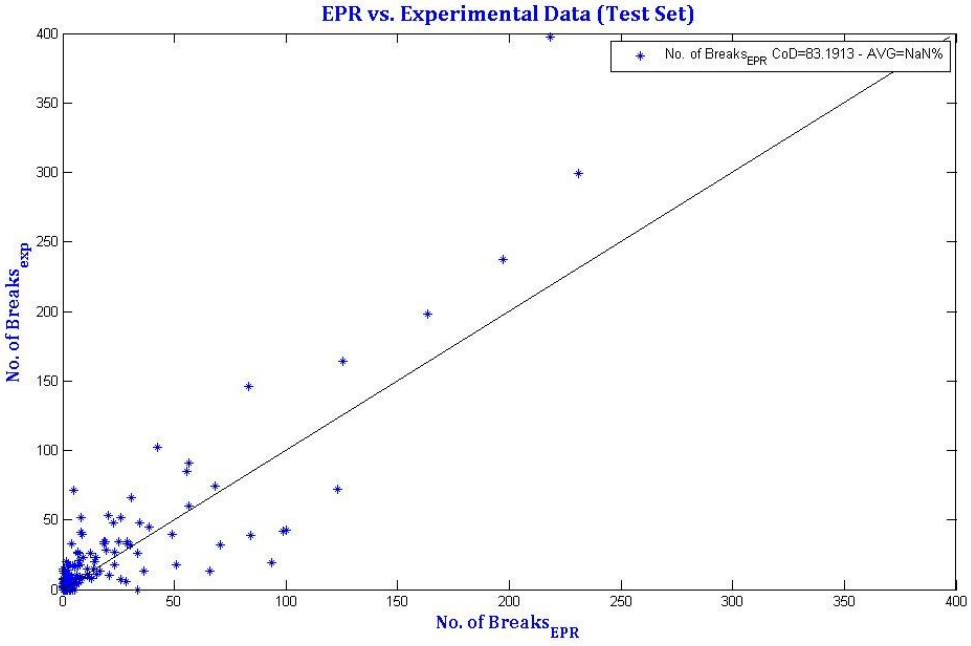


Figure C - 14 Scatter Plot of Model #7 for Testing for Montreal Dataset

$$No.of\,Breaks = +0.008929\frac{L^{1.5}A}{D^2}\ln\left(\frac{1}{L^{0.5}}\right) + 0.069455\frac{L^{1.5}M^{1.5}A^{0.5}}{D^2}$$

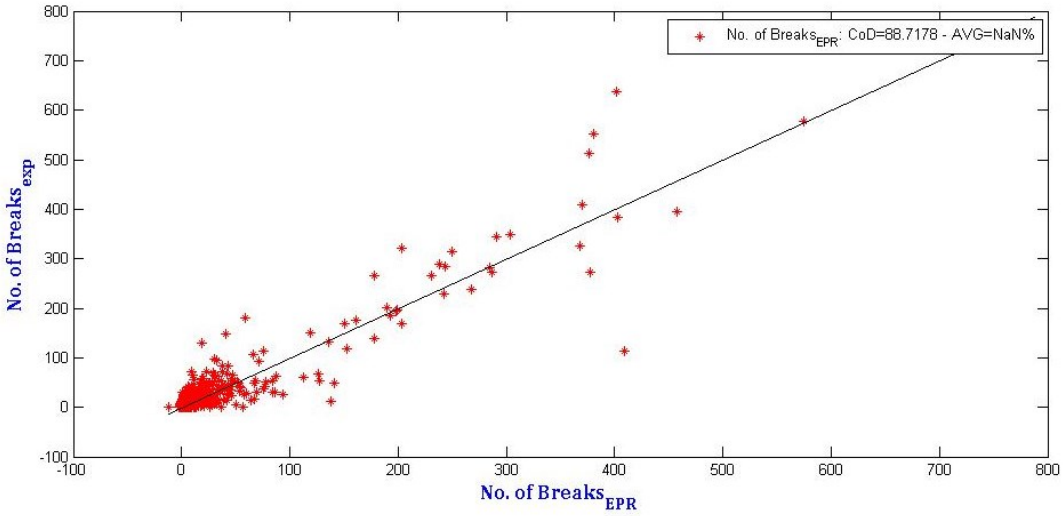


Figure C - 15 Scatter Plot of Model #8 for Training for Montreal Dataset

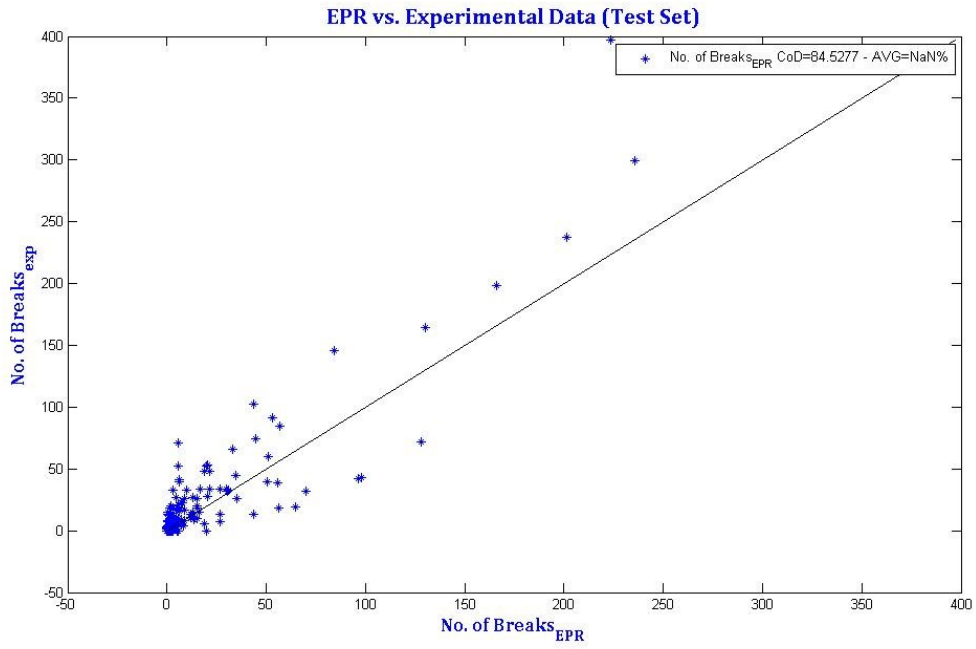


Figure C - 16 Scatter Plot of Model #8 for Testing for Montreal Dataset

$$No.of\,Breaks = +0.086502L^{0.5} + 0.00051089\frac{L^{1.5}A^{1.5}}{D^2}\ln\left(\frac{1}{L^{0.5}}\right) + 0.021313\frac{L^{1.5}M^2A^{0.5}}{D^2}$$

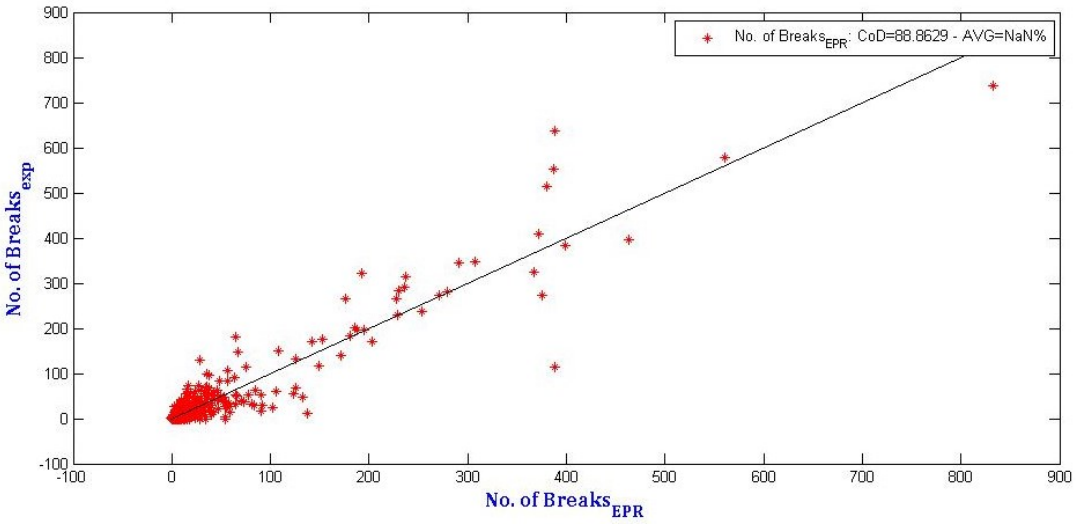


Figure C - 17 Scatter Plot of Model #9 for Training for Montreal Dataset

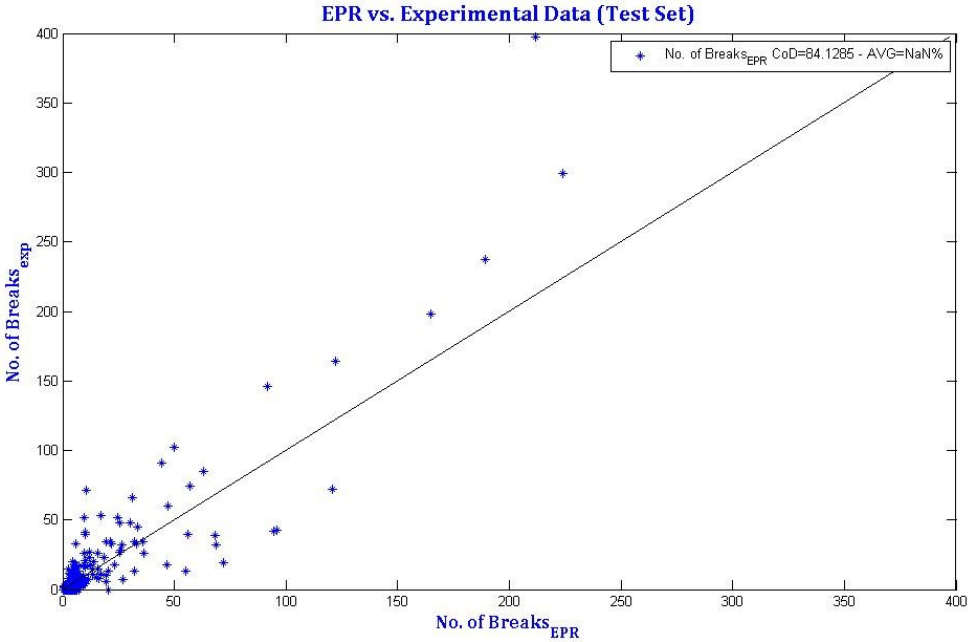


Figure C - 18 Scatter Plot of Model #9 for Testing for Montreal Dataset

$$No.of\,Breaks = +0.00044077L + 0.017413\frac{L^{1.5}M^2A^{0.5}}{D^2} + 8.8604e-005\frac{L^{1.5}A^2}{D^{1.5}M^2}\ln\left(\frac{D^{1.5}}{L}\right)$$

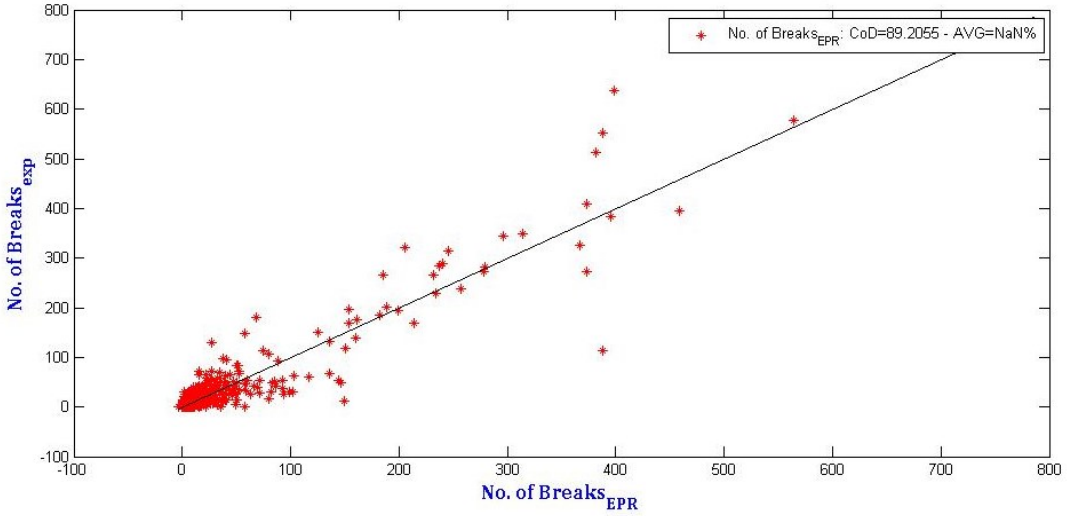


Figure C - 19 Scatter Plot of Model #11 for Training for Montreal Dataset

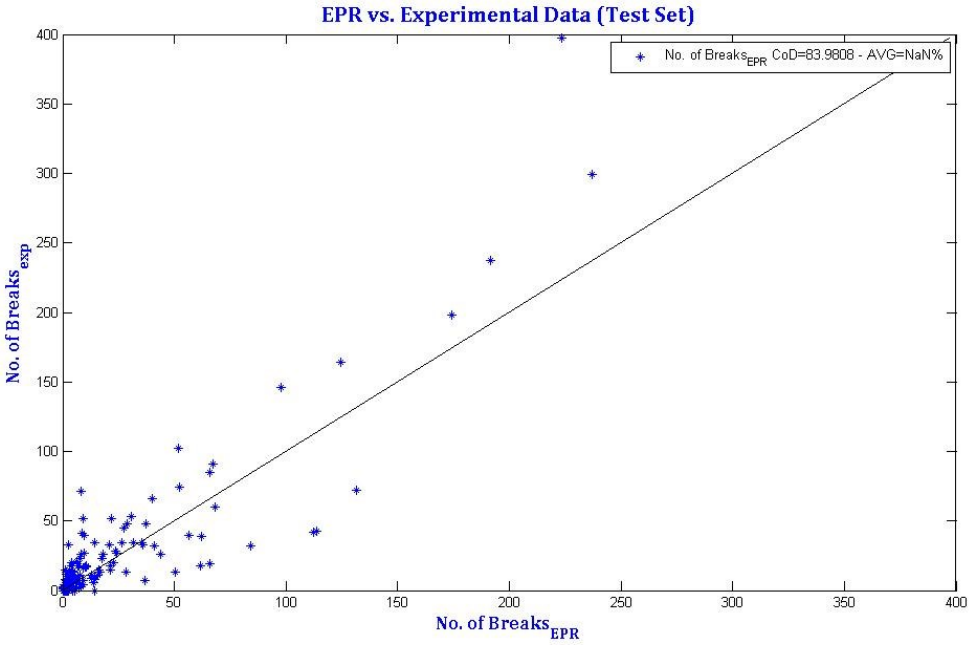


Figure C - 20 Scatter Plot of Model #11 for Testing for Montreal Dataset

$$No.of\,Breaks = +0.00057323L + 0.049651\frac{L^{1.5}M^{1.5}A^{0.5}}{D^2}\ln\left(M^{0.5}\right) + 8.8156e-005\frac{L^{1.5}A^2}{D^{1.5}M^2}\ln\left(\frac{D^{1.5}}{L}\right)$$

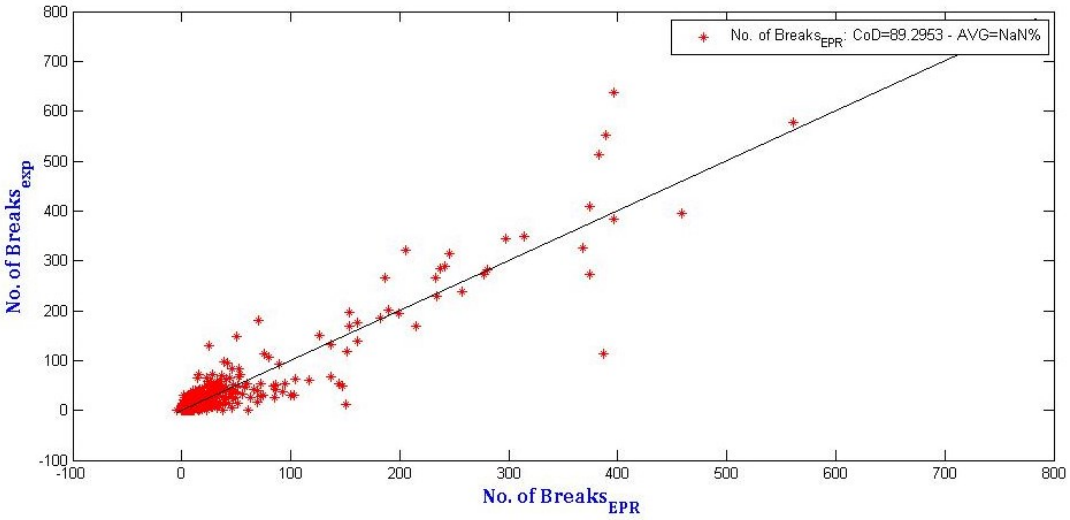


Figure C - 21 Scatter Plot of Model #12 for Training for Montreal Dataset

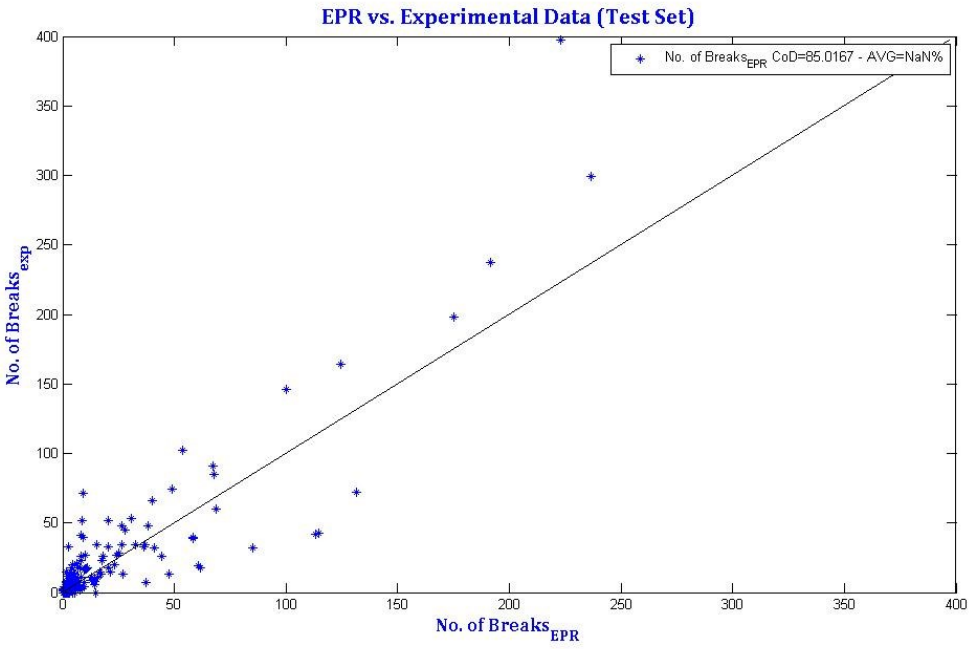


Figure C - 22 Scatter Plot of Model #12 for Testing for Montreal Dataset