Anneé : 2007

N° d'ordre : 4035

Université des Sciences et Technologies de Lille

École Doctorale "Sciences Pour l'Ingénieur"

Laboratoire de Mécanique de Lille (UMR 8107)

Thèse

pour obtenir le grade de

Docteur de l'Université des sciences et technologies de Lille

Génie Civil

Présentée et soutenue publiquement par

Ahmad ASNAASHARI

le 19 Oct. 2007

Modélisation de la défaillance des réseaux d'eau:

Approches Statistique, Réseaux du Neurones et Survie

Directeur de la thèse :

M. SHAHROUR Isam, Professeur, Université des Sciences et Technologies de Lille (USTL)

Membres du jury :

M. CARLIER Eric	Professeur	École polytechnique universitaire de Lille	Président
M. NAJJAR Yacoub	Professeur	Université de l'Etat du Kansas (USA)	Rapporteur
M. DIAB Youssef	Professeur	Université de Marne-la-Vallée	Rapporteur
M. OUAHSINE Abdellatif	Professeur	Université de Technologie de Compiègne	Examinateur
M. TAHON Christian	Professeur	Université de Valenciennes	Examinateur
M. LALLAHEM Sami	Ingénieur R&D	Airele Company - Roost Warendin	Examinateur

University of Science and Technology of Lille

Doctoral School of Engineering Science

Laboratory of Mechanics of Lille (UMR 8107)

A Thesis

submited in conformity with the requirments

for the degree of Doctor of Philosophy

Graduate Department of Civil Engineering

by

Ahmad ASNAASHARI

19 Oct. 2007

Water Pipelines Failure Modeling:

Statistical, Artificial Neural Networks and Survival Modeling

Thesis Supervisor:

M. SHAHROUR Isam, Professor, University of Science and Technology of Lille (USTL)

Examination Committee:

Mr. CARLIER Eric	Professor	École polytechnique universitaire de Lille	Chairman
Mr. NAJJAR Yacoub	Professor	Université de l'Etat du Kansas (USA)	Evaluator
Mr. DIAB Youssef	Professor	Université de Marne-la-Vallée	Evaluator
Mr. OUAHSINE Abdellatif	Professor	Université de Technologie de Compiègne	Examiner
Mr. TAHON Christian	Professor	Université de Valenciennes	Examiner
Mr. LALLAHEM Sami	Engineer, PhD	Airele Company - Roost Warendin	Examiner

DEDICATION

To my wife, Zeinab,

who has dedicated tremendous patience, sympathetic understanding, encouragement and support to my doctoral study,

and my son, Kyann,

who has given me so much happiness in my daily life.

AKNOWLEDGMENTS

At first, I would like to express my deepest gratitude to my advisor, Professor Isam SHAHROUR, for his continuous advice, guidance, support, supervision, and encouragement throughout my doctoral study and research at University of Science and Technology of Lille. His ideas and comments helped me to achieve the results that can be seen in the thesis.

I would also like to thank my committee members, Professor Eric CARLIER, Professor Youssef DIAB, Professor Yacoub NAJJAR, Professor Abdellatif OUAHSINE, Professor Christian TAHON and Dr. Sami LALLAHEM for their kind review of this document and their valuable remarks.

Thanks to Professor Jean Pierre HENRY and Dr. Mohammad HAJI SOUTOODEH, for their involvement in my education at University of Science and Technology of Lille.

I also want to thank all the people at the Labroatory of Machanic of Lille (LML) and department of Civil Engineering at École polytechnique universitaire de Lille for creating a warm and friendly atmosphere. I especially appreciate (without any order): Ali Zaoui, Laurent Lencelot, Hussein Mroueh, Marvan Sadek, Farzam Zoueshtiagh, Moussa Naït.

I am grateful for the help and friendship of all my fellow graduate students friends and specially (without any order): Raed Jafar, Ali Mohammad Ajorloo, Yousef Parish, Hussein Ghoreishi, Mohssen Dayyan, Mehdi Mehdizadeh, Parvin Shakerifard, Issa Mussa, Eddy Eltabach, Rami alabdeh, Loay khalil, Ahmad Al Quadad, Lia Tataie, Ahmad Arab, Fathi Baali, Fethi Benkhenafou, Bian Hanbing, Sebastien Burlon and Bassem Ali.

The development of prediction model required considerable input from the engineers, crew chiefs, and other staff members of the Sanandaj Water and Wastewater Utility (SWWU) whose patience with the process and contributions are much appreciated. Extensive cooperation was also received from the Iranian National Water and Wastewater Engineering Company staffs. Without their valuable work and support, the research would not have been possible. Special mention should be made of the staff of the failure and repairing offices who coordinated the process of data collection and provided many useful inputs: Dr. Motiee H., Soltani S., Pakrouh S., Rezvani M., Moradhaseli M. Karbaschi S. Farhad M. and Rahagh F.

Two months spent at the Centre for Failure detection in Water and Wastewater Network and Preventative Maintenance, Department of Seine Saint Denis (DEA93), were interesting and fruitful. I would like to thank Didier Lesage, Marc Bazot, Manu Jenin, and J.M. Delattre for their help, guidance, valuable opinions and inspiration. It has been a great pleasure to be a part of the “diagnostic” group.

My most sincere gratitude goes to the staff of le Service de Coopération et d’Action Culturelle (SCAC) de l'Ambassade de France en Iran and CROUS de Lille. Their help was greatly appreciated and they are generously thanked.

My special and deepest thanks to my wife, Zeinab, for her great patience and dedication and to my son, Kyann, for his love and smile.

Finally, I would like to thank my parents, brother and sisters and all my family and friends for their loving support and encouragement throughout my graduate career. Although being far away, they have always been my greatest supporters.

Résumé

La défaillance des réseaux d'eau constitue un problème majeur en Iran, qui nécessite des investissements importants et l'élaboration d'une stratégie optimale pour la réhabilition des réseaux d'eau. Ce travail constitue une contribution à cet objectif. Il vise le développement des outils pour améliorer la gestion et la maintenance des réseaux d'eau. Il comporte la détermination des principaux facteurs affectant la défaillance des réseaux d'eau, l'élaboration d'un modèle de prévision fondé sur les Réseaux de Neurones Artificiels (RNAs), et le développement d'un modèle de survie. Ces approches ont été appliquées sur le réseau d'eau de la ville de Sanandaj en Iran.

Le travail de thèse a comporté différents parties, notamment : la collecte de données sur le réseau de la ville de Sanandaj (Iran), l'analyse spatiale et statistique de ces données, le développement d'un modèle basé sur les Réseaux de Neurones Artificiels et l'application de l'approche de survie.

L'analyse des données a permis la détermination de principaux facteurs à l'origine de la défaillance des réseaux d'eau. Deux modèles de régression (Multiple et Poisson) ont été employés pour la prévision du nombre de défaillances du réseau d'eau. Ces modèles ont été comparés à l'approche des Réseaux de Neurones Artificiels. La comparaison a montré tout l'intérêt d'utiliser cette dernière approche pour la prévision de la défaillance des réseaux d'eau. L'approche de survie a été utilisée pour étudier la durée de vie et étudier l'impact d'une intervention sur le réseau d'eau.

Mots clefs: Réseaux d'eau potable, Défaillance, Modélisation statistique, SIG, Réseaux de neurones, Analyse de survie, Prévision, Réhabilitation.

Abstract

A major challenge to Iranian water industry concerns the minimization of failures in water distribution system. This thesis constitutes a contribution for this objective. It includes a) assessment of the main indicators through statistical analysis; b) development of Artificial Neural Networks (ANNs) models for predicting pipes failure number; c) elaboration of a survival models for quantification avoided failure from network based on various rate of renewal. The use of theses approaches generates a quantitative picture of the condition and performance of mains network towards the optimization of the maintenance and rehabilitation programs. All neural networks and survival models were trained and tested on field data in Sanandaj city (Iran).

The methodology followed in this research includes field data collection, descriptive spatial and statistical analysis besides predictive modeling which incorporate Regression, ANNs and Survival models. Descriptive analysis of historical failure data based on statistical methods allowed the determination of factors affecting the evolution of water pipelines failure. Indeed, geostatistical analysis and spatial interpolation provide scientific bases for depicting spatial relationships and the strength of dependencies between failure incidents and environmental, hydraulic and other geographic covariates. Review of univariate statistical inferences, indices of bivariate relationship and multivariate data analysis assess the correlation between the affecting factors and identify the important variables for the occurrence of failures on the water mains. Two regression models (Multiple and Poisson) were used for the prediction of the number of failures in water mains. Artificial Neural Networks (ANNs) models were also developed to predict the number of failures in water mains. Comparison of ANNs and regression approaches reveals that the use of ANNs model in pipeline failure studies provides better prediction. Finally, four survival models were developed to simulate time to failure in water mains, and 3 stratified failure dataset.

Keyword: Water network, Pipeline, Failure, Statistical analysis, GIS, Neural Network, Survival analysis, Prediction, Rehabilitation.

Table of content

List of tables

List of figures

List of abbreviations

Symbol	Meaning
AC	Asbestos Concrete
AG	Age of water mains
ANNs	Artificial Neural Networks
ASCE	American Society of Civil Engineers
ASE	Average Square Error
ASTM	American Society for Testing and Materials
AWWA	American Water Works Association
AWWARF	American Water Works Association Research Foundation
AWWSC	American Water Works Service Company
CDF	Cumulative Density Function
CI	Cast Iron
CPHM	Cox's Proportional Regression Model
CSIRO	Australia's Commonwealth Scientific and Industrial Research Organization
CV	Censored Value
CWWA	Canadian Water and Wastewater Association
DF	Degrees of freedom
DI	Ductile Iron
DP	Depth of water mains
DR	Diameter of water mains
EPA	Environmental Protection Agency

FR	Failure Rate
GIS	Geographic Information System
GLZ	Generalized Linear Models
GPS	Global Positioning System
HN	Hidden Node
ID	Identification Number
KM	Kaplan Meier
LL	Logarithm of water pipelines length
MARE	Mean Absolute Relative Error
MLR	Multiple Linear Regression
MLP	Multi Layer Perceptron
MP	Maximum Pressure in water pipelines on failure time
MR	Multiple Regression
MT	Material of water mains
NBS	National Bureau of Standards
NF	Number of Failure in water mains
NPF	Number of Pervious Failure
NRCC	National Research Council Canada
NWWEC	National Water and Wastewater Engineering Company
PE	Polyethylene
PHM	Proportional Hazards Models
PRM	Poisson Regression Model
PVC	Polyvinyl Chloride

SDD	Standard Distance Deviation
SDE	Standard Deviational Ellipse
SQL	Structured Query Language
SSE	Sum of Squares due to Error
SWWU	Sanandaj Water and Wastewater Utility
TIN	Triangulated Irregular Network
TK	Thickness of water mains
TL	Traffic Load category on water mains
TPs	Thiessen Polygons
UFW	Unaccounted For Water
UTM	Universal Transverse Mercator
UV	Ultra Violet
VMR	Variance Mean Ratio
UWRAA	Urban Water Research Association of Australia
WHO	World Health Organization
WIN	Water Infrastructure Network
WRc	Water Research company
WSAA	Water Services Association of Australia

General Introduction

Introduction and Problem Statement

Water loss due to leakage is a major challenge for water utilities. It frequently reaches of 30% or even 40% from water supply. Since leakage rate increases with mains failure, the water system managers are highly concerned by the minimization of water pipelines failure. The control of failure in water pipes constitutes a major challenge for the sustainability and the environmental protection.

The urban water supply is based on a large and complex infrastructure that has been expanded and developed during the last century. While getting older, water supply assets, primarily pipes, are exposed to the deterioration process and consecutive pipe failure. It is common for cities to have scores, hundred, and even more than a thousand water pipelines breaks each year. Nowadays, most Iranian water utilities observed high rates of failure in water lines. In 2006, there were 229561 reported breaks along the Country's 96,788 kilometers of water lines (NWWEC 2006). This represents, on average, of more than 230 breaks for every 100 kilometer. Therefore, the failure rate of the Iranian water pipelines system is nearly four times the maximum failure rate that has been reported in the literature (McDonald et al., 1994). In study area, the average ratio of pipe breaks per 100 km has been reported as 67 which is considered high and indicates the network is in poor condition. This deterioration not only manifests itself in increased operating and maintenance costs, water losses, frequent service disruptions and a reduction in the quality of water supplied (Kleiner, 1998) but also includes enormous hidden costs (AWWA, 2001). Regardless of environmental and social costs, the cost of repairing and maintaining Iranian's existing water network is estimated at over $50 million each year (NWWEC, 2006).

To ensure that authorities can manage their water distribution systems to provide an adequate supply of safe water in a cost-effective, reliable and sustainable manner, it is essential that they develop a clear understanding of water pipelines deterioration processes (Canadian

InfraGuide, 2002a). An accurate quantitative picture of the condition and performance of system will allow utilities to implement efficient proactive pipe failure management strategies to minimize the overall economic, social and environmental costs of water pipelines network operation.

To date, few standardized techniques are available for Iranian water utilities to evaluate distribution systems and to develop proactive procedures for determining rehabilitation and replacement needs (NWWEC, 2006). Most of Iranian water authorities have adopted some from of subjective ranking system such as work crew opinion to prioritize pipeline rehabilitation. A smaller number of water utilities have completed statistical analyses to predict pipeline failure and incorporate the results within future planning.

In addition, water utility operators manage and operate distribution systems in a reactive mode by responding to emergency breaks and water pipelinesleaks. Experience has shown that a significant number of water line repairs are performed on an unscheduled basis. In this time of budget cuts and limited resources, the ability to optimize the use of maintenance dollars by employing predictive models in the planning stages is rapidly becoming a reality of underground infrastructure management (Crane A.I., 1994). Planned maintenance for a facility in need of repair can yield significant savings over unscheduled or emergency repairs (Mays, W., 2001). The key is to enable planners to predict accurately which components are in the most urgent need of repair, and when others will need repair.

To achieve this aim, methods have been developed to obtain information as to which pipelines are most likely to fail, and when these failures are probable to occur. Predictive modeling includes a collection of techniques that can be used to determine the likelihood of failure or failure rate, for a particular entity. These modeling techniques range from very basic selection rules to complex analyses including spatial and statistical methods together with artificial neural network and survival analysis.

Objectives

The key objectives of this thesis are:

I) To establish a reliable prediction model for water pipelines distribution failure,

II) Elaboration of a preventive maintenance strategy for water pipelines based on scenarios for prioritization of future water pipelines replacement and rehabilitation.

Methodology and Plan of Work

The methodology followed in this research consists of six steps: data collection, descriptive spatial analysis, statistical analysis, regression analysis, ANNs and Survival modeling. Application on the Sanandaj city - Iran was included:

- A literature search to obtain published information on the water pipelines failure analysis and modeling,
- Collection of case study water pipelines failure information through a literature survey and combination of different database included water distribution failure database, customized ArcView/GIS, hydraulic model (Epanet) and deep interview with engineers and crew,
- Elementary analysis of historical failure data based on statistical methods to determine factors which affecting progression of water pipelines failure as well as application of spatial analysis includes clustering and spatial interpolation methods to provide scientific reasons for depicting spatial relationships and the strength of dependencies between failure incidents and environmental, hydraulic and the other geographic covariates,
- Review of univariate statistical inferences, indices of bivariate relationship and multivariate data analysis to assess correlation between the affecting factors and identify the important variables for the occurrence of failures on the water pipelines as well as fitting two regression model namely Multiple and Poisson,

- Application of Artificial Neural Networks (ANNs) models to predict number of failure in water mains,
- Implement of non-parametric and parametric survival models for *"time to failure"* of water pipelines to quantified various rate of renewal over the mains network on the percentage of failures which avoided from this network.

Layout of Thesis

Following the objectives and methodology, this dissertation is organized in five chapters. The first chapter presents briefly the literature on existing water pipelines failure models as well as the deterioration process in mains distribution system. Chapter two provides data requirements and issues related to data acquisition through focuses on the description of the study region. It provides information about identifying the most influential factors which affect the pattern and trend of water pipelines failure. Integrated statistical and spatial approaches have been presented to achieve this goal. Chapter three focuses on the univariate, bivariate and multivariate statistical analysis. Factor analysis was also conducted to discover underlying determinant factors and recognize the relative relationships among variables. Indeed, two regression model namely Multiple and Poisson were fitted to predict the number of failures. Chapter four presents the use of ANNs model for prediction water pipelines failure. Finally, in chapter five, non-parametric and parametric survival models for time to failure of water pipelines are implemented.

1. Literature Survey on Water Pipeline Failure

1.1 Introduction

The distribution systems of public drinking water supplies include the pipes and other conveyances that connect treatment plants to consumers' taps. They constitute a significant management challenge from both an operational and public health standpoint. Furthermore, they represent the vast majority of physical infrastructure for water supplies, such that their repair and replacement represent an enormous financial liability (EPA, 2005).

There are poor quality water distribution networks all over the world and the situation is becoming worse and worse due to inefficient design, poor construction work, improper or unqualified material, improper bedding, aged pipelines, poor network management and maintenance, surrounding environment and breaks from unexpected elements, for example damages from nearby underground constructions (Zhang, 2006). Even when water pipelines are properly installed, the pipes will deteriorate over time (Kleiner, 2005). As water pipelines deteriorate both structurally and functionally, their breakage rates increase, network hydraulic capacity decreases, and the water quality in the distribution system may decline. Kleiner and Rajani (2001) classified the deterioration of pipes into two categories. The first is structural deterioration, which diminishes the pipes structural resiliency and its ability to withstand the various types of stresses imposed upon it. The second is the deterioration inner surface of the pipes resulting in diminished hydraulic capacity, degradation of water quality and reduced structural resiliency in case of severe internal corrosion. Both categories of deterioration contribute to diminish the reliability of the distribution network.

A number of professional organizations such as ASCE, NRCC, AWWA, WIN and AWWARF have studied deterioration process related to drinking water. A summary of the extensive reviews are provided in the next sections.

1.2 Failure on Water Pipelines

Water pipelines failures take place in multiple stages, rather than in a single episode, as shown in Fig. 1.1 The recent work at the National Research Council Canada (NRCC) has shown that the failure process is much more complex than expected (Makar, 2001).

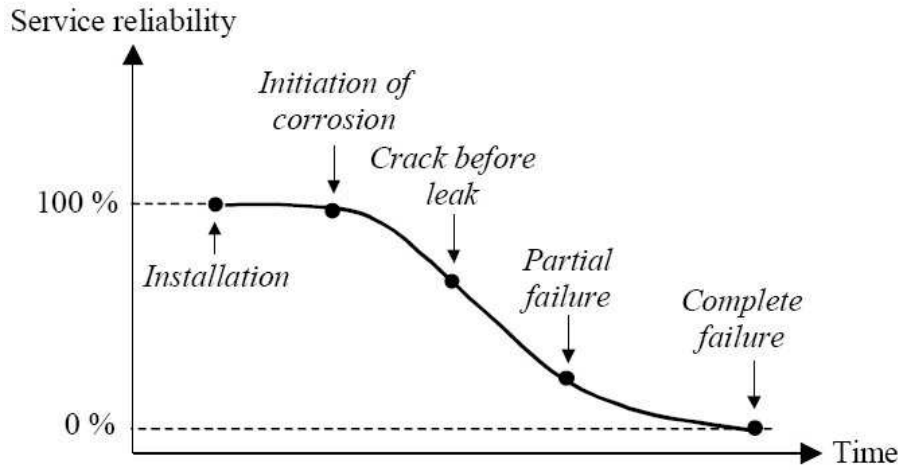

Fig. 1.1 Water main failure development (Misiunas, 2005)

According to Fig. 1.1, the following steps can be identified:

- *Installation*: The new intact pipe is installed,
- *Initiation of corrosion*: After the pipe has been operating for some time, the corrosion processes start on the interior or exterior (or both) surface of the pipe,
- *Crack before leak*: Cracks, corrosion pits and graphitization are typical products of the corrosion process. In some cases cracks can be initiated by mechanical stress,
- *Partial failure*: Eventually, developing corrosion pits and cracks reduce the residual strength of the pipe wall below the internal or external stresses and the pipe wall breaks. As a consequence, the leak or burst will be initiated depending on the size of the break. In some cases the size of the failure is not big enough to be readily detected,
- *Complete failure*: The complete failure of the pipe can be caused by a crack, corrosion pit, and already existing leak/burst or a third party interference. Such a failure is usually

followed by water appearing on the ground surface or a considerable change in the hydraulic balance of the system.

Not all pipes will have a failure sequence as shown in Fig. 1.1 Makar et al. (2001) have explained that stress corrosion cracks are likely to be active cracks, i.e. develop with time. The evidence of a multi-event cracking is presented, indicating that there can be a substantial time interval between the initial and subsequent cracks (Makar et al. 2001). According to Saegrov et al. (1998), the temporal development of the failure is influenced by the material of the pipe. Steel and ductile iron pipes are likely to leak before they break. Cast iron and larger diameter prestressed concrete pipes typically break before they leak. Plastic and PVC pipes can do either, depending on the installation and operational conditions. The deterioration mechanisms in plastic pipes are not well known since they are likely to be slower and plastic pipes have been in use only for the last 30–40 years.

1.2.1 Why and when do pipes fail?

Pipe breakage is likely to occur when the environmental and operational stresses act upon pipes whose structural integrity has been compromised by corrosion, degradation, inadequate installation or manufacturing defects.

Buried water pipelines are designed to withstand certain design loads. Generally, these loads include earth load, truck/live load, working pressure, and water hammer pressure. The pipe material and wall thickness are chosen to withstand these loads. Pipes located in regions prone to freezing temperatures sometimes experience an additional load (frost load) caused by frost heaving of surrounding soil. Similarly, wide and rapid temperature variations in the soil pipe-water environment lead to additional thermal stresses on the pipe. Leakage in pipes and bad construction practices around the pipe lead to pipe bed disruption and thereby making it prone to breakage due to beam action (Agbenowosi, 2001).

By considering the structural failure of a main, there are two types of stresses that can cause a burst: Longitudinal stresses and transverse stresses. Longitudinal stresses generally cause the main to fail through the creation of circumferential cracks. The actions that create these longitudinal stresses in mains include thermal expansion or contraction, beam action and

internal pressures. Transverse stresses can split into two types, namely hoop stresses and ring stresses. Hoop stresses are created from the internal pressure of the water inside the main. Ring stresses are associated with external forces including the earth load of soil covering the mains traffic load and forest penetration (Savic, 1997).

In addition to the increased loads on the pipe, the pipe's structural integrity is jeopardized temporally by corrosion at a rate dependent on the pipe material type; characteristics of the surrounding soil; and the hydraulic and chemical properties of the water flowing in the pipe. Corrosive soils accelerate the development of corrosion pits on the pipe outside surface. Corrosive water accelerates the graphitization and the eventual reduction in pipe wall thickness from the inside of the pipe (O'Day, 1982).

1.2.2 Bathtub curve

The failure function for most pipelines can best be described in terms of the "Bathtub" curve (Fig. 1.2). It is commonly assumed that three general classes of pipeline failure may occur. Early-life failures, generally attributed to design errors or manufacturing/assembly problems.

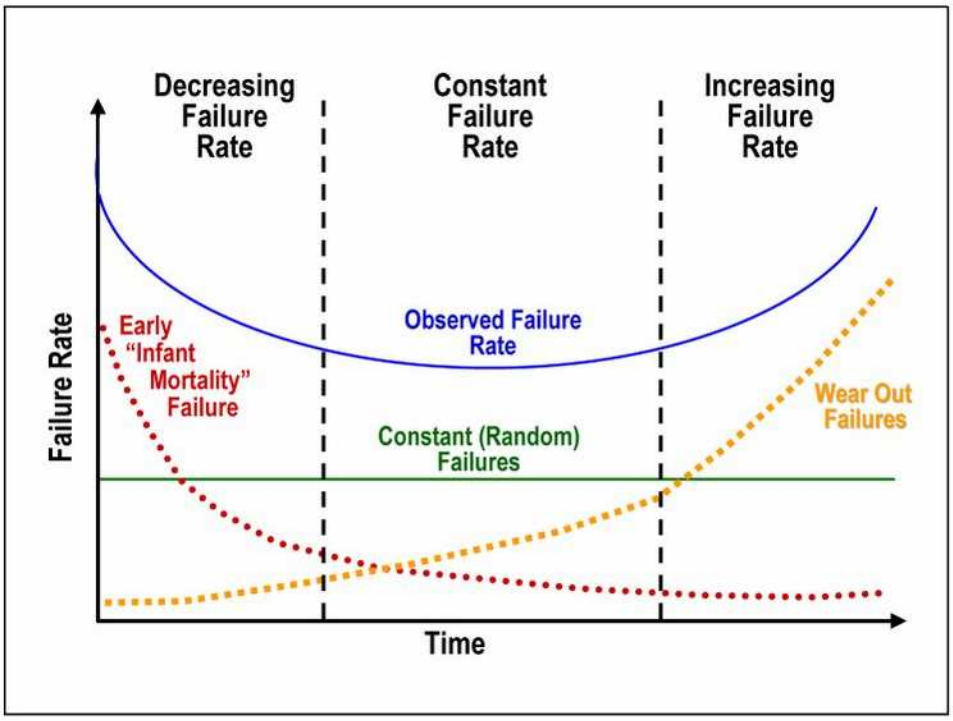

Fig. 1.2 Typical "Bathtub" curve for water pipelines deterioration (Macmillan, 1986)

After all components settle, the failure rate is relatively constant and low (random failures). Then, after some years of operation and accumulation of damages, the failure rate again begins to increase exponentially (so-called wear-out failures), until pipeline will completely have failed. The combination of these three influences provides a basis for understanding the traditional "Bathtub curve" for the time-dependence of the hazard rate.

For example, new polyethylene pipelines often suffer several small failures right after they were installed (Fig. 2.12). The early life failures are frequently attributed to design errors, poor material selection and problems associated with manufacturing or assembly process. Then, as the pipeline reaches a particular age, it becomes more prone to breakdowns, until finally, the pipe will have failed.

1.2.3 Individual pipe failure probability

Pipe break rates in a distribution system can be determined from historical break/repair data. Here, the probability of failure of an individual pipe is given by:

$$p_i = 1 - e^{\beta_i} \quad \text{and} \quad \beta_i = \lambda_i L_i \tag{1.1}$$

where:

β_i = Expected number of failures per year for pipe *i*

λ_i = Expected number of failures per year per unit length of pipe *i*

L_i = Length of pipe *i*

Section 1.7 will explain more probabilistic methods in detail.

1.2.4 Consequence of failure

The costs of a water main failure event may be classified into three categories: (a) direct, (b) indirect, and (c) social costs. While direct costs are relatively easy to quantify in monetary terms, indirect costs may require much more effort, and social costs are often the most difficult to describe and assess (Rajani and Kleiner, 2002). One study estimated that these indirect costs could equal 20% to 40% of the repair costs (AWWSC, 2002).

Strictly speaking the magnitude of failure consequence is a random value because no two failures have the same consequences. The failures of small distribution mains are usually repaired with little effort and typically collateral damage is relatively small. The failures of large transmission mains are relatively rare, and because only a few water utilities attempt to assess total failure damage, there are currently insufficient data to assign probability distributions to failure costs. More research is required to gain a better understanding of the true magnitude of indirect and social consequences of all failure types.

1.2.5 Water leakage

It is recognized that leakage of water reticulation pipes is a problem worldwide. Water leakage is a costly problem, not only in term of wasting a precious natural resource but also in economic terms. The primary economic loss due to leakage is the cost of raw water, its treatment and transportation. Leakage inevitably also results in secondary economic loss in the form of damage to the pipe network itself, e.g. erosion of pipe bedding and major pipe breaks, and in the form of damage to foundations of roads and buildings. Besides the environmental and economic losses caused by leakage, leaky pipes create a public health risk, as every leak is a potential entry point for contaminants if a pressure drop occurs in the system (Stewart et al. 1999).

1.3 Factors Affecting Pipelines Failure

The first step in understanding the water pipelines failure process is to analyse the factors which contribute to pipelines failure. Water pipelines are exposed to a variety of physical, chemical and loading factors in their operating environment (Boxall et al., 2001). These factors affect the breakage and deterioration rate of water mains. Kleiner and Rajani (2001) reported that these factors include operational, environmental, and physical characteristics. In addition, Best Practices (2003b) classified factors that contribute to water pipelines deterioration into 3 groups. Water pipelines breaks are caused when and where the loading on pipe exceeds the pipe strength (i.e. ability to resist loading). Many previous studies have investigated such causal factors by a number of authors (e.g. Morris, 1967; Shamir and Howard, 1979; O'Day and Kelly, 1982; Goulter and Kazemi, 1988). Several factors such as

pipe age, material and diameter, soil parameters, climate changes, pressure in the system and the type of environment of the pipe are the main factors that influence the frequency of pipe breaks in the supply system. Makar et al. (2001) worked on the cause of failure in gray cast iron pipes. They showed that corrosion, manufacturing defects, human error and unexpected levels of pipe loading play the role in large number of failure that occur each year. Morris (1967) suggested a number of possible causes for water pipelines breaks, but underlined that "the cause of water pipelines breaks cannot always be ascertained immediately". Most of time in root cause failure analysis, there are multiple causes of pipe strength deterioration. In effect, failure of a main often is the result of interacting forces. For example, corrosion may have weakened a pipeline to the point where excessive pressure (internal or external) will cause a break.

Overall, the most important variables describing the structural deterioration of water networks can be grouped into four categories (Aslani, 2003):

* Structural or physical variables,
* Internal or hydraulic variables,
* External or environmental variables,
* Operational or maintenance variables,

Table 1.1 provides a more extensive list of factors that contribute to pipe failure.

Table 1.1 Pipeline failures factors

Structural / Physical Indicators	Environmental / External Indicators	Hydraulic/ Internal Indicators	Operational / Maintenance Indicators
Aging water lines	Traffic loading	Higher operating pressure	Poor quality materials and fittings
Number of pervious breaks	External corrosion	Transit condition	Quality of installation and workmanship
----	Forest and cold weather	Internal corrosion	Third-Party Damage

1.3.1 Environmental indicators

Quite often a pipe failure is caused by a combination of some environmental forces. They can be induced by a number of different sources, such as pipe soil interaction, traffic or climate (Rajani et al. , 2004). Different environmental conditions lead to pipe deterioration with owns rate. The following indicators for the environment have been identified because they represent many of the critical environmental issues facing water pipelines failure in literature reviews. Indeed, soil type, soil moisture, groundwater presence, trench backfill material, and pipe bedding can be falled into this group (Best practices, 2003b).

Ground conditions

Stresses leading to pipe failure may be induced in pipes due to ground movement (Lackington (1980), Pascal and Revol (1994), Skipworth et al (2002)). Pascal (1994) cited an estimate that a quarter of the UK supply network was laid in highly aggressive and/or shrinkable soil and that there was strong evidence that pipe bursts caused by corrosion and fracture correlated with soil factors. In a study of cast iron pipes Tsui and Judd (1991) reported that 30% of the mains assessed failed due to corrosion, and that the majority of these occurred in highly aggressive soils, with prevalent shrink/swell characteristics. Jarvis and Hedges (1994) concluded that soil corrosivity maps provided a sound basis for partitioning pipes into areas of equal corrosion risk, and Grau (1991) provided worldwide reports of the use of soil maps for highlighting areas of high burst risk. It is useful to note that all of these studies appear to have considered soil properties with respect to burst rate in isolation of any of the other possible explanatory variables.

Francis (1994) suggested that the ground movement causing bursts was associated with traffic loading. However Pascal and Revol (1994) report that there was no association between bursts, traffic loading or the position of pipes, and Marshal (1999) reported that the response of fill to dynamic traffic loading was elastic with no permanent increase in external pressure on the pipe.

Traffic loading

European research has shown that the effects of vibration and high loading caused by heavy lorries is thought to be a major factor affecting buried pipelines and leading to pipe failure. In the Failnet approach, traffic is taken into account as a qualitative variable according to the number of vehicles per hour or the type of road. This analysis showed that failure rates increase with traffic load. Davies et al. (2001) considered traffic load as a parameter which affect the structural deterioration of rigid sewer pipes. They applied logistic regression for sewer condition and used 5 nominal categories for traffic load: (0) Urban road; (1) Main road; (2) Light road ; (3) Footpath/verge ; (4) Other. The study suggested that sewer location is an important variable in assessing the risk of sewer collapsing. Francis (1994) suggested that the ground movement causing bursts was associated with traffic loading. Instead, Pascal and Revol (1994) report that there was no association between bursts, traffic loading or the position of pipes. Also, Eisenbeis (1994, 1997) used land use over the pipe (i.e. no traffic vs. heavy traffic), as a variable in failure models.

External corrosion

The probability of failure due to external corrosion is a function of surrounding soil properties such as resistivity, pH, the presence of sulphate, microbiological influence as well as temperature, coating type and condition (Garry, 2000). However, the real world effects of these soi1 factors on the external corrosion rate of water pipelines are not well understood. Previous investigations into water pipelines corrosion have had mixed success in correlating the rate of external corrosion to specific soi1 properties (Weiss et al. 1982, O'Day 1982). Accordingly, several mechanistic approaches have been used to model corrosion (e.g. Romanoff, 1957; Rossum, 1969; Kumar et al., 1987). For modeling the change in pit depth with time, soil environment and age, Rossum (1969) developed a set of equations. Rossum's equation for the pit depth had the form:

$$p = f(\text{soil parameters}) * \text{time} * [(10\text{-}pH) / \text{soil resistivity}]\ N \qquad (1.2)$$

where: p = pit depth and N = parameter

His equations are partly based on the extensive data collection effort by the National Bureau of Standards (NBS). An analysis by NBS led to an equation of the form:

$$p = k\,(T)^n \qquad (1.3)$$

where: p = depth of the deepest pit at time T and k, n = parameters

The values of the parameters k and n were provided for the 47 different soil groups. Later, Rossum (1968) took advantage of these results in developing his equations.

Extreme weather condition

Main breaks are most likely to occur during extreme weather conditions. Rrigid weather is the most common time for main breaks, when both air and water temperatures can contribute to breaks. Hot, dry weather is the second most frequent time. Shifting ground and increased volume and pressure can stress water mains.

Kleiner and Rajani (2000) provided many references to reported observations on the influence of temperature and soil moisture on the frequency of water pipelines breaks. Rajani et al. (1996) showed that differential temperature change between pipe and soil, and also soil shrinkage due to dryness result in the development of stresses in the pipe.

The high breakage frequency of water pipelines during winter has been attributed to increased earth loads exerted on the buried pipes, i.e. frost loads. Cold temperatures frequently drive frost deeper into the ground, causing more water pipes to break. Since the pipe temperature drops in winter, pipes tend to contract. In this process, tensile stresses develop in the pipe because the pipe deformation is restrained by surrounding soil. Although compressive stresses are included during the warm season, pipes are more likely to break during winter since pipes with flaws or defects are much weaker in tension than in compression.

Utilities with cast iron pipe typically experience an increase in main failures with freezing temperatures. Although plastic pipes also are affected by a change in temperature due to their high coefficient of thermal expansion, it is less of an issue due to the flexibility of the pipe, and the phenomena of concern discussed here applies only to iron pipes.

Cohen and Fielding (1979) provided a simplified formula for the determination of the frost depth in a soil as a function of the freezing index. They further developed a modified Boussinesq equation relating the expected frost load with the frost depth. The results obtained from the modified Boussinesq equation compared very well with field measurements.

Rajani and Zhan (1996) described the mechanics and circumstances leading to generation of frost loads. They showed that dry soil (excepted after an extreme dry season) has low latent heat capacity and will therefore lead to deeper frost penetration. Additionally, Ann et al. (2005) reported that the number of pipe breaks dramatically increased due to a cold wave On 15 January 2001 in Seoul.

1.3.2 Structural indicators

The second group of variables and their parameters are the result of structural conditions. In fact, the failure of a water main is directly related to its structural/physical condition (Stephens et al. , 2003). Therefore, an indication of the physical condition of water pipes is an important input to the deterioration models. Unfortunately, due to lack of extensive pipe core sampling data, direct information on physical condition of water pipelines is limited. In Europe and US, there has been much research on the structural factors that contribute to pipe failure. Following is two more relevant indicators which were considered in the literature reviews.

Aging water lines

In water reticulation pipes the failure levels increase with age (WSAA, 1998). But some previous studies, O'Day et al. (1982), and Ciottoni (1983) presented that the rate of pipe failure was not as strongly correlated to age of pipe as expected. Boxall et al. (2001) have suggested that age alone is a poor indicator of the necessity for pipe replacement or rehabilitation. In addition, European research has also shown that pipe age is a fairly good indicator of pipe breaks in wastewater collection pipes. Herbert (1994) noted the usefulness of age as a measured, but concluded that it must be combined with knowledge of network condition and weak points to allow accurate assessment. From operational experience, it has been reported that certain pipelines that had already worked out their ‘rated useful service

life' were satisfactory. Although these studies and others note that age alone is a poor indicator of the likelihood of pipe failure, some studies have reported direct association between age and burst rate. For example, Kettler and Goulter (1985) found a strong correlation between the age of an asbestos cement main and its burst rate, and Pascal and Revol (1994) found that the number of breaks in cast-iron pipes increased with age.

Number of pervious breaks

Initial structural pipe condition can be represented by pervious number of breaks. Many researcher (Eisenbeis, 1994 ; Gustafson & Clancy, 1999) have shown that the breakage pattern strongly depend on the number of pervious breaks that pipes have experienced. Research in the US (Clark et al., 1999) has shown that generally, each time a pipe is repaired, the time to the next repair is increasingly shorter. Additionally, they found that after first failure, the number of failure events increased exponentially with time by using regression analysis. Similarly in a study restricted to pipes greater than 200mm diameter, Andreou and Marks (1986) presented that the time to next break decreased as each break occurred. The result of these analysis showed that the rate of deterioration was greater for pipes in the poor initial condition.

Goulter et al. (1988, 1990) showed that the probability of water-main breaks occurring is highest within a short time and a short distance from a previous break. That study measured the grouping, or clustering, of failures based on distance and time to the next failure. It was found that increasing the length of time and the distance that defined the dimensions on which the cluster was based did not cause a proportional increase in the number of failures that occurred in that cluster.

1.3.3 Hydraulic indicators

Higher operating pressures

A buried pipe has an inherent strength by which it can resist the internal and external forces: soil loading and internal pressure. The ability of a pipe to resist the stresses induced by internal pressure is a function of the tensile strength of the material and wall thickness

(Skipworth et al., 2002). As the pipe deteriorates with age, the strength of the pipe is reduced; making it increasingly vulnerable to loads that will eventually exceed the pipe's remaining strength value. It should be noted, however, that when decreasing the limit value of the load at which a pipe fails, pressure reduction only increases a pipe's lifetime for a finite period, and this will delay pipe failure but not eliminate its occurrence (Moglia et al. , 2006). Lambert (1998) has reported that high pressure is probably the most important factor in failure of pipelines. Especially in older systems, an increase in pressure even by a few metres, can result in a large number of bursts.

Moglia et al. (2006) used the Non-homogeneous Poisson statistical model for forecasting pipeline failures with pressure as one of the covariates. This means that the calculated failure prediction for a particular pipe will change with a change of pressure. This in turn allows for investigating the probable effects of reducing the operating pressure in a certain pipe or a region in the pipe network, such as a pressure zone.

Transit condition

During operations period, water pipelines rarely operate in steady state conditions. Any change of the water flow velocity in a pipeline causes pressure fluctuation (called "water hammer" or "surge"). As the velocity change is larger and faster so are the pressure changes. However, fast stoppage of high velocity flow may cause dangerous high/low pressure oscillating waves, exceeding the safe operating limits of the pipeline. Atmospheric pressure, common phenomena in the pressure oscillation, may damage the pipeline by cavitations and collapsing of the pipe due to the external pressure. The pressure surges can occur when water and air valves open and close during network operations as well as . These surges can be one of the factors in failure clustering, as valves are closed and opened during repair activities (Røstum, 2000). Additionally, The rapid filling of pipelines creates the potential for problems caused by the entrapment of air pockets within the pipes. These include water hammer caused by rapid explosion of air pockets (Whily and Streeter, 1993). In the summer time, due to increased demand and pressure surge in the mains, typically mains breaks occur. In 2004, Sinske token into account the air-pocket formations as a cause of failure in water network based on seven different types of air-pocket formations.

Internal corrosion

Service mains are subject to corrode from the inside as a result of contact with stagnant water. Interior corrosion is a concern because it weakness the pipe and increases the risk of a rupture, and because rust deposits built up and clogged the main. The amount of internal corrosion depends not only on the type material used to construct the service pipe but also depends on the characteristics of the water (e.g. pH, alkalinity, bacteria and oxygen content) being distributed. Overall, the internal corrosion of water pipelines is a better understood phenomenon than external corrosion. Tools such as Baylis curves and Langelier's equation have been developed which can be used to determine whether water will be corrosive to a pipe (Garry, 2000). More studies have been performed in water quality impacts which associated with this kind of degradation (Benjamin et al., 1996). But there is not enough study about distribution system pipes failure due to internal corrosion.

1.3.4 Operational indicators

In addition to Environmental, Hydraulic and Structural variables, Operational aspects such as human error, poor pipes manufacturing and inadequate control during production process as well as using bad initial material are also important factors in water pipelines failure.

Poor quality materials and fittings

The quality of pipe materials and fitting is an important factor which could increase the interruptions in the water supply. Poor quality controls during manufacturing of pipes and also using the improper initial material have created this problem. The types of defects can also occur during the pipe manufacturing process. Makar et al. (2001) after a three year investigation by the National Research Council Canada, has presented in addition to corrosion, manufacturing defects and human error play a role in the large number of gray cast iron pipe failures.

Bad storage and transport practice

Failure is also caused by UV degradation weakening the pipe, caused by bad storage practice (WHO, 2001). When plastic pipelines are subjected to long-term exposure to ultraviolet (UV)

radiation from sunlight, they suffer surface damage. According to the ASTM specification, if plastic pipe is stored outdoors, it may require protection from weathering in accordance with manufacturers' recommendations. And in warm climates, the covering should allow air circulation in and around the pipe. Ductile Iron pipe is not vulnerable to effects of exposure to sunlight or weathering.

Additionally, if pipes are not loaded and supported properly prior to being transported long distances, cracks can occur due to a phenomenon called "transit fatigue". Transit fatigue occurs when pipe flexes in a certain manner repeatedly over long periods of time during transport, resulting in cracking of the pipe wall. Fortunately, these defects are typically discovered during the hydrostatic pressure testing that occurs prior to the pipeline being placed in service; however, some can remain and grow during pipeline pressure cycles until failure occurs.

Quality of installation and workmanship

Human error plays a role in the large number of pipe failures that occur each year. Starting with an inappropriate design, there are several practices during and after the construction that can contribute to the failure of the pipe system (Karney, 1992). Poor transportation, movement and installation techniques can promote corrosion followed by the failure of the pipe. Accidentally removed coating exposes the pipe to extensive corrosion. Another possible cause of the failure may sometimes be linked to its installation (Boxall et al., 2001). Failure due to improper bedding and poorly controlled the fill used around a main and likewise poor pipes manufacturing and inadequate control during production process also important factors.

Third-party activity

Another possible cause of the failure is a third party damage (or nearby excavation). Excavations in the vicinity of pipelines disturb bedding conditions or hit on the line, resulting in pipe failure. For instance, Research in the U.K. (WRc, 2001) shows that work on adjacent services (e.g. gas, electricity) can cause pipe failure. The range of excavation damage runs from damage to the external coating of the pipe, which can lead to accelerated corrosion and the potential for future failure, to cutting directly into the line and causing leaks or, in some

cases, catastrophic failure. Mostly, the absence of utility maps increases the rate of failure by the excavation of contractors. The third party activity failure model takes into account such factors as: pipeline diameter, wall thickness, location, and depth of cover (Mather, 2001). The results of past research have demonstrated that the deeper a pipeline was buried, the less likely it was to be affected by third party activity. In other work, Greenwood (2002) evaluated the relationship between gas pipeline accidents caused by third party interference and wall thickness. Less thickness correlated with high frequency of failures.

1.3.5 Other factors

In water pipelines failure analysis it should considered another factors which can be affect the rate of failure. Most of them have been indicated in the references.

1.4 Common Water pipelines Failure Modes

O'Day et al. (1982) classified water pipe breakage types into three categories: (1) Circumferential breaks, caused by longitudinal stresses; (2) longitudinal breaks, caused by transverse stresses (hoop stress); and (3) split bell, caused by transverse stresses on the pipe joint. This classification may be complemented by an additional breakage type i.e., holes due to corrosion (Rajani and Kleiner, 2001).

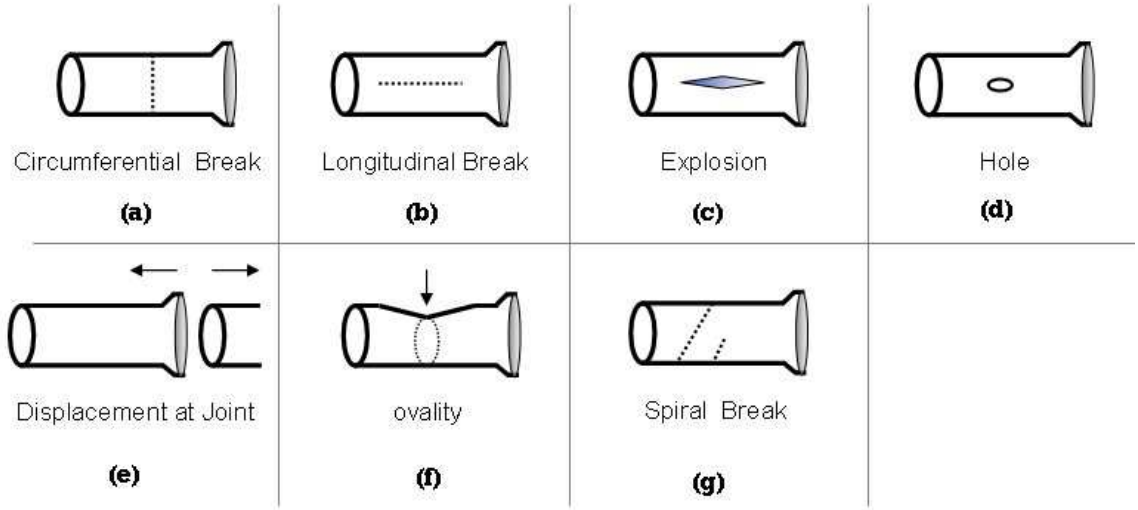

Fig. 1.3 Common failure modes for water pipelines in study area

Makar et al. (2001) investigated failure modes and mechanisms in gray cast iron pipe. They highlighted a number of modes and causes of pipe failures that have been encountered during a three year investigation by the National Research Council Canada . Overall, it is important for the type of failure to be identified so that the proper repair procedure can be undertaken. In the following several types of water pipelines failure will be explained.

1.4.1 Circumferential break

Circumferential or circular cracking (Fig. 1.3.a) is typically caused by bending forces applied to the pipe. Bending stress is often the result of soil movement, thermal contraction or third party interference (Misiunas, 2005). Circumferential cracking is the most common failure mode for smaller diameter cast iron pipes. In 2005, Hu and Hubble presented the statistics of asbestos cement water pipelines breaks from the city of Regina in Canada. Circumferential breaks were shown to be the predominant failure mode comprising 90.9% of all pipe failures.

Circumferential breaks due to longitudinal stress are typically the result of one or more of the following occurrences: (1) thermal contraction (due to low temperature of the water in the pipe and the pipe surroundings) acting on a restrained pipe, (2) bending stress (beam failure) due to soil differential movement (especially clayey soils) or large voids in the bedding near the pipe (resulting from leaks), (3) inadequate trench and bedding practices, and (4) third party interference. The contribution of internal pressure in the pipe to longitudinal stress, although small, may increase the risk of circumferential breaks when occurring simultaneously with one or more of the other sources of stress (Rajani and Kleiner, 2001).

1.4.2 Longitudinal break

Longitudinal or split cracking (Fig. 1.3.b) is more common in large diameter pipes that runs along the length of pipes. Longitudinal breaks due to transverse stresses are typically the result of one or more of the following factors: (1) hoop stress due to pressure in the pipe, (2) ring stress due to soil cover load, (3) ring stress due to live loads caused by traffic, and (4) increase in ring loads when penetrating frost causes the expansion of frozen moisture in the ground (Rajani et al. 1996; Kleiner, 2001). Kottmann (1994) reported cases of longitudinal

breaks in large-diameter grey cast iron, asbestos cement, and PVC low-pressure pipes as a consequence of air pocket formations during warm temperature conditions.

1.4.3 Spiral break

This type of failure appears to be common in medium sized pipes (approximately 400-500 mm diameter). The failure started as a circumferential crack, but then a section of the pipe broke and the crack propagated down the pipe barrel in a spiral (Fig. 1.3.g).

Fig. 1.4 Spiral breaks on cast iron in study area (closely inspection)

Fig. 1.4 shows a section taken out of the pipe, with the right edge of the section showing the rusty fracture surface. The spiral fracture appears to take place as a form of transition between circumferential cracking of small pipes loaded in bending and the longitudinal cracking seen in large diameter pipes. While some spiral fractures have been associated with pressure surges, the pipe shown here failed in normal service due to manufacturing defects. In the work of Makar et al. (2001) on gray cast iron pipe, some medium (380-500 mm) diameter pipes experience a unique failure mode where the crack in the pipe appears to start in a circumferential fashion and then propagates down the length of the pipe in a spiral fashion. This failure mode has been seen in Des Moines, St. Louis and Ottawa. In the two former cases the failures were associated with pressure surges. The appearance of this failure mode also suggests that the failure is produced by a combination of bending forces and internal pressure.

1.4.4 Hole

In 2001, Makar reported failure by hole due to corrosion in metal water pipelines (pitting). Using incorrect backfilling material around the pipes (include rocks or stones) also fail the pipelines as a point (Fig. 1.3.d).

1.4.5 Displacement at joint

Displacement at joints or joint failure has been popular defect in water pipelines distribution. As reported by a number of authors, pipe joints fails through the combined action of pipe corrosion, soil movement, bad joint assembly practice and water pressure (Rajani, 1999; Rajani, 2000; Makar et al., 2001; AWWAFR, 2000). Their attempts showed that rigid joints in older pipe system are particularly susceptible to damage by soil movement. Long life and good performance for pipelines can be achieved by proper handling and installation.

De silva et al. (2001) reviewed condition of joints in Australian cities water and sewer pipeline system over 25 years of service. They classified joint types in typical pipes and their typical failure modes and also evaluated quantitatively the joints of AC and cast iron pipes.

1.4.6 Elliptical deformation

Because of flexibility of polyethylene pipes, buried PE pipe may deform slightly under earth and other loads to assume somewhat of an elliptical shape having a slightly increased lateral diameter and a correspondingly reduced vertical diameter (Fig. 1.3.f). Elliptical deformation (ovality) increases the pipe's failure potential. Practically speaking, this phenomenon cannot be considered negligible as it relates to pipe failure potential. Due to crews error in construction procedures and inadequate control, geometric stability will be lost.

1.5 Failure Management Cycle

Since pipe failure has become quite a common event in the urban water supply systems, failure management is a part of the everyday operation of pipelines and pipe networks (Misiunas, 2005). However, the number of pipe failure management techniques that are currently practiced by the water industry is not very large. On the contrary, a whole range of

methodologies have been described in the literature, indicating the clear interest in such tools. Depending on the timing of failure management activities with respect to the failure itself, two types of pipe failure management strategies can be defined: proactive failure management, when the pipe repair/replacement decisions are made prior to the failure event to prevent the failure, and reactive failure management, when the repair/replacement is performed only after the failure has occurred. Kleiner et al. (2000) described the pipe management cycle with the focus on the proactive part. A similar pipe failure management cycle can be defined including both proactive and reactive parts as shown in Fig. 1.5.

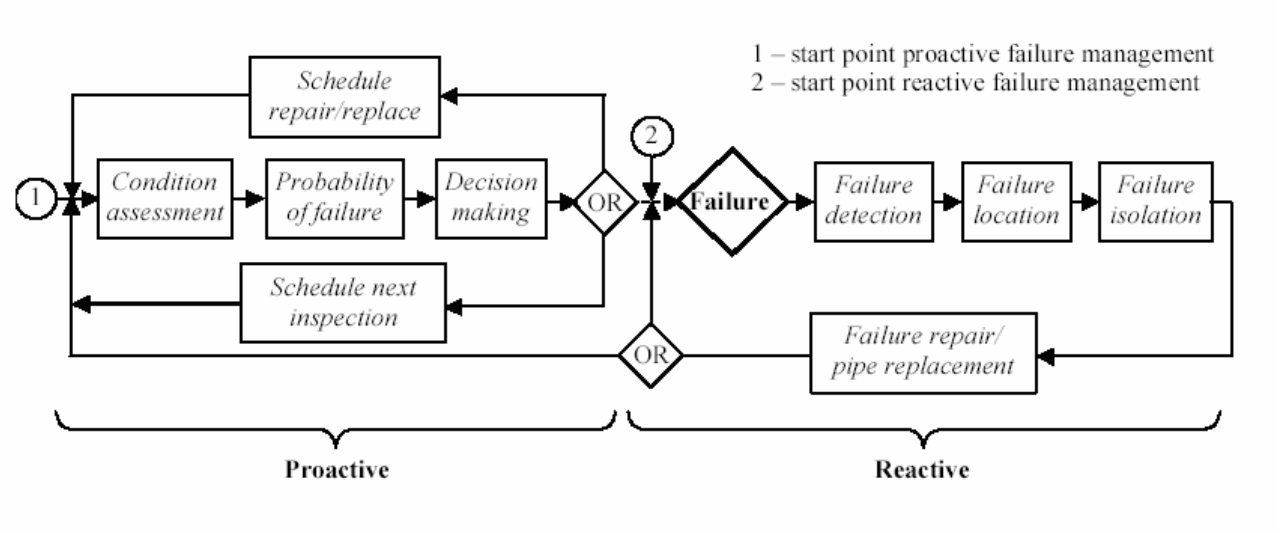

Fig. 1.5 The pipe failure management cycle (Misiunas, 2005)

1.5.1 Proactive failure management

In proactive failure management, the sequence starts at point 1 in Fig. 1.5 Condition assessment is a proactive technique that is used to evaluate the current state of the pipe (Misiunas, 2005). The results obtained from condition assessment are then used to estimate the probability of failure or the residual lifetime of the pipe. Depending on the estimated risk of failure, the decision is made whether the pipe needs to be repaired/replaced/rehabilitated. The rehabilitation time can be scheduled in the short or the long term. Alternatively, the time for the next inspection (condition assessment) is set. Proactive failure management is a part of an overall asset rehabilitation planning strategy. One of the main challenges in the process of the rehabilitation planning is, understanding the process of pipe deterioration. Ideally, if the

proactive failure management is efficient, all pipe incidents should be prevented. However, in case a failure occurs in a pipe, reactive measures have to be taken.

1.5.2 Reactive failure management

If the proactive pipe failure management is not implemented, a reactive management scheme has to be executed starting from point 2 in Fig. 1.5. Reactive approaches are quite simple in that a manager repairs a pipe only after it fails to meet its performance requirements such as hydraulic carrying capacity (i.e., experiences a break, low pressure, or excessive leakage) and water quality (e.g., excessive rust in distributed water). As a first step in the reactive management sequence, the failure has to be detected. After that, the actual location of the failure has to be identified and the damaged section of the pipeline/network has to be isolated. The repair of the failure or replacement of the broken pipe is the last step in the reactive management sequence. After the repair or replacement, the pipe management routine returns to the initial point. The benefit to this approach is that a pipe section realizes its full economic life. The disadvantage of this approach is that the cost of fixing a pipe after it fails is unplanned and may be more than fixing it prior to failure. In addition to the potential for increased and unplanned direct rehabilitation costs, there may be additional indirect costs due to customer service interruptions, damages to co-located utilities, damages to property, and traffic interruptions.

1.6 Quantifying Water pipelines Failure

In the most general sense, measurement is the foundation of scientific analysis, and it lies behind any quantitative analytical statement. Effective planning for the renewal of water distribution systems requires accurate quantification of the structural deterioration of water mains. However, exploration of quantitative model is more meaningful for utility asset managers. Direct inspection of all water pipelines in distribution networks is often prohibitively laborious and expensive. The application of physical models to assess the structural resiliency of each individual pipe is also not realistic in most case because accurate data are rarely available and are very costly to obtain. Using statistical methods to identify breakage patterns over time is an efficient and inexpensive alternate for measuring the

structural deterioration of pipelines. However, in the literature review, failure frequency has been expressed in point view of discrete (number of failure) or continuous sense (failure rate).

1.6.1 Number of failure (NF)

Using a number of measures such as the numbers of bursts on the mains can be applied to monitor the serviceability on mains network. Røstum (2000) in proposed Monte Carlo simulation based on the survival function to predict the expected number of failures within a given time horizon. In 1998, another research by WSAA has been developed Poisson process for the prediction of future numbers of failures of water mains.

1.6.2 Water pipelines failure rate (FR)

Statistics of the performance of water pipelines are typically expressed in terms of the annual number of breaks per hundred kilometers of pipe (# of breaks/100 km/year). The equation for calculating a rate of this type can be written as:

$$Failure\ rate = \frac{Number\ of\ \ failures\ \ per\ 100\ kilometres\ of\ \ water\ mains}{One\ yera} \tag{1.4}$$

The literature review shows that the failure rate is used in many of the proposed models for optimization of rehabilitation/replacement of water pipes (Kaara, 1984; Smith, 1994; Kleiner, 1997). Table 1.2 shows pipe failure rate in some different studies.

The maximum failure rate reported for a U.S. utility (4.3/100 km/year) was significantly lower than those reported for Australian (17.5/100 km/year) and UK (18.8/100 km/year).

According to a study done at the National Research Council (McDonald et al., 1994) based on the reported ratios of pipe breaks per 100 kilometres and the perception of water managers on the global state of their water pipe network, ratios of 40 breaks per 100 km and up are considered high and indicate a network in poor condition. Networks with ratios between 20 and 39 are considered as acceptable condition, while the ratio less than 20 indicates that the network is in good condition.

Table 1.2 Pipe failure rate for different case studies

Case study	Failure rate (per 100 kilometres /year)			
Germany	18			
Australia	35-44			
United State	16.7			
United Kingdom	18.8			
New York	5-6			
Lyon (France)	27.5			
Republic of Azerbaijan (Azervodokanal Association)	2000	2001	2002	
	83	113	85	
Canada (city of Kingston)	2001	2002	2003	
	15	13.6	10.5	
Canada (in four case study)	Chicoutimi	Gatineau	Saint-Georges	Calgary
	46	36	19	20
Canada (in 21 cities)	CI	DI	PVC	AC
	35.9	9.5	0.7	5.8
Moscow	CI	DI	PVC	Steel
	10	1.5	33.4	11.3
Seoul City	192			
Iran (in average)	**>200**			

Sources: Study by the WSAA 2000 , 2001; CWWA,1997; AWWARF,2000; Rajani et al. , 1994, OECD, 2003

As a common application of failure rate is prediction of it. There are many different approaches to predict failure rates and support rehabilitation planning, they typically incorporate one or all of the following major modeling techniques (Stone et al., 2000):

- *Probabilistic or statistical methods:* that estimate a pipe's condition, defined as a probability of failure, based on a statistical analysis of the historical performance (break rate or expected life) of like pipes in similar conditions (operational or environmental). Statistical models can also be used to predict future system requirements by assuming that past break patterns will continue into the future.
- *Deterministic methods:* that identify the best solution (i.e., pipe replacement date, least cost analysis, etc.) based, not on probability, but on a function of the initial pipe

conditions and an understanding of how it modifies given changes in operational conditions, environmental conditions, or time.

- *Heuristic methods:* that enable managers to apply expert judgment and weights to different decision criteria and to prioritize different rehabilitation.

1.7 Water pipelines Failure Modelling Approaches

The ability to predict failure probability (any other performance criteria) is highly desirable for activities such as investment planning and scheduling maintenance (Boxall et al., 2001). In literature review there are more research about water pipelines failure modeling but Pelletier et al. (2003) classified in three main categories: physical modeling, descriptive analysis, and predictive modeling. At a glance, Fig. 1.6 classifies the existing approaches in water pipelines failure modeling.

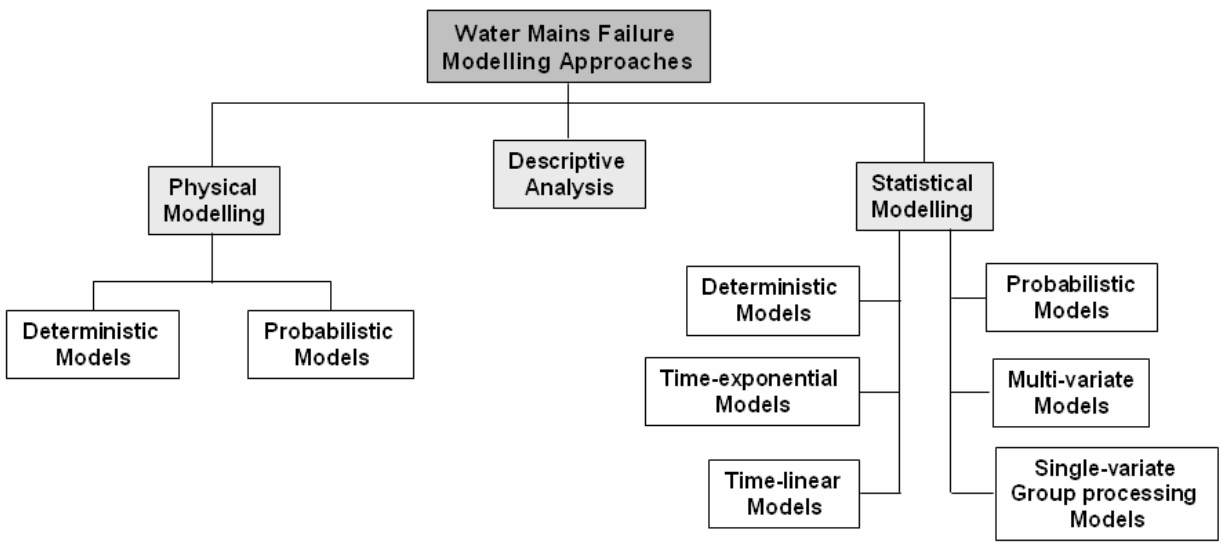

Fig. 1.6 Existing approaches in water pipelines failure modeling (Rajani and Kleiner, 2001)

1.7.1 Physical modeling

Physical/mechanical models of the degradation process employ engineering-based equations to derive structurally based estimates of pipe conditions (Melina, 2000). These models attempt to predict pipe failure by analyzing the loads to which the pipe is subject as well as the capacity of the pipe to resist these loads (Rajani and Kleiner, 2001). They consist of

evaluating of the scope and severity of corrosion on the internal and external pipe walls, and the estimation of resulting stresses from the loads applied to the water pipe (e.g. Doleac et al. 1980; Kumar et al. 1987; Rajani and Makar 2000; Makar et al. 2001).

Several components have to be considered in modeling this structural behavior. The residual structural capacity of water pipelines is affected by material deterioration due to environmental and operational conditions as well as quality of manufacturing and installation. This residual structural capacity is subjected to external and internal loads exerted by the soil pressure, traffic loading, frost loads, operational pressure and third party interference. Some models address only one or a few of the numerous components of the physical process that lead to breakage, while others attempt to take a more comprehensive approach. Initial efforts were aimed mainly towards development of deterministic models, while more recent models use a probabilistic approach to deal with uncertainties in defining the deterioration and failure processes. The models were classified as either deterministic or probabilistic, depending the approach taken to represent deterioration and failure processes (Rajani & Kleiner, 2001). The physical mechanisms that lead to pipe failure often require data that are not readily available and are costly to obtain. Thus, physical models may currently be justified only for major transmission water mains, where the cost of failure is significant, whereas statistical models, which can be applied with various levels of input data, are useful for distribution water pipelines (Rajani and Kleiner, 2001). In literature review, extensive efforts have been applied to model the physical processes of the degradation and failure of buried pipes. Two main groups of models have been identified: deterministic and probabilistic.

Deterministic models aim to predict the corrosion pit growth to estimate the remaining wall thickness and, consequentially, the service life of the pipe. Probabilistic models are designed to calculate the probability of the survival of the pipe over a certain period of time, predict the remaining lifetime or estimate the probability of failure. The main difference between theses models is that probabilistic models incorporate an uncertainty component which is ignored in the deterministic models. The type of data that is required for different methods is similar and includes pipe age, soil parameters, wall thickness and current depth of corrosion pits.

Physical deterministic models

Doleac et al. (1980) used the power function proposed by Rossum (1968) to relate corrosion pit depth with the pipe age to predict the remaining wall thickness of pit cast mains:

$$p = K_n K_a (10 - pH)^n \rho^{-n} t^n A^a \tag{1.5}$$

where:

p	=	average pit depth,
a, K_n, K_a	=	empirical constants derived from field or lab tests,
A^a	=	pipe surface area exposed to corrosion,
pH	=	soil pH,
ρ	=	soil resistivity,
n	=	soil aeration constant,
t	=	time (years).

Randall-Smith et al. (1992) proposed a linear model based on an assumption that corrosion pit depth has a constant growth rate (often referred to as corrosion rate), to estimate remaining service or residual life of water mains.

$$\rho = \left(\frac{t}{P_e + P_i} \delta \right) - t \tag{1.6}$$

where:

ρ	=	remaining life,
t	=	age of water mains,
δ	=	thickness of original pipe wall,
P_e	=	external pit depth
P_i	=	internal pit depth

Rajani and Makar (1999) described a methodology to estimate the remaining service life of grey cast iron mains by considering changes in the structural resistance of a pipe as a result of corrosion pits. They defined the “time of death” of an individual pipe segment as the time at

which its mechanical factor of safety fell below a minimum acceptable value set by the utility owner. They calculated the residual resistance of grey cast iron mains based on corrosion pit measurements while explicitly considering anticipated corrosion rates.

Physical probabilistic models

The probabilistic approach provides insight into the contribution of each parameter to the uncertainty of the results, which is ignored in the deterministic models. Several probabilistic physically based models have been proposed (Pandey (1998); Kiefner and Vieth (1989). They use the residual strength of pipelines.

Pandey (1998) presented a general probabilistic framework to estimate reliability by incorporating the impact of inspection and repair activities planned over the service life of a pipeline vulnerable to corrosion. The intent of this model was to schedule the optimal inspection interval and repair strategy while maintaining adequate reliability throughout the service life of the pipeline.

Based upon this method, European Union developed UtilNets application as a decision support system for rehabilitation planning and optimization of buried grey iron water mains. The system performs reliability-based life predictions of the pipes and determines the consequences of maintenance and neglect over time in order to optimize rehabilitation policy. The system gives a probabilistic measure of the likelihood of structural, hydraulic, water quality and service failure of pipe segments and of the entire distribution system.

1.7.2. Descriptive analysis

Descriptive analysis consists of calculating descriptive statistics to provide insight on breakage patterns and trends. Descriptive analysis can only be performed in cities that have comprehensive databases on the characteristics of their pipes and on pipe breaks. For this reason, there are very few case studies reported in the literature. Some cities often cited for participating in such studies are New York (O'Day et al., 1982; Male et al., 1988, 1990), Cincinnati and New Haven, Conn. (Clark et al., 1988; Goodrich, 1986; Kaara, 1984); Winnipeg, Man., Canada (Kettler and Goulter, 1985; Goulter and Kazemi, 1988; Jacobs and

Karney, 1994); and suburban Paris and Bordeaux, France (Eisenbeis, 1994). This type of analysis is limited by the challenges faced for constructing databases availability of personnel and resources, missing and conflicting data, non-computerized information (paper archives), and so on. As a matter of fact, building such databases has been a concern for many researchers (e.g. O'Day 1982; Clark and Goodrich, 1989; Habibian, 1992).

Pelletier et al. (2003) performed a basic descriptive analysis of the water pipe and breakage data over three case studies. Only the following six characteristics have been collected for all pipe segments in three municipalities: (1) pipe diameter; (2) length; (3)material; (4) year of installation; (5) type of soil; (6) land use above the pipe. Statistics on breakage rates are estimated by taking the ratio of the number of breaks on pipes in a given category and the total pipe length in that category in 1996.

1.7.3 Statistical modelling

The data required to build physical pipe deterioration models is often not available and its collection has a significant cost. As an alternative to physical modeling, statistical models have been proposed to explain, quantify and predict pipe breakage or structural pipe failures.

CSIRO researcher, Davis J. P., depicted that taking physical model across into the field is impossible. To overcome this, they developed a model that uses probability distributions to estimate the probable defect size along a pipe, and the probable loading conditions the pipeline experiences. The model preserves the details of physical degradation and failure mechanisms that occur in service, and can account for changes in operating loads and the surrounding soil environment. A review of statistical models that can be found in the literature is presented in Kleiner and Rajani (2001). Most of the available statistical models use the historical data of pipe failure to predict the future trends. These models can be classified broadly into deterministic, probabilistic multi-variate and probabilistic single-variate models that are applied to grouped data. Deterministic models use two or three parameter equations to derive breakage patterns, based on pipe age and breakage history. The division of pipes into groups with homogeneous properties (operational, environmental and pipe type) is often used, which requires efficient grouping schemes to be available. Probabilistic models are used to estimate pipe life expectancy or failure probability.

Deterministic models

The deterministic models predict breakage rates using two or three parameters, based on pipe age and breakage history. A variety of equations obtained through linear or exponential regression analysis factor have been identified (Shamir and Howards, 1979; Clark, 1982).

Time-exponential models

Shamir and Howard (1979) proposed the first attempt to statistically analyze break records. They used regression analysis to obtain a break prediction model that relates a pipe's breakage to the exponent of its age:

$$N(t) = N(t_0).e^{A(t+g)} \tag{1.7}$$

where t is the time elapsed (from present) in years; $N(t)$ is number of breaks per unit length per year (km^{-1} year^{-1}); $N(t_o) = N(t)$ at the year of installation of the pipe (i.e., when the pipe is new); g is the age of the pipe at time t; and A is coefficient of breakage rate growth (year^{-1}). Note that $N(t_o) \neq 0$, which means that on average a pipe is assumed to always have a breakage frequency, albeit very small in the beginning of its life. Shamir and Howard (1979) provided no details on the location of the study, the quality and quantity of available data or the method of analysis. They recommend that the regression analysis could be applied to groups of pipes that were homogeneous with respect to the factors influencing their breaks.

Walski and Pelliccia (1982) proposed to enhance the exponential model (Table 1.3) by incorporating two additional factors in the analysis, based on observations made by the US Army Corps of Engineers in Binghamton, N.Y. The first factor accounted for known previous breaks in the pipe, based on an observation that once a pipe broke it was more likely to break again. The second factor accounted for observed differences in breakage rates in larger diameter pit cast iron pipes.

Clark et al. (1982) proposed to further enhance the exponential model and transform it into a two-phase model. They observed a lag between the pipe installation year and the first break. Consequently, they proposed a model comprising a linear equation to predict the time elapsed to the first break and an exponential equation to predict the number of subsequent breaks.

Table 1.3 Deterministic time-exponential models (Rajani and Kleiner, 2001)

Researcher, year	Model	Notation	Data requirements
Shamir and Howard, 1979			
	$N(t) = N(t_0).e^{A(t+g)}$	t = time elapsed (from present) in years N(t) = N. breaks per unit length per year (km^{-1} $year^{-1}$) $N(t_0)$ = N(t) at the year of installation of the g = age of the pipe at the present time A = coefficient of breakage rate growth ($year^{-1}$)	Pipe length, installation date and breakage history; formation of homogenous groups essential according to criteria like pipe type, diameter, soil type, break type, overburden characteristics, etc.
Walski and Pelliccia, 1982			
	$N(t) = C_1.C_2.N(t_0).e^{A(t+g)}$	C_1 = ratio between {break frequency for (pit/sandspun) cast iron with (no/one or more) previous breaks} and {overall break frequency for (pit/sandspun) cast iron} C_2 = ratio between {break frequency for pit cast pipes 500 mm diameter} and {overall break frequency for pit cast pipes}	Same data as for Shamir and Howard (1979) plus information on the method of pipe casting and pipe diameter.
Clark et al., 1982			
	$NY=x_1+x_2D+x_3P+x_4I+$ $x_5 RES+x_6 LH+x_7 T$	x_i,y_i = regression parameters, NY = number of years from installation to first repair, D = diameter of pipe, P = absolute pressure within a pipe, I = % of pipe overlain by industrial development, RES = % of pipe overlain by residential development, LH = length of pipe in highly corrosive soil, T = pipe type (1 = metallic, 0 = reinforced concrete), REP = number of repairs, PRD = pressure differential, t = age of pipe from first break, DEV = % of pipe length in low and moderately corrosive soil, SL = surface area of pipe in low corrosivity soil, SH = surface area of pipe in highly corrosive soil	Time of installation, breakage history, type and diameter of the pipe, as well as information about operating pressures, soil corrosivity and zoning composition of area overlaying pipe. Additional types of data such as the type of breaks and pipe vintage required to enhance model.
	$REP = y_1 e^{y_2 t} e^{y_3^T} e^{y_4 PRD} e^{y_5 DEV}$ $SL^{y_6} SH^{y7}$		

Time-linear models

Kettler and Goulter (1985) suggested a linear relationship between pipe breaks and age (Table 1.4). Based on a relatively constant sample of pipes installed within a 10-year period in Winnipeg, Manitoba, they found a moderate correlation between the annual breakage rate and the pipe age (r^2 of *0.563* and *0.103* for asbestos cement and cast iron pipes, respectively). The application of this model is simple and straight forward, similar to the two parameter exponential model of Shamir and Howard (1979).

McMullen (1982) reported a regression model that was applied to the water distribution system of Des Moines, Iowa. They examined several models and the one that performed the best is given in Table 1.4. Their model predicts only the time to the first break of a pipe and thus can not be used as a full-fledged pipe break prediction model.

Jacobs and Karney (1994) applied a linear regression to 390 km of 150 mm cast iron water pipelines with about 3,550 breakage events recorded in Winnipeg. They divided the water pipelines into three age groups (0-18, 19-30, and >30 years) to obtain relatively homogeneous groups of water mains, and applied the equation provided in Table 1.4. Initially they applied this regression equation to all the recorded breaks and obtained coefficients of determination ranging from $r^2 = 0.704$ to 0.937 for the three age groups.

Table 1.4 Deterministic time-linear models (Rajani and Kleiner, 2001).

Researcher, year	Model	Notation	Data requirements
Kettler and Goulter, 1985			
	$N = k_0 . Age$	N = number of breaks per year k_0 = regression parameter	Same data as for Shamir and Howard (1979).
McMullen, 1982			
	Age=65.78+0.028 SR - 6.33pH- 0.049rd	Age = age of pipe at first break (years) SR = saturated soil resistivity (ohm-cm) pH = soil pH rd = redox potential (millivolts)	Data required typically not available; poradic data collection not expensive, however, continuous and extensive data collection program is costly; continuous monitoring of soil properties is important where ground water conditions have not reached steady state
Jacobs and Karney, 1994			
	$P = a_0 + a_1$ Length + a_2 Age	P = reciprocal of the probability of a day with no breaks a_0, a_1, a_2 = regression coefficients	Pipe length, age and breakage history; more data enables formation of homogenous groups.

Probabilistic models

A number of recent studies have been used a probabilistic approach to deal with uncertainties in defining the deterioration and failure processes. The probabilistic models were classified into two groups: probabilistic Multivariate Models and probabilistic Single-variate Group-Processing Models. The models that use probabilistic processes on grouped data to derive probabilities of pipe life expectancy, probability of breakage and probabilistic analysis of break clustering phenomenon fall under the class of the so-called probabilistic single-variate group-processing models.

Probabilistic single-variate group processing models

Herz (1996) proposed a lifetime probability distribution density function based on the principles that had originally been applied to population age classes or cohorts. The probability density *f(t)*, hazard *h(t)* and survival *S(t)* functions are given in Table 1.5.

Table 1.5 Probabilistic single-variate group models (Rajani and Kleiner, 2001)

Researcher, year	Model	Notation	Data requirements
Cohort Survival Model [Herz (1996); Deb et al., (1998)]			
	$N = k_0.Age$	f(t) = probability density function h(t) = hazard function S(t) = survival function t = useful lifetime of pipe a = ageing factor (year-1) b = failure factor (year-1) c = resistance time (years), i.e., pipe will not be replaced at age < c years	- pipe installation dates - pipe "time of death" - valid grouping criteria will enhance accuracy - alternative to "time of death": end of economic life (optimal time for replacement) requires break history
Bayesian Diagnostic Model [Kulkarni et al. (1986)]			
	Prob.[failure/specifiedcharacteristics]=	Pf, = system-wide probability of failure Pc/f, = probability of observing specified characteristics on a segment that failed Pc/nf. = probability of observing the same characteristics on a segment that has not failed	Grouping criteria ("sets of characteristics") such as pipe diameter length, age and type, soil characteristics, operating conditions such as pressure, etc.
Semi-Markov chain [Gustafson and Clancy (1999a)]			
	generalised gamma distribution for t_1 exponential distribution (identical for all t_i (i>1)	ti = time between the (i-1)th and the ith breaking pipe	- pipe breakage history - pipe type - other grouping criteria to enhance accuracy
Break Clustering [Goulter and Kazemi (1988);Goulter et al. (1993)]			
	m = m (s,t)	m = mean number of subsequent failures occurring in the cluster domain x = number of subsequent failures occurring in the cluster domain s = distance from the 1st break in a cluster t = time elapsed from the 1st break in a cluster	Pipe breakage history with the exact time and location of each break.
Data Filtering [Mavin (1996)]1			
	4 rules to filter pipe breakage data, based on calculating the probability of two consecutive breaks (Constantine and Darroch 1993), and discarding the second break if probability is low.		- pipe diameter - pipe material - traffic level - soil type

Gustafson and Clancy (1999) modeled the breakage history of water pipelines as a semi-Markov process. They developed an elaborate model to predict the inter-break times in water mains, based on historical data, but found this model inadequate for predicting future breaks.

Table 1.6 Probabilistic multi-variate models-proportional hazards and accelerated life (Rajani and Kleiner, 2001)

Researcher, year	Model	Notation	Data requirements
Proportional hazards (Marks et al., 1985)			
	$h(t,Z) = h_0(t)e^{b^T Z}$ $h_0(t) = 2\times10^{-4} - 10^{-5}t + 2\times10^{-7}t^2$	T = time to next break h(t, Z) = hazard function $h_0(t)$ = baseline hazard function Z = vector of covariates b = vector of coefficients to be estimated by maximum likelihood	- natural log of pipe length - operating pressure - percentage of low land development - pipe "vintage" (or period of installation) - pipe age at second (or higher) break rate - number of previous breaks in pipe - soil corrosivity
Andreou et al. (1987a, 1987b) ; Marks et al. (1987)			
	Early stage: same as Marks et al. (1985) described above Late stage: $h = \lambda = e^{b^T Z}$	h = hazard (constant at the late stage) Same as above	Same as above
Proportional hazards [Brémond (1997)]			
	$h(t,Z) = h_0(t)e^{b^T Z}$ $h_0(t) = \lambda\beta(\lambda t)^{\beta-1}$	t = time to (next) failure h(t) = hazard function λ, ß, = scale and shape parameters (respectively) of the Weibull distribution	- number of previous breaks - pipe diameter - ground conditions - traffic loading
Time dependent Poisson model [Constantine and Darroch (1993); Miller (1993); Constantine et al. (1996)]			
	$H(t) = \left(\frac{t}{\theta}\right)^{\beta}$ $\theta = \theta_0 e^{az}$	t = pipe age H(t) = mean number of failures per unit length at age t θ, ß, = scale and shape parameters, respectively θ_0 = baseline value a = vector of coefficients to be estimated by regression; Z = a vector of covariates affecting breakage rate.	- mean static pressure - overhead traffic conditions - pipe diameter - soil type
Accelerated life (Lei (1997); Eisenbeis et al.(1999))			
	$\ln(T) = \mu + x^T\beta + \sigma Z$ ==> $T = f(\mu,\sigma,Z)ex^T\beta$	T = time to (next) failure x = vector of explanatory variables Z = random variable distributed as Weibull σ = parameter to be estimated by maximum likelihood ß = vector of parameters estimated by max likelihood Z = random variable distributed as Gumbel (extreme distribution for minima)	- pipe age group and material - pipe diameter and length - pipe material was taken as stratification - traffic loading - soil acidity and humidity - number of previous breaks was taken both as a covariate and as a stratification variate

Probabilistic multivariate models

The probabilistic multivariate models can explicitly and quantitatively consider most of the covariates in the analysis. This ability makes them potentially more powerful and general for predicting the future breakage rates of water mains. It also reduces the need to pre-partition the data into groups, although often some level of partitioning may still be required. Kleiner, and Rajani (2001) classifies the sub model of this categories (Table 1.6).

Failure time prediction

Predictive modeling tool allows authorities to better target their maintenance and act on potential problems before being hit with the cost and disruption of a burst water main (Davis , 2003). The use of statistical analysis to predict failure time has a lot of benefit in the ranking or prioritizing process of water pipelines rehabilitation. The purpose of the statistical analysis is to determine if any combination of available data (e.g., pipe age, diameter, etc.) could be used to predict the time and probability of failures in water pipes. The final output of the model is a listing of main segments prioritized on the probability of failure for a given time period.

Failure time analysis models the probability that the pipe will fail before a certain time as some function of the independent variables (Lim et al., 1996). These probabilities for the first and subsequent failures in water pipes were modeled using various techniques. The Weibull regression model was determined to be the most appropriate model, based on the available data. The survival function for the Weibull distribution:

$$S(t|X) = \exp\left[-\left(t/\alpha(x)\right)^{\delta}\right] \tag{1.8}$$

where $\alpha(X) = exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k)$. A positive value of α_i increases this survival function, the probability of surviving beyond time t.

For the first failure, the estimate for the Weibull scale parameter $(\alpha(X))$ for each set of independent variables is:

$$\alpha(X) = exp[(3.987) + (-0.0002) * length + (0.0154) * diameter + (0.0161) * epoch + (-.0035) * press + (-0.0738) * steep + (-0.0763) * mat + (-0.1429) * soil\ ] \tag{1.9}$$

where:

length	=	the length of the pipe in feet,
epoch	=	the number of years between the installation date and 1996
press	=	the pressure in the pipe in psi,
steep	=	*1* if the pipe is on a steep slope, *0* otherwise,
mat	=	*1* if the pipe is made of galvanized iron, *0* otherwise,
soil	=	*1* if the pipe is located in TB or VL soil, *0* otherwise.

Spatial and statistics modeling tools

Spatial statistical methods incorporate spatial correlation according to the way geographical proximity is defined. Proximity further depends on the geographical information, which can be available at areal level or at point-location level. Areal unit data are aggregated over contiguous units (census zones) which partition the whole study region. Proximity in space is defined by their neighboring structure. Point-referenced or geostatistical data are collected at fixed locations (failure location) over a continuous study region. Proximity in geostatistical data is determined by the distance between sample locations.

Geographical data are correlated in space. Data in close geographical proximity is more likely to be influenced by similar factors and thus affected in a similar way. In the case of water pipelines failure, spatial correlation between breakage events and environmental factors and also break clustering was observed (Sundahl, 1997). The author compared number of breaks that occurred within a radius of 200 m and within a period of 2, 6 and 12 month from a previous break. He found that in the old part of the city, the spatial and temporal clustering of breaks was higher than in the newer parts.

Consideration of the clustering phenomenon in pipe breaks in Winnipeg (e.g. Goulter and Kazemi 1988; Goulter, Davidson and Jacobs 1993) showed that independent breaks as breaks that occur more than 90 days after and/or more that 20 m from a previous break. An independent break is often the first in a cluster of breaks. They applied a linear regression to 390 km of 150 mm cast iron water pipelines with about 3,550 breakage events recorded in Winnipeg. Initially they applied this regression equation to all the recorded breaks and

obtained coefficients of determination ranging from $r^2 = 0.704$ to 0.937 for the three age groups. A high correlation means that pipe breaks were uniformly distributed along the pipe. The addition of pipe age into the regression model improved the predictive power of the model marginally for relatively new pipes, and significantly for old pipes. The authors attributed this correlation with age to different manufacturing, installation and operation practices that were typical of different age groups of pipes. The authors further observed that these differences could be classified geographically and that the age (or rather the "vintage") of a pipe may be a convenient surrogate measure which may be gathered and managed in a Geographic Information System (GIS).

ANNs-based modeling

An artificial neural networks (ANNs) is a system composed of simple processing elements operating in parallel, whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes. The development of a neural network model requires the specification of a "network topology", a learning paradigm and a learning algorithm. Unlike the more commonly used analytical methods, the ANNs is not dependent on particular functional relationships, makes no assumptions regarding the distributional properties of the data. This independence makes the ANNs a potentially powerful modeling tool for exploring nonlinear complex problems (Olden and Jackson, 2002; Mas et al., 2004). According to published literature on ANNs various applications, its strength lies in its ability to handle non-linear functions, to perform model free function estimation, to learn from data relationships that are not otherwise known and, to generalize to unseen situations. ANNs have been shown to be universal and highly flexible function approximators or any data. Therefore, ANNs make powerful tools for models, especially when the underlying data relationships are unknown (Mas et al., 2004).

The use of neural networks has increased substantially over the last several years because of the advances in computing performance and the increased availability of powerful and flexible ANNs software. Recent ANNs applications include Rainfall-runoff modeling (Chen & Adams, 2006), Optimization for water and wastewater treatment (Legube et al. 2004), Demand forecasting (Elkateb et al., 1998), Forecasting chlorine residuals in a water

distribution system (Bowdena et al., 2006), Classification of buried pipe defects (Sunil & Fieguth, 2006), Spatial interpolation (Rigol et al., 2001), Drought forecasting (Morid et al. 2006, Mishra & Desai, 2006), Water quality modeling (Chau, 2006). Ahn et al. (2005) used an ANNs model to predict pipe breaks in Seoul city mains network. The ANNs model gave a good performance on detecting the pattern of pipe breaks basis on seasonal variation.

The ANNs model has been applied to the water distribution network of subdivision in Edmonton, Canada (Rajani and Kleiner, 2001). The model was trained with historical input data including temperature, rainfall, operating pressure, and number of breaks. However, some main physical and environmental factors were not included in the model, such as pipe age type, diameter, and soil properties.

Sunil and Fieguth (2006) combined neural networks and concepts of fuzzy logic for the classification of defects by extracting features in segmented buried pipe images. Among the comparison between back propagation neural networks and neuro-fuzzy projection network classifiers, they concluded that the proposed neuro-fuzzy classifier performs the best, with classification accuracies around 90% on real concrete pipe images. Moselhi and Shehab-Eldeen (2000) also developed the ANNs in the analysis and classification of defects in sewer pipelines. Their model was trained to classify four different types of defects including cracks, spalling, joint displacements, and reduction of cross sectional area.

Additionally, Al-Barqawi and Zayed (2006) used a supervised ANNs with the back propagation algorithm to develop the condition rating model for water mains. They put eight input factor to prediction model.

Bowdena et al. (2006) developed general regression neural networks for forecasting chlorine residuals in the Myponga water distribution system up to 72 h in advance. They demonstrated that ANNs not only is capable of forecasting chlorine residuals but it also provides better predictions for this case study.

Pijanowskia et al. (2002) and Mas et al., (2004) integrated ANNs and GIS to forecast land-use change, where GIS is used to develop the spatial predictor variables. Four phases were followed in their researches: (1) design of the network and of inputs from historical data; (2)

network training using a subset of inputs; (3) testing the neural network using the full data set of the inputs; and (4) using the information from the neural network to forecast changes. They concluded that ANNs constitutes a powerful alternative in spatial land-use change processes modeling, when more conventional models obtain poor performance. However, it is probably impossible to develop models of land-use processes, which present a high power of prediction because these processes depend upon very diverse factors from environmental to socio-economic and cultural that are changing over time. They recommended ANNs model for future prediction that helps the environmental planners and managers to develop policies aimed at controlling the adverse ecological and social effects of land-use changes.

1.8 Pipe Rehabilitation Planning

While water supply systems getting older, pipe rehabilitation planning is being given more and more attention both from the water industry and from the research community. The emphasis for urban water engineers now and in the future increasingly lies not in new installation but in the evaluation and rehabilitation of existing networks. It is estimated that while expenditures on new work are dropping rehabilitation work is increasing at an annual rate of 25 percent (Thomson J., 2006).

The literature review identified numerous methodologies for prioritizing water pipelines renewal programs, many of which relied to some extent on the availability of water pipelines break data (Davis, 2000; Lei and Saegrov, 1998; Kleiner, 2001). The main objective of rehabilitation planning is to ensure the required performance of the system and maximise the economic efficiency of the operation. There are three major performance indicators that have to be considered:

- Hydraulic performance,
- Water quality, and
- Reliability of the service.

Most of the existing rehabilitation planning methods require an understanding of the pipe deterioration process. Pipe deterioration modeling (asset modeling) was classified into two

groups: physical and statistical. Indeed, a third group of models was identified that represents an integrated approach based on the data mining techniques and hydroinformatics.

1.8.1 Hydroinformatics

Hydrinformatics is the discipline that provides a framework for development and application of advanced innovative techniques for water distribution network management. Savic et al. (1997) indicated that two major tools particularly suitable for water industry applications are geographic information systems (GIS) and data mining techniques, such as artificial neural networks (ANNs) and genetic algorithms (GAs). The main purpose of GIS is to collect, store and manage the accurate and comprehensive network data. Data mining also referred to as knowledge discovery in databases, data harvesting, data archeology, functional dependency analysis, knowledge extraction and data pattern analysis, is the automated way to analyse large volumes of data to identify trends and patterns that are important for operation, maintenance and rehabilitation of water supply system. With advanced SCADA (supervisory control and data acquisition) systems and large asset, customer and maintenance databases, water utilities are facing the challenge of efficiently extracting useful information from data. Data mining techniques can be used for different purposes. ANNs can be used for demand forecasting (Bougadis et al., 2005) and for scanning large amounts of data (both operational variables and historical records) to identify a failure event or to estimate failure patterns. GAs can be utilized for optimization of system design, operational decisions and maintenance plans (Savic et al., 1997).

1.9 Summary and Needs for Additional Research

In this short chapter, pervious laboratory researches, theoretical studies, and discussions on the behavior of deterioration and failure modeling of pipe assets in urban water supply systems, during the period 1974 to 2006, are summarized and reviewed. References were sought from the UK, Europe, Australia, Canada and the USA.

The first section describes the sequential process of water pipelines deterioration. This review is followed by a discussion of the analysis of factors which affect pipeline's failure and mechanism of breaks. Failure management cycle in the urban water supply systems is

explained. Section 1.6 shows how water pipelines failures can be quantified. A brief overview of existing water pipelines failures modeling approaches are presented in section 1.7. At the end of the chapter, pipe rehabilitation planning in water supply systems is discussed.

However, A pipe failure in a water distribution network is a complicated event, which usually results from a combination of several factors. Water network must be analysed individually to determine which variables are responsible for pipe failures. The main obstacles in developing a physical model for pipe failures are the lack of knowledge of the strength of the system and the external variables which act on each pipe. To overcome this difficulty, a statistical model based on analysis of historical failures could be used.

As noted in the literature review, there has been relatively success in developing models for deterioration process and detailed prediction of pipe failure. In the most cases, models have been correlated with a number of factors such as pipe material, age, diameter and loading conditions. However, the correlation data sets were quite low, bringing into question the validity of the overall approach. Therefore, the relationship between water pipelines failure and related factors requires more exploration, including whether statistical investigation, ANNs predictive model and survival analysis can be performed. Of course, integrated modeling between recently techniques to accurately quantifying in order to implement effective renewal plans for water distribution systems is still important.

Given that no universal rules exist for selecting pipes for replacement, this study sought to determine influential factors and time of failure relationships for the City of Sanandaj's water distribution system by analyzing the data available for the City's system. These results will then used to develop an evaluation process that would identify pipe segments with the highest priority for replacement.

2. Data Collection and Elementary Analysis

2.1 Introduction

Based on the research problem and the literature review, deterioration of water pipelines and prediction of future failure is an important issue in water network management and crucial factor in establishing the water pipelines renewal priorities. The aging of water supply infrastructure systems, coupled with the continuous stress placed on these systems by operational and environmental conditions, have led to their deterioration which manifests itself in the following (Kleiner, 1997):

- Increased rate of pipe breakage due to deterioration in pipe structural integrity. This in turn causes increased operation and maintenance costs, increased loss of (treated) water and social costs such as loss of service, disruption of traffic, disruption of business and industrial processes and disruption of residential life.
- Decreased hydraulic capacity of pipes in the systems, which results in increased energy consumption and disrupts the quality of service to the public (Adams and Heinke, 1987).
- Deterioration of water quality in the distribution system due to the condition of inner surfaces of pipes which may result in taste, odour and aesthetic problems in the supply water and even public health problems in extreme cases.

It has been reported that the distribution system often involves 80% of the total expenditure in drinking water supply systems (Clark et al., 1988). Given the reality of scarce capital resources, it is imperative that a comprehensive methodology be developed to assist planners and decision makers in finding the best (most cost-effective) rehabilitation policy that addresses the issues of safety, reliability, quality and efficiency.

This chapter is concerned with the comprehensive methodology for advancing research besides data collection in study area and elementary analysis of data. In first section research methodology is discussed. The proposed methodology is demonstrated through the

application to the pilot area in water network of Sanandaj city in Iran. Section 2.3 summarizes the approaches used by author to collect data requirements for the modeling. Section 2.4 describes preliminary statistical analysis to give insight on the impact of different risk factors on the structural deterioration of water pipes. In the rest of this chapter spatial overview is presented of existing data that were employed during the thesis. This section addresses a process for using integrated spatial and statistical analysis to discover not only the distribution of water pipelines failure in space but also indicate various spatial trends in failure. The last two section will be introduction to modeling approaches in chapters 3, 4 and 5 .

2.1.1 Methodology for present research

The following six-step process highlights the selected methodology in this research:

1. **Step one**: a literature search to obtain published information on the water pipelines failure analysis and modeling,
2. **Step two**: Collection of water pipelines failure information in selected area through a literature survey and combination of different database included water distribution failure database, customized ArcView/GIS and calibrated hydraulic model (Epanet), and deep interview with technicians and crew,
3. **Step three**: Elementary analysis of historical failure data based on statistical methods to determine factors which affecting progression of water pipelines failure as well as application of spatial analysis includes clustering and spatial interpolation methods to provide scientific reasons for depicting spatial relationships and the strength of dependencies between failure incidents, environmental and hydraulic variables, and the other geographic factors,
4. **Step four**: Review of univariate statistical inferences, indices of bivariate relationship and multivariate data analysis to assess correlation between the affecting factors and identify the important variables for the occurrence of failures on the water pipelines as well as fitting two regression model namely Multiple and Poisson,
5. **Step five**: Application of Artificial Neural Networks (ANNs) models to predict number of failure in all water pipelines and 3 range of materials,

6. **Step six**: Implement of non-parametric and parametric survival models for time to failure of water pipelines to quantified various rate of renewal over the mains network on the percentage of failures which avoided from this network.

2.1.2 Data collection plan

One of the most critical and expensive parts of any pipeline failure program is the collection of reliable data. We tried to have a short, but reliable, record period which is better than a longer, but less reliable period. With in the scope of this research, failures data in the central part of Sanandaj city (Iran) have been collected. The data collection methods used during the course of the study is illustrated in Fig. 2.1 It contains both map data (depicting location of failures) and attribute data (describing physical characteristics of each failure). The data gap was further reduced based on field notebook, census material, maps and as-built drawing as well as interviews with SWWU staff.

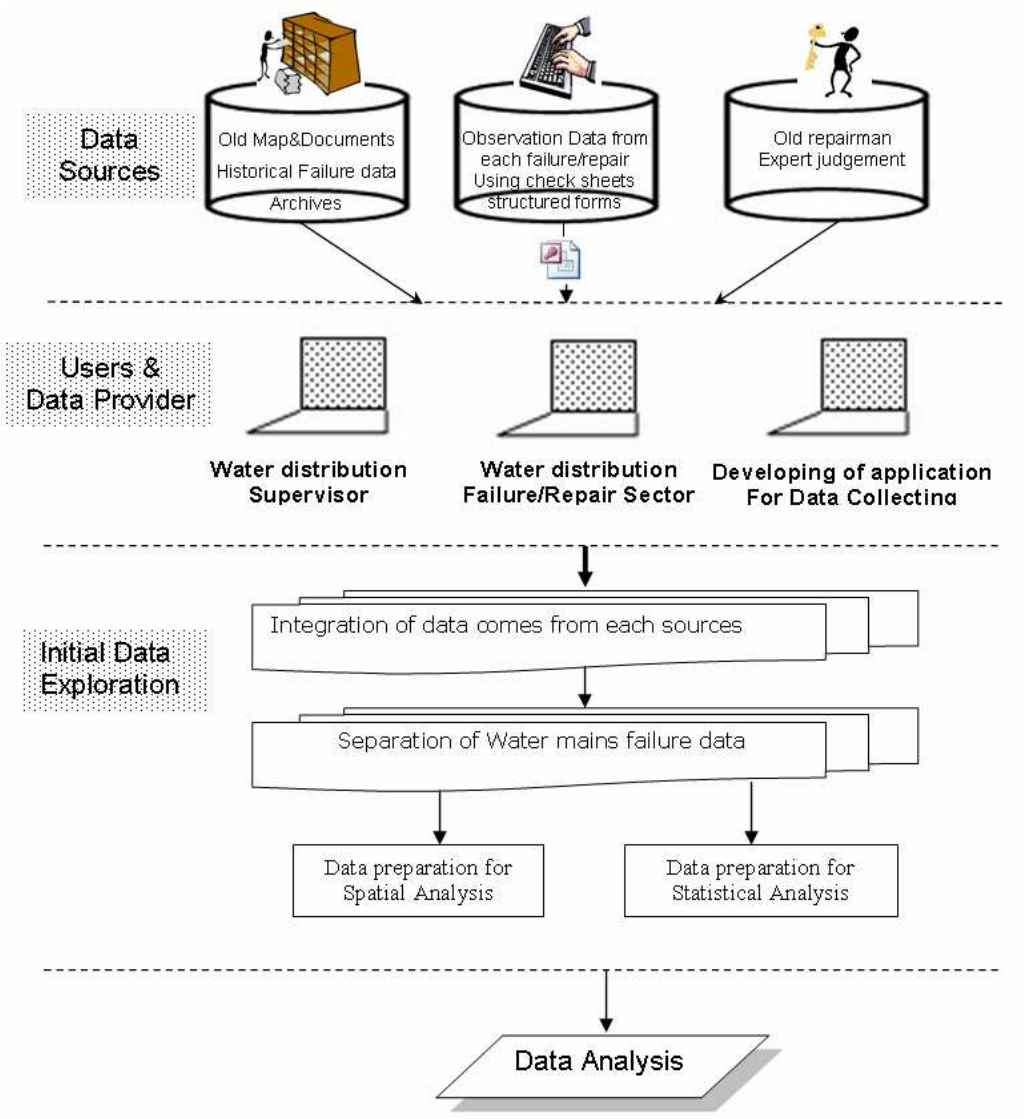

Fig. 2.1 Data collection process outline

2.2 Study Area Description

The area chosen for this study, Sanandaj city in the west of Iran (approximately longitude: 46.99 and latitude: 35.32), is an extraordinary case due to the unusually high number of mains breaks (in average: 67 failures/100 kilometer/year).

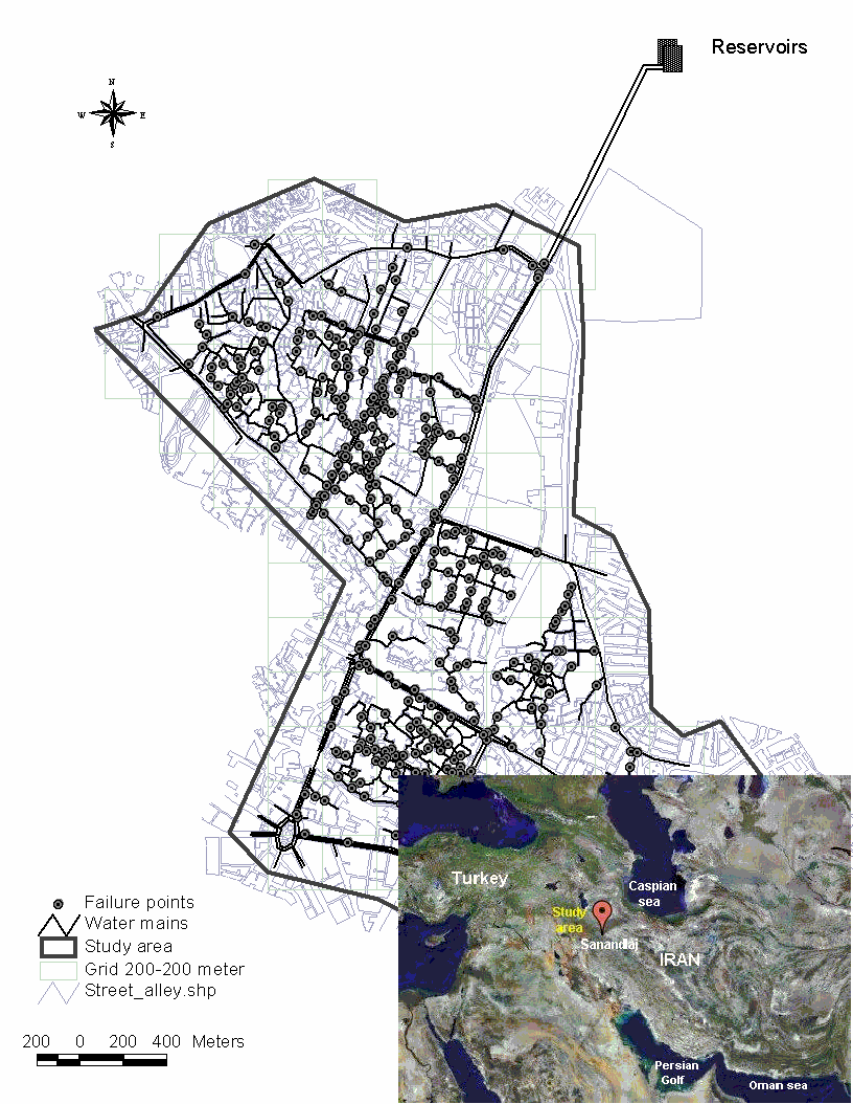

Fig. 2.2 Location of study area and selected part of water pipelines network

The city has a young distribution system, 70% of it built since 1960 and 30% since 1995. The SWWU operates a water distribution system covering approximately 90 km^2 and providing retail water service to more than 210,000 people and it maintains 250 kilometers of water mains, from 50 to 800 millimeters in diameter. The majority of pipe material in the past was cast iron , ductile iron , galvanized iron and asbestos cement. At recent years, since 1997, polyethylene pipes have been used quite extensively. Average water use was in the range of 35,000-45,000 m^3/day. Two major sources of drinking water, surface water and groundwater, supply the network. Treated water is delivered from the water treatment plant to the ground reservoirs by pumping. Water storage in the distribution system is provided by twelve

elevated storage facilities to maintain pressure levels in the gravity-pressure system. The pressure in the network is between 0.5 and 7.5 bar.

This work focuses on the pilot area which was found in a high-density residential and commercial area in the center of city. The pilot area serves approximately 11000 connections, corresponding to 554 mains segments. Its distribution system consists of nearly 56.7 km of water pipelines from 63 to 800 mm in diameter made variously of cast iron (13%), ductile iron (21%), asbestos cement (36%), and polyethylene (30%).

2.2.1 History of water pipelines failure in study area

The term "*failure or break*" in this study is taken to correspond to an entry on a water main section repair report sheet and constitutes a single repair event. A section of water main is defined by its "from and to" nodes. A node is a connection between two pipes (e.g., a tee or cross), a valve or a change in pipe characteristics (e.g., diameter, material).

Every year, the SWWU reports approximately 200 water pipelines breaks, most of which are minor. Water pipelines breaks may temporarily halt water supply to households and businesses in the surrounding areas. Breaks can also result in property damage, street and sidewalk closures, and traffic and business disruptions. Within the reference period 1995 to 2004, the water pipelines failure database contained 395 reported incidents.

Fig. 2.3 shows the annual number of breaks which presents the pilot area has experienced approximately 35 to 45 failures per year. From a material point of view, polyethylene and cast iron pipes had a higher number of failures than other materials. In ductile iron (DI), for the whole period, there is no overriding tendency and it is nearly horizontal, which indicates little deterioration rate. The blue line, asbestos cement, shows softly degradation of pipes. In this graph, it is apparent the cast iron deterioration was increasing until 2003 but has stabilized in the 2004. The annual variation of failure in polyethylene pipelines will be explained in Fig. 2.12. Finally, the thicker line which indicates the total annual number of failure, shows from one year to the next year there are variations. This graph indicates that the number of breaks has, in general, been increasing steadily.

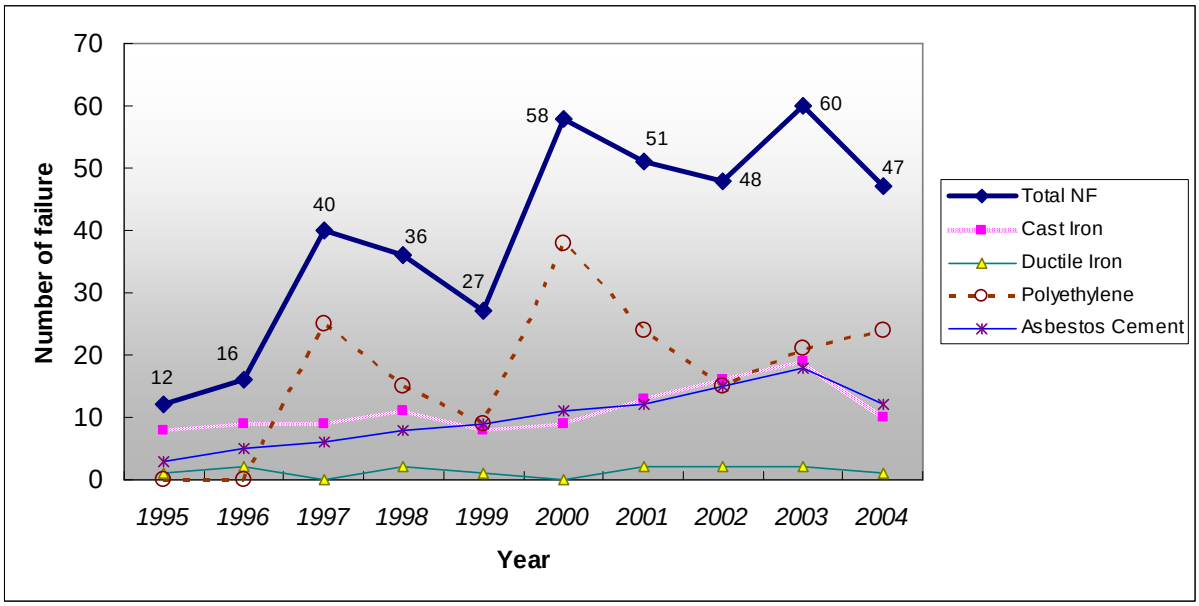

Fig. 2.3 Annual number of water pipelines breaks from 1995 to 2004

2.3 Description of Existing Data Sources

The results of the literature search and this work describe that difficulty in developing mathematical models for this type of problem was the lack of data on both the water pipelines network and the pipe breakage history. It constraints not only the type of models used but also their predictive capacity and their accuracy. Of course, the acceptance of poor and missing information is because the network is widespread and hidden, and because it is underground. Therefore, it is hard to measure data.

In the actual sampling area, data were available from multiple source records over the past years (Fig. 2.4). After most installation year in 1965, there was no recording system to pipeline's data and just the technicians who were working keep the data in the mind. Since 1995, the responsibility of water network in SWWU was changed and a paper based system developed for documentation of data in pipelines and failures. The network asset details of Sanandaj Utility's water network had been captured on an Autodesk system (AutoCad) and GIS since the early 1998s. After 2 years, the utility maintain computerized records of pipe breaks (failure database). These major sources of data considered for the analysis are briefly described below.

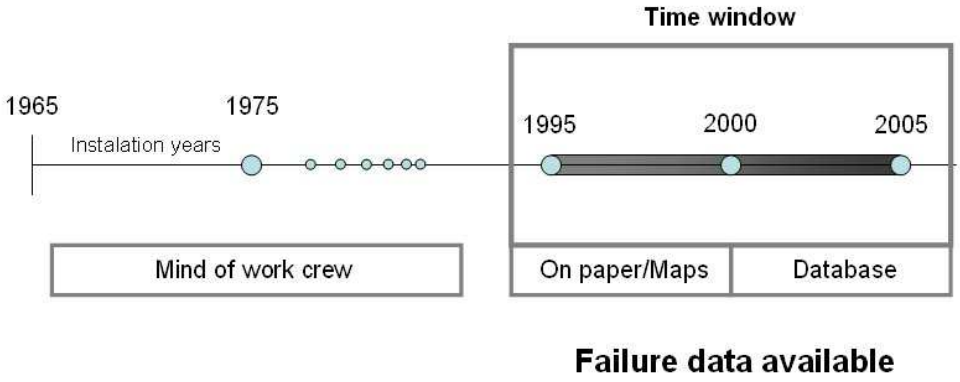

Fig. 2.4 Data availability on water pipelines failure in this research

2.3.1 Data quality

The accuracy of any prediction is directly proportional to the quality, accuracy and completeness of the supplied data. Good data, along with the appropriate model choice, usually results in good predictions. Consequently, considerable effort was required to perform quality assurance on the data.

Prior to analysis, quality control techniques were used to eliminate bad data, such as those pipes with failure dates before the installation date, and those pipes which could not be classified into categories of interest for this study. A lot of the data that has been recorded is of dubious quality, being recorded or inputted wrongly. Research in European countries has identified that a smaller amount of more accurate data can lead to better results than more complete, but uncertain data (Gat and Eisenbeis, 2000). The unreliable nature of the data creates several problems in modeling the deterioration process of pipes. The most obvious one is that with a lack of failure history it becomes very difficult to estimate the failure rates of pipes. This has been a major shortcoming in previous research and has resulted in the dominance of engineering judgment in the decision making process. Therefore, in this study 395 case of water pipelines failures were considered to the analysis and modeling.

2.3.2 Internally published reports and drawing

Documents reviewed during the study cited the SWWU's organization. The first data set includes internally published reports in technical and operation bureau. The contents of the reports have been reviewed and relevant information was summarized in database.

Table 2.1 Recommended items for water pipelines failure data collection

Field	Description	Purpose/Remarks
Water mains failure ID	A unique identifier assigned by the utility to track the incident and related data	Provides a common identifier for data analyses.
Water pipelines ID	A unique identifier to track the pipe on which the break occurred.	Provides a common identifier for data analyses.
Installation year	The year in which the pipe that failed was installed	Useful for analyses of main break trends.
Address	The nearest address to the location of the break	Spatial analysis of main breaks
Pipe material	The material of the pipe	Trend analyses of main breaks by pipe material
Pipe diameter	The nominal diameter of the pipe	Trend analyses of main breaks by pipe diameter
Length	The length of the pipe that failed. This length refers to the length recorded in a network inventory.	Needed for developing pipe specific replacement program.
Surface and traffic	Describes the surface under which the pipe is laid in order to estimate the traffic load the pipe experiences.	Helps to determine the cause of the main break. It will be used to examine trends in main breaks.
Depth of pipe	The distance from the ground surface to the top of the pipe.	Helps to determine the cause of the main break.
Type of failure	The type of water pipelines break, based on visual a observation	Allows for analyses of main break trends.
Probable cause of the failure	The most likely cause of the main break, based on a visual observation	Can be used to examine trends in main breaks.
Pipe wall thickness	Thickness of the pipe wall, typically from a pipe's product catalogue	Helps to determine the cause of the main break.
Maximum pressure in area of break	Describes the maximum pressure in the pipe that failed.	Useful for analyses of main break trends.
Date of excavation	The date on which the water pipelines was exposed to repair the main break	Temporal analyses of main break trends
Employee name	The name of the utility employee who provided the field data	Follow up of questions

A description of each data item requested is provided in Table 2.1.

For the city of Sanandaj, large amounts of data concerning the pipe inventory and failure history are available on the paper records such as repair Kardex cards and large-scale drawings, typically 1:2000 scale. These maps show the location of a pipe, diameter, material, and depth of burial. In most cases, the date of pipe installation is usually not included in the drawings but they collected by interview the old crew and customers. In addition, the Kardex cards contains descriptive information includes the failure location, dates of pipe repairs, type of repair, and some special notes on repair activities.

Another paper records was leakage data which are gathered by contractors personnel. There was one record for each leakage that has occurred in SWWU since the beginning of 1998. Each record contains the leak location, name of the street on which the leak occurred, leakage detection date, pipe material, pipe diameter, pipe installation date, and other relevant information about the leakage location. The advantage of using these leak records for creating an accurate pipe inventory arises from the fact that these records contain the information, which observed during the leak investigation.

2.3.3 Water network failure database

In urban water utility, a few cities in Iran have started to use computer based records during the last 5 years. Most water utilities already have procedure and forms in place for recording basic information related to breaks which defined by NWWEC. These records filled out by office personnel and field crews in responding to breaks in each component of network.

Initially, part of this research effort was devoted to preparing the software for collection mains failure in study area. It is a user-friendly application and was programmed into Access 97 to record data for each failure case. Shortage knowledge of SWWU's technicians and crews in English, all forms and interaction pages in database were converted in native language (Persian) which improved the quality and accuracy of data.

The database contains daily recorded information of contractors who repaired damaged water network components. This software was installed in Failure and Repair Division of the SWWU dating to 2000. The older breaks since 1995, i.e. those that occurred prior to the creation of the database, were entered manually into the database during this research.

In MS-Access, the electronic forms provide ways of collecting and documenting main breaks and pipe condition information using pick lists and standard input. Therefore, it reduces data entry errors, data redundancy and preserves data integrity.

The application has three basic tables, pre-constructed queries and macros, and customized menu items to store the data and report of query result. The pipeline's table includes *id*, *material*, *diameter*, *wall thickness*, *depth of burial*, *category of upper street*, *age*, *segment length* and *hydraulic pressure*. Failure attributes include information concerning positional and temporal data such as *X* and *Y coordinate (address)*, *cause*, *type* , and *time of failure*. Finally, administrative data will include *character of operators* and *repairing equips*, *security and classifying codes* , *cost data (equipments, material)*. The three tables have been joined via the common *id*. Fig. 2.5 illustrates the algorithm used for collecting the water lines failure.

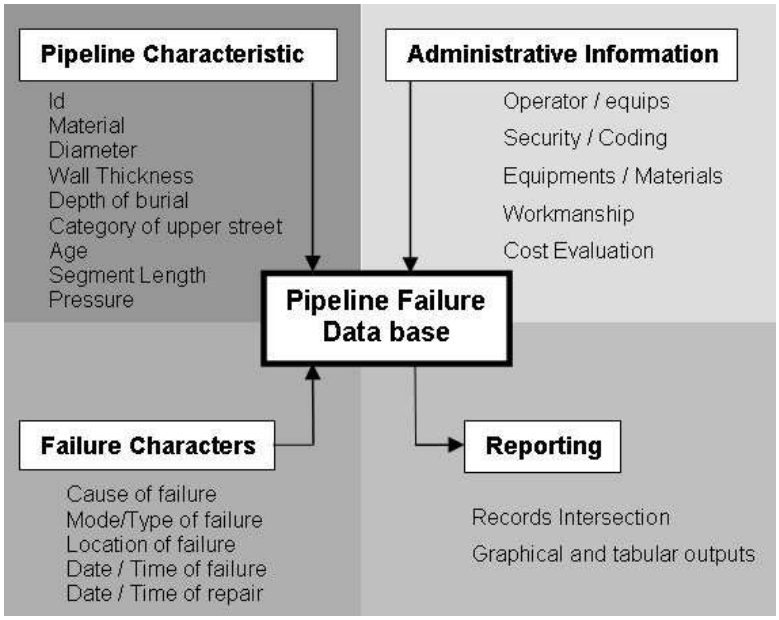

Fig. 2.5 Schematic description of water pipelines failure data collection application

Reporting process on pipe breaks in SWWU

When a break occurs, or water is appeared seeping from the ground, a local resident places a complaint by call. This complaint causes the SWWU water operators and equips go to the site and make an observation. Firstly, the result will come in a report paper sheet, which explains their investigation toward the leak, or break. SWWU's major intent for filling out the break

report is to keep track of the work SWWU employees and enter in the database. In study area, the following basic steps were typical of water utility's response to a mains break situation:

Report a Leak or break Customers call to SWWU about breaks or leaks 24 hours a day, including weekends and holidays. They give the street address or the SWWU costumer reference number.

Identification of problem A call from a customer of the system or an observation by a utility employee are typical examples of how a distribution system problem is first noted. Investigation of the problem will be initiated immediately or scheduled for the future depending on the perceived severity of the event.

Investigation of problem An experienced utility employee is dispatched to investigate the problem. The investigator identifies the nature of the problem (i.e., water pipelines break causing property damage, service line leak, etc.) where exactly the pipe is broken and initiates remedial action based on the severity of the incident.

Remedial action The solution to the problem may involve a temporary repair, a permanent repair or replacement, or notification of third party (e.g. for service line repair outside the utility's responsibility).

Follow-up Follow-up steps include recording crew times and other information related to the costs associated with the event, updating maintenance histories, alerting customers to the status,...

Recording data Each of these steps in the process generates data and they were recorded initially on the paper based forms and then transferred on the database.

The failure informing and repairing process stages are illustrated schematically in Fig. 2.6 and input requirements and the expected output data from the various stages can be seen.

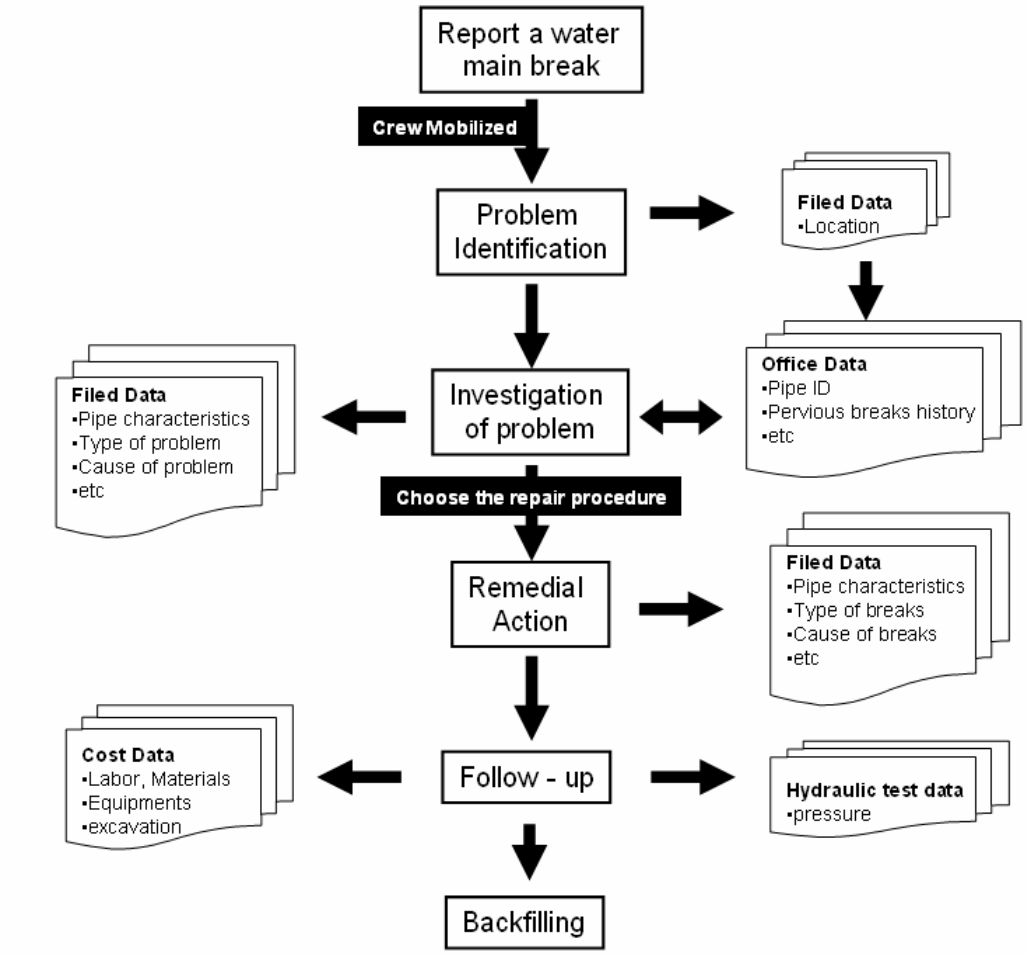

Fig. 2.6 Chain of events leading to a repair and data generated at each response step

2.3.4 GIS for water network and mains failure

Geographic Information System (GIS) has been defined as a system for capturing, storing, checking, integrating, manipulating, analyzing and displaying data which are spatially referenced to the earth (Musgrave, 1996). The backbone of both statistical and spatial data analysis in this research relies on the GIS-based user interface which was developed in ArcView 3.2a (Esri, 1999). Section 2.5 will illustrate the overlay, proximity and point pattern analysis of water pipelines failure via GIS.

In first stage, SWWU completed conversion of the paper version of the pilot water system to a geodatabase. The conversion was performed by digitizing the water system grid sheets into a

GIS. Further, failure data from related database was imported into GIS. This integration gives all statistical and customizing capabilities of Access and all the functionality of ArcView.

In terms of GIS, water pipelines failure is a phenomenon which can be expressed through occurrences identified as points in space. Generally, failure location was reported by a crew through measuring the distance from the reference point using a tape measure. The crew would estimate the distance of the failure location from a known reference, normally the nearest street, alley or intersection. Clearly the estimation was subject to errors but this work tried to minimize it. Technological advancement calls for a need to revolutionize this approach and there is a strong recommendation that global positioning system (GPS) receivers be adopted for capturing failure location in SWWU. Failure mapping starts with geocoding process that matches an address of breakage to a physical location (as a dot) along a street. Each failure point is converted manually into coordinate locations via street or address matching. At a scale of 1: 2000, the reference map provides a suitable level of detail to approximately locate property boundaries, streets and distribution mains features. Accordingly, the water pipelines network and failures were digitized visually against a backdrop of streets, property parcel and buildings. The process of identifying the coordinates of a failure point in ArcView 3.2a, is to use "*getx*" command for the x-coordinate and "*gety*" command for the y-coordinate. For means of the coordinate projection system, this chapter considered the UTM system. Both x and y are defined by distances in meters from an arbitrary reference point. These projected coordinates are added to the original data record and read directly into the spatial and statistical analysis.

When an address list is transformed into a set of coordinate points, corresponding failure attributed data are then imported from Access into ArcView. The platform in ArcView is connected with external databases using the SQL connection feature. In order to combine and join data to a point shapefile of locations according to each failure, a common field such as failure ID should exist in both tables. By joining this field between tables, data can be retrieved from each table and combined into another table. The data is then saved as a shapefile in GIS. In fact, each failure on the pipeline can be created by a common ID number in ArcView and Access databases. Therefore, the related non-spatial data about failure is stored in MS Access and graphical information is established in ArcView. Each record in the

tabular database is then connected with a failure point in GIS. Location of all of reported breaks for the 10-year data period were plotted on the map by small black dots (Fig. 2.2).

2.3.5 Water pipelines hydraulic model

To obtain the daily maximum pressure in water pipelines which the failure has been occurred on it, a calibrated hydraulic model was developed using the EPANET 2.0 software (Rossman, 2000). Data utilized for calibration were gathered during UFW study in the area by monitoring the actual pressure and demand on the site. In calibration phase, comparison of measured and simulated pressure at the specific location showed a mean relative different of 12.6% with range of differences being 0.6 to 24.6 % (Sanandaj's UFW report, 2000).

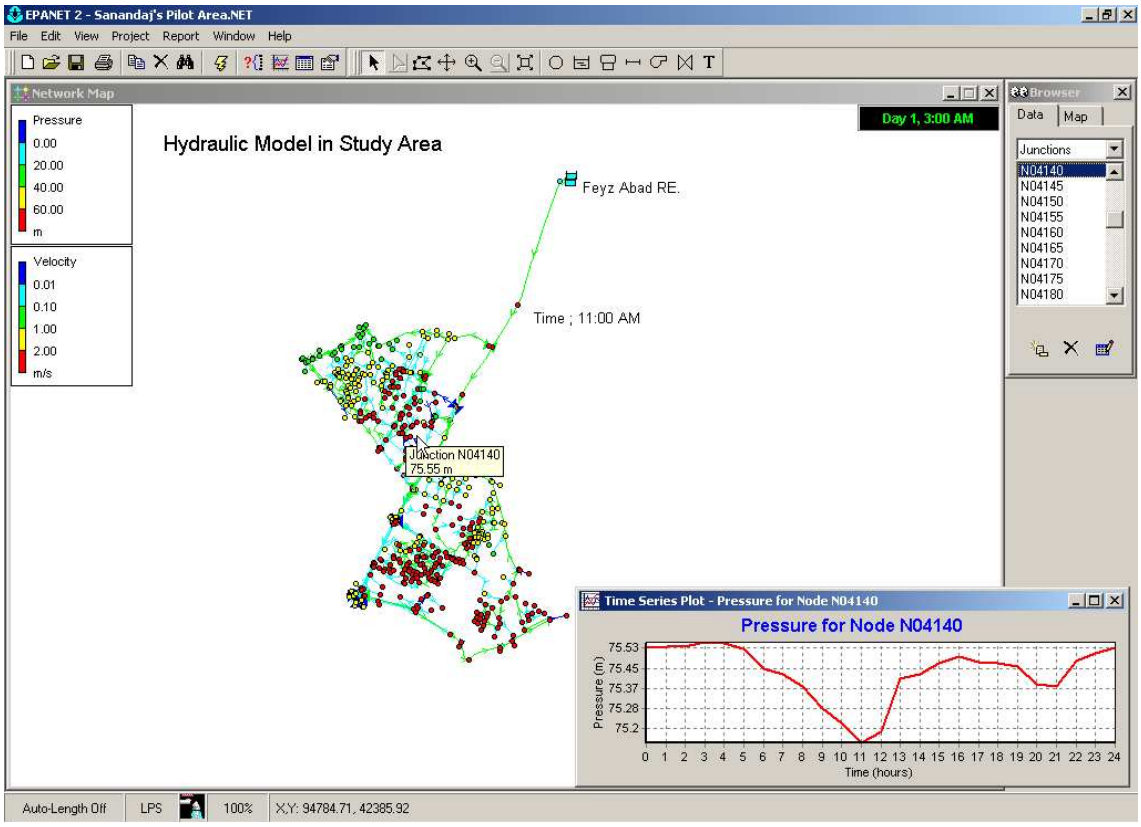

Fig. 2.7 Calibrated hydraulic model for water pipelines in study area

2.3.6 Interviews with technicians and supervisor

In addition to these data sources, interviews with laborers, supervisor, technicians and crew were undertaken for completeness, to ensure that no useful alternative data source was overlooked. During the data collection phases, not only in some cases there are a significant number of gaps in the failure data but also accessing to data was more difficult or expensive to obtain. For instance, pipe attributes like as location, material, diameter, year of installation and maintenance and past failures' data were in the mind of crew and technicians. In the absence of such data, an interview technique was used to treat these gaps and capture relevant information. The SWWU staff opinion has been also applied in verification of data.

2.4 Descriptive Statistics on Water pipelines Failure

Descriptive analysis organizes and summarizes the data and can be used to indicate various trends in failures and factors affecting pipe failures. Every effort to model the failure of a pipe network should begin with this basic analysis.

It should be noted that the deterioration processes of water systems are neither uniform nor identical. All the aforementioned studies show that water pipelines failure in general, is a result of various uncertain factors, some of which are site specific and hence they vary from one water distribution network to another. To assess the nature and frequency of failure as well as identifying the influential factors on it, the historical failure data of water pipelines in the City of Sanandaj (for the years 1995-2004) were collected and analyzed.

2.4.1 Factors affecting water pipelines breaks

In Europe, Australia as well as in US and Canada, there has been much research on the factors that contribute to pipe failures, with the goal of developing or improving predictive planning models. An extensive literature search and consultation with experienced water practitioners from SWWU, enable us to make the following categories for describing the deterioration of water pipelines in the selected network:

- Maintenance variables,
- Structural or physical variables,

- Environmental or external variables, and
- Operational variables

Therefore, identification of factors contributing to the occurrence of water pipelines failures in a given site was the first step to gain a better understanding of the failure mechanisms. In study area, the following factors were thought of as having the largest effect on the system of water pipelines deterioration and failures. This part addresses these factors which affecting pipe failure either time-dependent or static. For instance, pipe diameter or material are examples of static (i.e., will not change over time), while age of pipe is an obvious non-static factor.

Time dependent factors

Firstly, time-dependent factors which influence water pipelines breaks. The dataset from Sanandaj, Iran, present two factors depend on the time which deals with age of water pipelines and pervious number of failure.

Age

Generally, as pipe assets age, they tend to break more frequently. With respect to failure data, we found two behavior of failure in the material groups. In the plastic mains, polyethylene, the amount of failure was increased dramatically in the first 5 year after installation (Fig. 2.8).

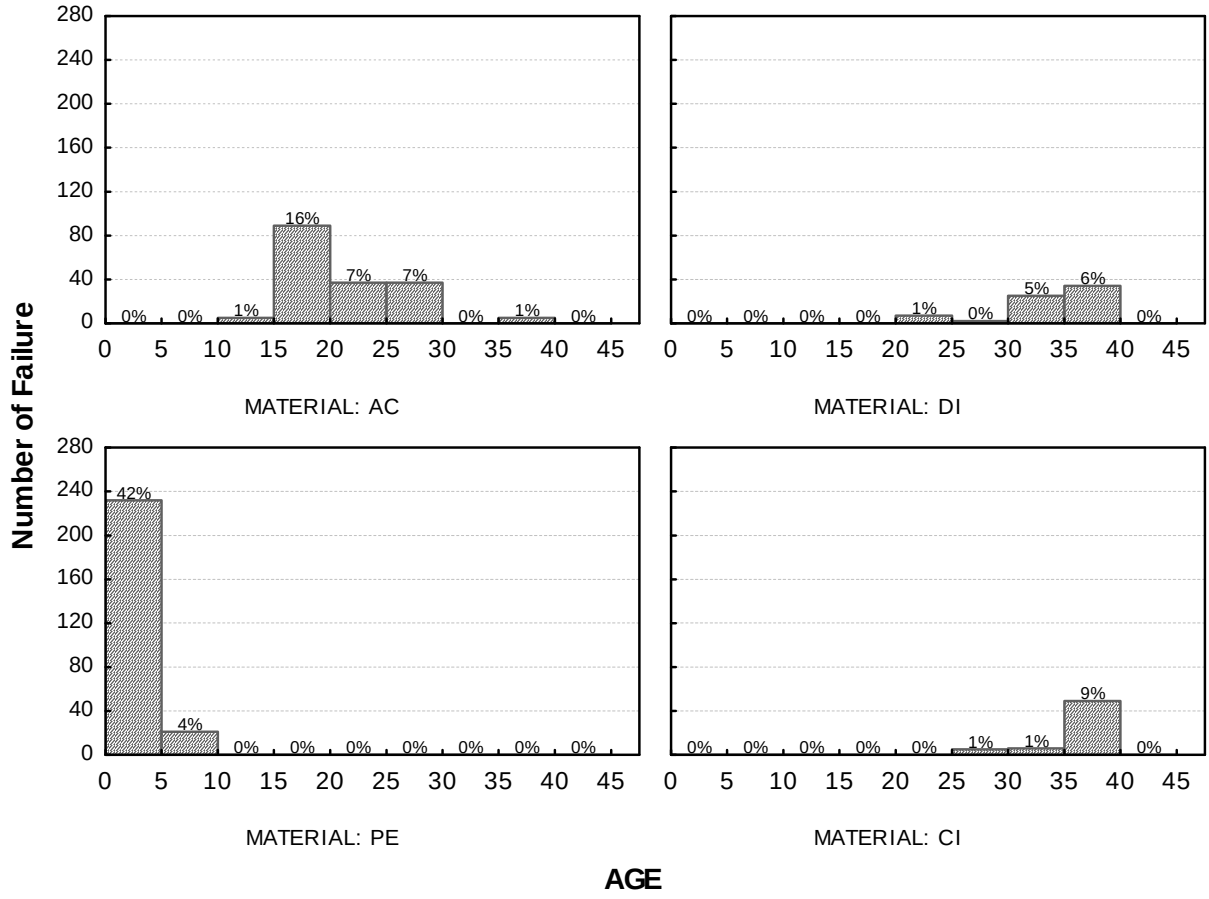

Fig. 2.8 Percentage of failures, by water pipelines material

In the second 5 year , the number of failures in water lines dropped off and then stabilized in recent years. In contrast, the non-plastic water pipelines showed a smooth increase in failure rates as pipes age. For example, cast iron (CI) material experienced 9% of total failure on the age of 35-40 years. Asbestos cement (AC) have experienced a increasing procedure until 20 years of old and after the number of failure was decreased. Fig. 2.8 looks specifically at the percentage of failure according to water pipelines age differentiated by the water pipelines material over the period 1995-2004. In database, the age variable is equivalent to the time between laying year and failure time.

Number of pervious failures (NPF)

For a repairable system, like water network, the time of operation is not continuous. In other words, its life cycle can be described by a sequence of up and down states. The system operates until it fails, then it is repaired and returned to its original operating state. It will fail

again after some random time of operation, get repaired again, and this process of failure and repair will repeat. Overall, if the times between failures tend to get shorter with age the item is said to be deteriorating. Alternatively, if the times between failures are increasing then the item is improving. Conceptually, the time-to-failure is decreased over the year. Figure below shows this interpretation.

The number of pervious failures (NPF) or cumulative number of breaks is a key factor in developing the prediction model (Pelletier, 2003). Clark et al. (1982) found that after first failure, the number of failure events increased exponentially with time. Goulter and Kanzemi (1988) observed the temporal and spatial clustering of water-main breaks, indicating that a previous break increased the likelihood of future breaks in its immediate vicinity. In Sanandaj dataset, we calculated the times between inter-failure. It was seen that the next failure times have been decreased (Table 2.2).

We also evaluated the percentage of failures which occurred after each 500 days from first break (Fig. 2.9). This graph shows that about 15.2 % of all subsequent breaks occurred within 500 days of pervious breaks. Additionally, we calculated the number of breaks in each statistical segment of water pipelines as well as their percentage. In study area there are 16 water pipelines (from 554 total mains) which had 4 failures or more and 332 segment registered no break at all within 1995 to 2004. The rest (206 mains) have failure between 1 to 3 . Fig. 2.10 shows in cast iron there is a water main with 9 failures. It means that this segment has priority for replacement. Meanwhile, polyethylene (PE) pipelines has maximum 6 failures at one segment. Ductile iron (DI) has minimum failure number on mains and present acceptable degradation. Most of Asbestos cement water pipelines had failure number between 1 to 3 .

Table 2.2 Time of subsequent failure in the water pipelines with more than 3 failure

Start time	Water pipelines ID	Material	Date of First failure	Time of subsequent failures (day)							
				2 th failure	3 th failure	4 th failure	5 th failure	6 th failure	7 th failure	8 th failure	9 th failure
01/01/1995	P03160	PE	11/10/2000	2	76	1087					
	P03165	CI	12/08/1997	272	545	1058					
	P04015	PE	28/10/1997	108	71	987					
	P04175	CI	16/08/1995	1665	1188	389					
	P04185	PE	28/07/1997	1206	833	572					
	P05009	CI	13/02/1995	1330	795	418					
	P07010	CI	13/06/1996	1646	630	437					
	P07110	CI	11/04/1999	637	941	110					
	P08135	CI	14/05/1996	1500	259	1161					
	P06030	PE	23/06/1999	96	97	367	1186				
	P14015	PE	28/08/1997	459	194	998	618				
	P04165	PE	24/08/1997	114	68	1056	923	233			
	P03520	CI	12/08/1995	869	248	257	858	223	389		
	P03500	CI	12/08/1995	1180	764	319	163	169	247	227	
	P03485	CI	14/04/1995	375	424	424	775	287	1	251	313

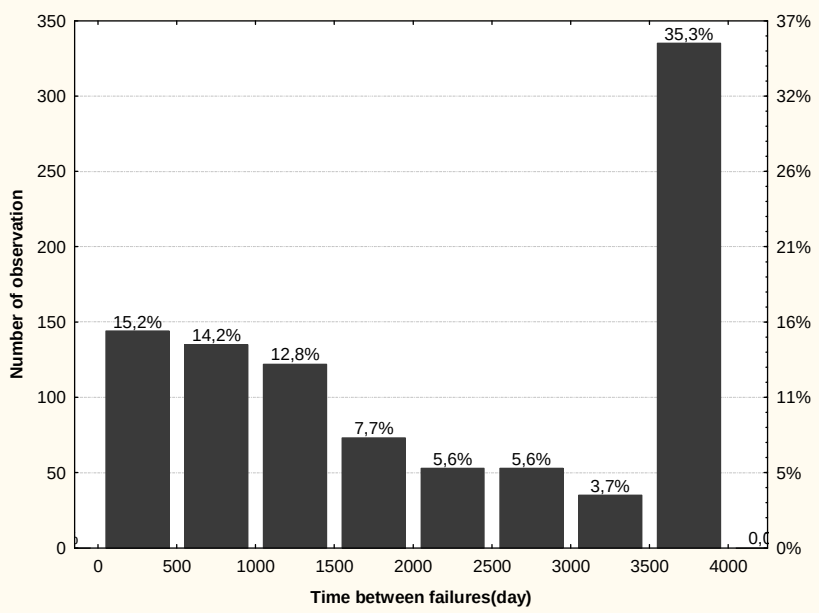

Fig. 2.9 The time between failures during study period

Fig. 2.10 shows the percentage of NF in each material of water mains.

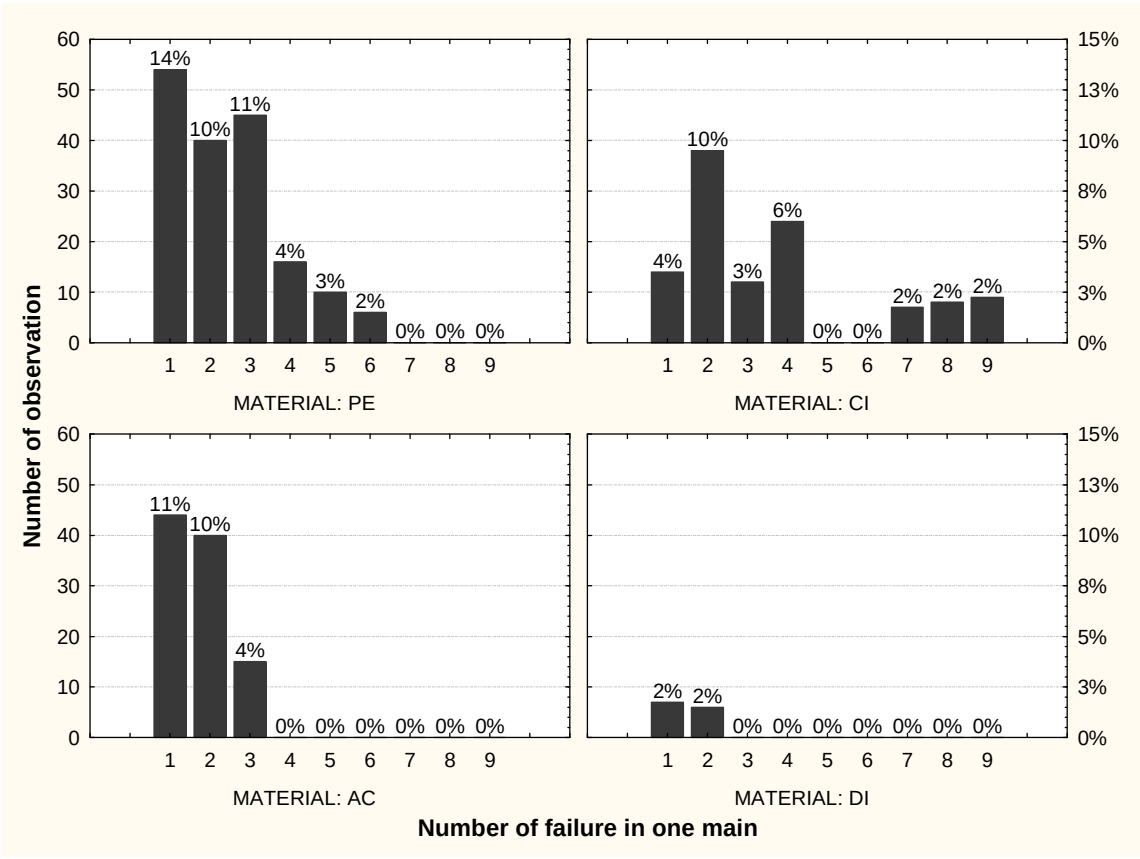

Fig. 2.10 Number of failure in each mains segment

From spatial point of view, section 2.5.3 evaluates the spatial pattern of failure in the study area. It shows 4 cluster of failure (Fig. 2.35) which report that approximately 16% of failures occurred within 10 meter of a previous failure, and 39% within 50 meters of another failure.

Static factors

These factors are static over time due to properties of the pipe and installation practice. They include pipe material, diameter, wall thickness, and depth of burial.

Pipe material

Similar to water pipelines age, the water pipelines material of construction information has been previously collected during system inventory updates. In the study area, pipe materials

are grouped into two categories: Rigid pipes (cast iron, ductile iron, and cement-based) and Flexible pipes (polyethylene). The material of a particular pipe depends on the year of installation and diameter. For large pipelines (with diameters over 200 mm) ductile iron (DI) and for small diameters polyethylene (PE) pipes were typically used. Cast iron (CI) and asbestos cement (AC) are still in service but no longer installed today, while mainly ductile iron (DI) and polyethylene (PE) are used for newer mains. Fig. 2.11 plots the number of failure in each material. Among the pipe materials used in study area, asbestos cement predominate for pipe dimensions of 100 to 200 mm. It represents about 36.1% of the total length of installed water mains. The majority of these pipes were laid in the 1970s and 1980s. After decades of service, the number of failures in AC pipes were increased greatly. But, most asbestos cement pipelines have exhibited a slightly reduced average failure frequency over the recently years (Fig. 2.8). Ductile iron pipe constitutes 20.8% of the mains network and the break rates for this material in 1995 and 2004 were 11.0 breaks/100 km/year. Cast Iron (CI) is the material that is most prone to failure. It accounts for about 13.4 percent of the pipes currently in use in the study area. These pipes yielded the highest failure rate of 147.4 breaks per 100 km of cast iron pipe in service per year. Lastly, polyethylene pipelines for period 1997-2004 have experienced 126.5 annually break per 100 kilometers of water mains. Overall it has been observed that before 1997 the failure rate for all four materials was very low and completely different than in years after 1997. This increasing after 1997 can be justified by the setting-up a paper based data collection system in 1997 and then in 2000 a computer data base. Table 2.3 summarizes different rate of failure in material categories of water pipelines in study area during 1995 and 2004.

Table 2.3 Water pipelines material in study area and failure rate

Pipe material	Length of pipe		Number of failure	Failure rate (per year in 100 km)
	km	%		
Cast Iron (CI)	7.6	13.4	112	147.4
Asbestos Cement (AC)	20.5	36.1	99	48.3
Ductile Iron (DI)	11.8	20.8	13	11.0
Polyethylene (PE)	16.9	29.7	171	126.5

More detail about failure behaviour of these materials were explained in Fig. 2.11

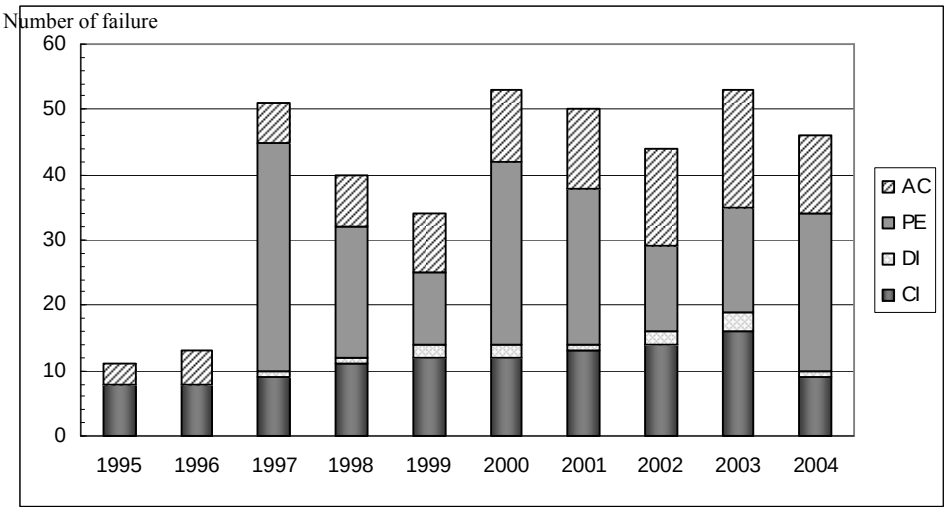

Fig. 2.11 Number of failure on each material by year

This initially statistical analysis indicates that almost half of all problems with PE pipe occur in the first year after installation. Fig. 2.12 represents the breakage pattern for polyethylene water pipelines during the last years. Since 1997 polyethylene pipes were laid, initial breaks have occurred in the early life of this pipe because of uneven setting and weak material. In December 2000 during a particularly cold period, the low temperature constitutes a major cause in the damage of polyethylene pipes (see Fig. 2.12).

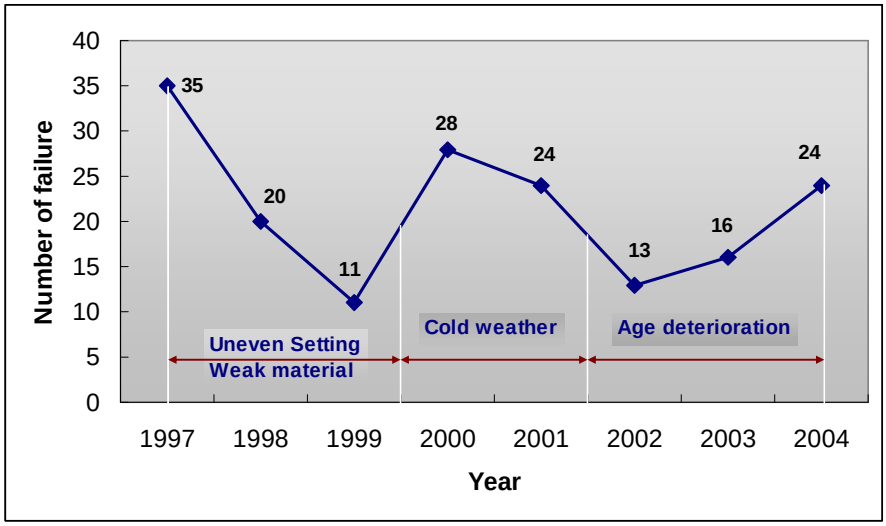

Fig. 2.12 Three phases in Polyethylene water pipelines failures in study area

Water main's diameter

Water distribution mains of various materials in study area are readily available in sizes ranging from 63 to 800 millimeter. Fig. 2.13 examines the data to see how failures were distributed among different line sizes.

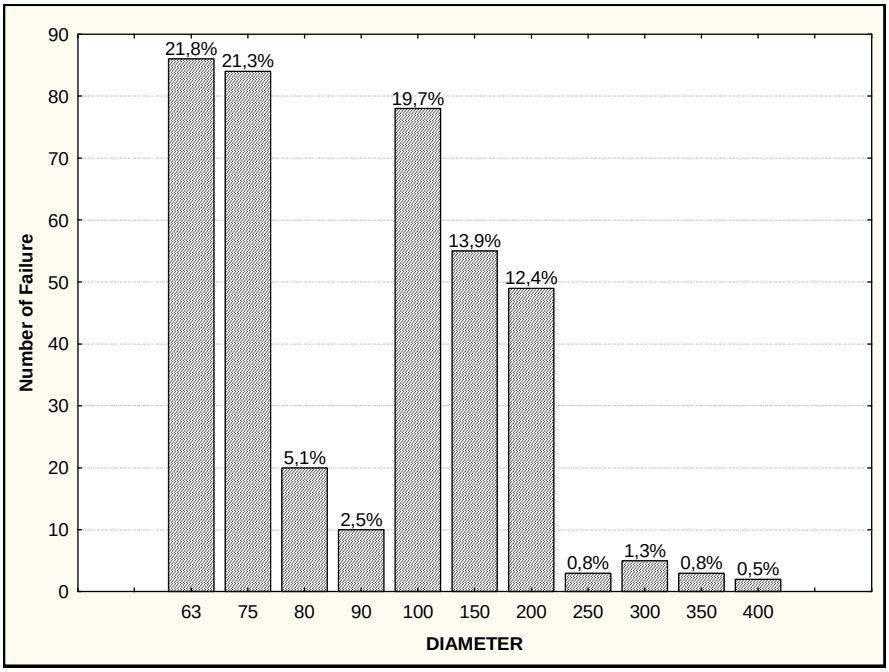

Fig. 2.13 Percentage of breaks in various water main's diameter (1995-2004)

About two-thirds of all failures occurred on the small diameter 63mm and 100 mm lines, which make up about 70.4 per cent of study area's mains system. Approximately, 26.3 per cent of operating failures occurred on 150 or 200 mm (6" and 8" nominal) diameter lines, which make up about 29.6 per cent of study area's water pipelines system. In other word, the frequency of breaks increases with a decrease in pipe diameter (Fig. 2.13). This finding agrees with that of previous researchers (e.g. Ciottoni 1983; Andreou and Marks 1986; Walski et al. 1986; Kettler and Goulter 1985). In contrast, for large diameter pipelines (diameter >400), no pipeline failures were reported. In the next chapter, the correlation of diameter and other variable will be explained.

Length

Currently, Sanandaj has approximately 56.7 kilometres of water pipelines within the pilot limits which the average length of a main segment is 149.66 m. As shown in Fig. 2.14, the length of a pipe has an effect on the number failures per pipe since we are measuring failures per pipe and not per pipe length. Maximum failure were recorded in the water pipelines in length category between 100-150 meter.

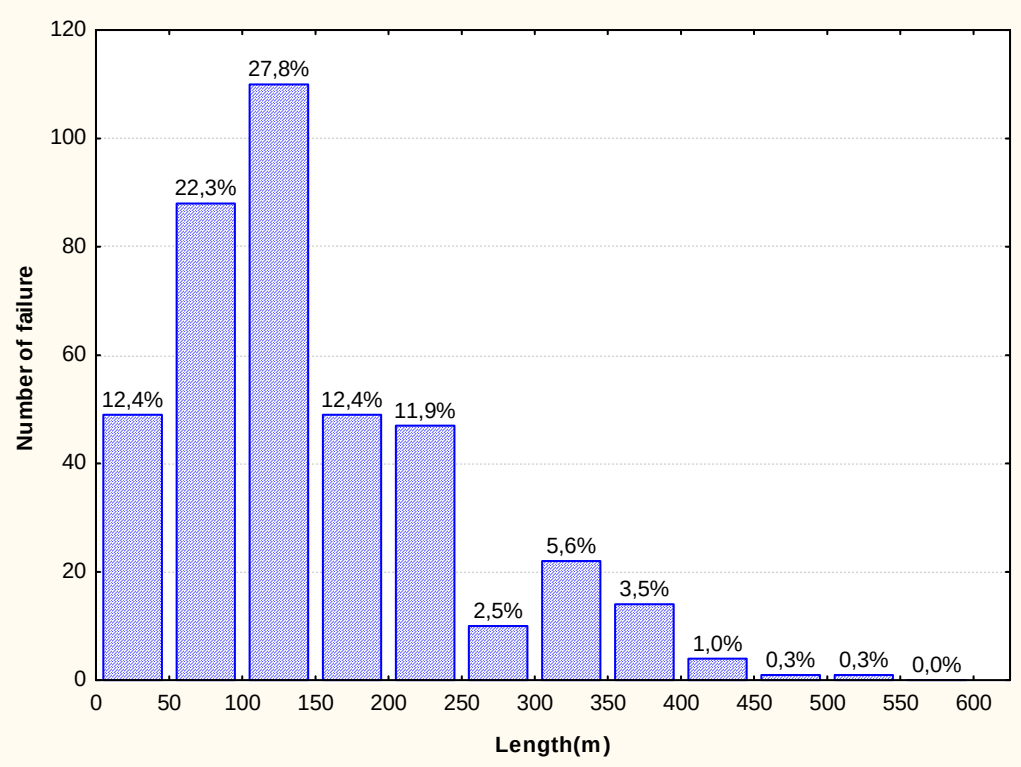

Fig. 2.14 Number and percentage of failure by length of pipelines

Pipe wall thickness

The ability of a pipe to resist the stresses induced by internal pressure and earth loads is a function of the tensile strength of the material and wall thickness (Skipworth et al, 2002). Fig. 2.15 plots the number of failures against the pipe thickness of four material. It is apparent that there is a general decline in the failure frequency when the thickness of pipelines is increased. The greater wall thickness provided an inherently more robust pipe, which was not as likely to fail structurally as a result of external and internal loading (Kettler and Goulter 1985).

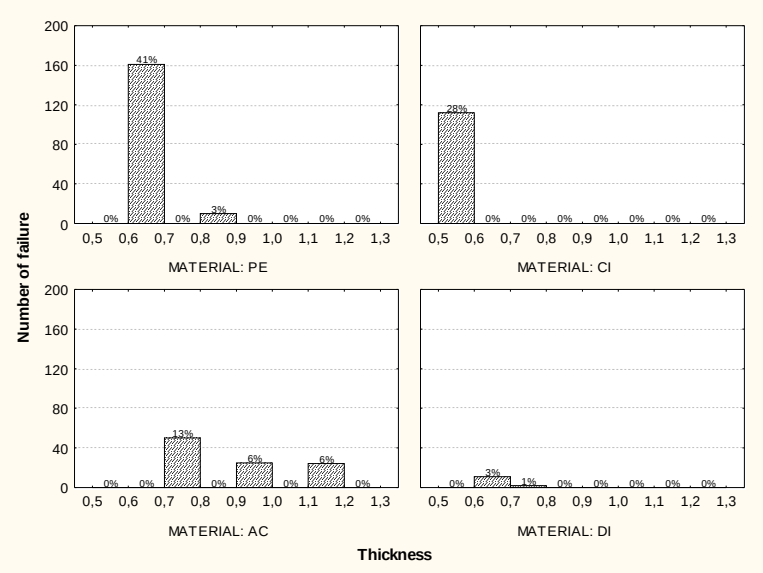

Fig. 2.15 Number of failure according to thickness of pipes by material

Depth of pipe burial

Pipelines tend to be buried with an increased depth of cover in locations where there is a perceived increased risk of damage caused by cold weather, traffic load and third party activity. In Western of Iran and in mountainous areas, such as Sanandaj, severe of winters causes 15.6 % of failures on the pipelines (Table 2.4 & Fig. 2.18). Burial depths in this area vary from 1 to 2 meter which depends on the diameter, ground condition of trenches and proximity to traffic routes. In some part of city the rock make difficulties in trenching and laying the pipelines. Usually, one meter is the minimum fill depth over the top of all water pipelines in the pilot area.

In addition, there is a relationship between depth of cover and causes of failure. In the most shorter of cover on the water mains, the traffic loads, excavation by third party and winter forest was reported as a probable cause of failure. Table 2.4 give the failure frequency as a percentage per depth of cover for cause of failure in terms of external interference (third party activity), traffic loads and winter forest.

Table 2.4 Percentage of failure frequency per depth of cover according to causes

Depth of cover (cm)	Length of water pipelines (km)	Number of failure	Cause of failure (%)		
			Traffic loads	excavation by third party	winter forest
100-150	20.2	222	7.2	23.5	8.1
150-200	36.5	173	6.4	14.7	7.5

With respect to material, polyethylene pipelines were laid less than 1.5 meter in the study area. The other is located in deeper trenches. The results presented in Table 2.4 clearly demonstrate the added benefits of burying pipelines deeper.

Water pipelines location

Pervious study have shown that traffic loads is a significant factor affecting pipe failure rates (EPA, 2005). Because of pilot location in the central part of city, most buried pipes are subject to large cyclic surface loads. In the case study dataset, traffic load was taken into account as a qualitative variable according to vehicles circulation or the type of road.

Table 2.5 Water pipelines length in each traffic load category by material

Traffic category	Length of water mains	Material (%)			
		AC	PE	DI	CI
Low load	29.5	11.3	15.2	1.2	2.8
High load	27.3	9.2	1.8	10.5	4.7

This classification relied on expert views and has been done by consulting the SWWU engineers. Therefore two traffic categories were created: low and high. Water pipelines located in the sidewalk, alley, and low traffic street fall into low traffic load and water pipelines located along main street were categorized as high loads of traffic.

Table 2.5 presents length of water pipelines in each category of traffic loads in term of pipeline material. As shown in Fig. 2.16, number of failure increase with traffic load in two material, cast and ductile iron. Polyethylene and asbestos cement pipelines experienced more

failure in the low traffic load. As illustrated in table 2.5, most PE and AC pipelines located in the area with low traffic. In contrast, DI and CI mains have been laid in the main street.

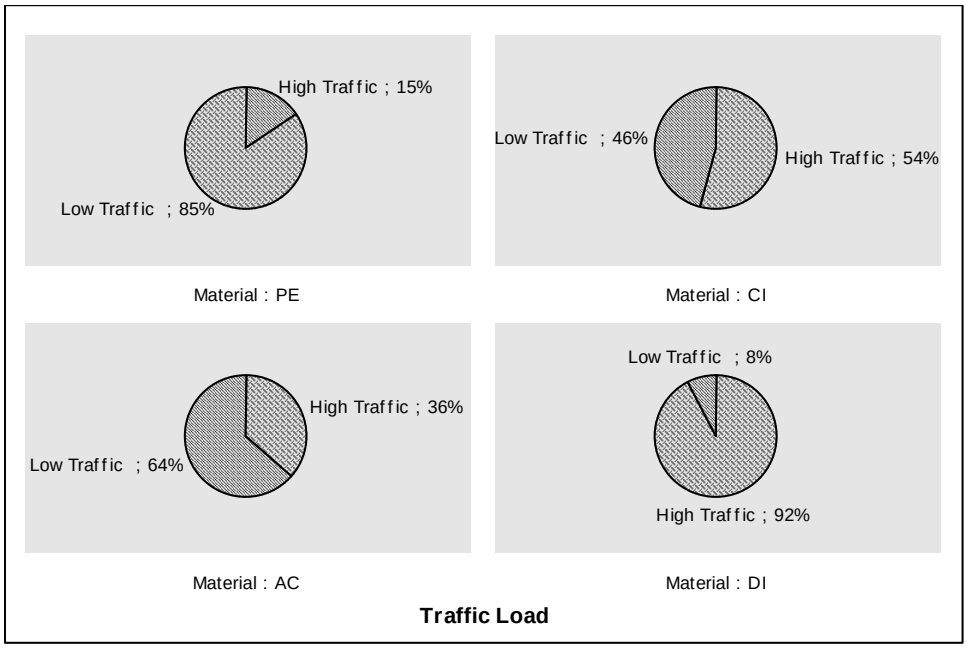

Fig. 2.16 Percentage of failure according to traffic load in terms of materials

Pressure

A variable that can influence the failure of some water pipelines is daily maximum pressure. Based upon SWWU's experience, water pipelines in higher-pressure areas are more likely to fail. Further, pressure in selected water pipelines varies widely throughout the day and night, as people use more or less water. This fluctuation puts a great deal of strain on the mains, which means they are more likely to burst. The water pipelines breaks during the period from 1995 to 2004 were analyzed and a total of 395 breaks were recorded for this period, of which 11.3 % caused by high pressure (Fig. 2.18). Daily maximum pressure for each water pipelines were calculated in hydraulic model and entered in dataset. The fitted graphs in Fig. 2.17 show the exponential increases of water pipelines failure in four materials.

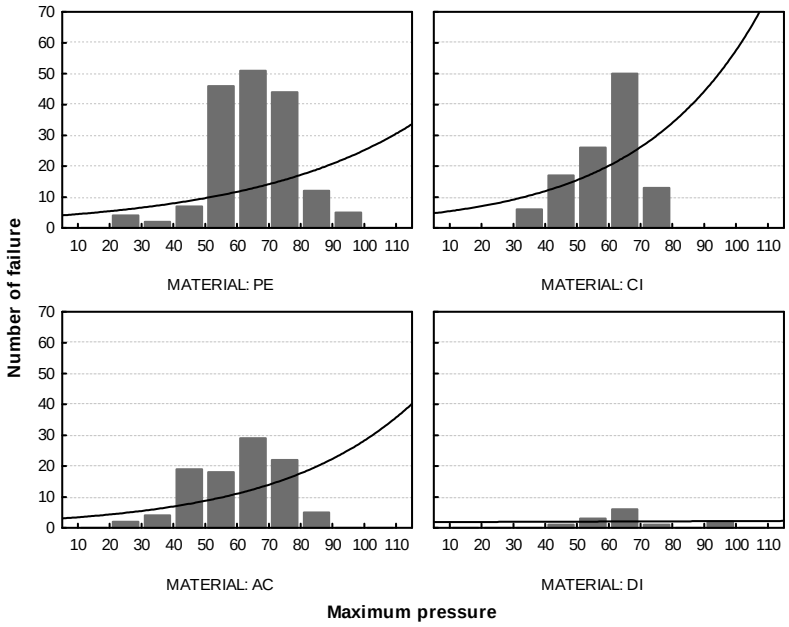

Fig. 2.17 Number of failure according to maximum pressure by material

In general, the water pressure is more dependent on the ground elevation. Put in other words, the lower the ground elevation causes the higher water pressure in pipelines. Due to geographical situation and topographical features, the study region is situated among a group of major and minor hills in near the mountains. Most of main distribution bear significant grater pressure and tolerate 7.5 bar during low demand condition. Inadequate pressure management in network cause more failures on supply and also increase leakage. An average of 11.3 percent of the water pipelines breaks from this problem demonstrates failure is related closely to the pressure in the system. In SWWU's experience, certain pipe material, such as polyethylene and cast iron are more susceptible to this than other. Subsequent to this, we developed the hydraulic model to estimate the pressure in more detail.

2.4.2 Summary of explanatory factors

While a large number of factors can contribute to the failure of water mains, 9 predictor variables were used in the statistical model and these summarized in table 2.6. The factors were identified in a more objective manner to use in statistical investigation. Many other factors can affect the rate of deterioration of water distribution systems and lead to their failure. But the problem is access to this data in the study area.

Table 2.6 Summary of variable in failure data analysis

Variables	Variable type	Symbol	Levels (for categorical variables)	Category/ Data source
Date of failure	Continuous quantitative	X_1	Day of failure reporting	Maintenance
X (coordinate of failure)	Continuous quantitative	X_2	Match failures address along street or intersection in GIS and find x and y in the UTM system	
Y (coordinate of failure)	Continuous quantitative	X_3		
Number of pervious breaks	Discrete quantitative	X_4	Experienced number of failure before	
Age of pipe	Continuous quantitative	X_5	the number of years between laying year and failure time	Structural or Physical
Pipe material	Nominal categorical	X_6	1– Cast Iron; 2–Ductile Iron 3–Asbestos Cement 4- Polyethylene	
Pipe diameter	Continuous quantitative	X_7	the diameter of the pipe in millimeter	
Length of pipe	Continuous quantitative	X_8	the length of the pipe in meter	
Pipe wall thickness	Continuous quantitative	X_9	the thickness of the pipe in millimeter	
Pipe location	Nominal categorical	X_{10}	0 if the pipe locates under pavement or roadway with light traffic, 1 otherwise	Environmental or External
Depth of pipe burial	Continuous quantitative	X_{11}	the depth of cover in meter	
Pressure	Continuous quantitative	X_{12}	the pressure in the pipe in meter of water column	Operational (Hydraulic Model)

2.4.3 Probable cause of failure

The causes of water pipelines deterioration are the focal point of this section. Exactly identification of causes for each failure is difficult and more expensive. In this research, the most probable cause of water pipelines failure was conducted through several individual investigations. Table 2.7 compiles data on water pipelines accidents based on their causes and mechanisms in each material.

Table 2.7 Incorporation of failure mode and causes

Water pipelines material	Cause of failure	Failure mode
Asbestos Cement (AC)	Traffic load	Displacement at joint
	Uneven setting	Longitudinal
Polyethylene (PE)	Weak material	Ovality
	Uneven setting	Joint imperfection
	Winter forest	Displacement at joint
	Traffic load	
	Overpressure	
Ductile Iron (DI)	Uneven setting	Holes
	Corrosion	Displacement at joint
Cast Iron (CI)	Corrosion	Holes
	Water hammer	

This table and Fig. 2.18 indicates that corrosion is a big problem for the cast iron (CI) pipes, as the breaks due to corrosion account for less than 6% of the total breaks for this period. At a glance, Fig. 2.18 shows the total number of failures that occurred in the period 1995-2004 differentiated by reported cause of failure. Uneven setting is the predominant cause of failure (21%). Traffic load caused the majority (about 19 %) of failures and weak material accounted for 15 %. The next largest cause was water hammer damage (12 % of all failures) and high pressure was responsible for about 11 % of total pipeline failures.

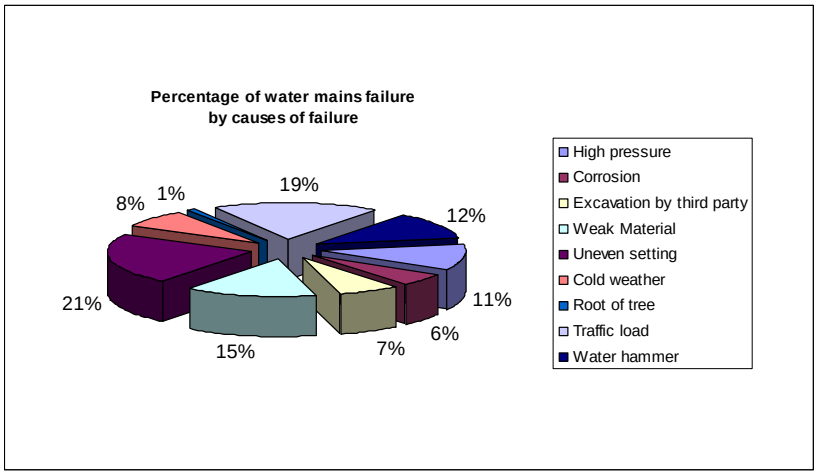

Fig. 2.18 Percentage of water pipelines failure by causes

2.4.4 Failure modes

Water pipelines in study region break in one of seven modes of failure: circumferential break, longitudinal break, explosion, hole, displacement at joint, ovality and spiral cracks (Fig. 1.3). Table 2.8 lists the different failure modes and their corresponding percentages which vary depending on the pipe's material. Among the 395 breaks, displacement at joint comprise the predominant failure mode (32%). An average of 25% of water pipelines breaks are longitudinal and other 44% are either circumferential, explosion, hole and spiral cracking. Just 7% of failures in polyethylene ovalized under the effects of earth and live loads (Fig 1.3 f.)

Table 2.8 Failure modes and their percentage

	Apply to water pipelines material & percentage			
Failure Mode	PE	AC	CI	DI
Circumferential	2 %	3 %	2 %	0 %
Longitudinal	13 %	7 %	4 %	1%
Explosion	3 %	1 %	0 %	0 %
Hole	6 %	1 %	4 %	2 %
Displacement at Joint	12 %	14 %	5 %	1 %
Ovality	7 %	0 %	0 %	0 %
Spiral cracking	0 %	0 %	13%	0 %

Fig. 2.19 indicates the percentage of failures over the period 1995-2004 differentiated by the seven failure modes categories.

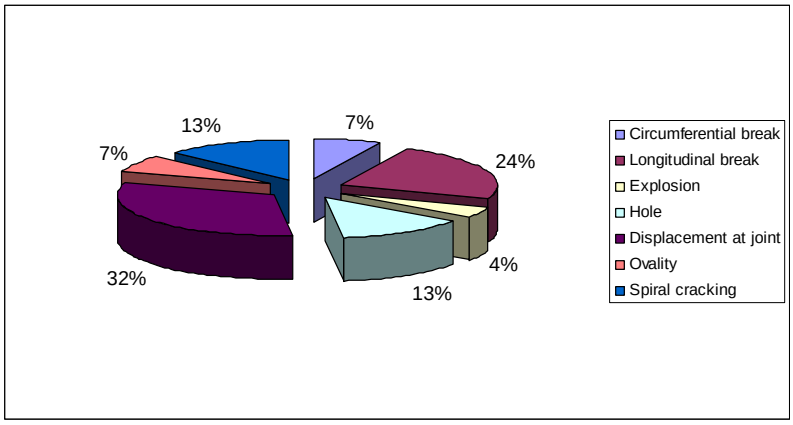

Fig. 2.19 Percentage of failure modes in selected water mains

Up to this point, we have dealt exclusively with what is commonly referred to as classical statistics. In this section, we conducted spatial analysis to determine water pipelines break pattern for pilot zone in the urban community of Sanandaj, Iran.

2.5 Spatial Analysis

The analysis strategy was established by expression of failures as a random distribution of points in space (Fig. 2.20). Spatial analysis involves the analysis of data representing geographical features which have a locational attribute such as absolute location (coordinates) or relative positioning (distance). Prior to point pattern analysis, a set of descriptive spatial statistics e.g. spatial measures of central tendency and dispersion were done. Additionally, spatial analysis applied to point data typically involves the analysis of point distributions and the relationship between point distributions and other spatial features. The objective of the analysis would be to determine if the point pattern is "regular", "random", or "clustered". Two basic techniques are used based on (*i*) counting of cases within small squares (quadrat analysis) and (*ii*) measuring distance to the nearest case (nearest-neighbor analysis).

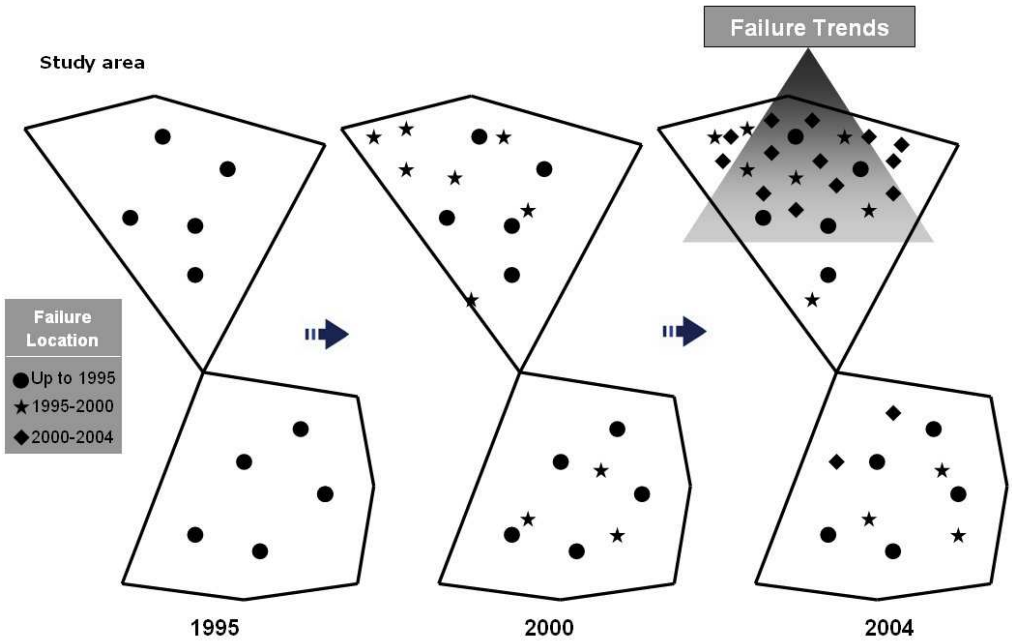

Fig. 2.20 Schematic description of temporal and spatial trends in water pipelines failures

All spatial statistical analysis was performed using the *CrimeStat III* software. It is a freely available package that provides a variety of tools for the spatial analysis of crime incidents or other types of applications involving point locations, such as the location of pipeline failures.

2.5.1 Mapping of failure point locations in GIS

GIS combines layers of spatially related information. Each layer comprises the pertinent map and its conjugate attribute data. This study used the point mapping method to displaying geographic patterns of failures. In the spatial database, failure mapping starts with geocoding process that matches an address of break to a physical location (as a point) along a street.

The water pipelines failure database is a GIS-based tool for integrating map and attributes data. In term of GIS, water main failure is a phenomenon which can be expressed through occurrences identified as points in space. Each failure point is converted manually into coordinate locations via address matching. The process of identifying the coordinates of a failure point in ArcView 3.2a, is to use "getx" command for the x-coordinate and "gety" command for the y-coordinate. In a project coordinate system, such as UTM, both x and y are defined by distances in meters from an arbitrary reference point.

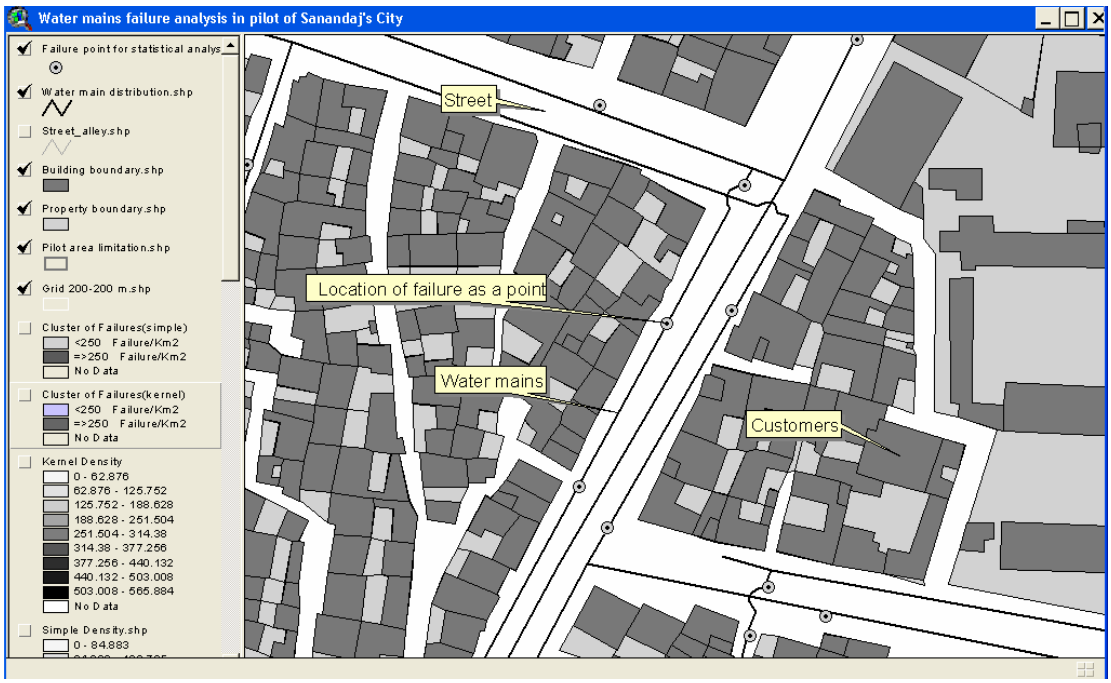

Fig. 2.21 Observed failure locations on water pipelines network and related GIS layers

When an address list is transformed into a set of coordinate points based on street reference layer, corresponding failure attributed data are then imported from Access into ArcView. The platform in ArcView is connected with external databases using SQL connection feature. In order to combine and join data to a point shapefile of locations according to each failure, a common field such as failure *id* should exist in both tables. By joining this field between tables, data can be retrieved from each table and combined into another table. The data is then saved as a shapefile in GIS. In fact, each failure on pipeline can be created by the common id number in ArcView and Access databases. Therefore, the related non-spatial data about failure is stored in MS Access and graphical information is established in ArcView. Each record in the tabular database is then connected with a failure point in GIS.

At a scale of 1: 2000, the reference map provides a suitable level of details to approximately locate property boundaries, streets and distribution mains features. Water pipelines network and failures are mapped against a backdrop of streets, property lines and buildings. It is composed of 5 distinct feature layers that contain more than 53.3 km of distribution mains and 4.4 km of transmission mains. Fig. 2.21 displays GIS layers in customized ArcView application and the location of failure points in the investigated area. In the map, each failure location is represented by a small black dot.

The analytical process manipulates both map and attributes-related data through the linkages that GIS establishes between them. Fig. 2.22 depicts spatial relationships and attributes database in the developed GIS database.

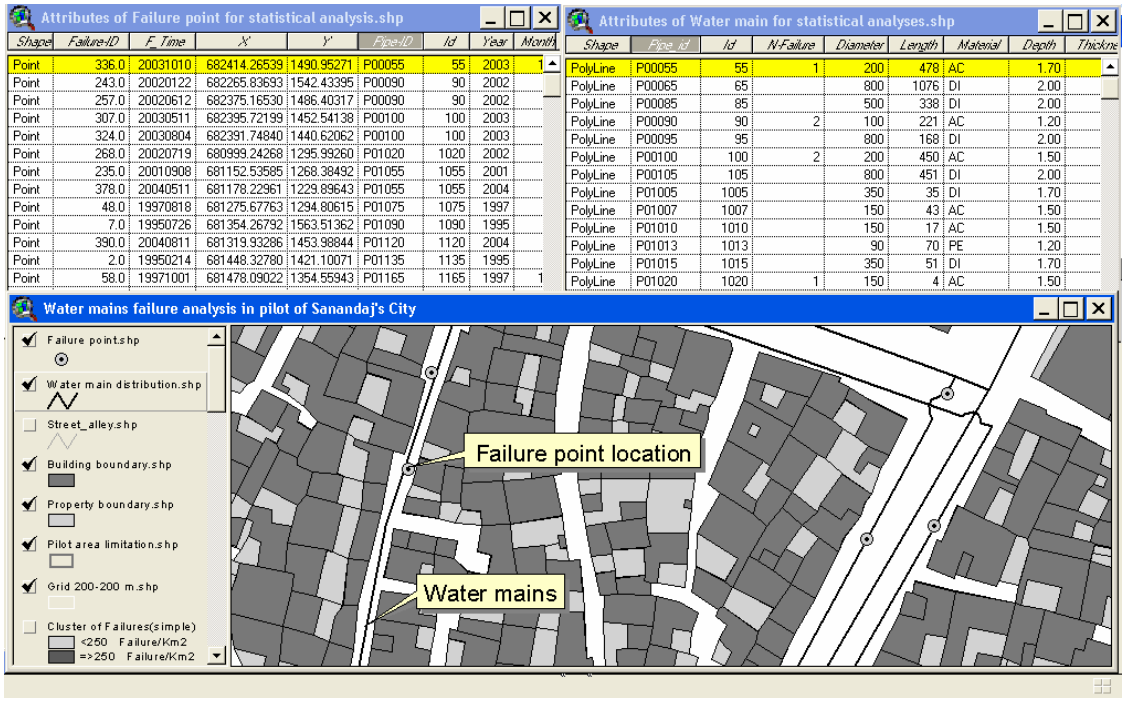

Fig. 2.22 Combination of layers in GIS for spatial analysis of water pipelines failure

2.5.2 Spatial statistical methods

Spatial statistical methods have been developed to facilitate the monitoring of geographic pattern in order to provide quick detection of emergent geographic clusters. Additionally, the geographical distribution of failure and its relationship to potential risk factors (referred to in this work as 'geographical distribution of traffic or hydraulic pressure') has been evaluated.

Several issues of major statistical journals have been devoted to spatial statistical methods in health and crime applications (Cressie, 1991). The study was undertaken in two steps: first by performing the point pattern analysis and then carrying out the spatial interpolation. This process creates a mathematical model which is used to estimate values across the raster surface. Various similar methods use location and values at corresponding sampling locations to estimate the variable of interest at un-measured locations.

Central tendency scores

Since points were used to indicate the spatial occurrence of pipeline failure, to get an idea about the overall pattern of failure distribution in the study area, some basic descriptive

statistics were generated. The measures of spatial central tendency, or centrographic measures, like mean and weighted mean center scores were determined for analysis of the failure distribution. The most basic form of statistical analysis for measurement of central tendency in geographical data is called Mean Center or balance point. It is concerned with the center of a geographic point dataset in which the points in the distribution represent occurrences of pipeline failure in different periods of time. By taking the *x* and *y* coordinates of each point, the bivariate mean center is determined using the expression:

$$(\bar{x}_{mc}, \bar{y}_{mc}) = (\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n}) \qquad (2.1)$$

where:

$\bar{x}_{mc}$, $\bar{y}_{mc}$	=	mean center coordinates
x_i , y_i	=	coordinates of each water pipelines failure location, and
n	=	Total number of failures.

The mean center computed from this formula (2.1), appears right in the middle of the geographic area (Fig. 2.23).

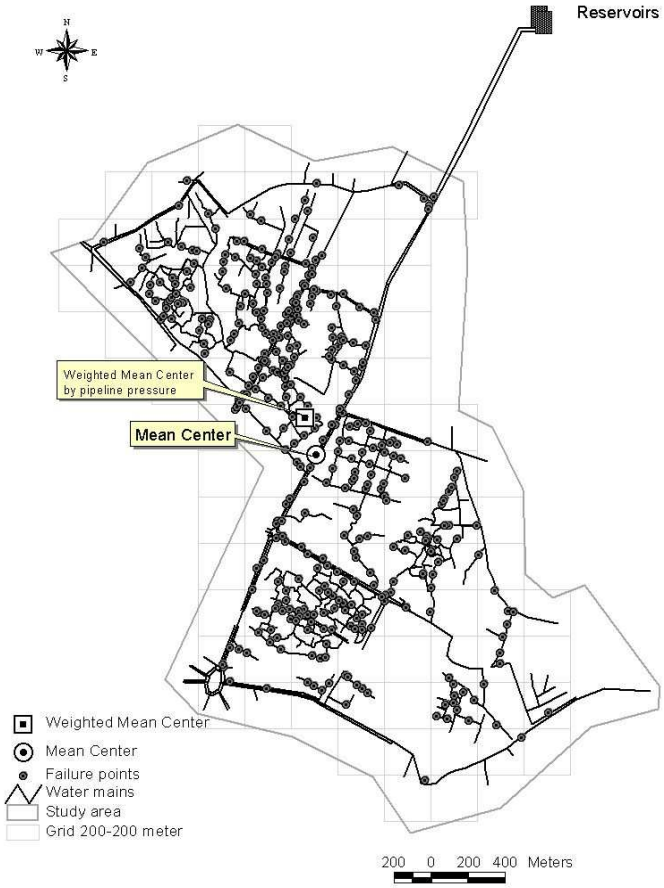

Fig. 2.23 The scatter plot of failure points, mean center and weighted mean center

Moreover, the weighted mean center is calculated by weighting each coordinate by another variable, as below:

$$(\bar{x}_{wmc}, \bar{y}_{wmc}) = \left(\frac{\sum_{i=1}^{n} w_i x_i}{w} \;,\; \frac{\sum_{i=1}^{n} w_i y_i}{w}\right) \tag{2.2}$$

where:

$\bar{x}_{wmc}$, $\bar{y}_{wmc}$ = weighted mean center coordinates,

w_i = weight,

w = total weight.

Descriptive statistic for failure data showed that high pressure in the pipeline is the common cause of failure in the case study (Fig. 2.18). By multiplying pressure in the coordinates of

each failure point, the weighted mean center would be mapped. Thus, water pipelines failure points affected by high pressure pull the weighted mean towards them. Fig. 2.23 shows that the computed weighted center is to the northwest of the mean center. It might show that there is more pressure in the northwest portion of the study area. According to the topography and general slope of the city, and to local expert opinion, this is correct.

Dispersion scores

Dispersion scores provide a unit measure of spread or variability of a failure distribution. This analysis addresses questions such as:

- Do the failures cluster about their central point or do they spread out around it?
- Where are the high or low concentrations of failure?
- Is there a pattern to the failure data?

In spatial statistical analysis, standard deviation is expressed as standard distance. While standard deviation indicates how observations deviate from the mean, standard distance indicates how points in a distribution deviate from the mean center. Standard deviation is expressed in the units of observation values, but standard distance is expressed in distance units. In terms of its application, standard distance is usually used as the radius to draw a circle around the mean center to give the spatial spread of the point distribution it is based on (Saraf et al., 2003). The radius equal to $SD_{x,y}$:

$$SD_{x,y} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x}_{mc})^2}{n-1} + \frac{\sum_{i=1}^{n}(y_i - \bar{y}_{mc})^2}{n-1}} \qquad (2.3)$$

where:

x_i , y_i = coordinates of each water pipelines failure location,

$\bar{x}_{mc}, \bar{y}_{mc}$ = coordinates of mean center, and

n = total number of failures.

The computed radius is equal to 720.28 m (Fig. 2.24). Essentially the average distance of points from the center provides a single unit measure of the spread or dispersion of water pipelines failure distribution.

Fig. 2.24 The circle represents the Standard Distance Deviation

While the standard distance deviation (SDD) is a good measure of the dispersion of the incidents around the mean center, it does not show the potential skewed nature of the data (anisotropy). The standard deviation ellipse gives dispersion in two dimensions and surrounds most of points. It can be a tool to explore variables that are affecting failure patterns.

As seen in Fig. 2.25, Standard Deviational Ellipse (SDE) is defined by three parameters: angle of rotation (θ), dispersion along major axis (δ_x), and dispersion along minor axis (δ_y).

The major axis defines the direction of maximum spread of the distribution and minor axis is perpendicular to it and defines the minimum spread

Fig. 2.25 Standard Deviational Ellipse around the mean center of failure locations

The angle of rotation, θ, which has the definition below:

$$\tan\theta = \frac{(\sum_{i=1}^{n} x_i'^2 - \sum_{i=1}^{n} y_i'^2) + \sqrt{(\sum_{i=1}^{n} x_i'^2 - \sum_{i=1}^{n} y_i'^2)^2 + 4(\sum_{i=1}^{n} x_i' \sum_{i=1}^{n} y_i')^2}}{2\sum_{i=1}^{n} x_i' \sum_{i=1}^{n} y_i'} \qquad (2.4)$$

$$x_i^{'} = x_i - \bar{x}_{mc} \quad , \quad y_i^{'} = y_i - \bar{y}_{mc} \tag{2.5}$$

The deviations along the x and y axes are defined as:

$$\delta_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i^{'}\cos\theta - y_i^{'}\sin\theta)^2}{n}} \quad , \quad \delta_y = \sqrt{\frac{\sum_{i=1}^{n}(x_i^{'}\sin\theta - y_i^{'}\cos\theta)^2}{n}} \tag{2.6}$$

Thiessen polygons

Conceptually, Thiessen polygons (TPs) is the simplest vector-based method in the interpolation process. It commonly applied in situations where attribute is categorical (Siska et al. , 2001) . This method assigns interpolated values equal to the values found at the nearest sample location. In our case, failures of water pipelines have been digitized as point data file. In the vector $\{(x_i , y_i) , z_i\}$, (x_i , y_i) is used to reference the location of point i while z_i is the measured attribute at point site i . Routinely, the location of failure events displays on map in Fig. 2.26 The location has been symbolized according to the cause of failure. Black dots designate locations with corrosion and square symbols represent weak material. It associates with 395 failures on water pipelines system during 10 years study.

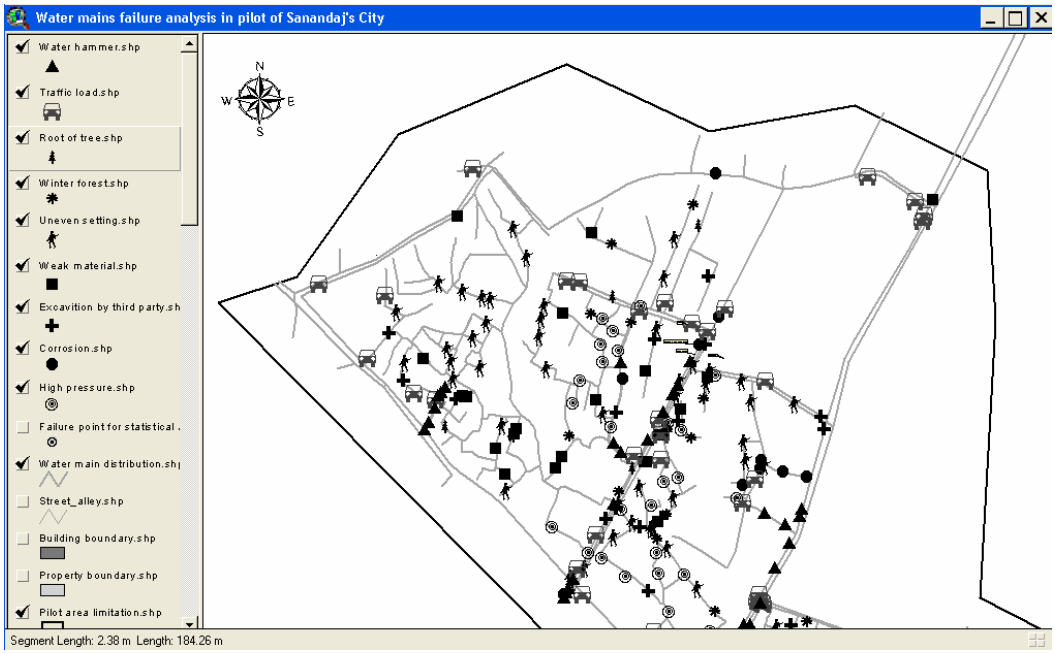

Fig. 2.26 A point data map depicting the 9 failure cause categories

TPs converts discrete data into continuous surface through spatial interpolation techniques.

Sanandaj's Water and Wastewater Utility has classified the causes of pipeline failures into 9 major categories. The following map describes the area of influence of a failure point regarding the cause of its failure. In other words, by using a mathematical process the catchments area for each failure point can been determined. Consequently, they can be assigned values that reflect the attributes of the regions they represent.

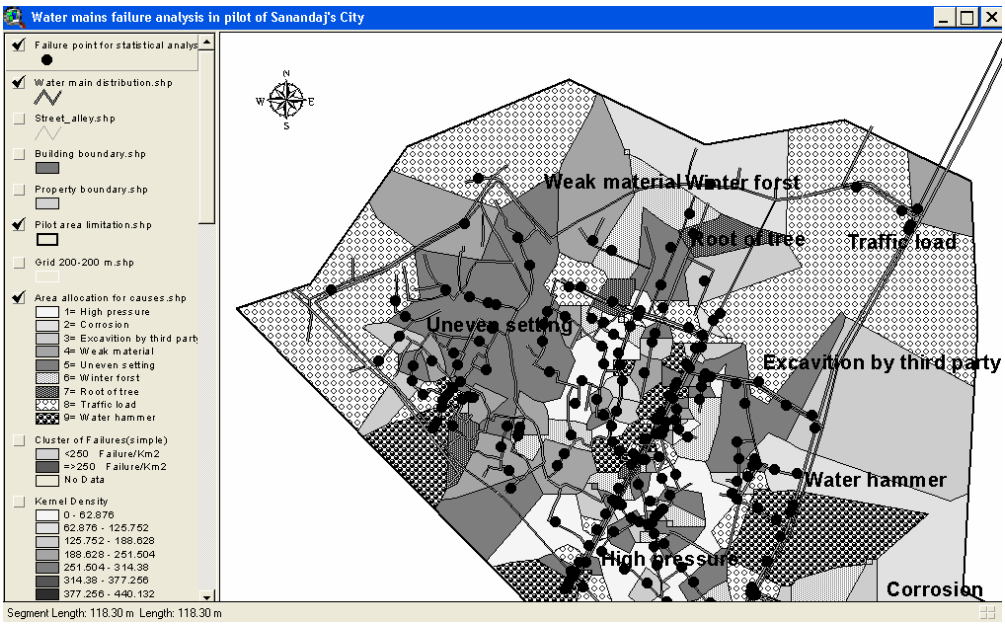

Fig. 2.27 Neighborhood interpolation by Thiessen Polygons for failure points

Our analysis indicates that ″traffic load″ is a major contributing factor to a large number of noted pipeline failures. As can be observed from Fig. 2.28, uneven setting is found to be the 2nd contributing factor to pipeline failures.

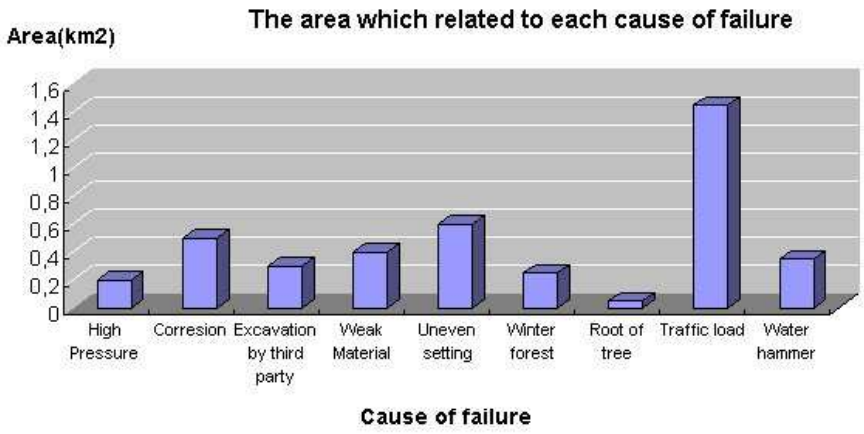

Fig. 2.28 Area size and cause of failure representation

With regard to this analysis, we can conclude that the irregular polygons show the spaced failure points are randomly distributed.

2.5.3 Point pattern analysis

Point pattern analysis is concerned with attempting to determine whether the distribution of points is random or whether it either clusters or is evenly distributed. This work was carried out two ways to assess the distribution of water pipelines failure : nearest neighbor analysis and quadratic analysis.

Nearest neighbor analysis

The nearest neighbor index (R) provides an approximation about whether points are more clustered or dispersed than would be expected on the basis of chance (Diggle, 1983). It compares the average distance of the nearest other point (nearest neighbor) with a spatially random expected distance. The expected distance is given by:

$$\overline{\mathrm{NND}}_R = \frac{1}{2\sqrt{Density}} \qquad \mathrm{Density} = \frac{n}{A} \tag{2.7}$$

Thus:

$$R = \frac{\overline{NND}}{\overline{NND}_R} \quad , \quad \overline{NND} = \frac{\sum_{i=1}^{n} NND}{n} \tag{2.8}$$

where:

n	=	number of failure points in the distribution,
A	=	the size of study area,
R	=	the ratio of the observed distance to the expected distance,
NND	=	distance between each point and its nearest neighbor,
$\overline{NND}$	=	the observed mean distance between nearest neighbor.
$\overline{NND}_R$	=	the expected value of the nearest neigbor distance in a random pattern

For these data, the mean nearest neighbor distance and expected mean nearest neighbor distance are calculated 36.2 and 55.9, respectively. Since the analyzed mean distance is shorter than the mean distance for the random pattern, the pattern is clustered. In fact, for this purpose, the R ratio is used. R values <1 indicate clustering, since the observed mean distance between neighboring points is less than that expected in a random pattern. The minimum value of *R* is zero, which occurs when all points are at a single location. The theoretical maximum of *R* is equal to 2.149, which occurs when points are maximally dispersed. Since the nearest neighbor index *R* is equal to 0.655, the distribution of pipeline failure is more clustered than random.

To help place confidence in the nearest neighbor index result, a test statistic can be calculated as follows (Cressie, 1991):

$$Z_n = \frac{\overline{NND} - \overline{NND}_R}{\sigma_{\overline{NND}}} \tag{2.9}$$

where the standard deviation of mean distance between nearest neighbors is:

$$\sigma_{\overline{NND}} = \frac{0.26136}{\sqrt{n \times (Density)}} \tag{2.10}$$

If values of the Z_n statistic are in the interval from -1.96 to +1.96, the pattern is random (at a confidence level of 95%). If the Z_n value exceeds +1.96, the pattern is regular. If the Z_n value

is lower than -1.96, the pattern is clustered (Rogerson et al., 2001). The value of $Z_n = -13.1$ was arrived at by using equation 2.8 which indicates a clustered pattern exists.

Quadratic analysis

Quadrat analysis is one of two most commonly used tools for analyzing the dispersion of points. At first, the study area was sub-divided into regular grid squares and the number of occurrences of water pipelines failure in each square is counted. The formula for determining the optimal quadrat size (length of the square side) is:

$$quadrat\ \text{size} = \sqrt{\frac{2 \times A}{n}} \tag{2.11}$$

where A is the size of the study area and n is the number of points.

The optimal quadrat size was determined to be 200 m. Fig. 2.30 depicts the classification of quadrat according to the number of points in each cell. Then, the variance/mean ratio (*VMR*) was used to compare the obtained empirical frequency with the theoretical frequency for the random pattern obtained from the Poisson distribution which the mean and variance are equal.

According to following expression, if the distribution is random the *VMR* is about 1.0. Larger values ($VMR > 1.0$) correspond to existence of "clumps" (spatial clusters). Smaller values ($VMR < 1.0$) correspond to a more-uniform-than-random distribution:

$$VMR = \frac{VAR}{MEAN} \tag{2.12}$$

Fig. 2.29 Number of water pipelines failure point in each cell

Thus :

$$\text{VAR} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad , \quad \text{MEAN} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.13}$$

where:

x_i = number of water pipelines failure points in cell *i*, and

n = total number of cells.

We obtain: *VMR = 5.9*; since *VMR* is greater than 1, the point pattern can be considered as more clustered than random. Rather than base a conclusion on variance/mean ratio, we can compare observed frequencies in the quadrats with random frequencies realized from a homogenous Poisson process. Therefore, chi – statistic (χ^2) test was calculated as follow:

$$\chi^2 = \sum \frac{(x_i - \bar{x})^2}{\bar{x}} \tag{2.14}$$

where:

x_i = number of points in each quadrat, and

$\bar{x}$ = mean number of points per quadrats.

The significantly large χ^2 value, 562.5, indicates that the distribution is not uniform and that there may be some underlying process causing the non-uniformity (clustering).

Convex hull

The convex hull is the smallest convex polygon containing the set of points in two and three dimensions. This polygon represents the minimum possible area that contains all points and can be imagined as a failure band stretched around the points. The generation of a minimum-bounding polygon (convex hull) provides additional insight into the spatial extent of identified clusters of failure points (Fig. 2.30). The appropriate number of clusters is determined in two partitions. Typically, partition with a large number of events indicates a high risk of water pipelines failure. The number of entities in a partition of a convex hull area (number of failure /area) is a reflection of the cluster degree of compactness.

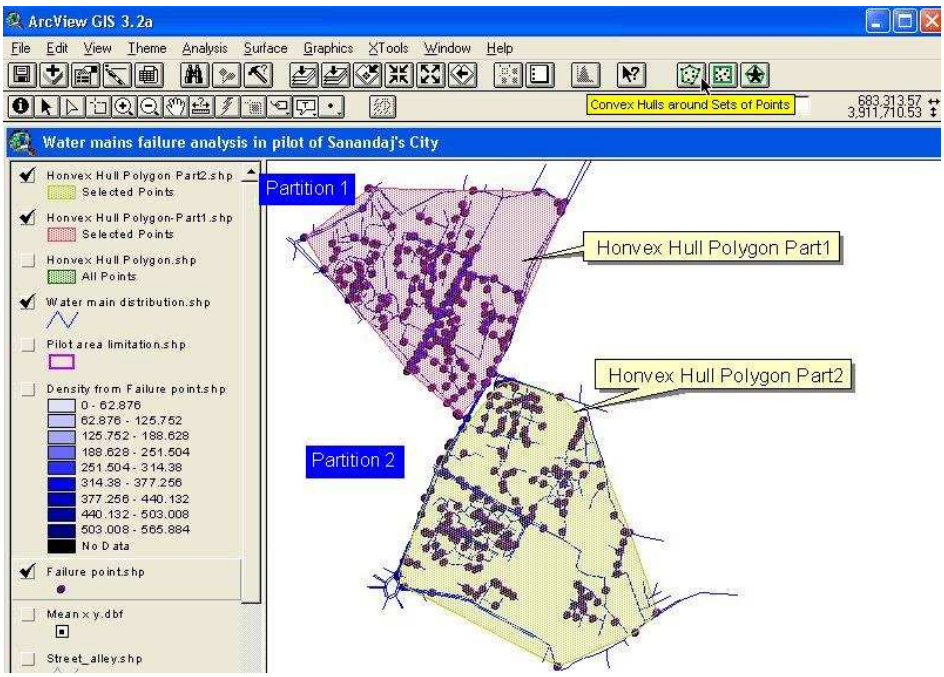

Fig. 2.30 The convex hull polygon around failure points

The circular spots in Fig. 2.30 represent the locations of failures that took place during study period.

Table 2.9 The mass of failures covering a unit of area

	Partition 1	**Partition 2**
Number of failure	209	186
Hull area (m^2)	1023980	1357693
Density by area (n/A)	**$2.04*10^{-4}$**	$1.36*10^{-4}$

Referring Fig. 2.30 and table 2.9, cluster partition 1 is associated with the most compact hull.

Density calculations

Continuous surface maps use a method to aggregate points within a specified search radius to create a smooth surface that represents the density of events across the area. It helps in identifying the location, spatial extent and intensity of failure hotspots. Results are visually attractive since it helps in invoking further enquiry and exploring the reasoning behind why

water pipelines failures are concentrated in specific areas. This work examines two methods: a simple density formula and a weighted "Kernel" procedure.

Simple density

The ArcView extension Spatial Analyst (version 2.0) was used to develop the digital density layer of the combined data points. A density surface is based on the division of the study area into square cells. A density value for each cell is calculated by counting the number of points within a defined search radius from the center of each cell (Fig. 2.31-a) and dividing by the search area. The density value (features per square km) is assigned to the cell. The search circle is then shifted to the next cell and the floating process is repeated until all of cells have been assigned a density value (Fig. 2.31-b). Accordingly, this process smoothes the density layer over the study area.

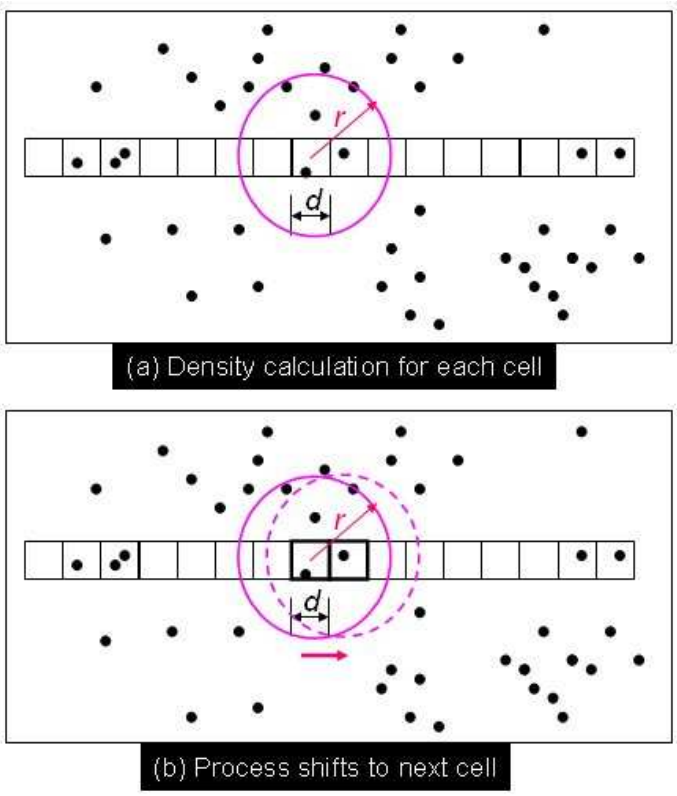

Fig. 2.31 Procedure used in density calculation

When doing density calculations, the recommended number of cells is between 10 and 100 cells per density unit (Mitchell, 1999). The density unit of features per square kilometer was

used herein. Using this recommendation, our study area is approximately equal to 100 cells (with dimensions of 100 × 100 meters) per square kilometer. A smaller cell size can typically produce smoother surface. The chosen search radius influences the appearance of the density surface. The larger the radius, the more generalized the patterns will be. The 60-meter search radius was selected in this study and the resulting density map is shown in Fig. 2.32. As illustrated in Fig. 2.32, the study area experienced an average density of 80 failures per square kilometers during the 10 year study period.

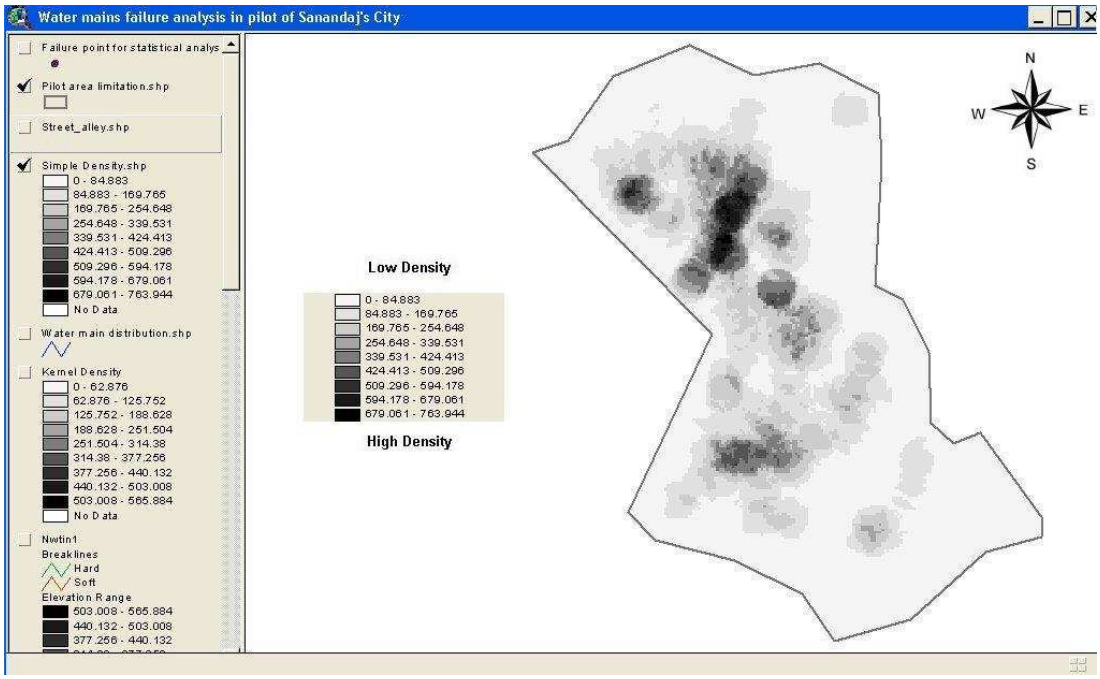

Fig. 2.32 Simple density estimation

Kernel density

Kernel estimate density is probably the most commonly used method as well as the most well understood statistically form of density estimation (Silverman, 1986). Our idea in water pipelines failure applications is to use kernel density estimation to transform distribution of discrete points or events representing incidence of failure into a continuous surface of failure risk. Essentially, a moving three-dimensional function (the kernel) of a given radius or bandwidth visits each of the points or events in turn, and weights the area surrounding the point proportionately to its distance from the event (Fig. 2.33). The sum of these individual kernels is then calculated for the study region, and a smoothed surface is produced.

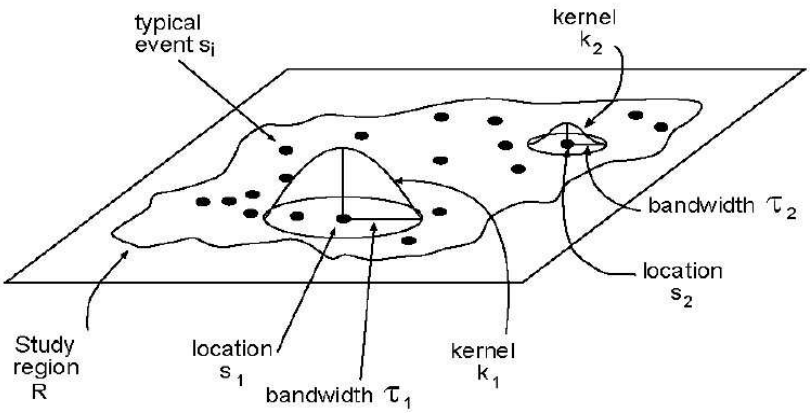

Fig. 2.33 Kernel Estimation of point patterns (Silverman, 1986)

There are a variety of different kernels. Fig. 2.33 demonstrates two such kernels, K_1 and K_2, however usually, only one form of kernel is used at any one time. The one used by ArcView (ESRI, 1992), and adopted in this study, is a quadratic kernel which has a property of being computationally simple, and is hence attractive for implementation within a GIS application where large data sets are not uncommon. It can be defined as:

$$k(s) = \begin{cases} \frac{3}{\pi}(1 - s's)^2 & \text{for } s's \le 1 \\ 0 & \text{otherwise} \end{cases} \qquad (2.15)$$

Using the above quartic kernel, if s represents a general location in R and $s_1,\ldots,s_n$ are the point locations of the n observed events whose underlying density we are estimating (Fig. 2.33), then the intensity $\lambda(s)$ at s is estimated by:

$$\hat{\lambda}_r(s) = \sum_{h_i \le \tau} \frac{3}{\pi\tau^2}\left(1 - \frac{h_i^2}{\tau^2}\right)^2 \qquad (2.16)$$

where h_i is the distance between the point s and the observed event location s_i and the summation is only over values of h_i which do not exceed τ. The parameter τ is the bandwidth and determines the amount of smoothing.

Fig. 2.34 depicts the density surface of failure locations. It shows failure incidents per square km based on kernel density calculation. The result of analysis in Fig. 2.34 indicates four

failure hot spot locations, signified by the darker pattern on the map. It has been reported in the literature (Bottom, S., 2002) that water pipelines failure tend to occur in cluster. This map shows also the variation and concentration of failure incidents across the study area. As a result, it categorizes hotspot areas into those where bursts have caused service problem (high-risk areas) and others where no service problems have occurred (low-risk areas). As mentioned in the literature review, the probability of failure decreases with time and distance following a previous failure (Goulter et. al. 1990). Due to the fact that densities can be converted into probabilities, accordingly, this kind of mapping allows users to evaluate the anticipated value of failure likelihood.

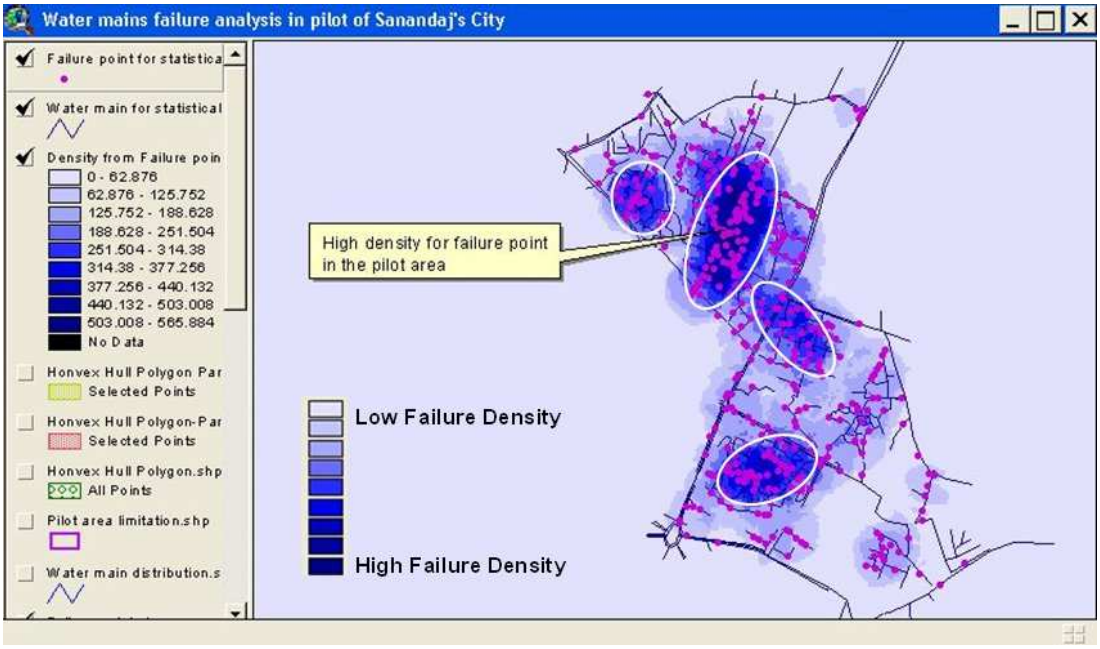

Fig. 2.34 Raster prediction map of failure density

TIN interpolation method

TIN or Triangulated Irregular Network has the ability to utilize stored GIS data to produce 3D surface model. Although water lines failure is not continuously distributed (failure occurs at separate points in geographic space), values between known points can be estimated to construct a continuous surface representation. In vector GIS, the TIN model represents a surface as a set of contiguous, non-overlapping triangles. Within each triangle the surface is represented by a plane. The triangles are made from a set of points called mass points, which

corresponds to the number of failure points in each cell of quadratic analysis. They are represented by a sequence of three nodes. Each cell has an x , y coordinate and a surface, or z-value that present the density of failure.

The central part of study the region shown in Fig. 2.35 indicates high density of water pipelines failure. Interpolated surface is shown thematically by shading each cell with dark or light color depending upon whether that cell is estimated to have a lower failure value (lighter shades) or higher failure value (darker shades).

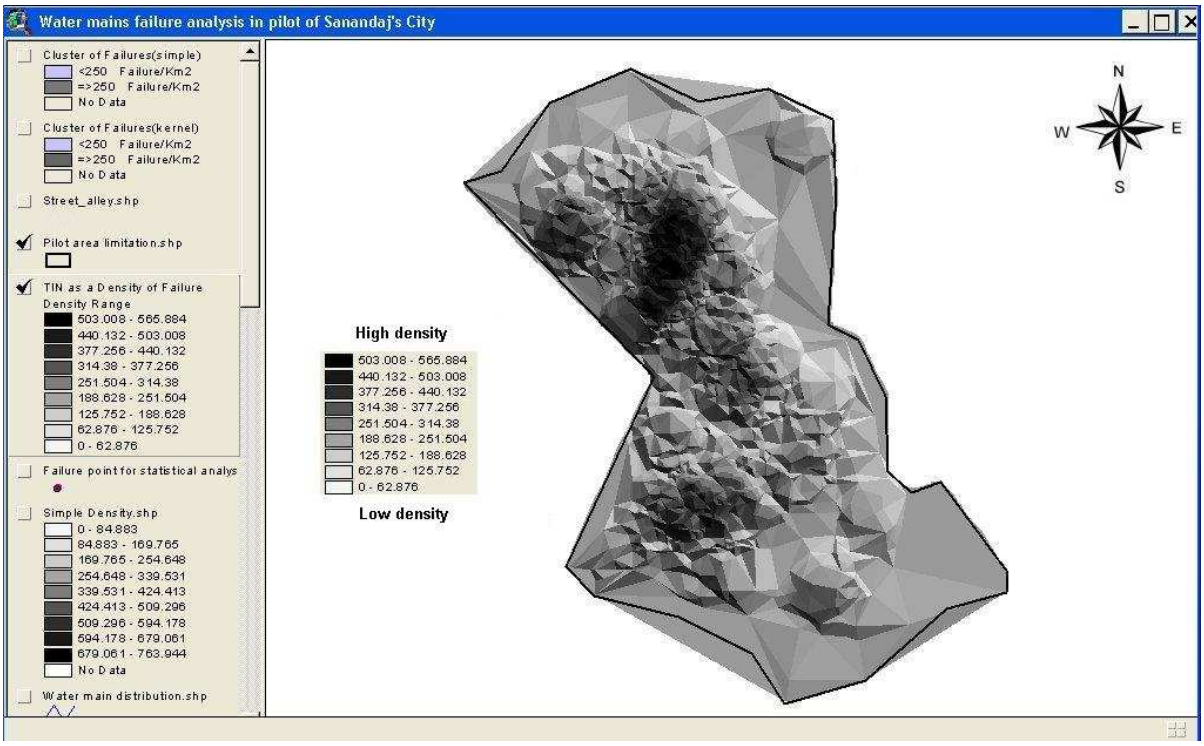

Fig. 2.35 A TIN-based failure density surface

Density of failure can also be represented using a 3-dimensional plot as shown in Fig. 2.36.

Again, x and y are used herein to define a location in the coordinate system, while the z-axis denotes the total failure density over the study period (i.e., 1995 to 2004). As noted in Fig. 2.36, maximum density of failure occurs in the central business district (i.e., along the major transportation arteries "Shohada Street").

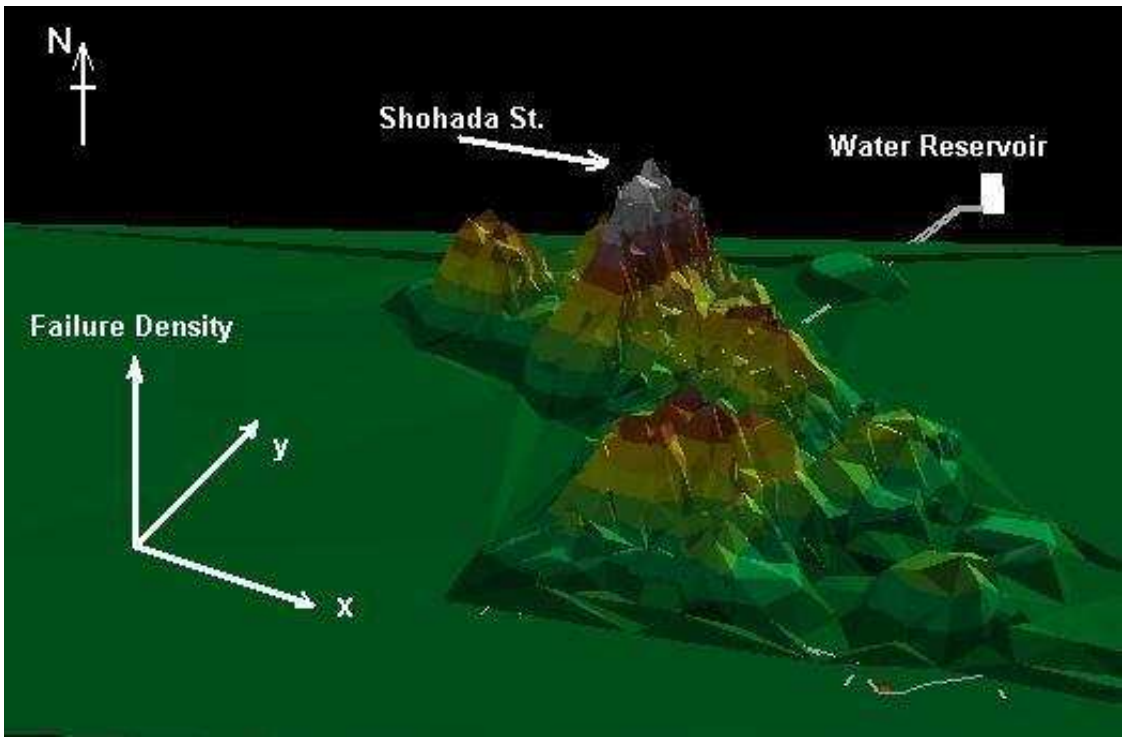

Fig. 2.36 Three-dimensional representation of water pipelines failure density in study area

A significant density increase is observed in the southwest part of the city and a small peak is also noted in the northwest region. The result of this analysis can be applied by cluster replacement scenarios as below.

2.6 Cluster Replacement Scenarios

Because of relatively high setup costs, pipeline renewal is often more efficient if it is done in clusters rather than on an individual pipe-by-pipe basis (Moglia et al., 2006). Clusters of pipes can be chosen for instance on the basis of high density of water pipelines failure leading to high risk (Fig. 2.35 & 2.36). When a high risk area was identified through this spatial analysis, the area is analyzed in more detail so that a decision can be made to establish exactly which pipes to replace. From point of maintenance view, the decision to replace a cluster of pipes is done in a sequential process as described in the work of Moglia et al. (2006).

2.7 Concluding Remarks

This chapter includes two main analyses using data on water pipeline failures in Sanandaj city in Iran. Preliminary statistical analysis gives insight on the impact of different risk factors on the structural deterioration of water pipes. In the rest, geostatistical analysis were addressed a process for using integrated spatial and statistical analysis to discover not only the distribution of water pipelines failure in space but also indicate various spatial trends in failure.

The results of descriptive analysis organize and summarize the data and indicate factors affecting pipe failures in study area. It was the first step to gain a better understanding of the failure mechanisms. Nine accessible factors were identified in a more objective manner to use in statistical investigation. Time dependent factors such as age of pipe and number of pervious failure (NPF) and static factors including pipe length, diameter, thickness, depth, material, pressure and traffic category were analyzed. Indeed, seven most probable cause and mode of water pipelines failure besides their corresponding percentages which vary depending on the pipe's material were identified.

Through the spatial and proximity analysis of water pipelines failure, spatial distribution of historical breaks was established by using appropriate interpolation methods. The failure density is calculated using raster and vector format. In this case, the simple kernel function is used. Interpolated surface estimates show how the intensity of the failure point pattern varies over the study area. It is useful for modeling the likelihood of incidents as well as the relationship between incidents and the underlying risk variables. This study indicates that a significant number of failures appear in geographic clusters. Notably, it shows a point distribution with a strong concentration in the downtown area. These areas are highly vulnerable to future failures because of traffic load.

In conclusion, analyses conducted via the point mapping, cluster identification and spatial interpolation techniques constitute powerful tools that can help in understanding the water pipelines failure pattern problem. The obtained understanding can provide a reliable method of predicting areas of the city that will need water main system replacement in the future. Therefore, planners and engineers with rational and objective justifications will implement an effective preventative maintenance strategy.

3. Statistical Data Analysis and Regression Modeling

3.1 Introduction

Although the analysis in chapter 2 clearly defined certain trends in the past water main breaks throughout the pilot area of Sanandaj city, in order to obtain results beneficial to the SWWU it would be necessary to conduct a more in-depth analysis. This chapter applies statistical approach for a sound analysis of water pipelines failure data in study area.

The goals of analysis in this chapter are two folds. At the beginning, we explore what factors is most probable to affect failure for water mains. This part determines whether a range of variables relating to pipeline construction, hydraulic, operational, and environmental conditions and other associated characteristics were likely contributors to the deterioration of water mains. Then, the relationship between these indicators and failure trends were modeled by two regression models: Multiple and Poisson regression. Fig. 3.1 illustrates several steps in modeling process. Two main other approaches, ANNs and Survival modeling, which have been taken to achieve the objective of thesis will be described in chapter 4 and 5.

According to diagram in Fig. 3.1, we firstly undertook a series of univariate analysis to explore the data for each indicator alone. Non-normal distribution of the data, particularly in form of large skewness, can result serious errors in analysis and incorrect conclusions. Therefore, normality test has been performed for all indicators, hence, the results of this test will be the base for bivariate and multivariate analysis. Secondly, bivariate analysis has been employed to investigate the relationship between the number of failure in one hand and the selected possible indicators in the other hand. The indicators that have significant relationship with water pipelines failure were selected as determinant indicators. These determinant indicators were used as an input for the prediction models. The degree of relationship between two sets of variables was measured by correlation coefficient (R^2). To avoid choosing independent variables which are so highly correlated, multicollinearity analysis among predictor variables was done. Thirdly, multivariate exploratory techniques mainly factor

analyses were applied in this context to discover underlying determinant factors water pipelines failure in study area. Factor analysis also provides information on the relative relationships among variables. Fourthly, modeling with Multiple-linear regression and Poisson regression has been used to examine the relationship between finding influential indicators and number of failure (NF) as response variables.

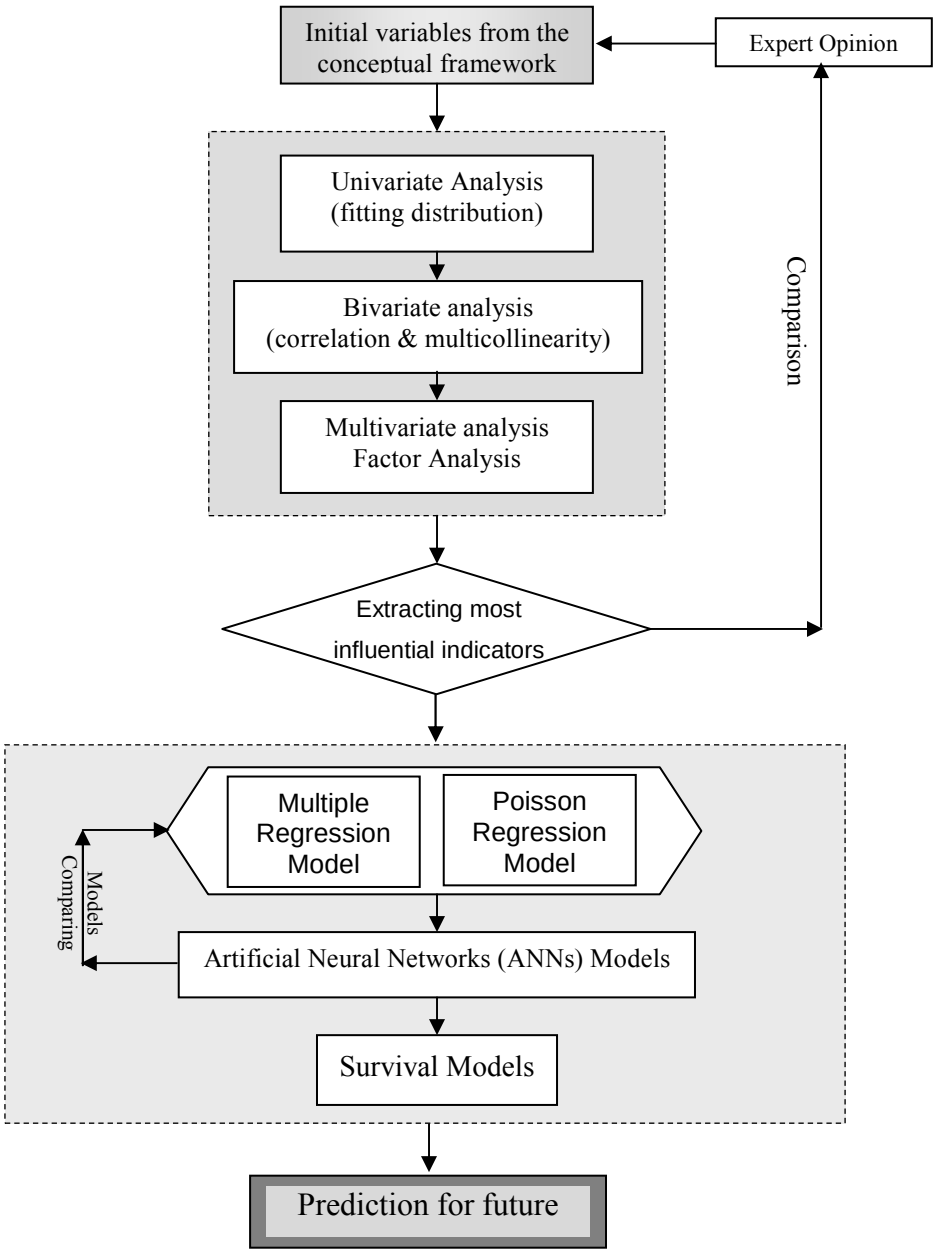

Fig. 3.1 Analysis plan and methodology

3.2 Exploring the failure data

As noted in the literature review, there has been not only relatively little success in developing statistical models for detailed prediction of pipe failure but also most water utilities have adopted subjective manner. In this thesis, we developed prediction models for water pipelines failure in Sanandaj city-specific conditions. Subsequent to this, the initial part explores data through univariate procedure.

3.2.1 Univariate data analysis

In any data analysis it is always a great idea to do some univariate analysis before proceeding to more complicated models. Univariate analysis explores each variable in failure data set, separately. Inspection of the distributions of variables was critically important when using the generalization of the linear regression model such that dependent variable whose distribution follows several special members of the exponential family of distributions. For example, we considered the number of water pipelines failure that occurs in a certain time interval.

In fact, when the assumption of normality is violated, interpretation and inference may not be reliable or valid. Therefore, normality test for each indicator has been performed using fitting empirical distribution. The most common of these tests are graphical presentation of variable distribution. Results from normality test for all the variables gave evidences that some of the indicators have non-normal distribution with mostly a positive skewness. Slight deviation from normality typically did not have significant effect on the statistical analysis. As a first improvement, data transformations have been performed using *log (10)* of the indicators values in order to minimize the skewness and produce a normally distributed data.

To determine how well a specific theoretical distribution fits the observed data, we used the Probability-Probability (or P-P) plot. In this plot, the observed cumulative distribution function was plotted against a theoretical cumulative distribution function in order to assess the fit of the theoretical distribution to the observed data. It should be approximately linear if the specified distribution is the correct model.

For instance, the logarithms of water pipelines length fall nearly along a line in this plot, and we would infer that they are well modelled as a normal distribution. But water pipelines

length do not come closer to a straight line in this plot, at the ends of data there is some deviation from line fit, such as an *S* shape along the diagonal line (Fig. 3.2). Thus, we inferred that they are poorly modeled as a normal distribution. Another variable, maximum pressure of water mains, follow the normal distribution. As shown in Fig. 3.2, all point fall onto a diagonal line (with intercept 0 and slope 1), then we didn't need to transform the data to bring them to normal distribution pattern. Age of water pipelines is also variable which follow moderate normal distribution.

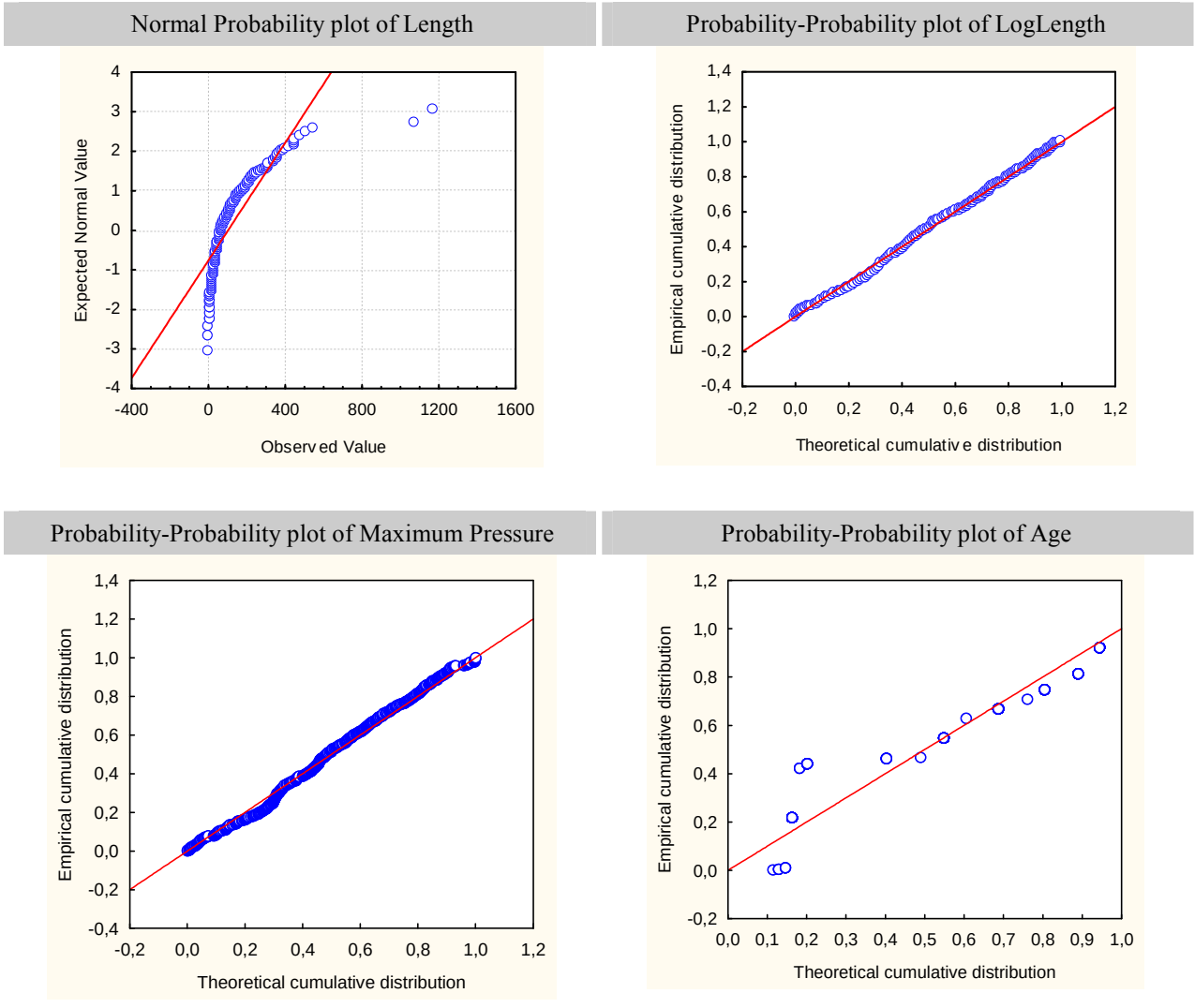

Fig. 3.2 Normality test of water pipelines length, age and maximum pressure through P-P plot

3.2.2 Bivariate analysis

A series of bivariate analysis have been taken to explore the concept of association between two indicators. A major consideration in any model is that the independent variables are statistically independent. Non-independence is called multicolinearity which means that there is overlap in prediction among two or more independent variables. This can lead to uncertainty in interpreting coefficients as well as an unstable model that may not hold in the future. Generally, it is a good idea to reduce multicolinearity as much as possible. To evaluate this matter, the most widely used type of correlation coefficient (pearson's *R*) were measured. It presents the values of two variables that are correlated to each other:

$$R_{xy} = \frac{\mathrm{cov}(x,y)}{s_x s_y} \qquad \mathrm{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \tag{3.1}$$

The correlation is high if it can be approximated by a straight line. However, several authors have offered guidelines for the interpretation of a correlation coefficient. Smith (1986), for example, suggested the following interpretation for values of R^2 between 0.0 and 1.0:

Table 3.1 Interpretation of the size of a correlation

Correlation Coefficient	Interpretation for correlation
$R^2 \geq 0.8$	strong correlation exists between two sets of variables
$0.2 < R^2 < 0.8$	correlation exists between the two sets of variables
$R^2 \leq 0.2$	weak correlation exists between the two sets of variables

To address this issue, the water main break data were analyzed for correlation between independent factors. Graphically, Fig. 3.3 presents the relationship between the thickness of water pipelines in one hand and diameter and depth of pipe burial in the other hand. It is obvious that there is a moderate correlation between the mains thickness and their diameter and depth with R-square equal to 0.26 and 0.31, respectively. It can be just explained the bigger and longer mains have more thickness. On the other hand, correlation between respond variable (NF) and independent variables has been examined. Graphs in Fig. 3.4 shows variation of failure frequencies according to history of failure and water pipelines diameter.

These correlations shows that more pervious failure on pipelines accelerates the coming failure ($R^2=0.86$). Further, failure tend to occur in small mains.

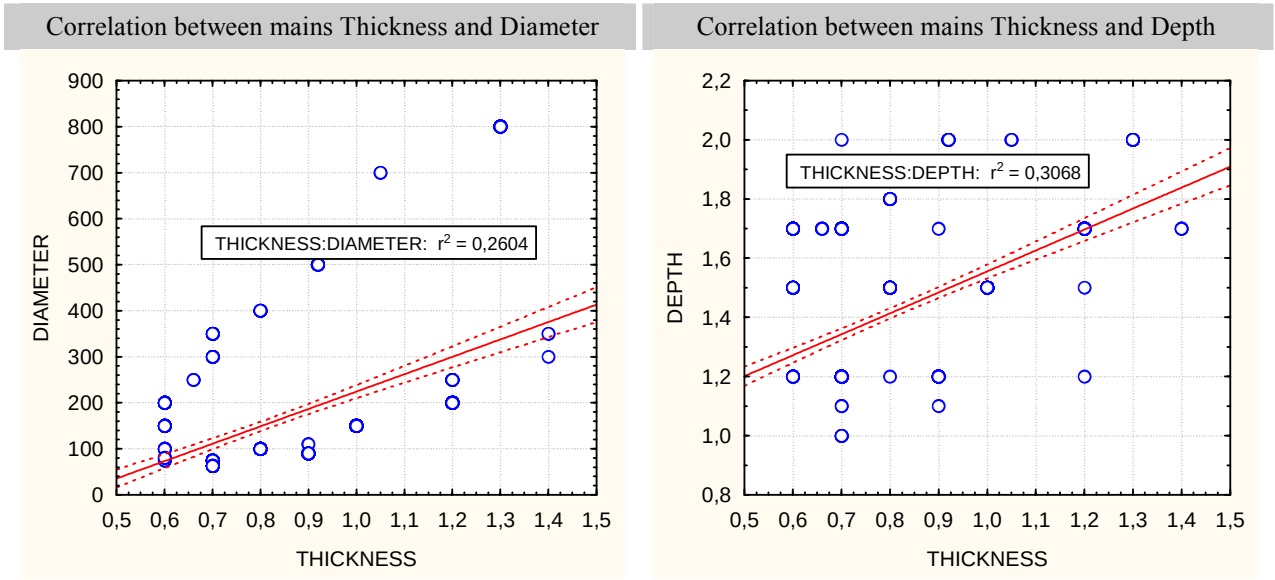

Fig. 3.3 Bivariate correlation among thickness with depth and pipes diameter

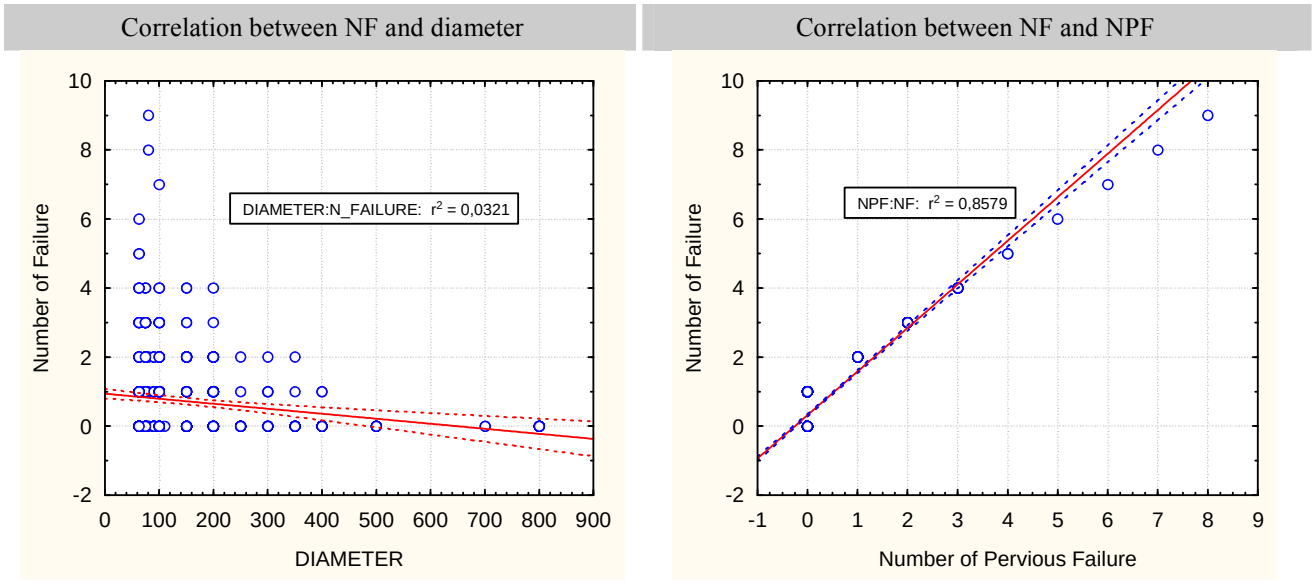

Fig. 3.4 Bivariate correlation among number of failure with diameter and pervious failures

Moreover, correlations among pipe breakage and all known physical, environmental and operational factors were computed. From the result of this analysis, the predominant factors that influence the pipe breaks were identified. Table 3.2 summarize correlations among the

nine variables included in the analysis. Abbreviations are as follows: *NF = number of failure*, *DR = diameter*, *LL = log-length*, *DP = depth*, *TK = thickness*, *AG = age*, *MP = maximum pressure*, *MT = material*, *TL = traffic load* and *NPF=number of pervious failure* . The item Number of failure in water pipelines shows relatively strong correlations with natural logarithm of length, diameter, thickness, depth, material and number of pervious failure. All bold values indicate a statistical significance of $p<0.05$ values with superscript **ns** is not statistically significant correlated. The length of a pipe has an effect on the number failures per pipe since we are measuring failures per pipe and not per pipe length. The highest number of failures is found in pipes with small diameters and reduced wall thickness as well as pipes laid in less deep. In contrast, correlation coefficient equal to 0.76 show that the previous failures of a pipe is a significant factor for the occurring future failures. Additionally, several factors such as maximum pressure, age and traffic load haven't correlated significantly with failure. Table 3.2 illustrates the statistical correlation coefficients.

Table 3.2 Correlations among 10 water pipelines failure indicators

	Correlations									
Variable	NF	DR	LL	MP	DP	TK	AG	MT	TL	NPF
NF	1.00									
DR	**-0.18**	1.00								
LL	**0.36**	0.18	1.00							
MP	0.03^{ns}	0.26	0.05	1.00						
DP	**-0.15**	0.82	0.20	0.14	1.00					
TK	**-0.24**	0.51	0.14	0.09	0.55	1.00				
AG	0.06^{ns}	0.64	0.21	0.02	0.83	0.26	1.00			
MT	0.09^{ns}	-0.41	-0.22	0.07	-0.74	-0.65	-0.64	1.00		
TL	0.01^{ns}	0.53	0.15	0.02	0.59	0.32	0.58	-0.41	1.00	
NPF	**0.76**	-0.20	0.35	0.03	-0.20	-0.20	0.02	0.03	0.01	1.00

(Correlations are significant at $p < 0.05$ ns: is not statistically significant correlated)

The subsets of predictor variables that best predict a response variable is shown in bold font. Therefore these five indicators should be considered as the main driving forces behind failure increasing in Sanandaj city. In order to establish a simple and not complicated model, the first

5 indicators that have significant correlation with the *NF* will be the main focus in the further analysis. The other 4 indicators have insignificant correlation with *NF*. It doesn't mean that they have no contribution to the failure frequencies but their contribution is low compared to the first 5 indicators and it needs to more examination such as evaluation by pipe material. Since each material has own behavior in water pipelines failure (Fig. 2.3), we also measured correlation through indicators for each material. Table 3.3 describes the correlation coefficients among failure indicators by material. Depends on each material category, the factors were varied to correlate with number of failure (*NF*).

Table 3.3 Correlations among nine water pipelines failure indicators (by material)

	Correlation			
Material	Asbestos Cement	Polyethylene	Cast Iron	Ductile Iron
Variable	NF_{AC}	NF_{PE}	NF_{CI}	NF_{DI}
NF	1.00	1.00	1.00	1.00
DR	**-0.23**	-0.07 ^{ns}	**-0.34**	**-0.32**
LL	**0.37**	**0.53**	**0.42**	**0.28**
MP	0.14^{ns}	**0.14**	0.07 ^{ns}	-0.10 ^{ns}
DP	**-0.19**	**-0.27**	**-0.37**	**-0.27**
TK	**-0.20**	-0.08 ^{ns}	-	**-0.29**
AG	**-0.25**	-0.10 ^{ns}	-0.05 ^{ns}	**-0.32**
TL	-0.11 ^{ns}	**0.21**	-0.07 ^{ns}	0.06 ^{ns}
NPF	**0.69**	**0.87**	**0.93**	**0.74**

(Marked correlations are significant at $p < 0.05$ ns: is not statistically significant correlated)

It is obvious from the correlation matrix that the correlation of *NPF* and *LL* have positive and *DP* and *DR* have negative correlation with *NF* in all material categories. Moreover, the insignificant indicators in table 3.2 were significant in different material. For instance, *MP* and *TL* in polyethylene and *AG* in asbestos cement and ductile iron water pipelines are significant. Then, it was concluded that in future modeling, 9 indicators must be considered.

3.2.3 Multivariate exploratory techniques- Factor analysis

Multivariate statistics provide the ability to analyze failure data set where there are many independent and dependent variables which are correlated to each other to varying degrees. For analysis involving multivariate statistics, factor analysis has been done.

Exploratory Factor Analysis is a technique which allows us to reduce a large number of correlated variables to a smaller number of "super variables" . It does this by attempting to account for the pattern of correlations between the variables in terms of a much smaller number of latent variables or Factors:

$$Factor_1 = \beta_1 Variable_1 + \beta_2 Variable_2 + ... + \beta_n Variable_n \quad (3.2)$$

where: β = the weights of a variable on a factor and are called factor loading

Factor analysis boils down a correlation matrix into a few major factors so that the variables within the same factor are more highly correlated with each other than with variables in the other factors. It is assumed that the observed variables are correlated or go together because they share one or more underlying causes. Factors that emerge in the analysis of change will show which variables tend to change together over time, and in which direction change takes place. Variables with factor loadings of 0.6 or greater ($\beta \geq 0.6$) are considered in interpreting each factor, with particular emphasis given to items with loadings greater than 0.6 (McDade & Adai, 2001). This analysis was performed for variables shown in table 3.2 by using STATISTICA software (Version 7.0) produced by Statsoft Inc., USA.

Factor analysis involves a two-step process. Initially, the elements are resolved into their principal components via principal components analysis. Determining the principal components requires transforming the data into orthogonal variables using the eigenvectors of the matrices of the original variables (Trost and Oberlender, 2003). Each principal component is a linear transformation of the original variables. Because the principal components are orthogonal, no interdependence or multi-collinearity exists in the transformed data. Once the principal components are determined, a factor rotation is performed. Factor rotation involves rotating the principal components about the axes of the original variables. The factor rotation preserves the orthogonality of the principal components, but a new transformation matrix is

formed with each rotation. Different methods exist for performing factor rotations. A preferred method, known as the method of maximum variance (varimax normalized), results in a series of rotations wherein each rotation creates a new variable or factor such that the maximum remaining variance in the data is explained by that variable (Trost and Oberlender, 2003). An important consideration during factor analysis concerns the number of factors to resolve during the analysis. The number of factors can range from one to the total number of original variables, which are 9 factors. Typical rotational method using varimax normalized has been employed to obtain a clear pattern of loadings, that is, factors that are somehow clearly marked by high loadings for some variables and low loadings for others. Several guidelines were followed to assist in determining how many factors to be included in the factor analysis. One of the most common guidelines is the minimum eigenvalue criterion. Essentially, this method involves ranking the eigenvalues of the principal components of all the variables from greatest to least, then counting the number of eigenvalues greater than one. Another important consideration in deciding the number of factors relates to the interpretability and meaningfulness of the resultant factor groups.

Applying factor analysis to 9 dominant variables in the water pipelines failure on Sanandaj city, it is noticed that only 2 factors were successfully extracted with eigenvalue more than 1.0. Table 3.4 presents these two factors with their eigenvalues and total variances (accounted for and cumulative) corresponding to the principal components.

Table 3.4 Eigenvalues and total variances for new factors

Factor Number	Eigenvalue	% of the Total variance	Cumulative Eigenvalue	Cumulative %
Factor 1	3.968628	44.09587	3.968628	44.09587
Factor 2	1.384831	15.38701	5.353459	59.48288

Extraction Method: Principal components analysis

From the second column of the table above (Eigenvalue), we find the variance on the new factors that were successively extracted. In the third column, these values are expressed as a percent of the total variance. As we can see, factor 1 accounts for 44.1 % of the total variance and comprised of variables like diameter, depth, thickness and ages (Table 3.5). It means that factor one explained more than 44.1 % of the actual rate in water pipelines failure of study

area, factor 2 accounts around 15.4 %. Although there are four more factors, they are all having variances less than 10 %. Put in other word, the first two eigenvalues cumulated around 59.9 percent of the total variance, while the other 4 factors explained less than 40.1 percent of the rates in the failure. The plot magnitude of eigenvalues versus the number of factors (Skree test) was proposed by Cattell (1966). It retains factors which are above the inflection point of the slope. Accordingly, the number of factors is chosen where the plot levels off to a linear decreasing pattern. The Skree plot is shown in Fig. 3.5, which also includes the percentage variances explained by two selected factors.

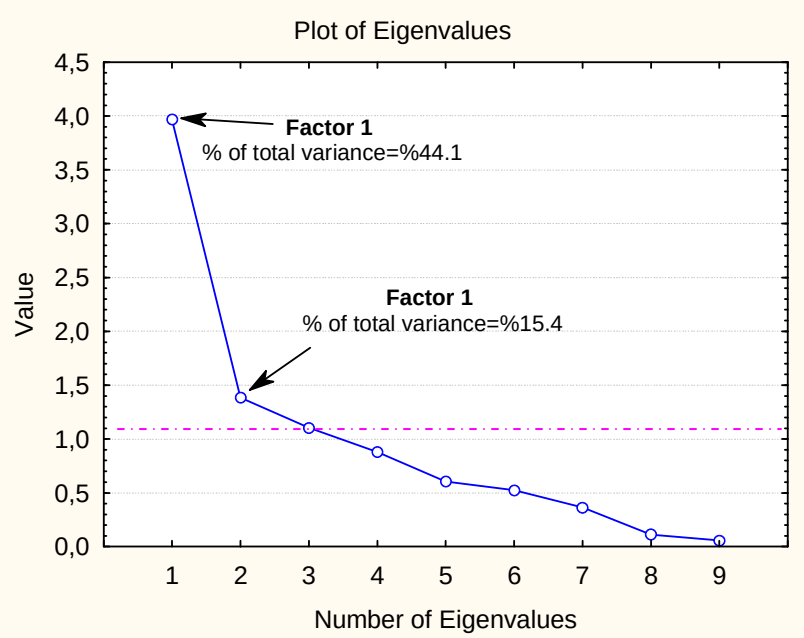

Fig. 3.5 A line graph of the eigenvalues for factor analysis

Therefore, the approach employed in this analysis is in selecting only these two factors with eigenvalues greater than 1 out of 6 factors. Table 3.5 presents the rotated factor-loading matrix for two factors. In fact, a factor loading is the degree to which every variable correlates with a factor. It identifies the major contributing elements to each of factor groups. Then, titles can be given to each factor group based on the perceived relationships among the primary indicators in each factor. In factor loading, a positive loading (e.g. 0.95) will indicate a positive relationship with the factor, whereas one with a negative sign (e.g. - 0.79) will suggest an inverse relationship. It can be also seen from this table that most of variables

associated with each factor are well defined and contribute very little to other factors, which help in the interpretation of results.

Table 3.5 Factor loadings matrix for water pipelines failure parameters

Variable	Factor 1	Factor 2
Diameter	**0.834754**	0.099547
Log Length	0.261178	**-0.744284**
Material	-0.790907	0.055027
Depth	**0.954680**	-0.000303
Thickness	**0.676034**	0.204554
Max pressure	0.130955	-0.039516
Age	**0.827456**	-0.165356
Traffic Load	-0.695220	0.091706
NPF	-0.143863	**-0.861711**

(Extraction Method: Principal components analysis ; Rotation method: Varimax with Kaiser normalization)

Further, it can be inferred that the first factor (F1), which explains % 44.1 of the total variance, is related to the variables diameter, depth, thickness and ages of water mains. While these parameters are positively loaded with this factor. Factor 2 (F2), on the other hand, explains % 15.4 of the total variance and is negatively loaded with logarithm of mains length as well as number of pervious failures in water mains. Factor 3 to factor 5 was strongly associated with an individual variable (unique factors) and they have a moderate to weak influence on the remaining indicators. Thus, they were considered as respective variables. Factor 6 (F6), explains 5.6 % of total variance and is loaded positively with water main's diameter and depth of burial. This factor may be termed as Geometry factors.

The factor loadings shown in Table 3.5 are represented by two-dimension scatter plot in Fig. 3.6. Each indicator is represented as a point.

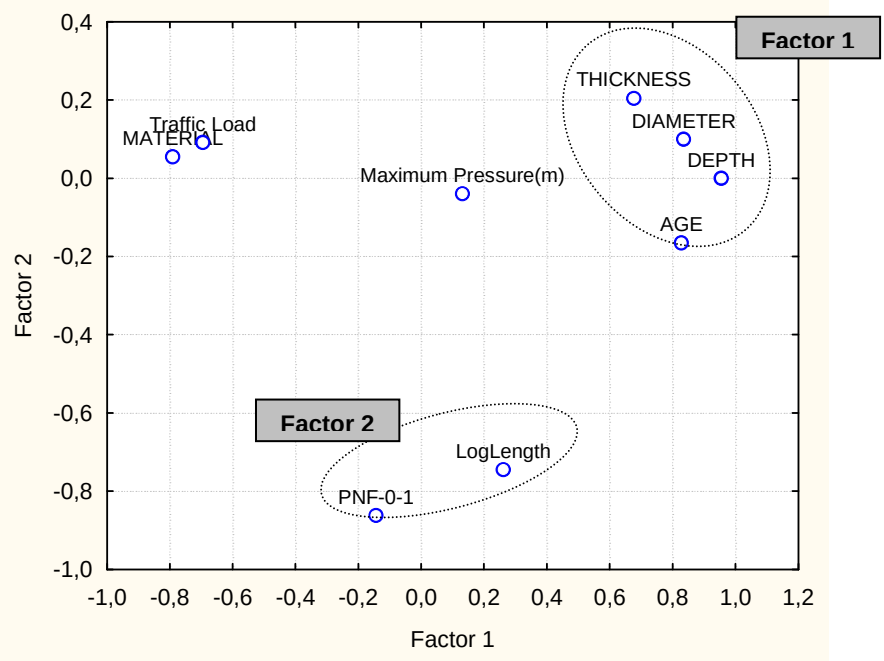

Fig. 3.6 Plot of the two-factor rotated solution for Factor 1 against Factor 2

(Rotation: Varimax normalized ; Extraction: Principal components)

Based on the above discussions, it may be concluded that the analysis must be done for each material of water mains. Overall, Factor analysis is not a simple procedure and it used routinely with many (e.g., 50 or more) variables. In this case study, factor analysis detected simple structure in a few number of variables which affect in failure process.

3.3 Multiple Linear Regression (MLR)

Multiple regression is a linear transformation of the *X* variables such that the sum of squared deviations of the observed and predicted *Y* is minimumized. The prediction of *Y* is accomplished by the following equation:

$$Yi = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} \qquad (3.3)$$

The "*b*" values are called regression weights and are computed in a way that minimizes the sum of squared deviations:

$$\sum_{i=1}^{N}(Y_i - Y_i^{'})^2 \qquad (3.4)$$

The most natural use of multiple regression is when all the variables concerned are continuous. Since this study include two categorical variables 2 two and more level in a multiple regression prediction model, additional steps are needed to insure that the results are interpretable. These steps include recoding the categorical variable into a number of separate, dichotomous variables. This recoding is called "*dummy coding*".

The age , thickness, diameter, depth, maximum pressure of water pipelines and number of pervious failure (NPF) in this survey would represent continuous and pipeline's material and location are categorical variables. The analysis uses of multiple regression to predict of a continuous *Y* with several continuous *X* variables in addition to categorical variables through the dummy coding.

In general form, a regression model where the j^{th} predictor variable is a classifier with *k* level can be interpreted as follows, provided the j^{th} variable in converted to dummy variables:

$$Y = b_0 + b_1 x_1 + ... + \sum_{u=1}^{k_j - 1} b_{ju} D_{ju} + b_p x_p \qquad (3.5)$$

where *Y* is the outcome variable, *b* is a regression coefficient, *D* is a dummy variable for a classifier variable of k levels and x is a non-classifier predictor variable.

Since categorical predictor variables cannot be entered directly into a multiple regression model, a categorical variable with *k* levels will be transformed into *k-1* variables each with two levels. For instance, the material of water pipelines in this study has 4 levels and water pipelines location has 2 levels. Then, three dichotomous variables for material and one for location and history of failure in water pipelines could be constructed. These variables contain the same information as the single categorical variable. They can be directly enter into the regression and also neural network model.

The simplest of dummy coding is for pipe location, it has two level, *1=high traffic street* , *0= light traffic street*. It is converted to one dichotomous variables which called *TL* . If water pipelines location = 1, then *TL* would be coded with a 1. If water pipelines location = 0, then *TL* would be coded with a 0. The dummy coding is represented below.

Table 3.6 Dummy coding for traffic category

Level of Categorical variable	Dummy Coded Variables
	TL
High traffic street	1
Light traffic street	0

This is a nominal variable with two levels, we needed one dummy code to distinguish pipes are located in traffic zone or not. Another categorical variable which has four subgroups is water pipelines material. In the same manner, three dummy coded contrasts would be necessary to use them in a regression analysis. The following table presents data and dummy coded.

Table 3.7 Dummy coding for material category

Level of Categorical variable	Dummy Coded Variables		
	M1	M2	M3
Asbestos Cement (AC)	1	0	0
Ductile Iron (DI)	0	1	0
Cast Iron (CI)	0	0	1
Polyethylene (PE)	0	0	0

Then, the material predictor with three dichotomous variables has been put into a multiple linear regression.

To make a regression model with the whole of sample (including all the materials), Number of Failure (*NF*) is the dependent variable and the independent variables are *DR*, *LL*, *DP*, *TK*, *AG*, *MP*, *MT*, *TL* and *NPF*. The prediction formula has been calculated as below:

$$FailureFrequency = 0.031Diameter + 0.116Thikness + 0.159M1 + 0.034TrafficLoad + 0.339M3$$
$$0.126LogLength + 0.041Max\,\mathrm{Pr}essure + 0.262M2 + 0.643PerviousFailure$$

To measure the overall goodness-of-fit in this regression model, the so-called coefficient of determination, $R^2=0.63$ was calculated, as shown in Fig. 3.7.

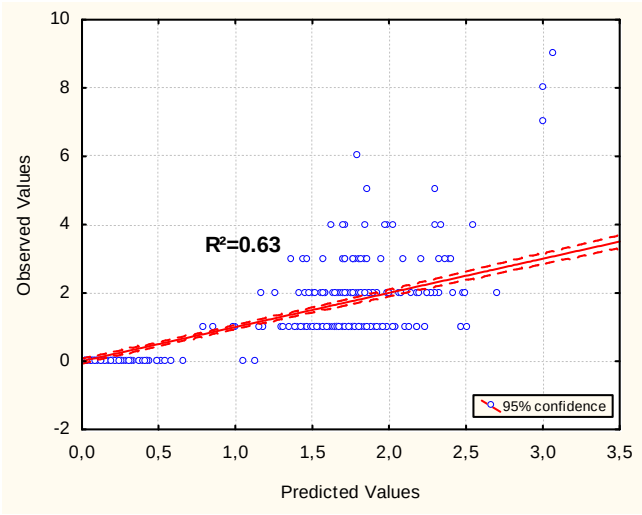

Fig. 3.7 Predicted values verso observed value with $R^2=0.63$

Unfortunately, a high R^2 value does not guarantee that the model fits the data well. There are many statistical tools for model validation, but the primary tool for most process modeling applications is graphical residual analysis. In the regression analysis, we always assume that the error term satisfies: (i) normally distributed with mean 0, (ii) the variance is constant, (iii) errors are independent. In the below, we will control these assumptions.

Estimates for goodness-of-fit

Model validation is possibly the most important step in the model building sequence (Gregory, 2003). Since the model was computed, we tried to evaluate that the model met the assumptions of the regression approaches? To response these question, residual plots were developed for validation of regression model.

Residual analysis

Rather than checking assumptions such as normality on the response variables directly in multiple regression, we checked the normality assumption on the random errors. Because residual analysis is applied to verify the prediction model in multiple regression. Fig. 3.8 depicts the histogram and normal distribution of standard residuals which obtained by dividing the residuals by their standard errors. Also, the assumption that the errors were normally distributed was checked by normal probability plot of the standardized residuals. As

shown in Fig. 3.8 , the points in the normal probability plot deviate from a straight line. Thus, there is statistical evidence against the assumption that the random errors are an independent sample from a normal distribution. Therefore, the non-random structure in the residuals suggest that the model fits the data poorly. Hence, we must examine an alternative solution for modeling.

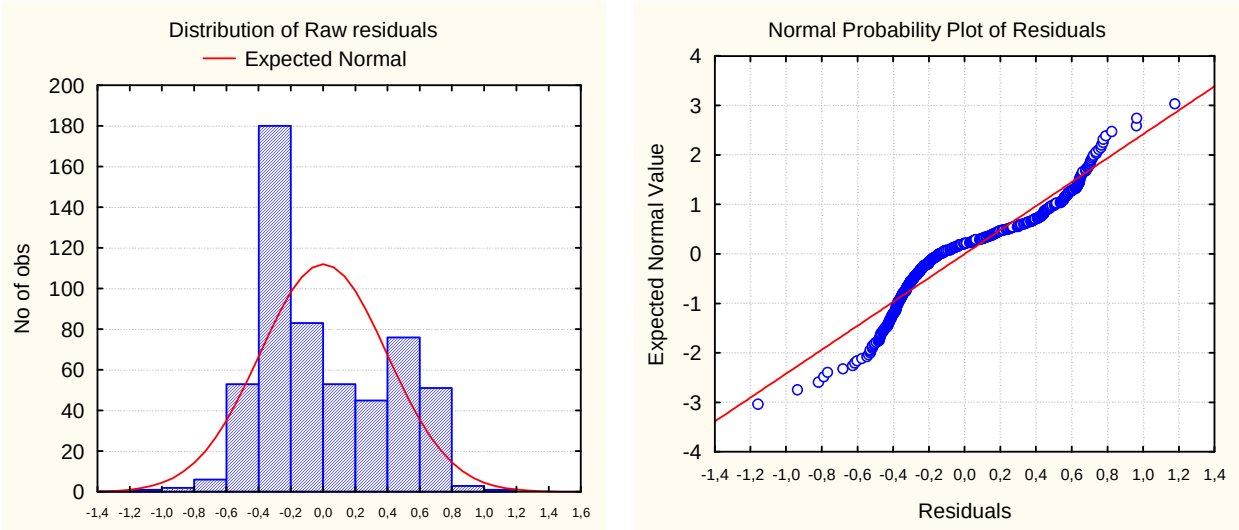

Fig. 3.8 Normal distribution and histogram of standardized residuals and probability plot

Recall that another assumption necessary for the validity of regression inferences is that the error term have constant variance for all levels of the predictor variables. For do this, we also plotted the residuals on the vertical axis against the predicted value on the horizontal axis. This plot does not suggest any systematic deviations (nonconstancy of the variance) nor that the variance of the error terms significantly varies with the level of the predicted values.

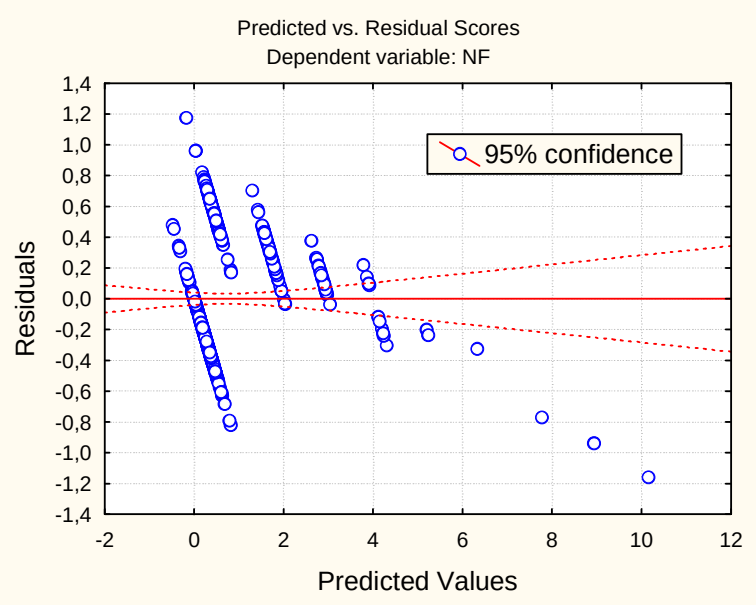

Fig. 3.9 Residual plot against predicted value

As a result, this unequal variances in error terms means that the model fit data inadequately. Henceforth, we tried to model failure via an alternative approaches: Poisson Regression.

3.4 Poisson Regression Model (PRM)

The Poisson regression model (PRM) is similar to regular multiple regression except that the dependent variable Y , number of failure, is a count that follows the Poisson distribution. Thus, the possible values of Y are the nonnegative integers: 0, 1, 2, 3, and so on. Further, it is assumed that large counts are rare. Accordingly, there were two major reasons why this research issue cannot be addressed via straightforward multiple regression techniques (as available in Multiple Regression):

- First, the dependent variable of interest has a non-continuous distribution and was not normally distributed. Thus, the predicted values should also follow the respective distribution.
- Second, reason why the multiple regression model might be inadequate to describe a particular relationship between a water main's condition is that the effect of the predictors on the dependent variable may not be linear in nature. For example, the (average) condition status of water main which has 1 failure as compared to the (average) condition

status of mains which have 2 failures is not markedly different. However, the difference in condition status of pipes with 6 and 8 failures is probably greater. Because of this, some kind of a logarithm function should be adequate to describe this non-linear relationship. Due to limitations 1 and 2 above, in this part Poisson regression can be used to predict failure frequencies with discrete distribution as well as nonlinearity relationship to some covariates.

3.4.1 Fitting a poisson probability distribution

To start modeling, we evaluated the probability distribution of dependent variable. By generating a histogram and fitting a distribution to data, an overall feel for the data were got (Fig. 3.10). This histogram shows the following:

- A "skewed right" (non-symmetric) distribution
- Non-negative value (failure data was non-negative)

In consequence, we could not assume that the probability distribution of Number of failure was normal. By applying the graphical presentation of data and the Kolmogorov-Smirnov test as a goodness-of-fit test it was concluded that the normal distribution is significantly different from the observed data. This distribution deviates grossly from a bell-shaped normal distribution and therefore, we rejected it as a model for the number of failure.

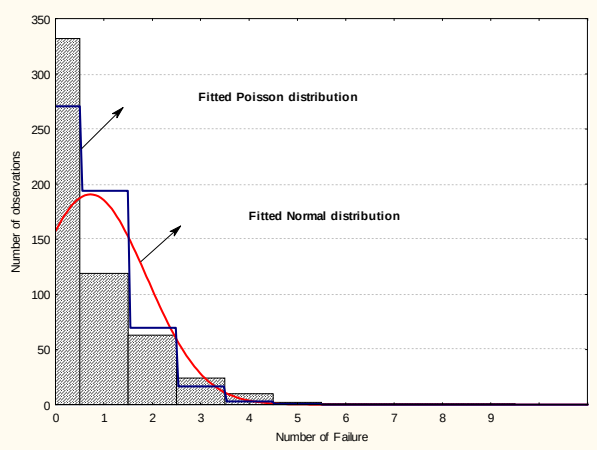

Fig. 3.10 Histogram, fitted Normal and Poisson distribution on failure's number

Fitting distribution was accomplished by the method of maximum likelihood and observed and expected frequency. Graphically, the fit between observed values and theoretical Poisson distribution defined by mean = 0.72 , as shown in Fig. 3.10. Additionally, failure numbers were distributed following the normal distribution with the mean=0.72 and the standard deviation= 1.34. Table 3.8 illustrates two distribution for observed frequency.

Table 3.8 Observed and expected number of failure by fitted Poisson and Normal distribution

Category	Observed Frequency	Poisson distribution		Normal distribution	
		Expected Frequency	Observed-Expected	Expected Frequency	Observed-Expected
0	332	270.5774	61.4226	148.5522	183.4478
1	119	193.8975	-74.8975	181.9670	-62.9670
2	63	69.4741	-6.4741	149.2319	-86.2319
3	24	16.5952	7.4048	60.7389	-36.7389
4	10	2.9731	7.0269	12.2349	-2.2349
5	2	0.4261	1.5739	1.2145	0.7855
6	1	0.0509	0.9491	0.0591	0.9409
7	1	0.0052	0.9948	0.0014	0.9986
8	1	0.0005	0.9995	0.0000	1.0000
9	1	0.0000	1.0000	0.0000	1.0000

The Poisson distribution models the probability of y events (i.e. failure in water mains) with the formula (Gregory, 2003):

$$\Pr(Y = y|\mu) = \frac{(\mu)^y \times e^{-\mu t}}{y!} \qquad y = 0, 1, 2, \ldots , \mu > 0) \qquad (3.6)$$

where:

y = number of failures in time (t)

μ = the mean incidence rate of a failure per unit of time

e = is the base of the natural logarithm (2.71828...).

The Poisson distribution has the property that its mean and variance are equal.

Chi-squared goodness-of-fit test

A chi-squared test was used to test the hypothesis that observed failure data follow a Poisson distribution. In this regard, we had two hypothesis:

- The null hypothesis is $H_0 : X \sim$ Poisson
- The alternative hypothesis is $H_1 : X$ does not follow a Poisson distribution.

Therefore, the basic questions that need to be addressed was: does the assumption of a Poisson distribution seem appropriate as a model for these data? To test this hypothesis, we used the chi-squared statistic which defines as:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \qquad (3.7)$$

where:

O = observed frequency
E = expected frequency

Using the Poisson distribution with μ = 0.72 we can compute p_i, the hypothesised probabilities associated with each class. From these we calculated the expected frequencies (under the null hypothesis) which shown in Table 3.8. The chi-squared statistic was calculated 63.32 . If we look up 63.32 in tables of the chi-squared distribution with df =2, we obtain a *p-value* < *0.001*. By conventional criteria, this difference is considered to be statistically significant. Thus, we conclude that there is a little or no real evidence to undermine the veracity of Poisson distribution. For this, we also evaluated the fitting of distribution by a graphical examination, Probability-Probability plot.

Probability - Probability (P-P) Plot

The Probability-Probability (P-P) plot is a graphical technique for assessing whether or not the failure data set follows Poisson distribution. The data were plotted against a theoretical

distribution in such a way that the points should form approximately a straight line. The correlation coefficient associated with the linear fit to the data in the probability plot is a measure of the goodness of the fit. According to this data, a straight line to the points on the probability plot, the probability plot has a correlation coefficient of 0.9 and 0.53 for Poisson and normal distribution, respectively. In consequence, Poisson probability plot indicates that the Poisson distribution does in fact fit these data better than normal distribution.

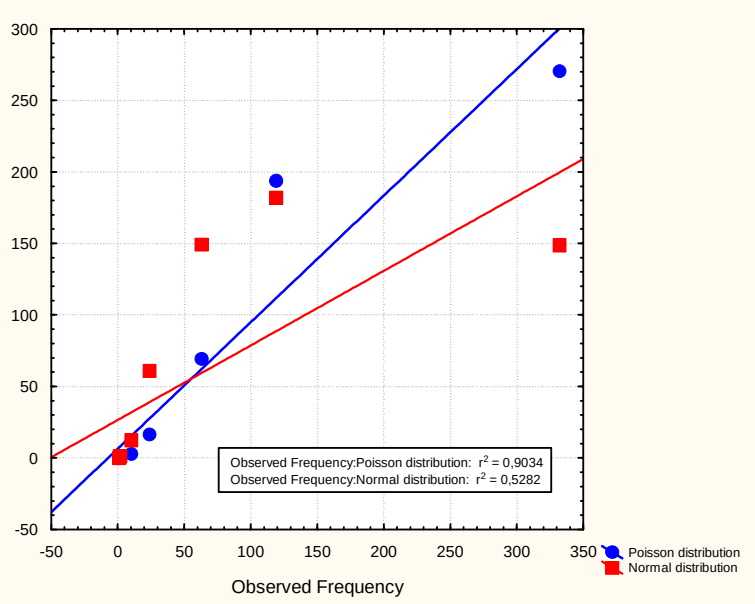

Fig. 3.11 The Poisson probability plot for number of failures in water mains

3.4.2 Estimated poisson regression model

Since the response variable is in the form of counts data, a Poisson regression model was fitted. Further, to model the non-linearity of failure number related to the predictors, based upon an assumption of Poisson distribution for the dependent variable values, we used the *log* link function. Hence, a Poisson distribution with a logarithmic link function was used in the GLZ . So the logarithm of the response variable is linked to a linear function of explanatory variables such that:

$$\mathrm{Log_e(Y)} = \mathrm{int}\,ercept + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_mX_m \qquad (3.8)$$

So, the regression equation is given by:

$$Y = \exp(\text{int}\,ercept + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_mX_m) \quad (3.9)$$

Using this notation, the Poisson regression model for water pipelines failure is written as:

$$\text{Predicted number of failures} = \left| \begin{array}{l} \textbf{exp}\ [\ \text{intercept} + b_1 \times \text{Diameter} + b_2 \times \text{Loglength} + b_3 \times \text{Depth} + \\ b_4 \times \text{Thickness} + b_5 \times \text{Maximum pressure} + b_5 \times \text{Age} + b_6 \times \text{M1} + \\ b_7 \times \text{M2} + b_8 \times \text{M3} + b_9 \times \text{NPF} + b_{10} \times \text{Traffic load}\ ] \end{array} \right| \quad (3.10)$$

Where:

$b_1, b_2, \ldots, b_{10}$ = the regression coefficients (are unknown parameters that are estimated from the selected dataset) and "log" means natural logarithm.

The maximum likelihood method was used to calculate the parameters of Poisson regression model. All statistical analysis was based on analytical sub-routines in STATISTICA Ver. 7 (StatSoft 2004). Thus, the estimated Poisson regression model is:

$$\text{Predicted number of failures} = \left| \begin{array}{l} \textbf{exp}[\ -3.73974 - 0.00317 \times (\text{Diameter}) + 1.21935 \times (\text{Loglength}) \\ + 2.18492 \times (\ \text{Depth}) - 2.09377 \times (\text{Thickness}) + 0.00662 \times \\ (\text{Maximum pressure}) - 0.06569 \times (\text{Age}) + 0.82601 \times (\text{M1}) + \\ 0.23961 \times (\text{M2}) + 1.53406 \times (\text{M3}) + 0.40147 \times (\text{NPF}) - 0.07057 \\ \times (\ \text{Traffic load})] \end{array} \right| \quad (3.11)$$

Put in other word:

$$\text{Predicted number of failures} = \left| \begin{array}{l} (e)^{-3.73974} \times (e)^{0.00317\ \text{Diameter}} \times (e)^{1.21935\ \text{Loglength}} \times (e)^{2.18492\ \text{Depth}} \\ \times (e)^{-2.09377\ \text{Thickness}} \times (e)^{0.00662\ \text{Maximum pressure}} \times (e)^{-0.06569\ \text{Age}} \times \\ (e)^{0.82601\ \text{M1}} \times (e)^{0.23961\ \text{M2}} \times (e)^{1.53406\ \text{M3}} \times (e)^{0.40147\ \text{NPF}} \times \\ (e)^{-0.07057\ \text{Traffic load}} \end{array} \right| \quad (3.12)$$

This equation is useful for estimating the number of failures in water pipelines by the related variables. Obviously, the direct interpretation of regression coefficients is difficult because the formula for the predicted value involves the exponential function.

Adequacy of the model

In order to assess the adequacy of the Poisson regression model, we firstly checked graphical representation of predicted value against observed data. As can be seen from Fig. 3.12, the R^2 value is 0.74, indicating that there is a moderate fit of data using the Poisson model. Needless to say, by comparing the R^2 value for two recent models, it can be concluded that the Poisson regression model ($R^2=0.74$) works better than multiple regression ($R^2=0.63$).

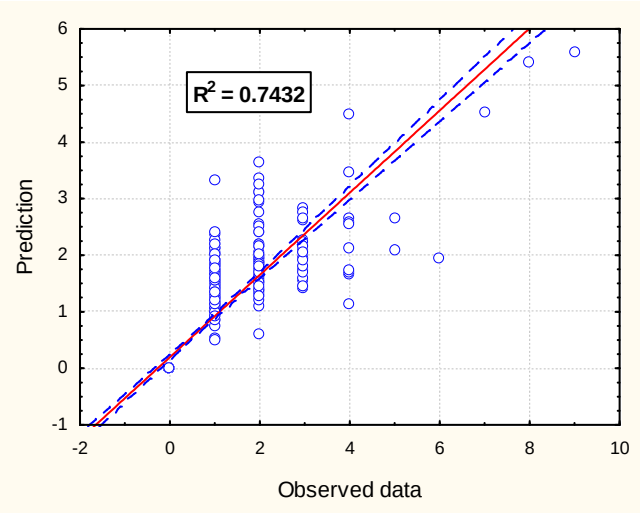

Fig. 3.12 Predicted number of failure data against observed data

Secondly, to evaluate the goodness of fit, a common statistic that was computed is the so-called Deviance statistic. Deviance serves the same purpose as sum of squares in multiple linear regression since the full model always has zero deviance. In this analysis the deviance explained by the regression is 86.7 at 543 degrees of freedom (df). Entering the table for values of χ^2 at 543 *df*, we got p-value less than 0.0001. This p-value is quite small and shows that the model does not fit very well. Table 3.9 contains information on assessment of fit.

The next step is to assess the fit of the respective model by carrying out a diagnostic plot of deviance against fits. In fact, Fig. 3.13 is a plot of the residuals against the predicted values

that showed more or less a horizontal band. The residual plot has several deviance residuals larger than 2 in absolute value, which Fig. 3.14 shows some real outliers. The black solid line is the lowest fit for the deviance plot and the grey dotted line is for zero. As we can see, the black line does not agree with the grey dotted-line for predicted value. This indicates non-equal variances of residuals and perhaps dependence on the predicted values, thereby, confirming the inadequacy of the model.

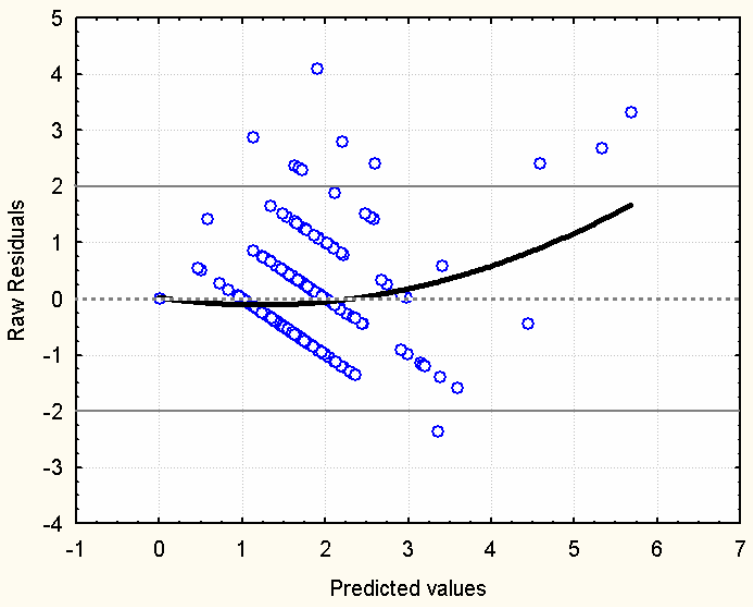

Fig. 3.13 Plot of residuals against fitted values

Test for over-dispersion in poisson regression

Deviance and Pearson Chi-Square divided by the degrees of freedom are used to detect over-dispersion or under-dispersion in the Poisson regression. For Poisson distribution the mean and the variance are equal, which implies that the deviance and the Pearson statistic divided by the degrees of freedom should be approximately one. Values greater than 1 indicate over-dispersion, that is, that the true variance is bigger than the mean, values smaller than 1 indicate under-dispersion, the true variance is smaller than the mean. For the developed Poisson regression model, these values are less than 1, giving evidence of under- dispersion. Therefore, the model does not fit the data well which contain in Table 3.9.

Table 3.9 Goodness-of-fit statistics (Poisson distribution & Log link function)

Statistics	Degrees of freedom (df)	Statistics	Statistics/df
Deviance	543	86.689	0.159
Scaled deviance	543	480.643	0.885
Pearson Chi^2	543	97.935	0.180
Scaled P. Chi^2	543	613.45	1.130

Model checking with observational statistics

Predicted values and residual statistics for each combination of predictors in the model were calculated. Figure below plots the Pearson Chi-square values (contributions to Chi-square) for each case against the predicted values. In this graph, the Outlier shown on the right side of the graph (e.g. the 540th and 551th data point) has a large Chi-square value and thus is the largest contributor to the lack of fit for this model.

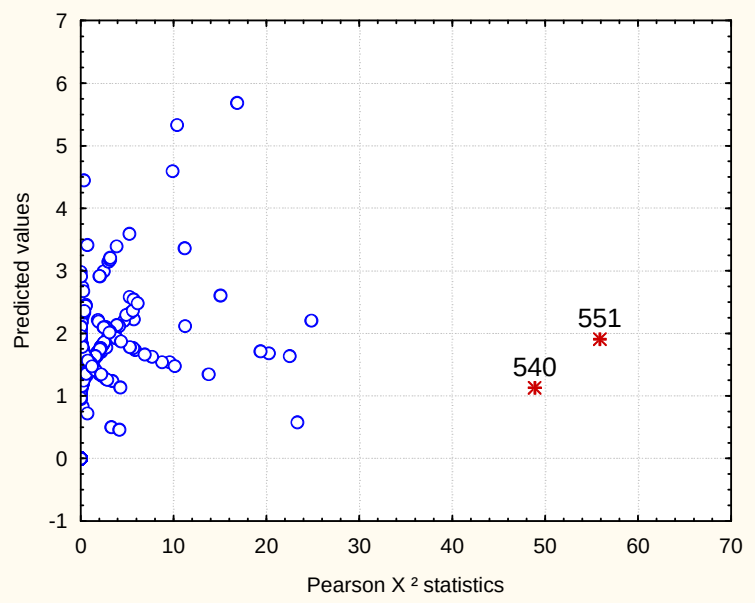

Fig. 3.14 Chi² statistics by predicted values

3.5 Concluding Remarks

This chapter demonstrates mathematically that some factors like diameter, length, depth, thickness, age, material and pervious failure are statistically significant which play a role in the failure of urban water mains. Such information should support the development of predictive programs along water mains.

Second part of this chapter established two regression models for water pipelines failure on the selected area. The water distribution network in Sanandaj city was examined using different regression models: Multiple and Poisson regression. Both models have been employed to analyze water pipelines failure for 10 years data. Since two models impose some special requirements, then it was seriously and deeply assessed. In addition, for evaluate the pre-defined underlying relationship between dependent and independent variables in these approaches, the initial assumptions have been examined. In spite of good fitting by multiple regression ($R^2_{adj} = 63\%$), violation of this model from assumptions leaded to erroneous estimation of failure frequency. Thus, multiple regression fits the data inadequately. To overcome these challenges, an alternative model, Poisson regression, was used on the data set. It fits not only better than Multiple regression model (with $R^2_{adj} = 74.4\%$) but also consider the initial assumption. To assess the adequacy of model, three goodness-of-fit tests were given for the overall fit of a model: Pearson, deviance and Graphical representation. The significant p-value for deviance statistic in the Poisson model and diagnostic plot of deviance against fits confirmed the adequacy of model. But there was also evidence of underdispersion for Poisson model which can be interpreted as an adequacy of model. Overall, by comparing prediction performance via these statistic indicators, this study demonstrates that Poisson regression could not be an alternative method for analyzing water pipelines failure frequency. It seems to handle this type of data better than Multiple regression but it is not sufficient. In conclusion, applying more accurate model such as ANNs has been justified. Neural networks have abilities to adapt data that has been presented to them in the form of input-output patterns. This characteristic of neural networks has earned them the title of "dynamic regression" when compared to rigid regression methods such as Multiple or Poisson regression.

4. Artificial Neural Networks (ANNs) Modeling

4.1 ANNs Modeling of Water pipelines Failure

In this chapter Artificial Neural Networks (ANNs) were developed to predict number of failure (NF) in each water pipelines over the selected area. According to pipe material, 4 models have been constructed separately for whole of water pipelines and 3 models for the failures in metallic, cement and plastic water mains. Based on the characteristics of the available failure data, static back propagation neural network was identified as the most suitable ANNs for developing their prediction models. These model have been successfully used in the past to solve problems that require the computation of a static function (e.g. Najjar et al., 1996). Statistical accuracy measures such as coefficient of determination (R^2), Sum Squared Error (SSE) and the Mean Absolute Relative Error (MARE) on both training and testing data sets were used to filter out the most optimal networks. Graphical performance comparison of each model against observed data in terms of coefficient of determination were also examined.

4.1.1 ANNs background and theory

Artificial neural networks mimic the ability of the human brain in predicting patterns based on learning and recalling processes (Najjar et al., 1997; Al-Barqawi and Zayed 2006). They are considered very powerful predictive modeling technique. ANNs are used here as analysis and predictor tool for exploring and modeling the relationship between the input variables (e.g. material, diameter, etc.) and the predicted variable (Number of failure in mains network).

ANNs consist of a number of artificial neurons variously known as “processing elements”, “nodes” or “neurons”. Processing elements in ANNs are usually arranged in layers: an input layer, an output layer and one or more intermediate layers called hidden layers. Each layer consists of individual neurons such as that depicted in Fig. 4.1.

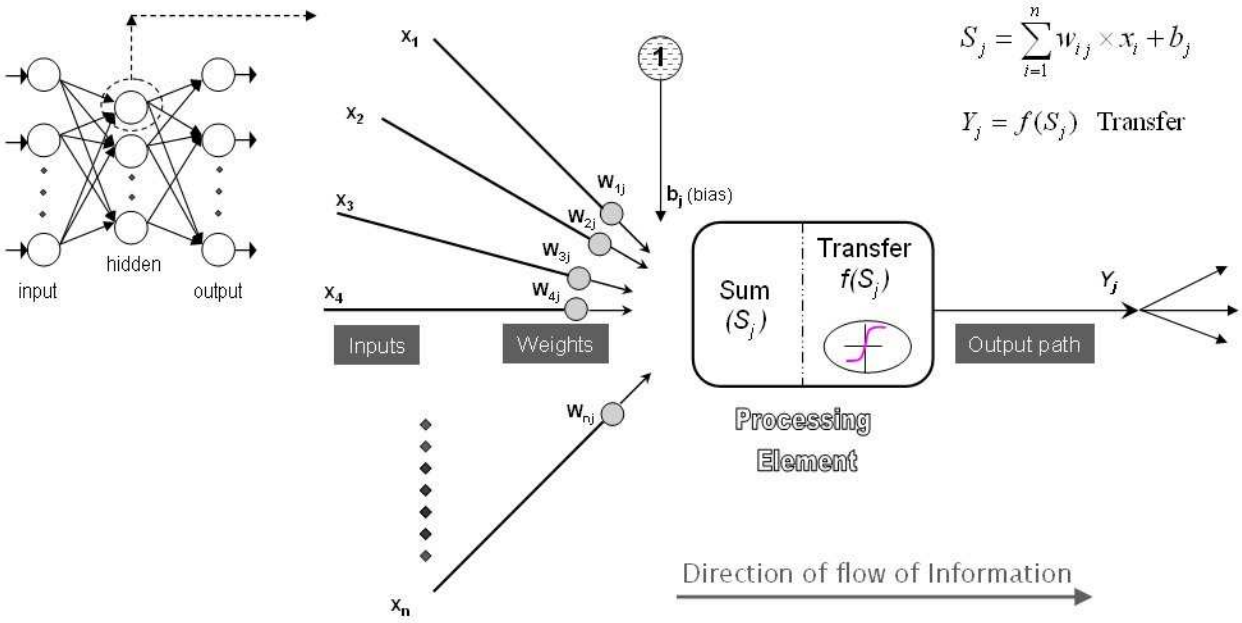

Fig. 4.1 A basic artificial neuron

For a given artificial neuron there is n inputs with signals X_1 through X_n and weights W_1 through W_n . The products of multiplying each input and connection weights, are simply summed, fed through a transfer function to generate a result, and then outputted. The output propagates to the next layer (through a weighted synapse) or finally exits the system as part or all of the output.

In this analysis, all input and output variables were normalized according to the following relation:

$$\overline{X}_{Normalized} = \frac{(X_{Actual} - X_{Minimum})}{(X_{Maximum} - X_{Minimum})} \qquad (4.1)$$

Normalization has been found to more effective in achieving faster training by preventing larger number from overriding smaller one (Najjar et al., 1997). By applying the sigmoid function, the data was normalized between 0 and 1 :

$$Y_j = F(S_j) = \frac{1}{1+e^{-S_j}} \qquad [\ 0 < F(S_j) < 1\ ] \qquad (4.2)$$

The output Y_j passes as a signal to the output node (k). The net entering signal of an output node is:

$$S_K = \sum_{i=1}^{n} \overline{X}_i \times w_{ik} + b_k \qquad (4.3)$$

The incoming signals of the output node (S_k) is transformed using the sigmoid type function to scale the output (Y_k):

$$Y_k = F(S_k) = \frac{1}{1+e^{-S_k}} \qquad (4.4)$$

The scaled output is descaled to produce the target output according to the following formula:

$$Y_k = \overline{Y}_k (Y_{\max(k)} - Y_{\min(k)}) + Y_{\min(k)} \qquad (4.5)$$

4.1.2 ANNs application to water pipelines failure

Since the failure behavior of water pipelines is very complicated, to date a comprehensive fundamental theoretical model has not been produced. Therefore, a reliable empirical method for predicting failure number based on historical data remains the preferred approach. However, the complexity of failure processes means that even traditional methods, such as regression analysis, are handicapped in producing sufficiently accurate models. Artificial neural networks (ANNs) have the ability to derive highly complex relationships and associations from historical data.

Over the last few years, the use of ANNs has increased in many areas of geotechnical engineering. The literature reveals that ANNs have been used successfully in failure prediction and classification in water and sewer pipelines. For instance, the ANNs model has been applied to the water distribution network of a subdivision in Edmonton, Canada (Rajani and Kleiner, 2001). The model was trained with historical input data including temperature, rainfall, operating pressure, and number of breaks. Ahn et al. (2005) has been used ANNs model for predicting water pipe breaks in service pipes and mains in Seoul city (Korea). They observed that the prediction model performed well based on pipe characteristics and water and soil temperatures. Jafar et al. (2005) modeled the failure in water network using ANNs and compared the results with multiple regression. Based on failure data from the city of

Waterloos in France, they found better performance of ANNs modeling. Moselhi and Shehab-Eldeen (2000) deployed the ANNs in the analysis and classification of defects in sewer pipelines. The ANNs was trained to classify four different types of defects including cracks, spalling, joint displacements, and reduction of cross sectional area. However, these models cannot be applied in other networks such as this case study.

4.1.3 ANNs modeling steps

The ANNs modeling process in this research has followed through five steps as diagram below:

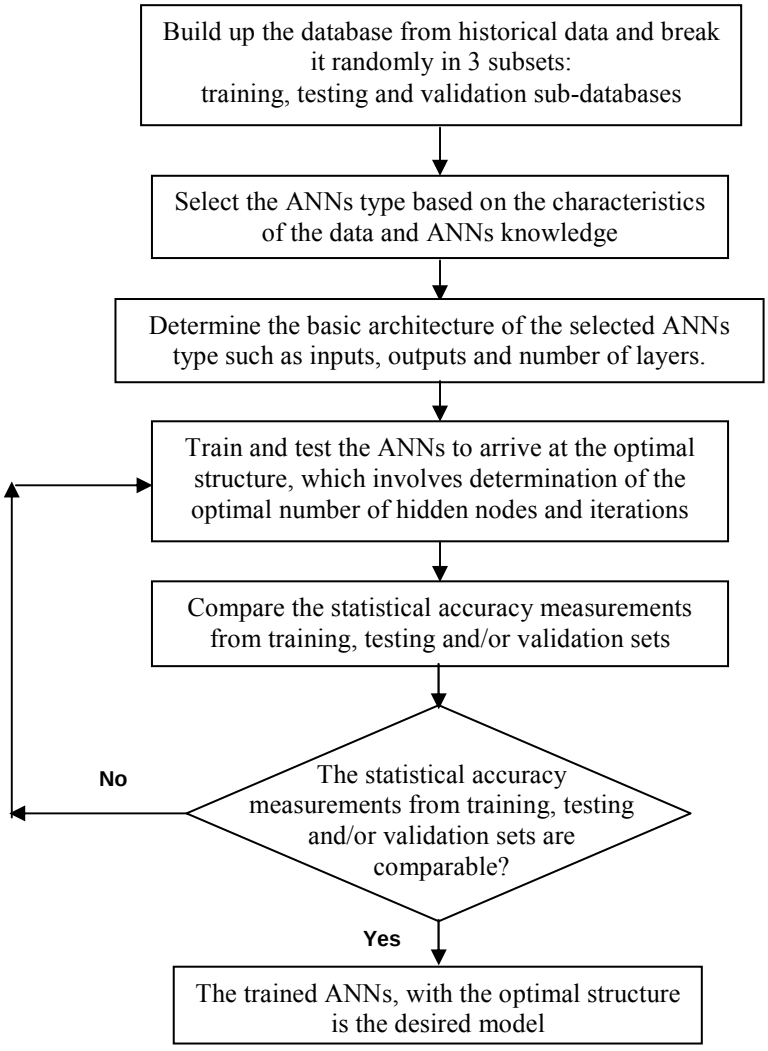

Fig. 4.2 ANNs modeling steps

4.1.4 The ANNs software

The ANNs software used in this work was developed by Professor Najjar from Kansas State University (Najjar, Quick Manual, 1999). TR-SEQ1, source coded in C^{++} language, is a three-layered ANNs training program (1 input layer; 1 hidden layers; and 1 output layer) that is capable for performing simultaneous sequential training and testing. It is a comprehensive, powerful and less time consuming package and characterized by intelligent problem solver that can guide step by step through the procedure of creating a verity of different networks and choosing the network with the best performance.

The required parameters are specified in two main input files: *STP.dat* and *SPEC.dat*. The first file includes all failure records which involve both input variables and observed output for training and testing cases. The input data was normalized and sigmoid function was used. Second file is the specification file which describes the architecture of desired network as well as some information about input and output variables. After training, the program produces five output variables, namely: *result.dat*, *stp.out*, *trhist.out*, *trnet1.out* and *trnet2.out*. One supplementary file was used to validation phase.

In order to produce graphs, all results were imported into the Microsoft Excel and STATISTICA spreadsheet. By taking full advantage of both applications, we plotted all kind of graphs for the results of this analysis and interpretation of results.

4.1.5 Determination of model architecture

Determining the network architecture is one of the most important and difficult tasks in ANNs model development (Maier and Dandy 2000). It requires the selection of the optimum number of layers and the number of nodes in each of these. There is no unified theory for determination of an optimal ANNs structure. It is generally achieved by fixing the number of layers and choosing the number of nodes in each layer. There are always two layers representing the input and output variables in any neural network. Choosing the number of middle layers (hidden) is the most crucial decision in creating the ANN structure. It has been shown that one hidden layer is sufficient to approximate any continuous function provided that sufficient connection weights are given (Hornik et al. 1989). Hecht-Nielsen (1989) provided proof that a single hidden layer of neurons, operating a sigmoidal activation

function, is sufficient to model any solution surface of practical interest. In Geotechnical modeling, one hidden layer is commonly used by Najjar et al. (1997) and Shahin et al. (2001). Increasing the number of parameters of an ANNs by adding hidden neurons or layers, complicates network training. Moreover, a large number of parameters increases the chance of overtraining occurring.

Accordingly, in this study, for an adequate description of the underlying relationship between number of failure and input variables, a neural network with one hidden layer as well as input and output layers has been chosen. Since, our goal is to create a model that correctly maps the input to the output using historical data, we selected the Multi-Layer Perceptron (MLP) networks. This model will then be used to produce the output (number of failures) when the desired output is unknown. The architecture of the statically designed ANNs model for this study is shown in Fig. 4.3.

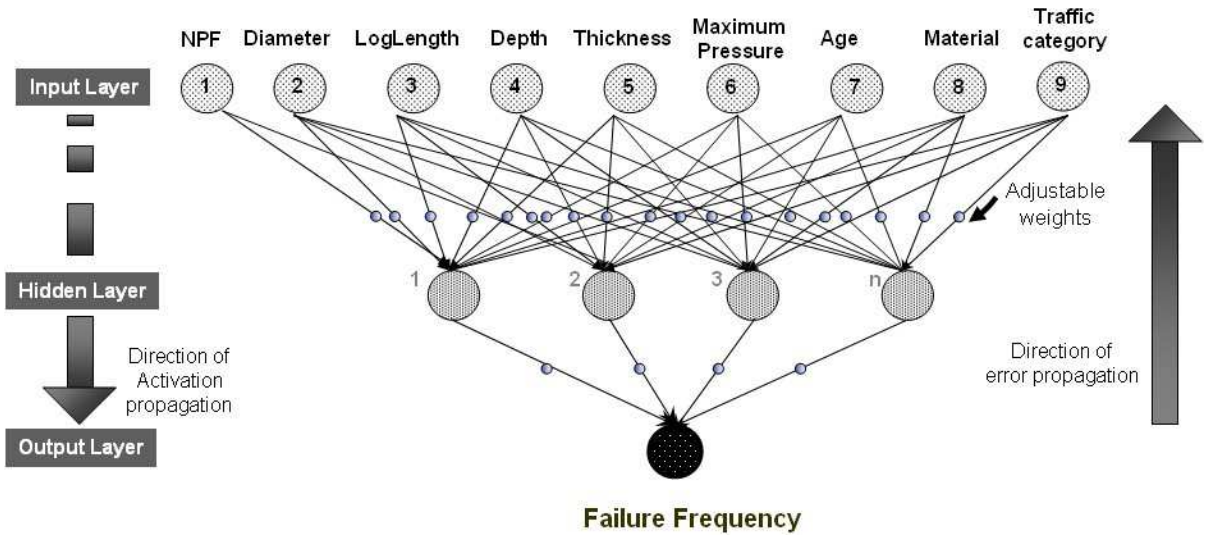

Fig. 4.3 Architecture of designed neural network for prediction of water pipelines failure

An important step in developing ANNs models is to select the model input variables that has the most significant impact on model performance (Faraway and Chatfield 1998; Kaastra and Boyd 1995). A good subset of input variables can substantially improve model performance. Based on a priori knowledge of water pipelines failure, data availability in this case study, as well as the results of correlation and factor analysis in chapter 3, nine input variables (9 nodes) and one output node were selected for the current work. The 9 inputs are:

- Water pipelines diameter
- Logarithm of each mains segments
- Cover of soil upper mains
- Thickness of pipe
- Maximum hydraulic pressure
- Age of water mains
- Water pipelines material
- Traffic category and NPF

The network was used to predict the following output:

- Number of failure in each water pipelines

As more applications, the number of input and output neurons was fixed and then number of neurons within hidden layer will be optimized. So, the best configuration will be found by trial and error for determining the exact number of neurons in a hidden layer.

With respect to Fig. 4.3, the neural network model developed for this study can be expressed in the following compact form:

$$\{Number\ of\ \ failure\} = ANN_{9-NH-1}\begin{Bmatrix} Diameter & LogLength & Thickness \\ Age & Depth & Max\,\mathrm{Pr}\,essure \\ TrafficCategory & Material & NPF \end{Bmatrix} \quad (4.6)$$

The 9-NH-1 label stands for the architecture of the selected neural network. Numbers 8 and 1 denote the number of input and output parameters, respectively. *NH* is the optimal number of hidden nodes which needs to be determined through trial and error in stage two. In the training phase, we started the number of hidden node for first trial by formula below:

$$HN = \frac{Number\ of\ training\ case\ \ -\ Number\ of\ output}{Number\ of\ \ input - Number\ of\ \ output + 1} \quad (4.7)$$

Through this equation, we calculated 24 neurons for a hidden layer in dataset for global model in Sec. 4.2.

The optimal number of hidden nodes

Determining the optimal number of hidden nodes has always been a question that is raised in neural networks applications and there is no direct and precise way of determining the best number of nodes in each hidden layer. Several rules-of-thumb were developed by many researchers regarding the approximate determination of required number of hidden nodes in a hidden layer from the knowledge of the number of nodes in both the input and output layers (Najjar et al., 1997). For single hidden layer networks, there are a number of rules-of-thumb to obtain the best number of hidden layer nodes. One approach is to assume the number of hidden nodes to be 75% of the number of input units (Salchenberger et al. 1992). Another approach suggests that the number of hidden nodes should be between the average and the sum of the nodes in the input and output layers (Berke and Hajela 1991). A third approach is to fix an upper bound and work back from this bound. Hecht-Nielsen (1989) and Caudill (1988) suggested that the upper limit of the number of hidden nodes in a single layer network may be taken as *(2I+1)*, where *I* is the number of inputs. The best approach found by Nawari et al. (1999) was to start with a small number of nodes and to slightly increase the number until no significant improvement in model performance is achieved. Another way of determining the optimal number of hidden nodes that can result in good model generalization and avoid over fitting is to relate the number of hidden nodes to the number of available training samples (Maier and Dandy, 2000). For instance, Wanas et al. (2001) showed, empirically, that the best performance of a neural network occurs when the number of hidden nodes is equal to *log(T)*, where *T* is the number of training samples.

In this study, we used the iterative method, starting from an initial guess (according to equation 4.7), for determining the required number of hidden nodes in a hidden layer and online monitoring of accuracy measures on the testing datasets. This was done by varying the number of initial hidden nodes until the network was able to best learn the patterns involved in the testing datasets. For each set of hidden neurons, the network was trained in batch mode to minimize the average square error (ASE) at the output layer. In order to check any over-fitting during training, a threefold cross-validation was performed by keeping track of the efficiency of the fitted model. The training was stopped when there was no significant improvement in the model's efficiency, and the model was then tested for its generalization properties.

Generalization of neural networks

To reach the best generalization and overcome the problem of data over fitting through cross-validation technique, the dataset should be split into three parts:

- The training set is used to train a neural net and adjust the connection weights. The error of this dataset is minimized during training.
- The testing set measures the ability of the model to generalize, and the performance of the model using this set is checked at many stages of the training process and training is stopped when the error of the testing set starts to increase. The testing set is also used to determine the optimum number of hidden layer nodes and the optimum values of the internal parameters (learning rate, momentum term and initial weights). Finally, a test set was applied to checking the overall performance of a neural net.

- The validation set is used to determine the performance of a neural network on patterns that are not trained during learning.

Najjar et al. (1999) recommended that test and training data sets must be selected randomly. In this work, we adopted a randomly grouping the data into 3 folds (training, testing and validation) by generation random number using the JavaScript *Math.random* function. Accordingly, the entire database (554 datasets) in global model was divided randomly into training, testing and validation sub-databases at the ratio of about 51% : 26% : 23%.

Training the neural network

The process of optimising the connection weights is known as “training” or “learning”. This is equivalent to the parameter estimation phase in conventional statistical models (Maier and Dandy, 2000). The MLP neural networks learn using an algorithm called back-propagation (Rumelhart and McClelland, 1986) which is the predominant method of supervised training. With back-propagation, the input data is repeatedly presented to the neural network. With each presentation the output of the neural network is compared to the desired output and an error is computed. This error is then fed back (back-propagated) to the neural network and used to adjust the weights such that the error decreases with each iteration and the neural

model gets closer and closer to producing the desired output. This process is known as "training".

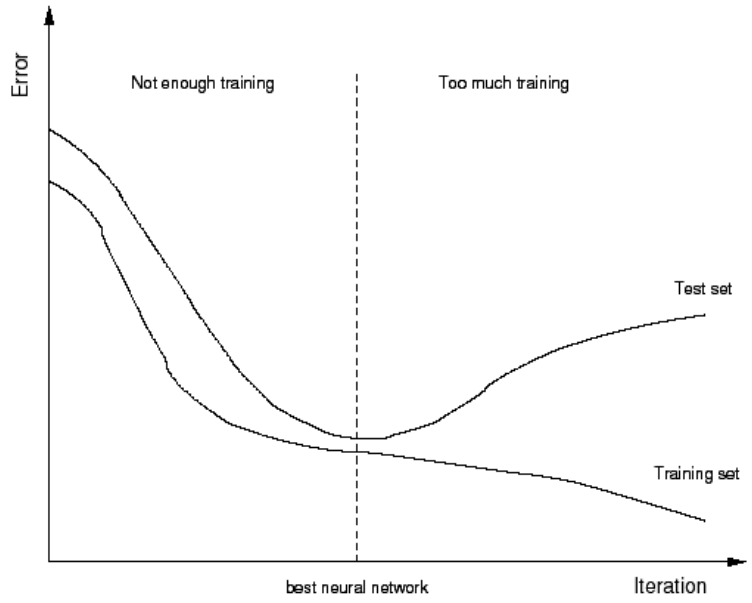

Fig. 4.4 Generalization versus training error (Moody, 1992)

Fig. 4.4 shows a typical error development of a training set (lower curve) and a testing set (upper curve).

The learning should be stopped in the minimum of the testing set error as well as when the error of the testing set starts to increase. At this point the net generalizes best. When learning is not stopped, overtraining occurs and the performance of the net on the whole data decreases, despite the fact that the error on the training data still gets smaller.

Model validation

Once the training phase of the model has been successfully accomplished, the performance of the trained model should be validated (Shahin et al., 2001a). The purpose of validation phase is to ensure that the model has the ability to generalize within the limits set by the training data in a robust fashion, rather than simply having memorized the input-output relationships that are contained in the training data. The approach that is generally adopted in the literature to achieve this is to test the performance of trained ANNs on an independent validation set,

which has not been used as part of the model building process. If such performance is adequate, the model is deemed to be able to generalize and is considered to be robust.

Error evaluation and selecting the optimal ANNs model

The selection process of the desired optimal network model is composed of two consecutive stages, which indicate the values of efficiency. In the first stage, statistical accuracy measures such as coefficient of correlation (R^2), Sum of Squares due to Error (*SSE*) and the Mean Absolute Relative Error (*MARE*) on both training and testing data sets to filter out the most promising optimal networks. The values of both *SSE* and *MARE* close to zero indicate a better performing model. The values of R^2 range from 0 to 1, with higher values close to 1 indicating better model performance. These statistical indices can be expressed mathematically as illustrated in the following equations:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{4.8}$$

$$SSE = \frac{\sum_i (y_i - \hat{y}_i)^2}{N} \tag{4.9}$$

$$MARE = \frac{1}{N} \sum_i \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100\% \tag{4.10}$$

Where y_i , $\hat{y}_i$ and $\bar{y}$ stand for the observed, predicted, and mean value, respectively. N is the total number of data set.

Statistically, an optimal network is defined as the one with the best overall accuracy measures. In the second stage, the predicted and observed data graphical responses for both training and testing sets for the selected most promising networks. Based on the overall graphical evaluation of each model's performance, the absolute optimal network can easily selected.

4.2 ANNs Model for Total Water Mains

First model in this domain was related to an ANNs model for all failures on the total water pipelines in study area that named Global model. Total case in this dataset contains 554 water pipelines which has been divided randomly into three subsets: a training set: 279 cases (50%), testing set: 139 cases (25%), validation set: 136 cases (25%).

To obtain the best model, several neural networks varying with the number of hidden nodes (NH) were presented with the training sets to generalize the relationships between the input and output parameters. Since the number of hidden nodes is unknown, the trial and error process was begun by determining the hidden node by using equation (4.7). We calculated 21 hidden nodes for the first net configuration. By minimizing the $SSEN_{ts}$, the initial number of hidden nodes was selected for *Itr=2000* and *HN=2*. Then, the trial-and-error procedure was started with one hidden node, and then the number of hidden node was increased to 2 during the trials.

Using the training and then the testing dataset, the least error structure in the testing dataset was selected based on the statistical accuracy measures such as SSE, Mean Absolute Relative Error (*MARE*) and Coefficient of determination (R^2). Then, to decide which number of hidden nodes produces a good prediction network, all networks have been tested on sets that have never been used in the training process (validation dataset). The results are summarized in table 4.1, which are used to assess the model's performance.

Based on statistical accuracy measures such as *SSE*, R^2, *MARE* and graphical evaluation of testing datasets, the optimal network structure was found to be 2 hidden nodes and 2000 iterations. The corresponding accuracy measures of this network are $SSEN_{tr}$[1]$= 0{,}001221$, $R^2_{tr}= 0{,}91229$, $MARE_{tr}=60.8\%$ (for training datasets) and $SSEN_{ts}$[2]$= 0{,}001244$, $R^2_{ts}= 0{,}90048$, $MARE_{ts}=60.7\%$ (for testing datasets).

[1] tr= training subset
[2] ts= testing subset

Table 4.1 Statistical accuracy measures in each trial of finding optimal hidden nodes (Total mains model)

Trial	Itr	HN	$MARE_{tr}$	$MARE_{ts}$	R^2_{tr}	R^2_{ts}	$SSEN_{tr}$	$SSEN_{ts}$
1	2000	1	608.11	606.965	0.91204	0.90049	0.001225	0.001244
2	2000	2	608.099	607.008	0.91229	0.90048	0.001221	**0.001244**

In other word, the $SSEN_{ts}$ and $SSEN_{tr}$ parameter were used as a criterion for the selection of the appropriate neural network architecture. The values of *SSE* closer to "*0*" indicate a better fit. In Fig. 4.5 both of these parameters were plotted against the number of hidden nodes.

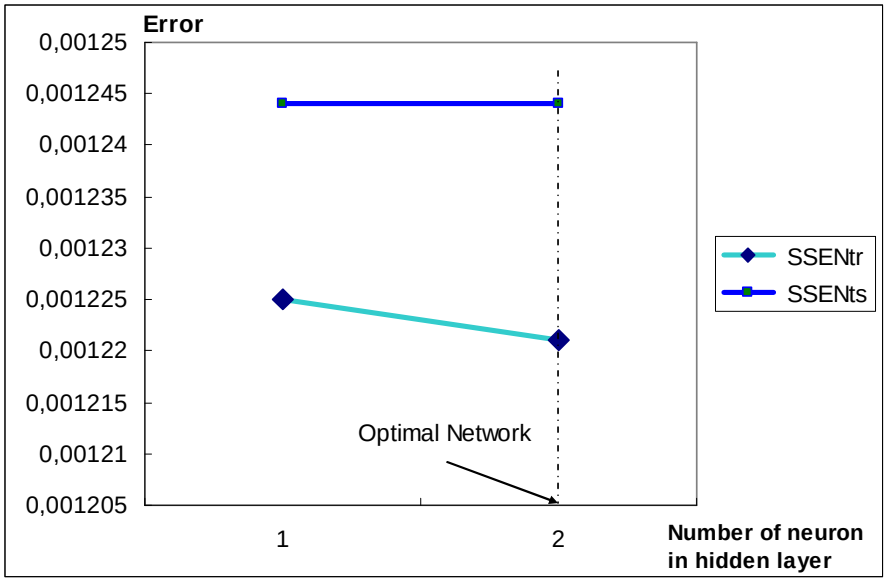

Fig. 4.5 Number of nodes in the hidden layer versus SSE for the testing and training stages

The optimal number of hidden nodes was selected when both of $SSEN_{ts}$ and $SSEN_{tr}$ errors have minimum values. Therefore, the network with two hidden nodes was selected as the most appropriate network for predicting of failures frequencies in each water mains.

Additionally, Fig. 4.6 depicts the comparison between observed values and predicted by the two different network architectures for testing and training date sets.

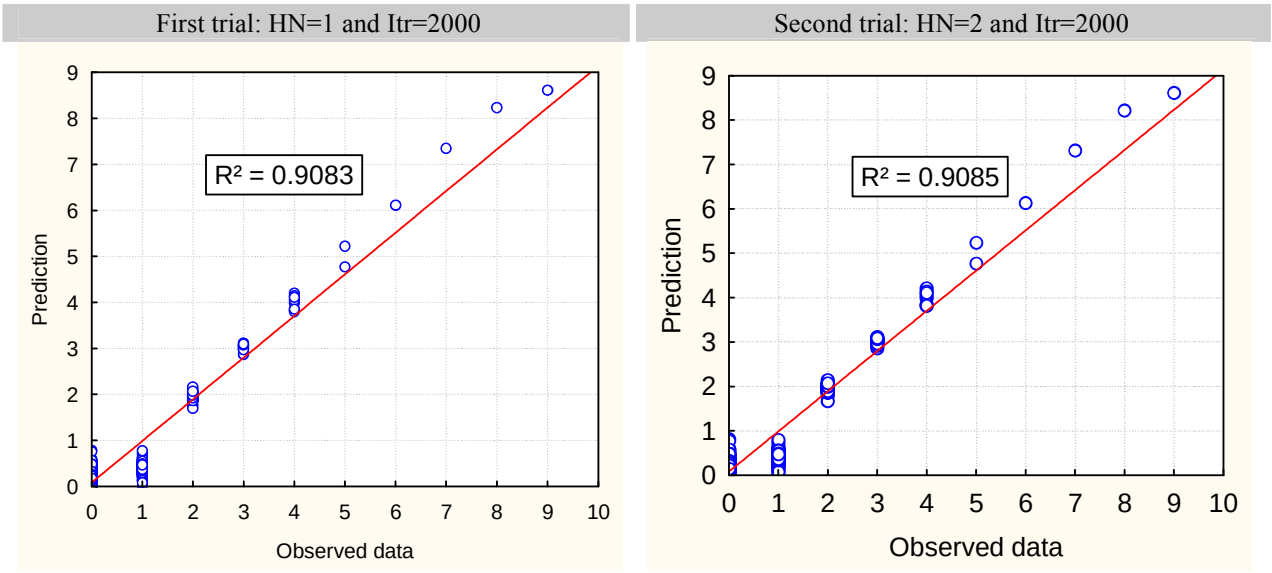

Fig. 4.6 Predicted versus observed values of failure frequency in the testing and training subsets

The solid line shown on the plot is a linear trend line fitted to the predicted values. Model performance was evaluated in terms of coefficient of determination (R^2) which reflect the overall error performances of the model. The result of the ANNs with two hidden nodes demonstrated a bit higher coefficients of determination ($R^2 = 0.91$).

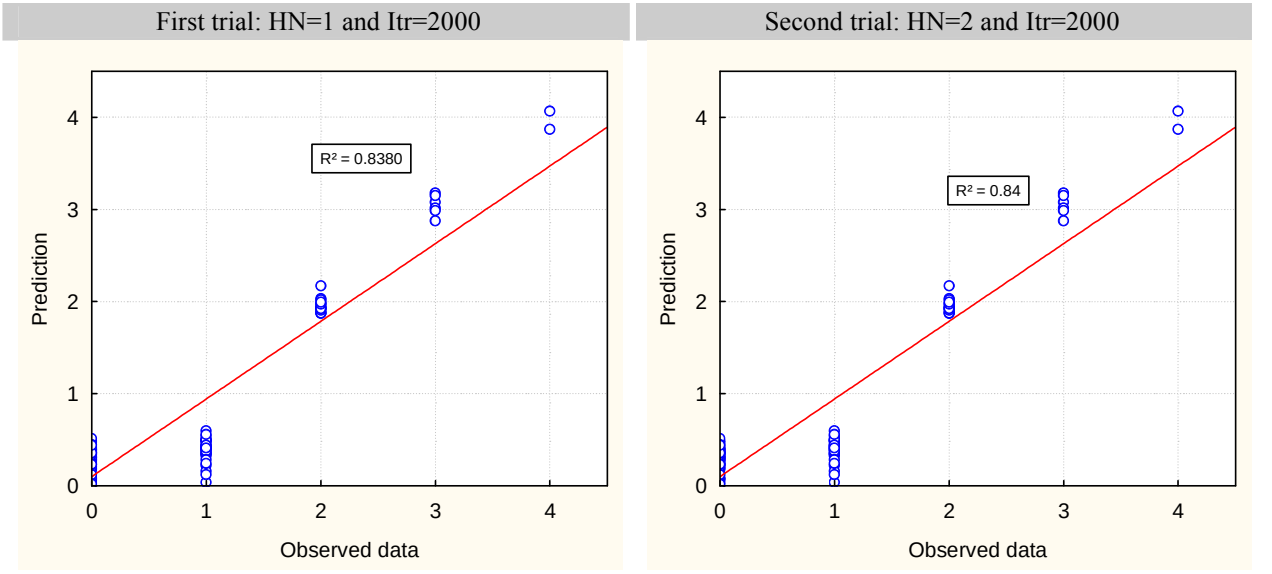

Fig. 4.7 Predicted against observed failure frequency during optimization process on validation cases

To validate the proposed ANNs models, randomly selected validation datasets were used to evaluate the prediction accuracy. The high value of coefficient of determination *(R^2=0.84)*

during the validation phase confirms that a network with 2 nodes in hidden layer produces the most accurate prediction results. Fig. 4.7 shows the scatter plot for predicting failure number through two proposed models. It also depicts that the predicted values from the neural networks with two hidden nodes matched the observed values better than those obtained from another network's architecture. Then, the network with 2 hidden nodes was selected for prediction of failure in the study area.

In summary and based on Fig. 4.6 and Fig. 4.7, the Global ANNs model for all of failure on total water pipelines found that the best network consisted of 2 hidden nodes with correlation coefficients equal to *0.91* and *0.84* for the training and testing, and validation sets, respectively.

4.2.1 Prediction with the global ANNs model

Since the data collection program was installed in SWWU in 2000, we have had access to more reliable data. Failure information for the period 2000 through 2004 was then extracted for comparing the observed and predicted value. Fig. 4.8 shows the correlation between the predicted and observed values of failure frequencies for the period 2000-2004. R^2=0.78 and the narrow bound of 95% confidence in this graph indicates the ANNs model predicts well. Otherwise, a very wide interval for the fitted coefficients could indicate that we should use more data when fitting before we can say anything very definite about the coefficients.

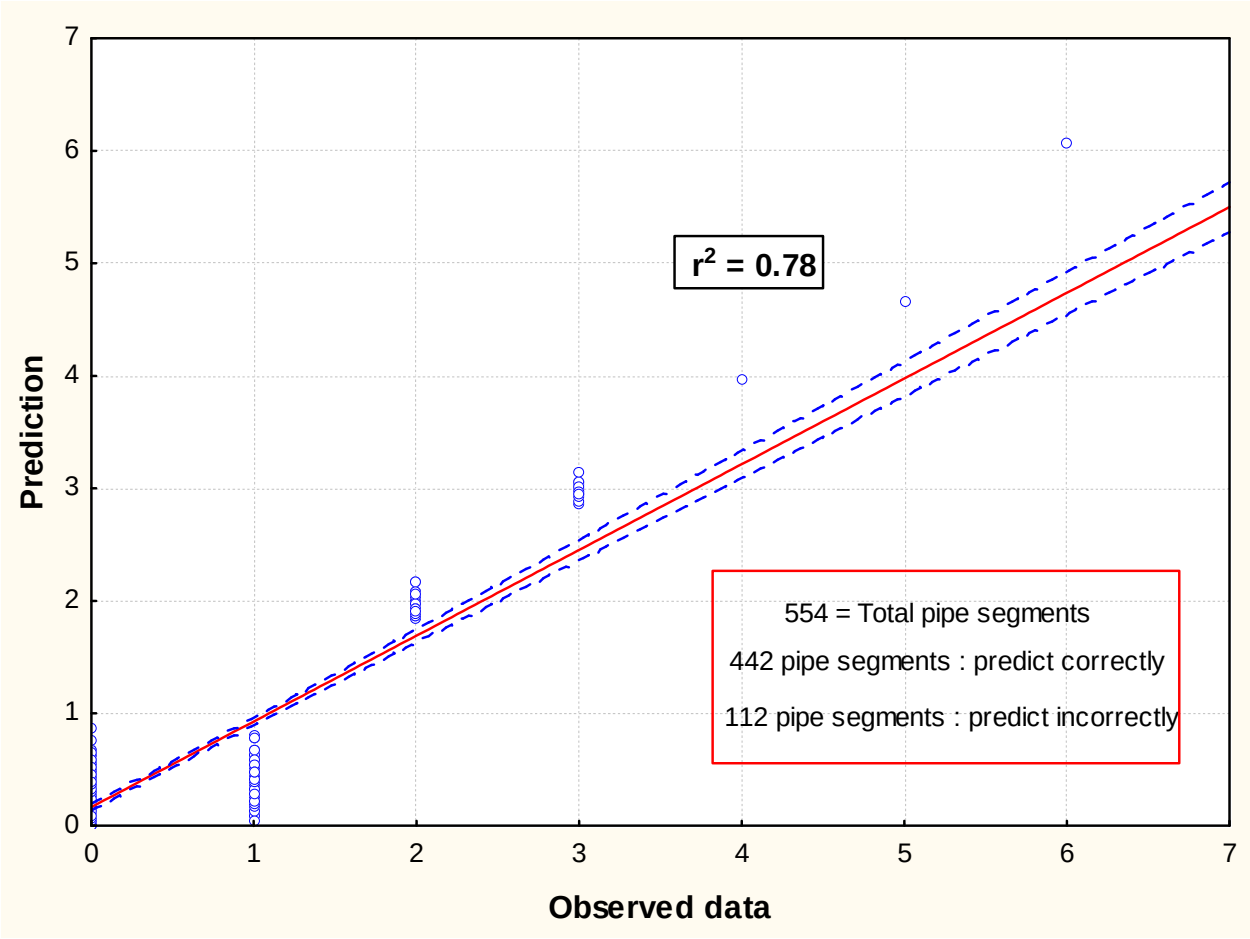

Fig. 4.8 Comparison of predicted and observed values of NF (during 2000-2004)

4.2.2 Sensitivity analysis

Neural Networks offer an interesting ability to test the sensitivity and significance of the various input variables to produce the outputs. This can be carried out by manipulating the connection weights and biases of the designed network (Hajmeer et al., 1997). This gives some information about the relative importance of the variables used in a neural network. It often identifies variables that have low significant effect on the accuracy of the network. Key variables with high sensitivity can improve model's performance significantly.

While the selection of input variables is a critical part of neural network design, we conducted Sensitivity Analysis, which rates the importance of input variables with respect to the global model. Table 4.2 compares the ranking of the nine most influential variables among input

variables for Global model. It also shows the rank order for each input, which puts the input variables into order of importance.

Table 4.2 The influence ranking of input variable on the output in Global model

Input variable	Rank
NPF	1
Material	2
LogLength	3
Age	4
Traffic category	5
Thickness	6
Diameter	7
Maximum pressure	8
Depth	9

Considering to table 4.2, number of pervious failure with rank of "1" identifies the most influential variable affecting model output and depth with rank of "9" indicates that it has no positive effect on the model.

In this thesis, we also arranged all water pipelines into small groups based on three existence pipe material. For each group we developed the ANNs model as below.

4.3 ANNs Model for Metallic Water Mains

In this part we adjusted the model for two cast and ductile iron material which represent the metallic water pipelines in study area. The failures in cast iron and ductile iron were considered in this model. By trial and error, the optimal network was determined for two hidden node. Table 4.3 illustrates the statistic indices for the best network.

Table 4.3 Error evaluation for finding the hidden neurons

Trial	Itr	HN	$MARE_{tr}$	$MARE_{ts}$	R^2_{tr}	R^2_{ts}	$SSEN_{tr}$	$SSEN_{ts}$
1	5300	3	581.648	557.94	0.94976	0.95207	0.000774	0.001146
2	5300	3	581.798	557.341	0.94743	0.95474	0.000811	**0.001113**
3	5300	3	581.565	557.843	0.94843	0.95208	0.000795	0.00118

In Fig. 4.9 both of error evaluation parameters, $SSEN_{ts}$ and $SSEN_{tr,}$ were plotted versus the number of hidden nodes. The optimal number of hidden nodes was selected when the error of the testing and training set starts to increase. Considering to value of *SSE* for testing and training, 0.001113 and 0.000811 respectively, 2 hidden nodes were identified as the best number. Therefore, the network with "8-2-1" architecture best predicts the failures frequencies in both cast and ductile iron water mains.

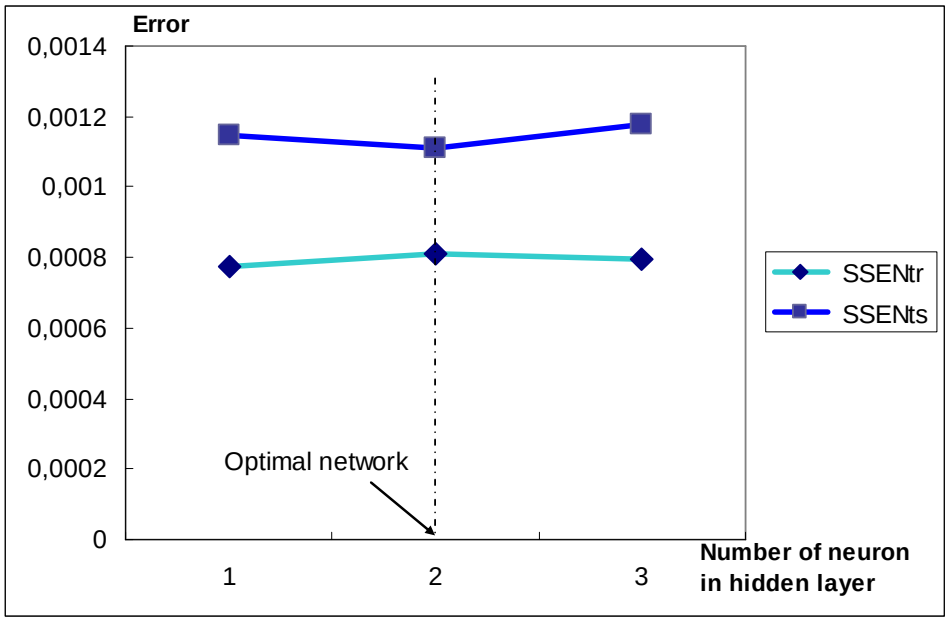

Fig. 4.9 Number of nodes in hidden layer versus SSE for the testing and training stages

The graphs in Fig. 4.10 show predicted values for the number of failures in each mains against observed values for the training and testing subset besides validation cases. The poor performance of this model during testing and training may be due to over fitting of the model as the number of cases in the training data is limited.

By using trained neural network for metallic pipelines, we made predictions on data over 2000-2004. This model with R^2= 0.86 appears to fit the data well (Fig 4.11).

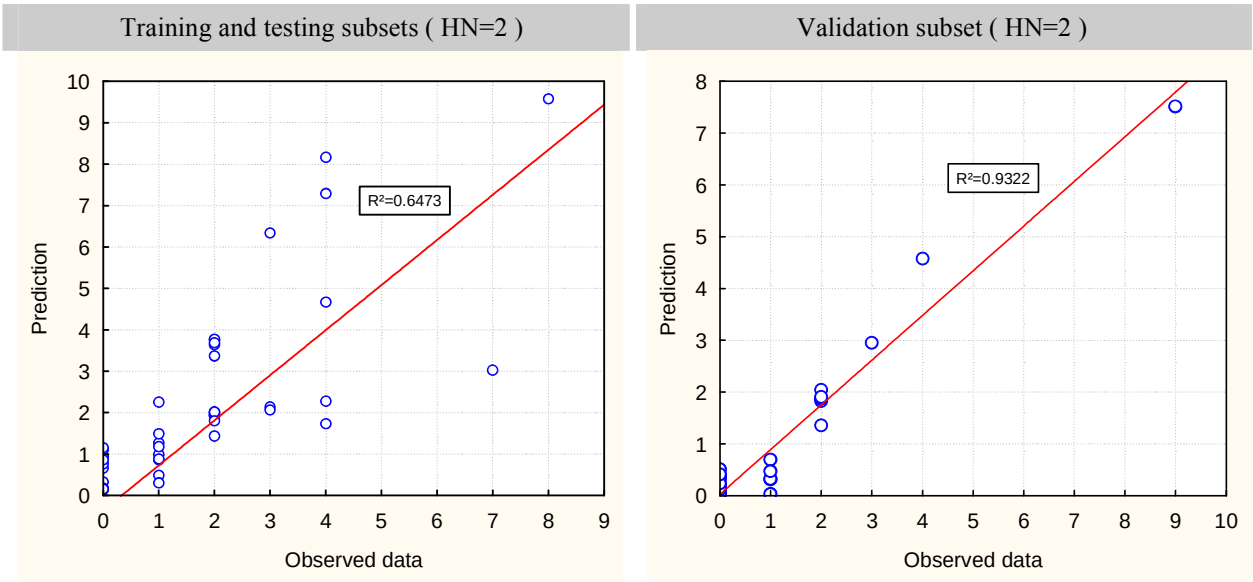

Fig. 4.10 Predicted versus observed values of failure frequency in the testing and training subsets

4.3.1 Making prediction for 2000-2004

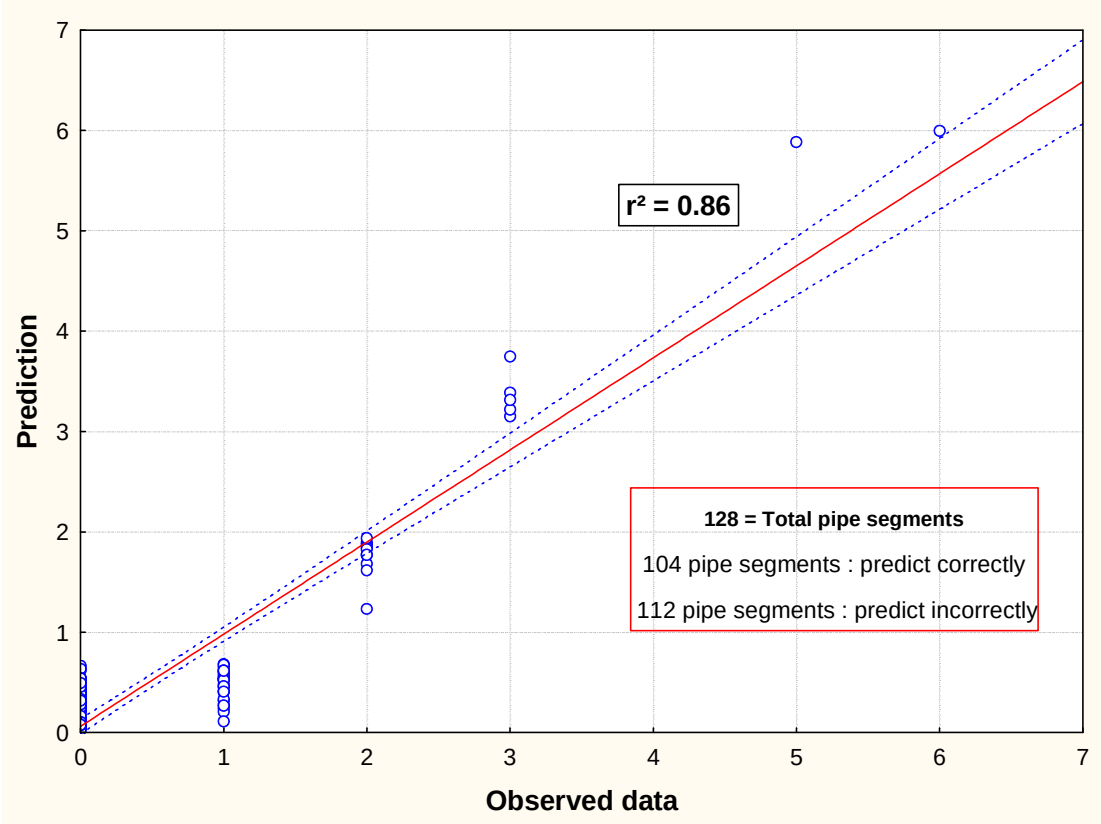

Fig. 4.11 Comparison of predicted and observed values of NF (during 2000-2004)

4.4 ANNs Model for Cement Water Mains

Asbestos cement (AC) pipes account for approximately 36.1% (20.5 kms) of the total length of mains in the selected area with a corresponding average break rate of 48.3 breaks/100 km/year. Here ANNs are used to model the number of failure in asbestos cement material. The history of failure in 173 segments was randomly separated into three groups of 87 (about 50% of the total sample), 43 (about 25% of the total sample) and 43 (about 25% of total sample) as training, testing and validation samples, respectively.

A similar approach was repeated for this network to determine the number of hidden layer neurons. We carried out trial and error procedure with HN= 9 as an initial guess which was determined by equation 6. Consequently, a network configuration of "8-1-1" was selected for the prediction in AC water pipelines failure. Table 4.4 reports the error evaluation for this optimization process.

Table 4.4 Statistical accuracy measures in each trial of finding optimal hidden nodes (AC mains model)

Trial	Itr	HN	$MARE_{tr}$	$MARE_{ts}$	R^2_{tr}	R^2_{ts}	$SSEN_{tr}$	$SSEN_{ts}$
1	3500	2	609,86	620.78	0.83652	0.77906	0.00695	**0.008876**
2	3500	2	611.68	622.042	0.80887	0.75706	0.00816	0.009839

Fig. 4.12 shows the comparison between actual values and predicted values in two different network structures for the testing and training data sets. The solid line shown on the plot is a linear trend line fitted to the predicted values. Model performance was also evaluated in terms of coefficient of determination (R^2). The result of the ANNs with one hidden node demonstrated high coefficients of determination ($R^2 = 0.82$).

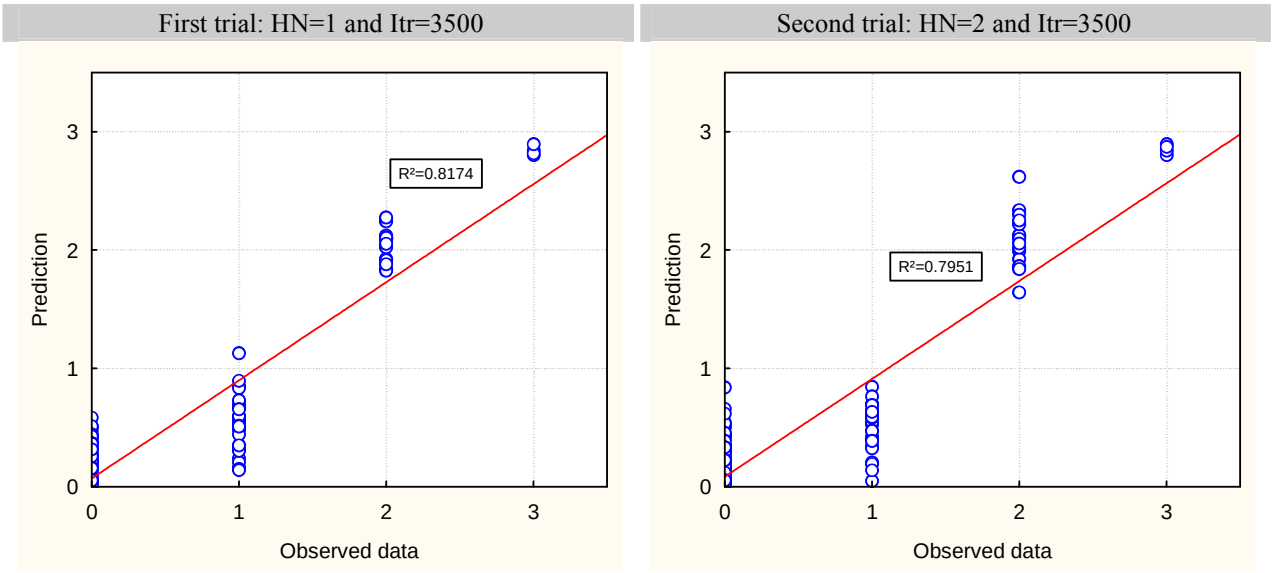

Fig. 4.12 The correlation between predicted and observed values during testing and training

For the validation cases, the coefficient of determination value was established equal to 0.68 and *0.66*. These coefficient also confirm that network with one hidden node fits better.

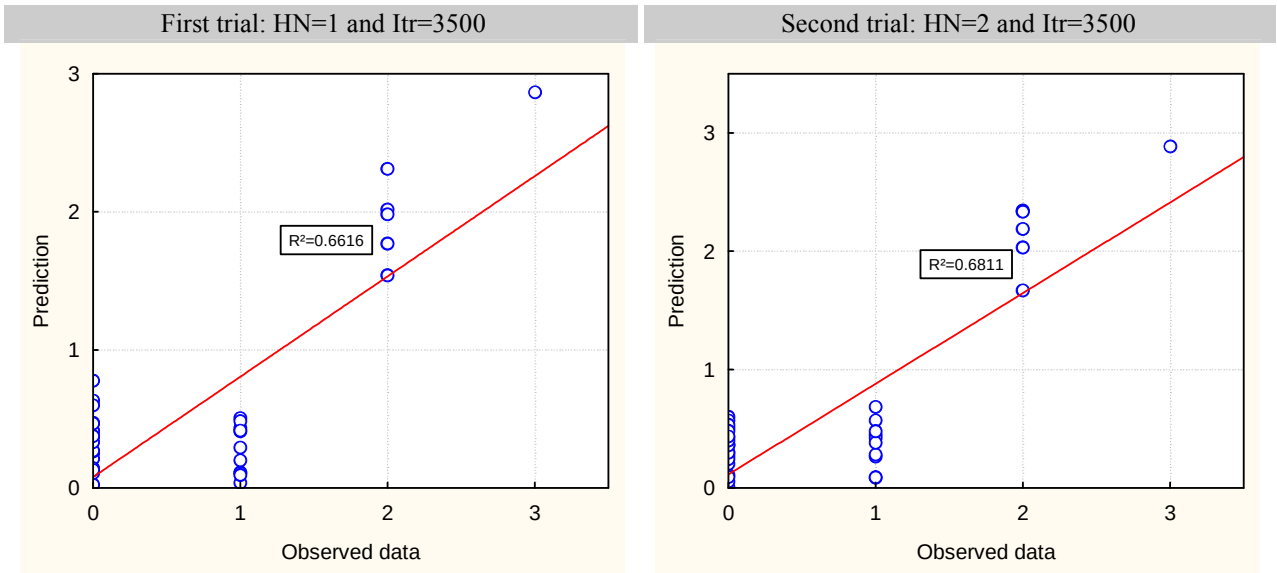

Fig. 4.13 Scatter plot of predicted values versus actual values of failures number on each AC water pipelines during optimization process on validation cases

To increase predictive performance, the training, testing, and validation sets were randomly drawn from the whole data set five times. Then, all networks have been trained and tested. The best results were obtained for 90 cases in training (52%), 45 cases for testing (26%) and 38 cases (22%) for validation. After training and testing, according to SSE error for the testing phase, the finalized network consists of one nodes in the hidden layer [NN(8:1:1)].

4.4.1 Making prediction on data over 2000-2004

The best model [NN(8:1:1)] was considered to predicting the number of failure (NF) for Asbestos Cement water pipelines during 2000-2004. Fig. 4.14 compares the prediction value by the selected ANNs model and observed value in the period 2000-2004. This graph shows R^2 along with the 95% confidence bounds. Clearly, $R^2=0.59$ shows that the ANNs models are able to make reasonably good forecasts in this type of pipelines.

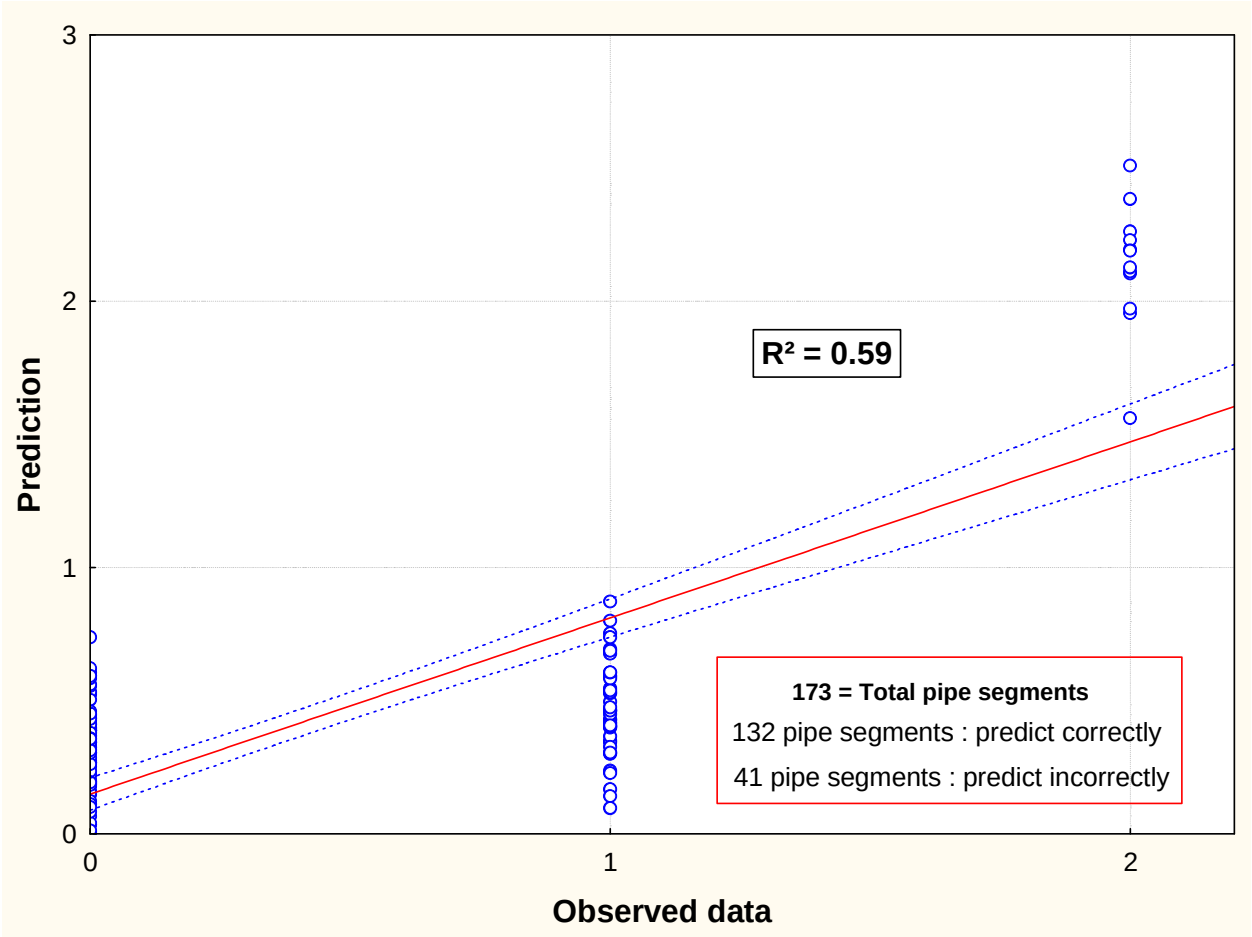

Fig. 4.14 Comparison of predicted and observed values of NF (during 2000-2004)

4.5 ANNs Model for Plastic Water Mains

Since the newer material like polyethylene have not been in the ground long enough to have collected significant amounts of historical data to support accurate ANNs predictions, but approximately 43% of mains failures occurred on polyethylene lines which comprise about 30% of the study area's total pipeline system. For this type of material, the network was developed and evaluated as per the methodology described earlier for models.

In the first stage, the entire database (253 samples for failure in polyethylene) was divided into training, testing and validation sub-databases at the ratio of approximately 50% : 25% : 25%. The number of hidden layer neurons was evaluated in the range of 1-11. Statistical accuracy measures, SSE, MARE and R^2 for testing and training dataset were illustrated in Table 4.5. Fig. 4.15 provides the plot of errors versus number of hidden nodes.

Table 4.5 Statistical accuracy measures in each trial of finding optimal hidden nodes (PE mains model)

Trial	Itr	HN	$MARE_{tr}$	$MARE_{ts}$	R^2_{tr}	R^2_{ts}	$SSEN_{tr}$	$SSEN_{ts}$
1	8200	11	634.328	622.751	0.88521	0.87689	0.002	0.002465
2	9100	7	633.793	621.698	0.88859	0.89738	0.001939	0.002005
3	9100	7	633.885	621.742	0.88807	0.89708	0.001948	0.002013
4	6100	7	633.822	621.588	0.88893	0.89872	0.001932	**0.001988**
5	9200	5	634.646	622.786	0.88495	0.88213	0.001997	0.002336
6	7100	8	633.968	621.749	0.88857	0.89825	0.001936	0.001989
7	9100	7	634.431	622.166	0.88612	0.89417	0.001976	0.002067
8	7400	13	633.898	621.868	0.89125	0.89628	0.001884	**0.002024**
9	9100	9	634.46	622.225	0.88599	0.89316	0.001978	0.002095
10	10000	10	634.526	622.27	0.88562	0.89419	0.001984	0.002064
11	8100	11	634.482	622.286	0.88487	0.89303	0.002	0.002098

According to the statistical accuracy measure for $SSEN_{ts}$, as noticed in Table 4.5 and Fig. 4.15, this model has two possibility for model topology which means that desirable network can predict by 4 or 8 hidden nodes. For different number of hidden nodes (4 and 8 nodes) the network was trained against different training, testing and validation sets. We compared the predicted and observed value for testing and training data with HN= 4 and 8. Additionally, correlation of observed and the predicted number of failure for the validation dataset were examined. According to the statistical accuracy measure for SSE_{ts} and Fig. 4.16 and 4.17, the

best results were obtained when 8 neurons were employed in hidden layer. Accordingly, this model [NN(8:8:1)] can most accurately predict the output variable.

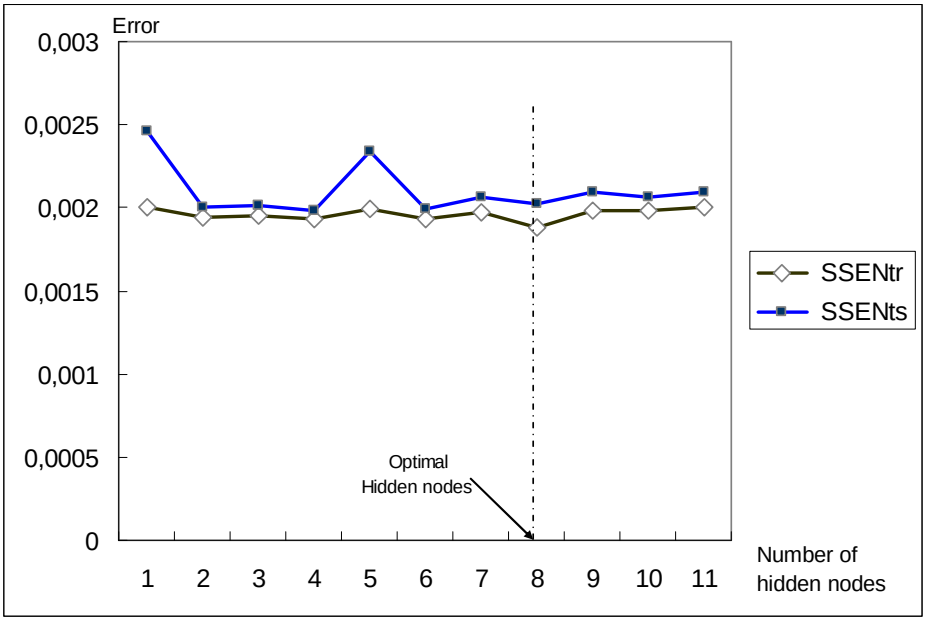

Fig. 4.15 Number of nodes in hidden layer versus SSE for testing and training stage

The graphs in Fig. 4.16 represent the comparison between actual values and predicted by two different network structures for testing and training date sets.

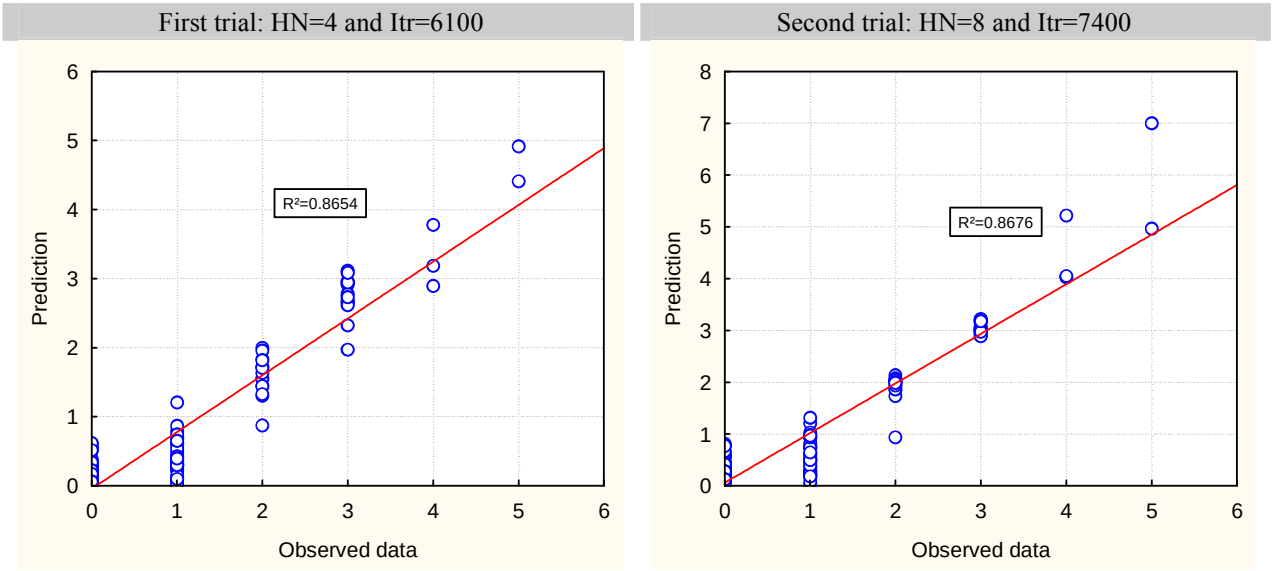

Fig. 4.16 Predicted versus observed values of failure frequency in the testing and training subsets

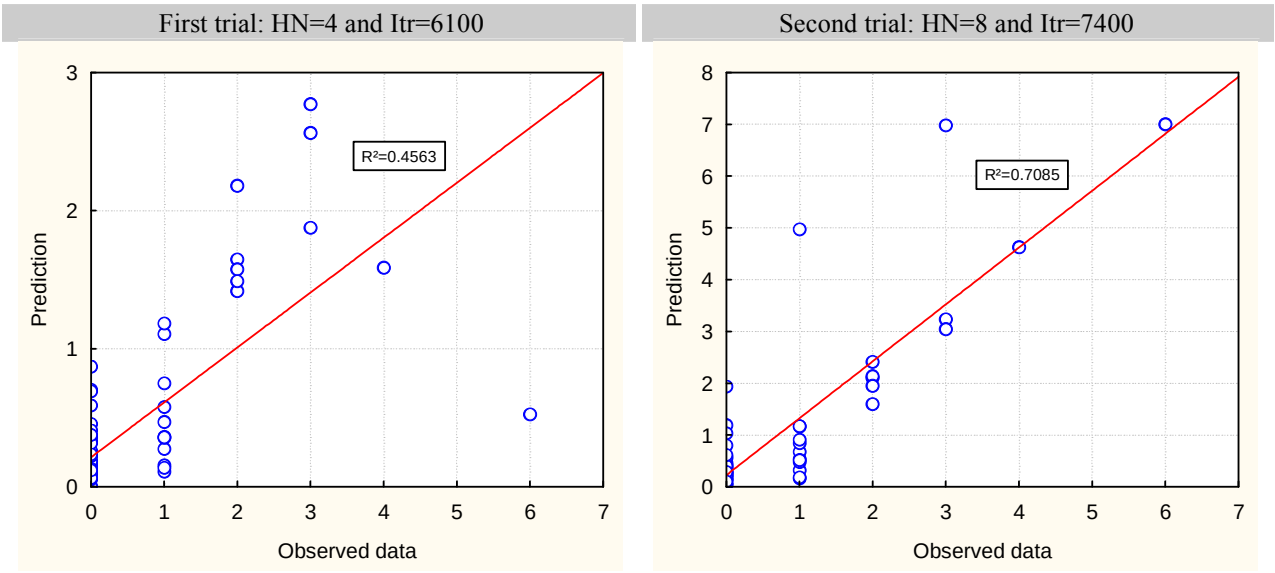

Fig. 4.17 Predicted against observed failure frequency during optimization process on validation cases

It was founded that the value of $R^2 = 0.45$ and 0.71 for network with 4 and 8 hidden nodes, respectively. It confirms that ANNs with 8 hidden nodes have a good agreement with data.

4.5.1 Making prediction on data over 2000-2004

Comparisons between the observed and predicted 5th-year data depicts that the predicted values were in close agreement with the observed values as shown by the relevant statistical parameter (correlation coefficient = 0.75). Fig. 4.18 plots the fitted line based on observed dat against predicted value.

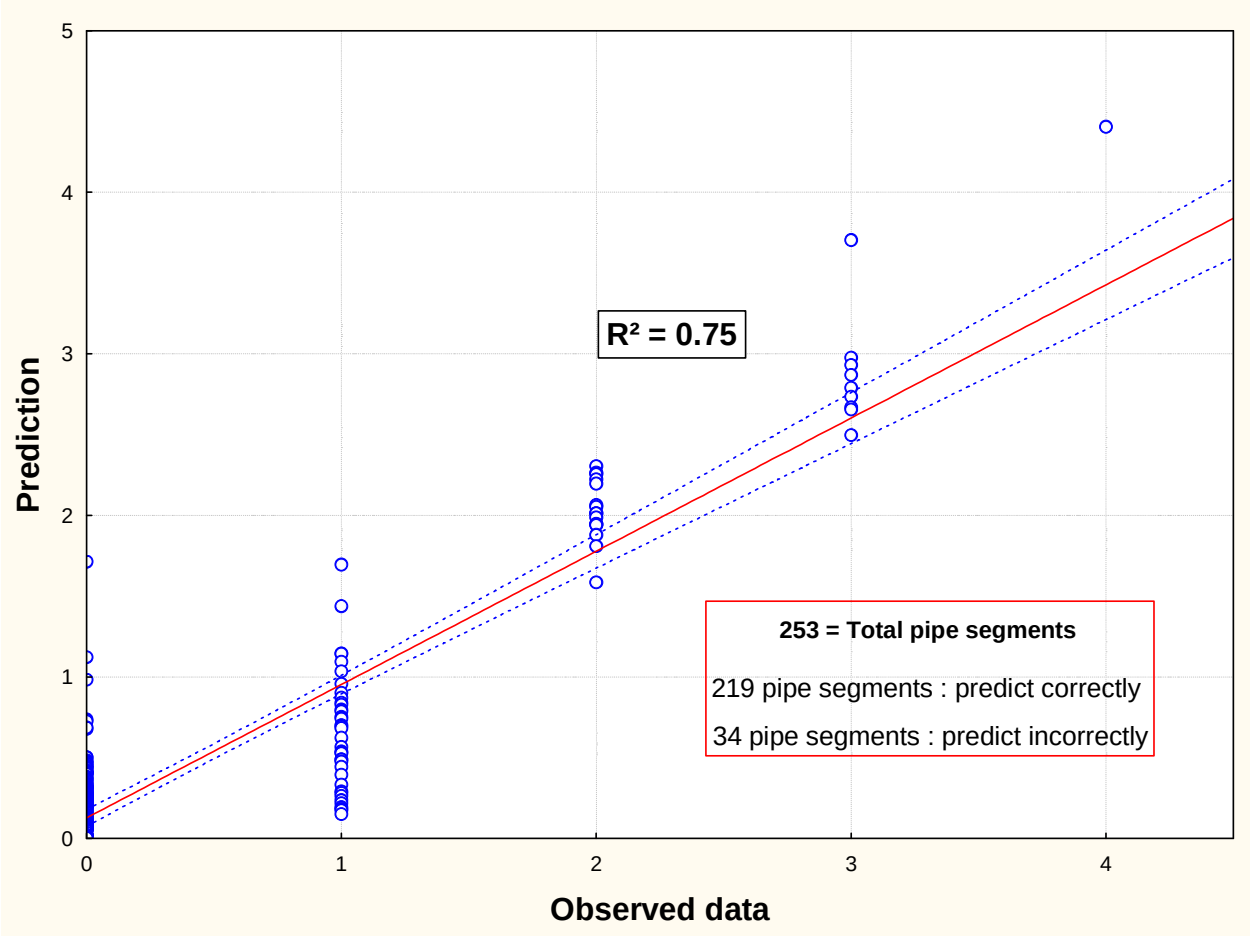

Fig. 4.18 Comparison of predicted and observed values of NF (during 2000-2004)

4.6 Regression Models versus Artificial Neural Network

In chapter 3, Multiple and Poisson regression model has been employed to analyze total water pipelines failure data within 1995-2004. This part compared the prediction accuracy of ANNs with the Regression models on the global model. Though various measures can be used for comparison purposes, for simplicity, only coefficient of correlation (R^2) was considered here.

Table 4.6 Comparison of alternative models

Method	Coefficient of correlation (R^2)
ANNs	**0.78**
Poisson regression	0.74
Multiple regression	0.63

It is evident from table 4.6 that the ANNs performs better than the traditional regression models for prediction of water pipelines failure frequency.

4.7 Concluding Remarks

The purpose of this chapter is to develop failure frequency prediction model by using ANNs approach. The models forecast individual pipes failure number which is one major technical indicator for defining annual rehabilitation program and prioritize pipes to be rehabilitated. Based on the available failure data in the study area, back propagation neural network with one hidden layer was identified for developing prediction model. The ANNs models were developed in 3 consecutive stages. In the first stage, the ANNs architecture was determined based on problem characteristics and ANNs knowledge, and the input and output categories were determined through statistical analysis. In the second stage, the neural network was trained and tested on actual data to find the optimal number of hidden nodes and iterations for the ANNs architecture determined from stage one. In the third stage, the best performing ANNs from the first two stages was re-trained on all observed data to enhance the prediction accuracy and to arrive at optimal model.

In this case study, we trained four ANNs models based on total failure data and 3 separated failure data according to pipe materials. Comparison between forecasted and observed failures for last 5 years show that the designed neural network for total water pipelines , with $R^2 = 0.78$, give satisfactory prediction. Further, three models for metallic, cement and plastic pipelines had correlation coefficient equal to 0.86, 0.59 and 0.75, respectively. In conclusion, stratification of material did not improve the results except in metallic pipelines.

Finally, comparison of the Global ANNs model and the corresponding multiple linear regression as well as poisson regression for failure showed that the ANNs models outperformed its counterpart in prediction accuracy. Prediction accuracy, flexibility for use, and capabilities for investigation associated with the developed ANNs models support the conclusion that ANNs provides an attractive and powerful tool for prediction of failure in water distribution.

5. Survival Analysis of Water Pipelines Failure Time

5.1 Introduction

This chapter comprises an application of survival analysis in Sanandaj's water pipelines failure that can be used to assist water managers in identifying efficient pipe maintenance strategies. Using parametric and non-parametric survival models, the expected number of failures during a given time period is computed. Survival analysis characterizes the distribution of the survival time for different groups of pipes, to compare this survival time among different type of materials. Then, four models were developed to simulate time to failure in all water mains, and 3 stratified failure dataset: metallic, cement and plastic water mains. The various models were calibrated on the historical failure data collected over the period 1995 - 2001, then they were tested on the more reliable data since 2002. These models determine the Benefit Index curves, i.e. "impact of the various rate of renewal over the mains network on the percentage of failures which avoided from this network". In this chapter, we describe the methodology followed for the survival analysis in our research. Then models and results will be illustrated. The analysis discussed in the following sections were performed using three statistical software: Statistica, SAS and EGRET package.

5.2 Principal of Survival Analysis

There are several questions an investigator might wish to ask in relation to survival data. First, it may be of interest to estimate the survival time distribution for a group of individuals. Among other things, this allows one to calculate the risk that the event will occur within a given interval and compute derived quantities such as the median residual lifetime, i.e., the time to occurrence of the event for an individual that has survived (i.e., not experienced the event) until the beginning of the interval. A second objective might be to compare survival time distributions among two or more groups, e.g., individuals subjected to different treatments following the diagnosis of a disease. Finally, one might wish to quantify the effects

of one or more independent variables (covariates) on survival times in an effort to develop a model to describe or predict survival times in a population.

Here, survival analysis focuses on the lifetime of a pipe which is predominantly used for rehabilitation planning. The pipe lifetime is treated as a random variable and a standard statistical distribution is then fitted to a collection of similar pipes. In effect, the purpose of survival analysis is to model the underlying distribution of the failure time variable and to assess the dependence of the failure time variable on the independent variables (Rostum, 2000). Such analysis has been performed for several European water networks and North American networks (Eisenbies, 1994 ; Le Gat, 2000 ; Lei and Saegrov, 1998). In 1972, D. R. Cox introduced the Proportional Hazards Model in order to estimate the effects of different covariates on the time to failure of a system. Kaara (1984) and Andreou (1986) introduced the use of proportional hazards model for analyzing failures in water distribution networks.

Fig. 5.1 illustrates the failure history in water pipelines with the failure times t_1, t_2, …, t_i. Each pipe has a vector of covariates or explanatory variables Z ($Z=[z_1,z_2,z_3,..z_p]$) which incorporate in time of failure. We are interested in modeling the relationship between the failure history and the covariates Z such as regression models . Two general classes of regression models are considered in order to relate the hazard function or intensity function to the covariates.

As can be seen in Fig. 5.1, failure data are available from a particular starting time (e.g. 1995) during the follow-up period, 10 years. The problem in this analysis is that we do not have the complete mains failure history in study area. Before 1995, the failure data have not recorded. In survival analysis we called this left-censored failure data. Additionally, in the future failure data will be recorded but these data not included in the analysis. This means that the data is also right censored

5.2.1 Censored observations

In general, censored observations arise whenever the dependent variable of interest represents the time to a terminal event, and the duration of the study is limited in time. In pipeline failure analysis, censored observations occur because by the end of the study period, some pipelines

will not be failed and we don't know how long it will function properly thereafter, and thus, that observation is censored. For water pipelines network, we have two types of censoring:

- Left censoring
- Right censoring

By left censoring we mean that there is a period of time after installation when no data is recorded. When a case is right censored, the dependent variable is known to be greater than a specific value, but its true value is not known (i.e. pipe has not failed by the time the maintenance record ends). More accuracy is achieved by including cases in which the event has not happened yet (right censored data). If the event has occurred the censoring value, CV is set equal to 1, else CV=0 (right censored).

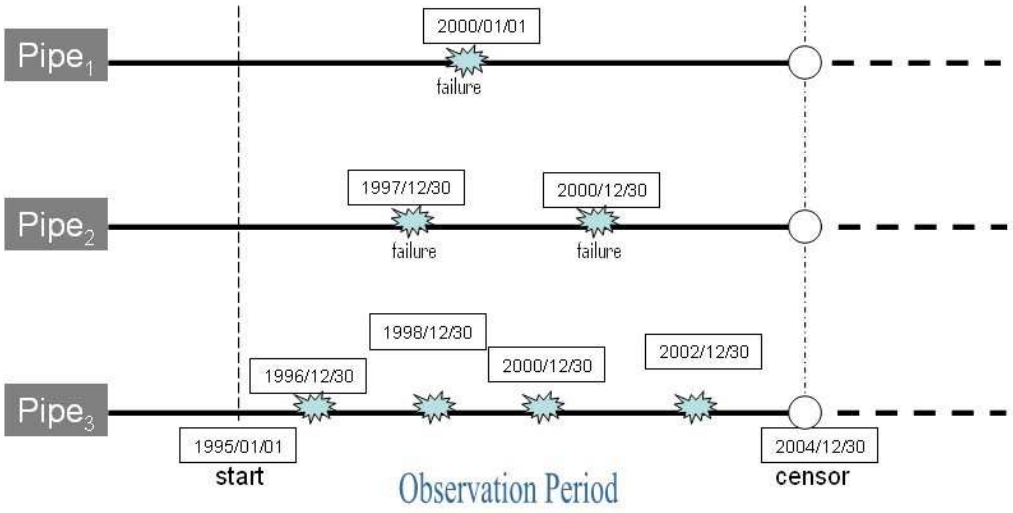

Fig. 5.1 Availability of failure data in water pipelines and times of failure

The initial goal in survival analysis is to characterize the distribution of the survival time for a given population, to compare this survival time among different groups. There are several parametric and nonparametric tests to compare two survival distributions. At first step in this work, failure data was analyzed by means of a life-table, or Kaplan-Meier curve, which is the most common method to describe survival characteristics.

5.3 Non-Parametric Survival Model

The non-parametric approach, Kaplan-Meier (1958), used to estimate the survivor function and cumulative hazard function based on water pipelines failure data that were multi censored. For our model building, we first needed to get the data into the proper format for survival analysis by using the STATISTICA 7.0 commands. Here, the total number of times between failures was 949 which conclude 395 uncensored records (41.8 %) and 554 censored (58.2%). These number are equal to total number of failure and number of water pipelines in study area, respectively. To perform the survival analysis, dataset must be probably prepared (table 5.1). In survival modeling, an individual can fail only once: a new statistical individual is therefore created after each break. With 395 breaks between 1995 and 2004 and 554 statistically defined segments, for all water pipelines 949 statistical individuals has been created. Survival analysis requires the creation of two variables, time of between failures, and the censored variable, "CV" to record the right censorship status. This is a binary variable indicating whether the observation is censored. After a break, the newly introduced statistical individual has a new date of installation that is the date of the latest break now considered to be the date of installation. For instance, records of 4 water pipelines in survival data set are shown as below.

Table 5.1 Preparing dataset for survival analysis

Pipe ID	Number of failure	Date of failure	Left time	Right time	NPF*	CV**
P03060	0	----	1995/01/01	2004/12/30	0	0
P03075	1	1997/09/12	1995/01/01	1997/09/12	0	1
P03075			1997/09/12	2004/12/30	1	0
P03110	2	1997/08/02 2000/12/19	1995/01/01	1997/08/02	0	1
P03110			1997/08/02	2000/12/19	1	1
P03110			2000/12/19	2004/12/30	2	0
P03145	3	1997/12/04 2000/11/09 2003/07/11	1995/01/01	1997/12/04	0	1
P03145			1997/12/04	2000/11/09	1	1
P03145			2000/11/09	2003/07/11	2	1
P03145			2003/07/11	2004/12/30	3	0

** NPF= Number of pervious failure*

*** CV= Censored variable; CV=1 if the statistical individual experiences a failure*

We assumed that before the first failure all the pipes are safe, and so the proportion surviving is =1. Hence, if we denoted the start time of the study as t_0 then we have $S(t_0) = 1$.

To start survival analysis, the Kaplan Meier (KM) survival curves were plotted based on rate of failure for the different material (Fig. 5.2). Any jumping point is a failure time point.

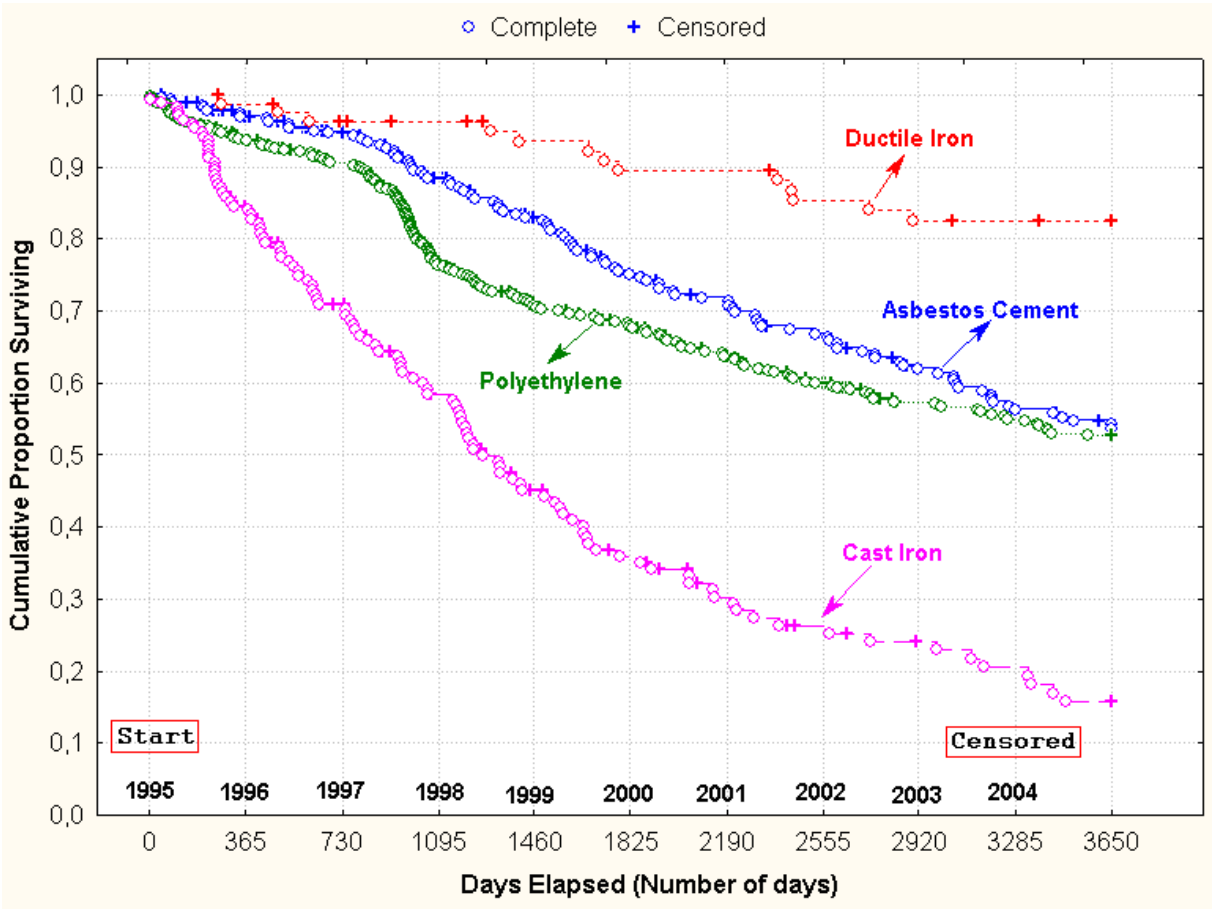

Fig. 5.2 Survival curves of failure rats in four material groups

The red dashed lines, showing the influence of the time factor on the respective breakage rates, are nearly horizontal, which indicates little deterioration rate in ductile iron pipelines. It can clearly be seen that the ductile water pipelines in study area have been deteriorating in a relatively steady rate. In contrast, the cast iron (CI) water pipelines have been deteriorating quite fast. The survival function drops off sharper than other curves within the time span covered by the study. After ductile iron, Asbestos cement water pipelines had fast survival declines with time which is plotted in blue dotted-lines in Fig 5.2 Polyethylene water

pipelines had especial behavior. Between 1997-1998, this group has a rapid drop in failure rate. This indicates polyethylene water pipelines in the first years of life have more failure. Most of them were caused by two main reasons: weak material and poor installation practices.

Overall, as can be seen in this figure, the KM curve for ductile iron pipeline group lies above that for other material group and there is a big gap between these curves. Also they have slowly drop-off as compared to the other materials. Therefore, we could conclude that, somehow, ductile iron pipeline have a greater chance of survival.

With Chi^2=110.6 and df=3, p-value was calculated less than 0.0001. By conventional criteria, this difference is considered to be statistically significant.

Whereas the Kaplan-Meier method is useful for comparing survival curves in four material groups, parametric survival model allows analyzing the effect of several risk factors on survival. In the next step, we derived two parametric survival models (Cox proportional-hazards regression and Weibull) from the failure data in selected area.

5.4 Parametric Survival Model

Since the data were used in a study to predict failure times, this case study is a form of reliability analysis. Data in reliability analysis do not typically follow a normal distribution (Gregory et al., 2003); non-parametric methods (techniques that do not rely on a specific distribution) are frequently recommended for developing confidence intervals for failure data. One problem with this approach is that sample sizes are often small due to the expense involved in collecting the data, and non-parametric methods do not work well for small sample sizes. For this reason, a parametric method based on a specific distributional model of the data is preferred if the data can be shown to follow a specific distribution. Parametric models typically have greater efficiency at the cost of more specific assumptions about the data, but, it is important to verify that the distributional assumption is indeed valid. If the distributional assumption is not justified, then the conclusions drawn from the model may not be valid.

There is a large number of distributions that would be distributional model candidates for the data. However, we restricted ourselves to consideration of the following distributional models because these have proven to be useful in reliability (Gregory et al., 2003):

- Exponential distribution
- Weibull distribution

Here, the distribution of random variable of duration were expressed in 3 closely related way:

Distribution of lifetime	Survival function	Hazard function
f(t)	$S(t)=\int_t^{\infty} f(u)du = 1-F(t)$	$h(t)=\frac{f(t)}{S(t)}$

5.4.1 Fitting a theoretical survival distribution

As a first step in survival analysis, we determined a good distributional model for survival time of the water mains. Two mathematical functions, the survivor function S(t) and the hazard function h(t) establish the survival analysis. In this part, we specified the lifetime distribution through either the exponential or weibull survival time distributions:

Function	Exponential distribution	Weibull distribution
Distribution of liftime	$f(t)=\lambda e^{-\lambda t}$	$f(t)=\lambda\gamma^{\gamma-1}e^{-\lambda t^{\gamma}}$
Survival function	$S(t)= e^{-\lambda t}$	$S(t)=e^{-\lambda t^{\gamma}}$
Hazard function	$h(t)=\frac{\lambda e^{-\lambda t}}{e^{-\lambda t}}=\lambda$	$h(t)= \lambda \gamma\ t^{\gamma-1}$

λ is referred to as the *scale* parameter, while γ is referred to as the *shape* parameter.

To choose the best fitting distribution using chi² base test statistics, the parameters for that distribution and the goodness-of-fit chi² were calculated. Table 5.2 displays the parameter estimated for two distributions.

Table 5.2 Goodness-of-fit Chi-square for fitted theoretical survival distributions

Survival Distribution	Lambada	Variance Lambada	Sed. Err. Lambada	Log likelihood	Chi²	df	p-value
Exponential	0.000188	0.000000	0.000010	-1475.46	27.59037	10	0.002105
	0.000183	0.000000	0.000010	-1476.00	28.66654	10	0.001416
	0.000196	0.000000	0.000010	-1475.14	26.93558	10	0.002674
Weibull	0.000494	0.000000	0.000238	-1476.25	29.15970	9	0.000611
	0.000294	0.000000	0.000130	-1474.91	26.48197	9	0.001708
	0.000333	0.000000	0.000149	-1474.94	12.55033	9	**0.184**

The Chi² test is based on the comparison of the likelihood of the respective model with the null model. If this test is significant, we can conclude that the fitted distribution is significantly different from the observed data, and therefore, we reject it as a model for the survival times. From table 5.2 it can be seen that none of the different parameter estimates for the exponential distribution seems to fit the observed survival distribution. Additionally, the exponential distribution is used to model data with a constant failure rate (indicated by the hazard plot which is simply equal to a constant (Fig. 5.5). Since, in this study the failure rate is not constant and dependant on time, then the exponential function has been rejected.

Just we found that the only one yielding a non-significant fit is the Weibull distribution with weighted least squares parameter estimates. It appears that the third set of parameters provides a reasonable fit to the data; the Chi² test for that model is not significant (p-value = 0.184). Therefore, we concluded that the Weibull distribution with third set of parameters (weight 3) provides a good theoretical model for the data. Additionally, to check the adequacy of fit, survival functions were plotted in Fig. 5.3.

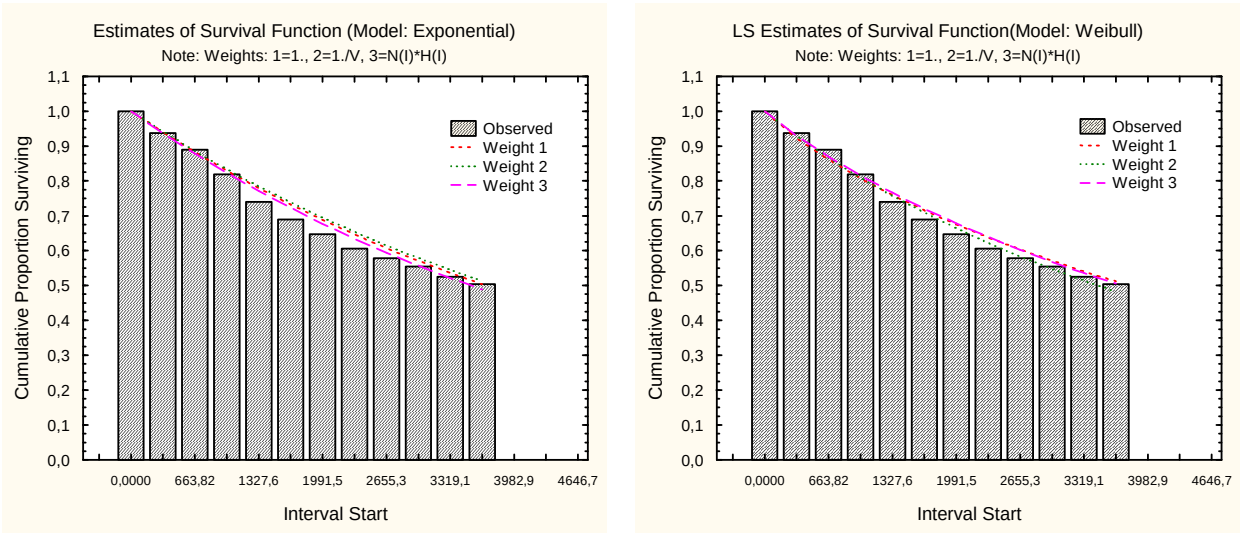

Fig. 5.3 Plot of Exponential and Weibull survival function

The probability density of two chosen distribution show decreasing over time (Fig. 5.4). It reflects the fact that probability of failure is greater in the earlier time intervals.

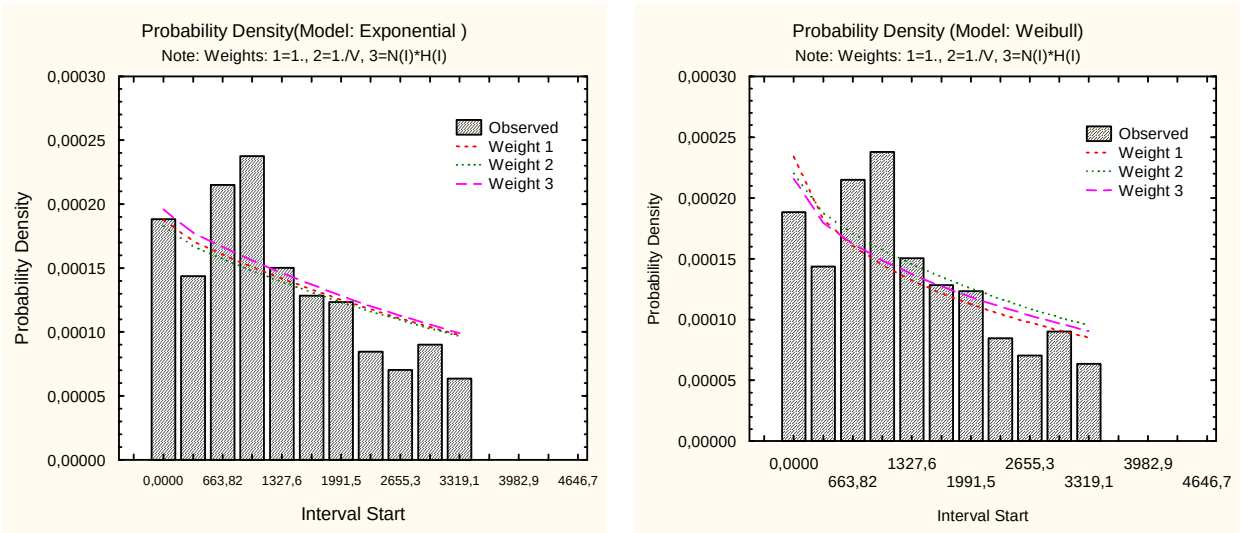

Fig. 5.4 Probability density of Exponential and Weibull survival function

Hazard function

The other important concept in survival analysis is the hazard rate. From looking at graphs with discrete time in Fig. 5.4 (time measured in large intervals such as 231 days) the hazard rate is the probability that a pipe will experience an event at time *t* while that individual is at

risk for having an event. For example, if a pipe had a hazard rate of 1.2 at time *t* and a second pipe had a hazard rate of 2.4 at time *t* then it would be correct to say that the second pipe's risk of an event would be two times greater at time *t*.

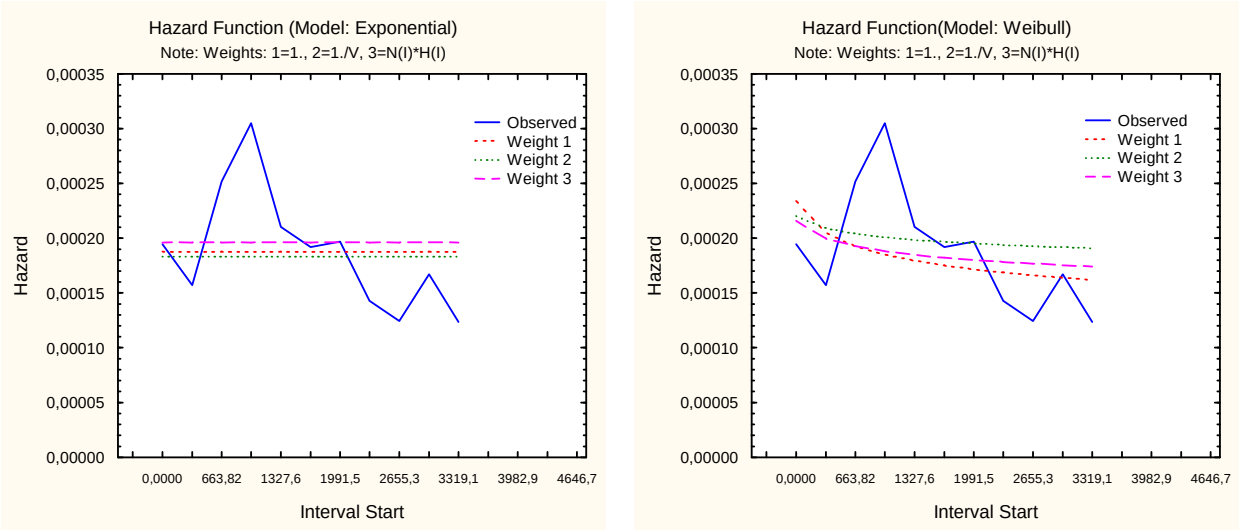

Fig. 5.5 Plot of hazard function for Exponential and Weibull distribution

Fig. 5.5 plots the hazard function in two exponential and Weibull model. Throughout Fig. 5.3 to Fig. 5.5, we concluded that the Weibull proportional hazard model is an appropriate for model of times to failure in water mains.

5.5 Proportional Hazards Models (PHM)

A common question in medical or engineering (failure time) research is to determine whether or not certain variables are correlated with the survival or failure times. There are two major reasons why this research issue cannot be addressed via straightforward multiple regression techniques: First, the dependent variable of interest (survival/failure time) is most likely not normally distributed - a serious violation of an assumption for ordinary least squares multiple regression. In this data set, survival times follow a Weibull distribution. Second, there is the problem of censoring, that is, some observations will be incomplete.

The proportional hazards models, proposed by Cox (1984), were used in this section. Non-parametric and parametric model formulation were examined. The hazard rate function was modeled with two known distributions: Cox PHM and Weibull PHM . The hazard rate

function of a PHM is the product of a baseline function, h_0, function of time, and of a parametric portion that takes the risk factors into account in a multiplicative manner. Given a set of k covariates x_i , the hazard function at time t is modeled by:

$$h(t|x) = h_0(t) * e^{\beta_1 x_1 + \ldots + \beta_k x_k} \quad (5.1)$$

where x_i = risk factor; and β_i = regression parameter of x_i.

One feature of a PHM is that even if the baseline function is not formulated, the relative importance of the risk factors (hazard ratio) can still be evaluated (Vanrenterghem A., 2007). One condition of PHM applicability is that risks remain proportional over time. For example, we suppose the hazard rate function of pipe A is as follows :

$$h_A(t|x) = h_0(t) * e^{\beta_1 x_{A1} + \ldots + \beta_k x_{Ak}} \quad (5.2)$$

The hazard rate function of the reference pipe R ($x_{Ri}=0$ for all i) is:

$$h_R(t|x) = h_0(t) * e^{\beta_1 x_{R1} + \ldots + \beta_k x_{Rk}} = h_0(t) \quad (5.3)$$

The ratio of the risks is:

$$\frac{h_A(t)}{h_R(t)} = e^{\beta_1 x_{A1} + \ldots + \beta_k x_{Ak}} \quad (5.4)$$

This represents pipe A's risk of breaking compared to pipe R's, the reference pipe. It is entirely dependent on the risk factors associated with pipe A. Then, if the baseline function is not specified, we will have the Cox proportional hazards models (CPHM). When the baseline function, h_0, is to be formulated, the Weibull model is one option. The function is then called a WPHM which is defined as:

$$h(t|x) = \lambda\gamma(\lambda t)^{\gamma-1} * e^{\beta_1 x_1 + \ldots + \beta_k x_k} \quad (5.5)$$

where λ and γ are scale and shape parameter, respectively.

5.5.1 Non-parametric Cox's hazard model

To determine the relationship between most influential independent variables and survival time, we used Cox's proportional regression model (CPHM). It does not make any

assumptions about the nature or shape of the underlying survival function. We estimated the regression coefficient for the independent variables in the prediction of survival times using the proportional hazard model. According to Eq. (5.1), we obtain the hazard ration :

$$Ln\left[\frac{h(t|x)}{h_0(t)}\right] = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \qquad (5.6)$$

We selected most 9 influential variables for the analysis and censored variable. By the estimation procedure, the log-likelihood of the regression model via Newton-Raphson iterations were maximized. The regression coefficient were obtained as:

$$Ln\left[\frac{h(t|x)}{h_0(t)}\right] = \begin{array}{l} 0.21 * NPF - 0.04361 * Age - 0.00541 * Diameter + \\ 2.02105 * Loglength + 0.99M1 + 1.136M2 + 2.206M3 \\ - 0.1838 * Depth - 1.1247 * Thickness + 0.084 * TL \\ 0.01052 * Maxpressure \end{array} \qquad (5.7)$$

The overall Chi² value for the model is significant (Chi² =385.625 ; df = 10; p = 0.00001). Then, we concluded that at least some of the independent variables are significantly related to pipelines survival. In table 5.3 significant risk factors appear in bold (p-value < 0.05).

Table 5.3 Statistical significant of each variable in CPAM for total pipelines

	Standard error	t-value	Wald Statistics	p-value
NPF	0.042314	4.96420	24.6433	**0.000001**
Age	0.018034	-2.44663	5.9860	**0.014425**
Diameter	0.002683	-2.01738	4.0698	**0.043664**
LogLength	0.180141	11.21926	125.8717	**0.000000**
M1 (AC)	0.388662	2.54962	6.5006	**0.010789**
M2 (DI)	0.705977	1.60898	2.5888	0.107632
M3 (CI)	0.616552	3.57835	12.8046	**0.000346**
Depth	0.677492	-0.25332	0.0642	0.800020
Thickness	0.776531	-1.52750	2.3333	0.126646
Traffic Load	0.13338	1.34192	1.8007	0.179633
MaxPressure	0.004175	2.65635	7.0562	**0.007903**

Note: Highlighted variables are statistically significant (p-value < 0.05)

Therefore, we concluded from table 5.3 that *NPF* , age, diameter, logLength, material (AC and CI) and maximum pressure are the most important predictors of hazard. In fact, through the use of a Cox proportional hazards model, six factors significantly entered the equation predicting survival time in the sample.

5.5.2 Parametric Weibull hazard model

In the second regression model of survival analysis, we considered accelerated failure time models which obtained by modeling the logarithm of failure time instead of the failure time itself (see table 1.5). This model specifies that the natural logarithm of the time to failure T is related to independent variable (risk factors) via a linear model. The model for the whole set of pipes is:

$$Ln(t) = \sum_{i=1}^{n} \beta_i X_i + \sigma W \qquad (5.8)$$

Where:

t = Time to failure σ = Scale on errors n = Number of influential factors

X_i = Risk factors W = Error vector

For a given failure on a specific pipe with related covariate values for this pipe, Eq. (5.8) can be rewritten as:

$$w(t) = \frac{\ln(t) - \sum_{i=1}^{n} \beta_i X_i}{\sigma} \tag{5.9}$$

Using the extreme value distribution with survival function (Klein et al., 1997):

$$S(w)=exp\ [-exp(w)] \tag{5.10}$$

By entering $w(t)$ into Eq. (5.10) , the survival function for each water pipelines is:

$$S(t,\beta,X) = \exp\left[- \exp\left(\frac{\ln(t) - \beta_0 - \sum_{i=1}^{n} \beta_i X_i}{\sigma}\right)\right] \tag{5.11}$$

As in logistic regression, parameter estimates in weibull survival models are obtained using maximum likelihood estimation. This analysis was performed with the statistical software SAS (SAS Institute, 1994). All four WPHM models loaded with 7 years of breaks is then used to validate the number of projected breaks for the next 3 years.

5.6 WPHM Model for All Water Mains

As illustrated before, all pipeline dataset was prepared for WPHM model. We used the Sanandaj failure data in the period 1995-2001 for determining the model parameters and next three years (between 2002-2004) for model validation. The results of validation are given in Fig. 5.6 which gives a comparison between the forecasts of the model and the observations. Visually we can say that the model gives good results.

The overall Chi-square value for the model is highly significant, then we concluded that at least some of the independent variables are significantly related to survival. Table 5.4 shows that four covariates logLength, diameter and material were entered in the model which significantly contribute to the prediction of time (p-value<0.05). These are mainly risk factors inherent to pipes (Length, diameter, material). Other variables were found not to significantly contribute to the prediction of time, and were not included in the model.

Table 5.4 Significant variables according to WPHM modeling (total water mains)

Risk Factors	p-value
Log Length	**< 0.0001**
Diameter	**< 0.0001**
M1 (AC)	**0.0097**
M3 (CI)	**0.0207**

The values of parameters β , σ and p-value for total water pipelines is illustrated in Table 5.5.

Table 5.5 Estimated model parameters for all water mains

Parameter	Value	p-value
β_0	9.9677	< 0.0001
β_1	-0.6282	< 0.0001
β_2	0.0093	< 0.0001
β_3	-0.5048	0.0207
β_4	0.5350	0.0097
σ	0.8855	

5.6.1 Benefit Index

One important issue in WPHM is the plot of the cumulative density function (CDF) obtained from the water pipelines dataset against the theoretical or predicted cumulative density function of the Weibull distribution. Findings of researches in number of European utilities indicated that this plot can be interpreted as an indicator for rehabilitation. In effect, this index aims at assessing number of failures avoided, if a defined percentage of pipes with the highest breaks probabilities were rehabilitated (Eisenbeis et al., 2002 ; Le Gauffre et al., 2000). The percentage of pipes that would have been replaced is plotted against the percentage of actual breaks avoided. It is assumed that the pipes chosen first for replacement would be the ones with the highest expected number of breaks for the next 3 years. Fig. 5.6 shows the benefit index curve for all water pipelines with WPHM modeling. It is apparent from graph that 20 % of breaks (for the next 3 years) are avoided if 7 % of the pipes are replaced. We assume the water pipelines most at risk are replaced first.

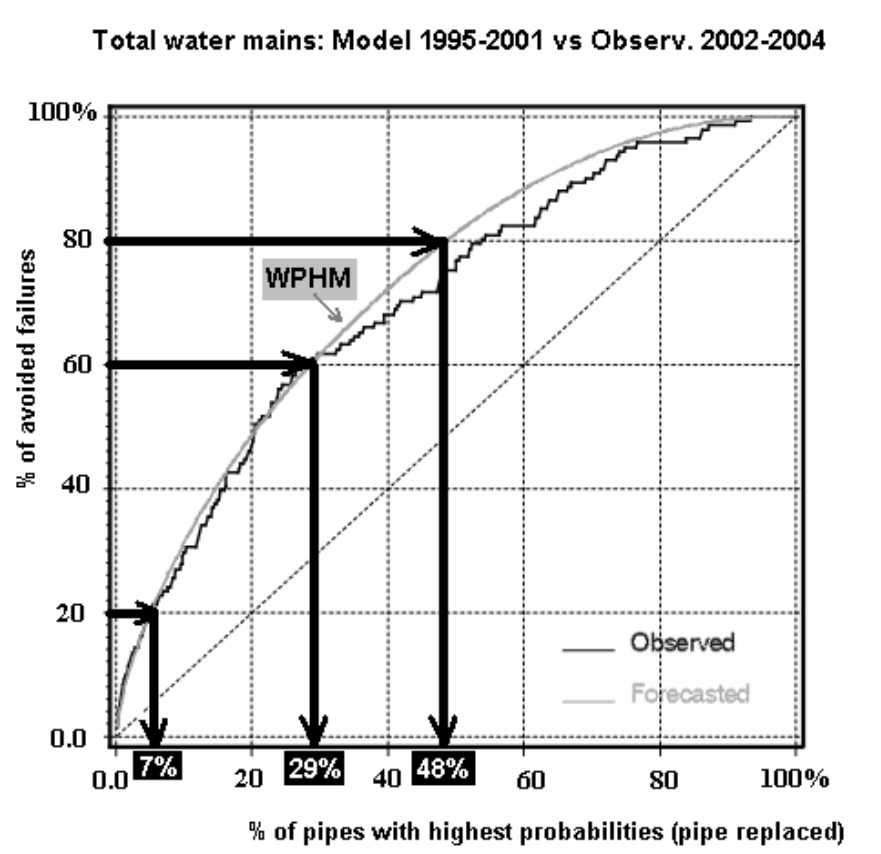

Fig. 5.6 Global model testing on data 2002-2004 (Benefit Index)

Note that a replacement rate of 1% means that a pipe is replaced every 100 years. Indeed, 29% of the highest failure risk pipes can avoid almost 60 % failure and 48 % allow to avoid almost 80%. Different values of avoided failure can be estimated by this graph.

Since the material predictor is not a function of time, then it may be too complicated to model the hazard ratio for that predictor as a function of time. Meanwhile, the pattern of failure in material categories are different. Therefore, we stratified the dataset by material instead of including it as a predictor in the model. So, three groups were created, metallic, cement and plastic water mains. Following are three WPHM models.

5.7 WPHM Model for Metalic Water Mains

This model was developed for metallic water pipelines which include both cast and ductile iron in study area. The model was tested on the data of metallic water pipelines in study area

over of the period 2002-2004. The results of test are given in Fig. 5.7 which presents a comparison between the forecasts of the model and the observations. Visually we can say that the model gives good results. The overall Chi-square value for the model is significant, then some of the independent variables are significantly related to survival. The analyses yielded significant results for loglength, diameter. In contrast, the NPF and traffic load were not statistically significant. Table 5.6 gives the list of significant indicators and their p-values.

Table 5.6 Significant variables according to WPHM modeling (metallic water mains)

Risk Factors	p-value
Log Length	**0.0034**
NPF	0.2083
Diameter	**0.0002**
Traffic category	0.1880

Table 5.7 illustrates the values of β and σ in WHPM model.

Table 5.7 Estimated model parameters for Metallic water mains

Parameter	Value	p-value
β_0	8.0245	< 0.0001
β_1	0.1003	0.0034
β_2	0.1975	0.2083
β_3	0.0023	0.0002
β_4	0.2024	0.1880
σ	0.0750	

Fig 5.7 is presenting Benefit Indices for WPHM using metallic water pipelines data since 1995 to 2001. Based on this graph, 20 % of breaks (for the next 3 years) are avoided if 5 % of the pipes are replaced. Some more forecasting are shown in table 5.8. These can be used to estimate the impact (benefit) of an action on the network. Indeed, the model makes it possible to calculate the number of the failures which could have been avoided if a certain number of sections (having the risk more high) had been rehabilitated.

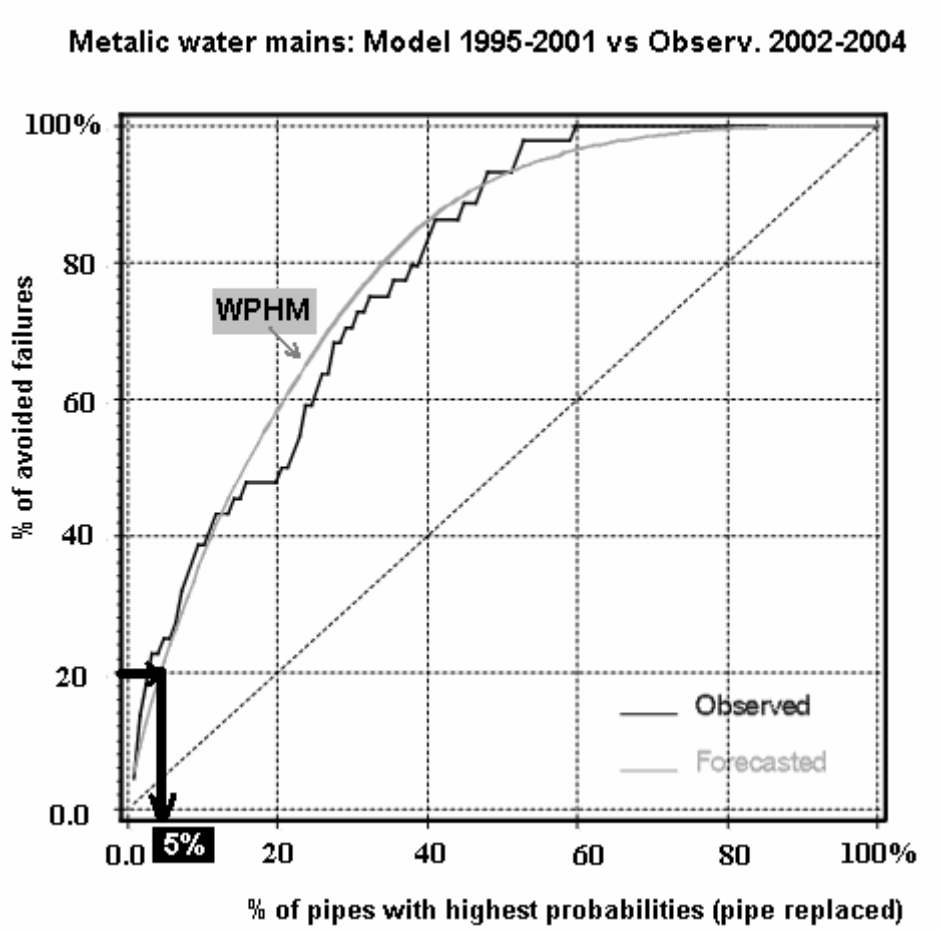

Fig. 5.7 Test of Metallic model on data 2002-2004 (Benefit Index)

Table 5.8 presents different percent of failure avoided according to pipe replacement.

Table 5.8 Results of benefit indices: comparison of failures observation against the forecasts

WPHM Model	
Period of Model	1995-2001
Compared to	2002-2004
% of pipes with highest predicted failure rate	% of avoided failures
40 %	11.2 %
60 %	21.7 %
80 %	36.2 %

As noted in table 5.8, the replacement of 40%, 60% and 80 % of cast and ductile iron which have the highest risk in failure, would have made it possible to avoid 11.2%, 21.7 and 36.2% of the failures, respectively.

5.8 WPHM Model for Cement Water Mains

This model relates to the network of asbestos cement which include 272 statistical individual. They account for 99 failure within 10 years. The chi² test made it possible to determine the influential factors of the model. Table 5.9 gives the list of these factors and the corresponding p-value. We found that loglength, diameter and age are statistically significant(p-value<0.05).

Table 5.9 Significant variables according to WPHM modeling (cement water mains)

Risk Factors	p-value
Log length	**0.0001**
NPF	0.3463
Diameter	**0.0115**
Age	**0.0099**

Table 5.10 presents the parameters of the model which were given on the data 1995 - 2001. This table also illustrates the values of the model parameters and p-value which makes it possible to measure the significantly of each parameter in the relation of survival. It is noted that the p-value lower than 0.05 are significant in the relation of survival.

Table 5.10 Estimated model parameters for Cement water mains

Parameter	Value	p-value
β_0	7.6763	< 0.0001
β_1	-0.4944	0.0001
β_3	0.0059	0.0115
β_4	0.8868	0.0099
σ	0.6201	

The model was tested on data within the period of 2002-2004. The graph in Fig. 5.8 demonstrates that the model gives good results. The benefit index graph also shows that

choosing 9% of the pipes with higher failure risk could allow to avoid 20% of the failure. Additionally, if the top of 20 % pipes (classified by predicted failure rate, highest first) had been rehabilitated, this should have resulted in a decrease in observed failures of 43%.

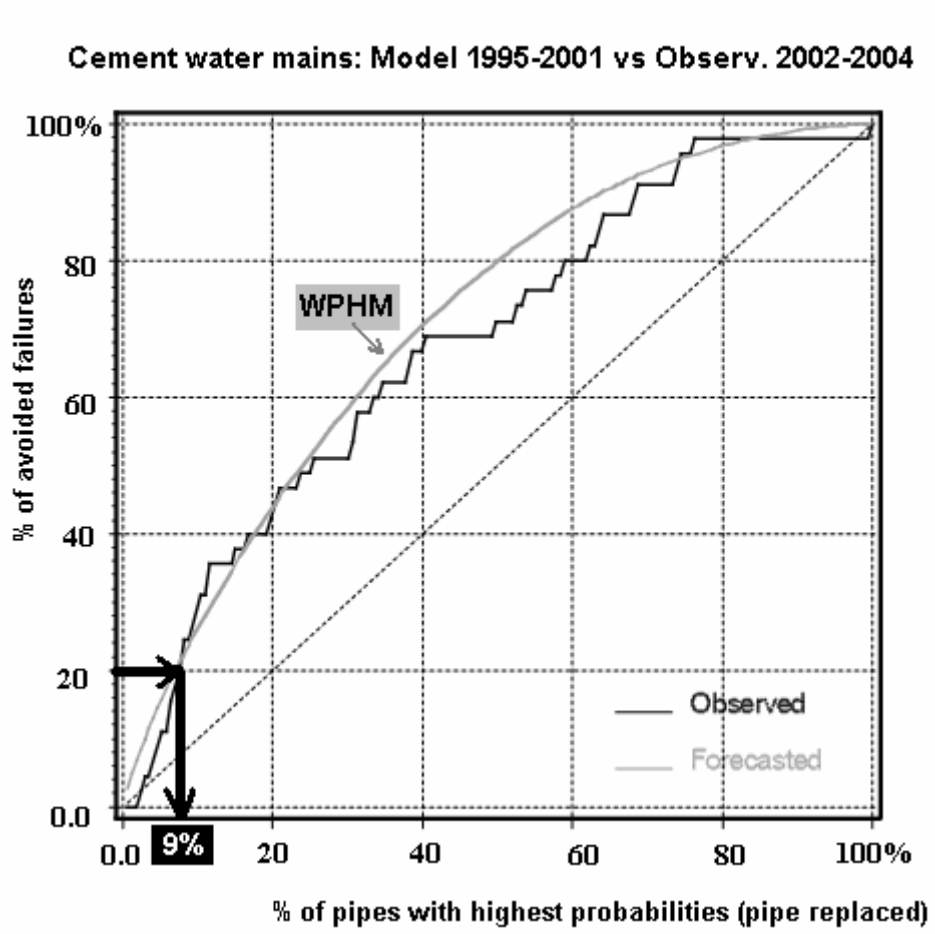

Fig. 5.8 Test of Cement model on data 2002-2004 (Benefit Index)

5.9 WPHM Model for Plastic Water Mains

The final model relates to plastic water pipelines in the selected network. Dataset for survival analysis of plastic water pipelines includes 426 total number of statistical individuals. 172 (40%) statistical individual have experienced breaks (uncensored) and the rest 254 (60%) are censored which means that they will be failed in the coming future.

Chi-squared test determines the influential factors of the model. Table 5.11 gives the list of these factors and the corresponding p-value. Accordingly, we can keep the logLength,

diameter and age of water pipelines in the fitted model. Table 5.12 gives the parameters of the model which were given on the data 1995 - 2001. The values of the parameter σ and β, as well as p-value are also given. It is noted that this value is lower than 0.05, which shows that these parameters are influential in the relation of survival.

Table 5.11 Significant variables according to WPHM modeling (plastic water mains)

Risk Factors	p-value
LogLength	**< 0.0001**
NPF	0.1904
Diameter	**0.0022**
Age	**0.0579**

As noted in table 5.11 and 5.12, the number of pervious failure in the survival model of plastic water mains, is not significant (p-value = 0.1904).

Table 5.12 Estimated model parameters for plastic water mains

Parameter	Value	p-value
β_0	13.5915	< 0.0001
β_1	-2.3065	< 0.0001
β_2	-1.0962	0.1904
β_3	0.0721	0.0022
β_4	2.9938	0.0579
σ	1.2833	

For plastic pipelines, the Benefit Index curve was developed in Fig. 5.9 through validation of model with failure data in the period of 2002-2004. This shows the percentage of failure avoided in polyethylene for various rate of replacement. By reading of the Benefit Index curve, it is indicated that 20% of failures (the next 3 year) are avoided if only 4 % of pipes with highest failure risks are replaced. If lower replacement ratios are targeted, for instance, 1 or 2 % , 4.2% and 8.5% of the breaks can be avoided, respectively.

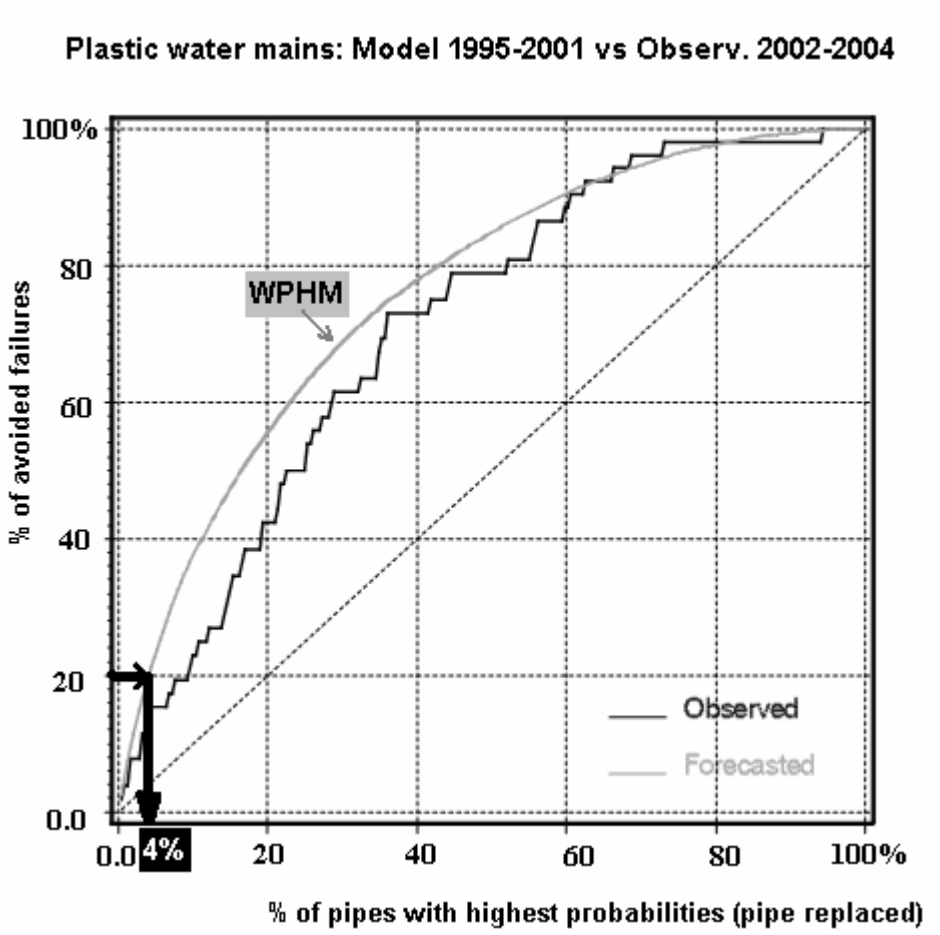

Fig. 5.9 Test of plastic model on data 2002-2004 (Benefit Index)

5.10 Model Comparison

To establish a priority list for water pipelines rehabilitation project according to pipe material, we compared four WPHM models discussed above in table 5.13.

Table 5.13 Percent of pipes that will be rehabilitated

Model	% of avoided failure			
	20	40	60	80
Total water mains	7	14	29	48
Metallic water mains	5	11	22	36
Cement water mains	9	17	31	50
Plastic water mains	4	12	24	43

Since the cost of pipeline rehabilitation is large but the budget is limited, then this priority list is necessary, and SWWU can formulate strategies for repair, replacement, and rehabilitation. Table 5.13 demonstrates that the priority for rehabilitation project is firstly metallic water mains, then plastic and finally cement pipelines. Because with less percent of pipe rehabilitation in metallic and plastic water mains, we can avoided more failures in study area. As a result, the authorities can set up the strategy for the water pipelines rehabilitation by considering to the rehabilitation priority.

5.11 Concluding Remarks

In this chapter, time to failure analysis was conducted for water pipelines failure in Sanandaj city using parametric and non-parametric survival approaches. Firstly, the most commonly used technique, Kaplan-Meier (KM), were introduced. Different KM curves for material groups indicated that metallic, cement and plastic water pipelines have different pattern of degradation. The slope of the KM survival curve is an important statistic telling if failures are occurring rapidly or the failure rate is slow. The findings of this study show that the survival function in the ductile iron pipelines has a slower drop-off as compared to the other pipes. Therefore, we concluded that ductile iron water pipelines have a greater chance of survival. Also, as a result of KM graph, cast iron pipes are expected to be more prone to bursting than others. Based on result of KM survival function, it is highly recommended to look at the KM curves as an option for evaluation of failure rate in different pipes material.

Fitting a theoretical survival distribution in time to water pipelines failure and varying the failure probability over time in the other hand, demonstrated that Weibull distribution provides a better fit to historical failure data. The exponential distribution of life time, survival and hazard function was compared in this study to weibull model. The results of comparing the model with null hypothesis using Chi^2 based test statistics, concluded that the fitted weibull distribution is statistically significant.

This chapter also examines the impacts of potential risk factors on the survival rate of water pipelines by two proportional hazard model namely, Cox's (CPHM) and Weibull (WPHM) model. 10 years survival was evaluated using Cox proportional hazards model. Among nine covariates in developing Cox's model, number of pervious failure (NPF), age, diameter,

loglength, material and maximum pressure were statistically significant (p-value < 0.05). In contrast there were no statistically significant relationship between traffic load categories, depth and pipe thickness in Cox's model.

Secondly, Weibull model was chosen for prediction through Benefit Index curves. Thus models of survival were developed for the total water mains, metallic, asbestos cement and plastic mains. These four models were calibrated on the data over the period 1995 - 2001 and tested on the failures observed within the period 2002 - 2004. The tests of validation gave good results, which shows the relevance of this approach for the forecast of the failures in the water pipelines networks. All 4 WPHM models assessed the relative importance of the risk factors by chi-square test. The results showed that loglength and water pipelines diameter are significant in all models (p-value<0.05). The age risk factor was significant in non-metallic water pipelines such as asbestos cement and polyethylene. In contrast, traffic load categories, thickness, depth pressure and number of pervious failure have a weak influence on the risk in the models. Finally, Benefit Index curves were plotted for total water mains, metallic, asbestos cement and plastic mains. It assessed the percentage of failures avoided, if a defined percentage of pipes with the highest breaks probabilities were rehabilitated. Further, a priority list for water pipelines rehabilitation project according to pipe material has been established. This allows for implementation of pipeline rehabilitation project firstly on metallic water mains, then plastic and finally cement pipelines.

General Conclusion

This work included the use of Statistical, Artificial Neural Networks and Survival approaches for the prediction of water pipelines failure. It provides planners and engineers with tools for an effective preventative maintenance and rehabilitation program. The performances of these approaches were studied on 10 years collected data in city of Sanandaj (Iran).

Firstly, preliminary statistical analysis allowed a better understanding of the failure mechanisms. It gave insight on the impact of various risk factors on the structural deterioration of water pipes. Nine indicators were identified to be used as input parameters in the prediction of water main failure. They include age of pipe, number of pervious failures, pipe length, diameter, thickness, depth, material, pressure and traffic. Geostatistical analyses were used for the determination of both the distribution of water pipelines failure in space the spatial trends of failures considering risk factors.

Univariate statistical inferences, indices of bivariate relationship and multivariate data analysis were conducted for the determination of the correlation between the affecting factors and identify the significant indicators of water pipelines failure. Factor analysis was also conducted to identify determinant factors and recognize the relative relationships among input indicators. It was concluded that the analysis must be carried out for each material of water mains. Further, two regression models, namely Multiple and Poisson were fitted to predict the number of failures. Since the multiple regressions could not fit the data adequately, an alternative model, Poisson regression, was used. It allowed a better fitting of data togother with respect of the initial assumption.

Neural network approach was used to model the water pipelines failure. Four model were developed: a global model and a model for each material. Each model was trained and tested to find the optimal number of hidden nodes. Comparison between forecasted and observed failures for the last 5 years showed that the designed neural network for total water mains, with $R^2 = 0.78$, gave satisfactory prediction. Further, three models for metallic, cement and

plastic pipelines gave R^2=0.86, 0.59 and 0.75, respectively. Stratification of material did not improve the results except in metallic pipelines. Moreover, comparison of the ANNs model and the Poisson regression showed that the ANNs provides an attractive and powerful tool for predicting water main failure.

In third part, the survival approach was used in order to shed additional light on the pipeline failure process as well as to extract useful information in future planning. These analysis focuse on time to failure of water pipelines by using parametric and non-parametric survival models. Kaplan-Meier (KM) curves for material groups showed that survival function in the ductile iron pipelines has a slower drop-off as compared to the other pipes. Therefore, they have a greater chance of survival. In contrast, cast iron pipes were more prone to bursting than others. Asbestos cement water pipelines showed fast survival declines with time. Polyethylene water pipelines have a particular behavior : in the first years, they have more failures. Results from parametric modeling demonstrated that Weibull distribution provides a better fit of the evolution of water pipelines failure. Two proportional hazard model namely, Cox's (CPHM) and Weibull (WPHM) model were used to examine the impact of potential risk factors on the survival rate of water mains. Based on p-value, the significant factors were determined. Finally, Benefit Index curves were plotted for total water mains, metallic, asbestos cement and plastic mains. It assessed the percentage of failures avoided, if a defined percentage of pipes with the highest breaks probabilities were rehabilitated. In a nutshell, the use of this approach made it possible to establish a priority list for future water pipelines rehabilitation project in accordance with their material. Accordingly, it is suggested that implementation of pipeline rehabilitation project firstly on metallic water mains, then cement and finally plastic pipelines.

This work constitutes a starting point for future research for more comprehensive assessment of the state of Iranian water distribution systems. It also provides tools for improving rehabilitation strategies based on quantitative and predictive models.

Bibliography

Adams, B.J. and Heinke, G.W. (1987). Canada's urban infrastructure: rehabilitation needs and approaches, Canadian Journal of Civil Engineering, Vol. 14, No. 5, pp.700

Ahn J.C., Lee S.W., Lee G.S. and Koo J.Y. (2005). Predicting water pipe breaks using neural network. Water Supply. Vol. 5, No. 3-4, pp. 159–172 © IWA Publishing 2005.

Al-Barqawi Hassan and Tarek Zayed, (2006). Condition rating model for underground infrastructure sustainable water mains. Journal of Performance on constructed facilities. Journal of Performance of Constructed Facilities, Vol. 20, Issue 2, pp. 126-135.

Benjamin, M.M., Sontheimer, H. and Leroy, P. (1996). Corrosion of iron and steel. In: Internal corrosion of water distribution systems. 2nd edition. AWWA Research Foundation and DVGW Technologiezentrum Wasser, Denver, CO. pp. 46.

Berke, L., and Hajela, P. (1991). Application of neural networks in structural optimization. NATO/AGARD Advanced Study Institute, 23(I-II), 731-745.

Boxall, J. B., O'Hagan, A., Pooladsaz, S., Saul, A. J., Unwin, D. M. (2007). Estimation of burst rates in water distribution mains. Water Management, Vol. 160, Issue: 2, pp. 73-82.

Bougadis J. , Adamowski K., Diduch R. (2005). Short-term municipal water demand forecasting. Hydrological Processes. Vol. 19, Issue 1 , pp. 137–148

Bowdena G., Nixon J., Dandyc G.,Maier H., Holmes M. (2006). Forecasting chlorine residuals in a water distribution system using a general regression neural network. Mathematical and Computer Modeling. Vol. 44, Issues 5-6, pp. 469-484.

Cattell, R. B. (1966). The scree test for the number of factors. J. Multivariate Behavioral Research, Vol.1, pp. 245-276.

Chau K.W. (2006). A review on integration of artificial intelligence into water quality modeling. Marine Pollution Bulletin, Vol. 52, Issue 7, pp. 726-733.

Chen, J.; and Adams, B.J. (2006). Integration of artificial neural networks with conceptual models in rainfall-runoff modeling. Journal of Hydrology. Issue : 318(1-4), pp. 232-249.

Ciottoni, A.S. (1985). Updating the New York City water system. Proceedings of the Specialty Conference on Infrastructure for Urban Growth. New York, pp. 69-77.

Clark, R. M., Eilers, R. G., and Goodrich, J. A. (1988). Distribution system: Cost of repair and replacement. Proc., Conf. on Pipeline Infrastructure, B. A. Bennett, ed., ASCE, New York, 428–440.

Clark, R. M., and Goodrich, J. A. (1989). Developing a data base on infrastructure needs. J. Am. Water Works Assn., 81(7), 81–87. [ISI]

Clark, R. M., Stafford, C. L., and Goodrich, J. A. (1982). Water distribution systems: A spatial and cost evaluation. J. Water Resour. Plann. Manage., 108(3), 243–256. [CEDB]

Cohen, A., and M.B. Fielding, (1979). Prediction of Frost Depth: Protecting Underground Pipes. Jour. AWWA. 71(2):113-116.

Corne, S., Murray, T., Openshaw, S., See, L., and Turton, I. (1999). Using computational intelligence techniques to model subglacial water systems. Journal of Geographical System, 1, 37-60.

Crane I. A. (1994). Incorporating GIS and predictive modeling into facilities management. pp. 589-598. AM/FM International Proceedings.

Davis, J. P., Allan, I., Burn, S. & van de Graaff, R. (2003). Identifying trends in cast iron pipe failure with GIS maps of soil environments. Proc. Pipes 2003 – Back to Basics: Design & Innovation, Wagga Wagga, NSW, Australia, 21–23 Oct. 2003 (CD ROM).

Davis J. P., Clark B.A. , Whiter J.T. & Cunningham R. J. (2001). A statistical investigation of structurally unsound sewer. Proceeding of the International Conference on Underground infrastructure research. pp. 125-133.

Davis J. P. , Clark B.A. , J.T. Whiter , R.J. Cunningham , A. Leidi, (2001). The structural condition of rigid sewer pipes: a statistical investigation . Urban water 3. pp. 277-286.

Desilva D., Burn L.S. Eiswirth M. (2001). Joints in water supply and sewer pipelines: An Australian perspective. Proceeding of the Conference "Pipes Wagga Wagga", October 2001, Australia.

Diab Y., Morand D. (2001). The considerations of risks in the analysis of urban buried pipes behaviour. Proc. of the International Conference on Underground infrastructure research. pp. 133-138.

Dixon K., Blakey G. & Whiter J. (2001). GIS-based risk analysis of ferrous water pipelines. Proc. of the International Conference on Underground infrastructure research.

Doleac, M. L., S. L. Lackey and G. Bratton, (1980). Prediction of Time-to-Failure for Buried Cast Iron Pipes, AWWA Annual Conference Proceedings, Atlanta, pp. 31-38, GA., June.

Eisenbeis, P., P. Le Gauffre, and S. Saegrov, (2000). Water infrastructure management: An Overview of European Models and Databases, AWWARF Infrastructure Conference and Exhibition Proceedings, Baltimore, Maryland, 2000.

Elkateb, M. M., K. Solaiman, (1998). A comparative study of medium-weather-dependent load forecasting using enhanced artificial/fuzzy neural network and statistical techniques. Neurocomputing 23(1-3): 3-13.

Faraway, J. and Chatfield, C. (1998). Time series forecasting with neural networks: a comparative study using the airline data, Applied Statistics., vol. 47, pp. 231-250.

Francis, C. (1994). "Sieving the evidence on leakage." Water and Waste Treatment, DR Publications, London.

Goulter, I., Davidson, J., & Jacobs, P. (1993). Predicting water-main breakage rates. Journal of Water Resources Planning and Management, vol. 119, No. 4, pp. 419-436.

Goulter, I.C., and A. Kazemi, (1988). Spatial and temporal groupings of water main pipe breakage in Winnipeg. Can. J. Civ. Eng., 15, pp. 91–97.

Grau, P. (1991). Problems of external corrosion in water distribution systems. Water Supply Congress, International Water Supply Association, International Reports, 9 (3/4) 5-1, 5-45.

Gustafson, J. M., and Clancy, D. V. (1999). Modeling the occurrence of breaks in cast-iron water pipelines using methods of survival analysis. Proc., AWWA Annual Conf., American Water Works Association, Denver.

Kaastra, L. and M. Boyd, (1995). Designing a Neural Network for Forecasting Financial and Economic Time Series', Neurocomputing, 10(3), pp. 215–36.

Karney, B.W. and McInnis, D. (1992). Efficient calculation of transient flow in simple pipe networks, Journal of Hydraulic Engineering, Vol. 118, No. 7, July 1992, pp. 1014-1030.

Kettler, A.J. and Goulter, I.C. (1985). An Analysis of Pipe Breakage in Urban Water Distribution Network, Canadian Journal of Civil Engineering, 12, pp. 286-293.

Kiefner, J.F., and Vieth, P.H. (1989). Project PR-3-805: A modified criterion for evaluating the remaining strength of corroded pipe. Pipeline Corrosion Supervisory Committee of the Pipeline Research Committee of the American Gas Association.

Klein, J. and Moeschberger, M. (1997). Survival analysis: Techniques for Censored and Truncated data. Springer, New York.

Kleiner, Y., Rajani B. (1999). Using limited data to assess future needs. Journal of the AWWA, 91(7), pp. 47- 62.

Kleiner, Y., Rajani B. (2001). Comprehensive review of structural deterioration of water mains: statistical models. Urban water 3 (2001), pp.131-150.

Kottmann, A. (1994). Pipe Damage due to Air Pockets in Low Pressure Piping, Proceedings of 2nd International Conference on Water Pipeline Systems, Edinburgh, Scotland, pp. 11-16.

Kumar, A., M. Bergerhouse, and M. Blyth. (1987). Implementation of a pipe corrosion management system. Corrosion 87, Paper Number 312, National Association of Corrosion Engineers, Houston.

Kumar A., (2005). Comparison of neural networks and regression analysis : A new insight. Expert system with applications. 29 (2005), pp. 424-430.

Gat Y. and P. Eisenbeis, (2000). Using maintenance records to forecast failures in water networks. Urban Water, pp. 173-181.

Jarvis, M.G. and Hedges, M.R. (1994). Use of soil maps to predict the incidence of corrosion and the need for iron mains renewal. Journal of the Institution of Water and Environmental Management, Volume 8 Issue 1, pp. 68-75,.

Habibian, A. (1994). Effect of temperature changes on water-main break", Journal of transportation engineering 120 2 (1994), pp. 312–321.

Hajmeer, M., Basheer, I., Najjar, Y., (1997). Computational Neural Network for Predictive Microbiology II: Application to microbial Growth. International Journal of Food and microbiology 34, pp. 51-66.

Hau Y., B. Clark, C. Howes, S. Leidi, R. Cunningham, M. Matthews, (2005). A statistical investigation of factors affecting the deterioration of sewer joints. Conference for the Engineering Doctorate in Environmental Technology, Guildford, UK.

Hecht-Nielsen, R. (1989). Theory of the back propagation neural network. Proceedings of the Intemational Joint Conference on Neural Networks. Washington D.C., vol. 1, pp. 593-608.

Herbert, H. (1994). Technical and economic criteria determining the rehabilitation and/or renewal of drinking water pipelines. Water Supply, 12 (3/4, Zurich), pp. 105-118.

Herz, R. K. (1996). Ageing processes and rehabilitation needs of drinking water. Journal Water SRT, 1996. 45(5), pp. 221-231.

Hornik, K., Stinchcombe, M., and White, H. (1989). "Multilayer feedforward networks are universal approximators." Neural Networks, 2, pp. 359–366.

Hu, Y. and Hubble, D. (2005). Failure conditions of asbestos cement water pipelines in regina, Canadian Society of Civil Engineering (CSCE) 33rd Annual Conference, Toronto, Ontario, Canada, June 2-4, 2005.

Jacobs, P., and Karney, B. (1994). GIS development with application to cast iron water main breakage rate. 2nd Int. Conf. on Water Pipeline Systems, BHR Group Ltd., Edinburgh, Scotland.

Lambert, A. O. (1998). A realistic basis for objective international comparisons of real losses from public water supply systems. The Institute of Civil Engineers Conf., Water Environment 98 - Maintaining the Flow, London.

Lackington, D.W. and Large, J.M. (1980). The integrity of existing distribution systems. Journal of the Institute of Water Engineers and Scientists, 34, pp. 15-32.

Lei, J. and S. Saegrov, (1998). Statistical Approach for Describing Failures and Lifetime of Water Mains. Water Science and Technology, 1998. 38(6): pp. 209-217.

McDade, Thomas W., Adai Linda S. (2001). Defining the ''urban'' in urbanization and health: a factor analysis approach. Social Science & Medicine Journal No. 53 pp. 55–70. Elsevier Science Ltd.

McMullen, L. D. (1982). Advanced Concepts in Soil Evaluation for Exterior Pipeline Corrosion, Proceeding AWWA Annual Conference, Miami.

Maier H.R. and Dandy G.C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications. Environmental Modeling & Software, Vol.15, pp. 101-124.

Marshall, P. (1999). Evaluation of Long Term Performance: The Behaviour of Buried Pipes." UKWIR, Research No. 99/WM/20/12.

Makar J.M., Desnoyers R. and McDonald S.E. (2001). Failure modes and mechanisms in gray cast iron pipes. Underground Infrastructure Research: Municipal, Industrial and Environmental Applications, Proceedings, Kitchener, Ontario, June 10-13, 2001, pp. 1-10

Malandain J., Le Gauffre P., Miramond M. (1998). Organizing A Decision Support System For Infrastructure Maintenance: Application To Water Supply Systems, Proceedings. First International Conference on New Information Technologies for Decision-making in Civil Engineering. Montreal (Canada) 11-13 Oct. 1998, pp. 1013-1024. ISBN 2-921145-14-6.

Mas, J.F., Puig, H., Palacio, J.L., Sosa, Lopez, A. (2004). Modeling deforestation using GIS and artificial neural networks. Environmental Modeling & Software 19 (2004) pp. 461–471.

McDonald, S., Daigle, L., and Félio, G. (1994). ''Réseaux d'aqueduc et systèmes d'égouts.'' Rep., Infrastructures Laboratory, Institute for Research in Construction, National Research Council of Canada, Ottawa (in French).

Mishra A.K., Desaia V.R. (2006). Drought forecasting using feed-forward recursive neural network. Ecological Modeling Vol. 98, Issues 1-2, pp. 127-138.

Mitchell, A. (1999). The ESRI Guide to GIS Analysis, Volume 1: Geographic Patterns and Relationships. Redlands, CA: ESRI Press.

Moglia M. , Burn S. and Meddings S. (2006). Decision support system for water pipeline renewal prioritisation. ITcon Vol. 11 pp. 237-248 .

Moody, J.E., Hanson J.E., and Lippmann, R.P. (1992). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. Advances in Neural Information Processing Systems 4, pp. 847-854.

Morris, R.E. (1967). Principal causes and remedies of water main breaks. J. AWWA, 54: 782-798.

Moselhi, O., and Shehab-Eldeen, T. (2000). Classification of defects in sewer pipes using neural networks. Journal of Infrastructure Systems, ASCE, 6 (3), 97-105.

Najjar, Y., Basheer, I., Hajmeer, M. (1997). Computational Neural Network for predictive microbiology: Methodology. International Journal of food microbiology volume 34, pp. 27-49.

Najjar, Y. M., Basheer, I. A., and Naouss, W. A. (1996). On the identification of compaction characteristics byneuronets. J. Computers and Geotechnics, 18(3), pp. 167-187.

Nawari, N. O., Liang, R., and Nusairat, J. (1999). Artificial intelligence techniques for the design and analysis of deep foundations. Elec. J. Geotech. Eng., http://geotech.civeng.okstate.edu/ejge/ppr9909/index.html.

Olden, J.D., Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecol. Model. 154, pp. 135–150.

Pandey M.D. (1998). Probabilistic models for condition assessment of oil and gas pipelines. NDT&E International, Vol. 31, No. 5, pp. 349-358, 1998

Pascal, O. and Revol, D. (1994). "Renovation of water supply systems." Water Supply Congress, International Water Supply Association, 12 ((1/2) Budapest), 6-3, 6-7.

Pelletier Genevie`ve, Mailhot Alain, and Villeneuve Jean-Pierre, (2003). Modeling Water Pipe Breaks—Three Case Studies; Journal of Water Resources Planning and Management, Vol. 129, No. 2, March/April 2003, pp. 115-123.

Pijanowskia, B. C., Brown, D. G., Shellitoc, B. A., and Manikd, G. A. (2002). Using neural networks and GIS to forecast land use changes: a land transformation model." Computers, Environment and Urban Systems, 26(6), pp. 553-575.

Rajani, B. , Zhan, C. , Kuraoka, S. (1996). Pipe-soil interaction analysis of jointed water mains. Canadian Geotechnical Journal, v. 33, no. 3, June 1996, pp. 393-404

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

Rajani, B., Kleiner, Y. ; Sadiq, R. (2006). Translation of pipe inspection results into condition ratings using the fuzzy synthetic evaluation technique. Journal of Water Supply Research and Technology: Aqua, v. 55, no. 1, Feb. 2006, pp. 11-24

Rigol, J.P., C.H. Jarvis and N. Stuart, (2001). Artificial neural networks as a tool for spatial interpolation, Int. J. Geographical Information Science, 15, pp. 323-343.

Rossum, J. R. (1969). Prediction of pitting rates in ferrous metals from soil parameters. Journal of American Water Works Association 61: pp. 305–310.

Rumelhart D. E. and McClelland J. L. (1986). Parallel distributed processing: explorations in the microstructure of cognition, Vol. 1, MIT Press, Cambridge, MA.

Salchenberger, L. M., Cinar, E. M., and Lash, N. A. (1992). Neural networks: A new tool for predicting thriftfailures. Decision Science, 23, pp. 899-916.

Sundahl, A. (1997). Geographical analysis of water main pipe breaks in the City of Malmö, Sweden, J. Water SRT – Aqua, 46(1), pp. 40-47.

Savic Dragan A. and Godfery A. Walters (1997). Hydroinformatics, data mining and maintenance of UK water networks. Journal of Quality in Maintenance engineering. Vol. 46 , No. 6, pp. 415-425.

Sinske S. and Zietsman H. (2004). A spatial decision support system for pipe-break susceptibility analysis of municipal water distribution systems. ISSN 0378-4738, Water SA, Vol. 30 No.1.

Silverman B. (1986). Density estimation for statistics and data analysis (London: Chapman and Hall).

Shahin, M.A., M.B. Jaksa, and H.R. Maier (2001). "Artificial neural network applications in geotechnical Engineering" Australian Geomechanics, Vol. 36, No. 1, pp. 49-62.

Shamir U., Howard C. D. (1979). An analytic approach to scheduling pipe replacement, Journal of the AWWA, Vol. 71, No. 5, pp. 248-258, May 1979.

Skipworth, P. Engelhardt, M. Cashman, A. Savic, D. Saul, A. Walters, G. (2002). Whole life costing for water distribution network management. Thomas Telford Publishing, London, ISBN 0727731661.

Sunil K. Sinha and Paul W. Fieguth (2006). Neuro-fuzzy network for the classification of buried pipe defects. Automation in Construction, Vol. 15, Issue 1, pp. 73-83.

Stewart Burn, Dhammika De silva, Matthias Eiswirth, Osama Hunaidi, Andrew speers and Julian thornton (1999). Pipe leakage, Future challenges and solutions. Pipes wagga , Australia 1999.

Trost S. M., and Oberlender G. D. (2003). Predicting Accuracy of Early Cost Estimates Using Factor Analysis and Multivariate Regression. Journal of Construction Engineering and Management, Vol. 129, No. 2, ASCE, USA.

Tsui E. and G. Judd, (1991). Statistical Modeling of Water Main Failures. Urban Water Research Association of Australia: Sydney.

Walski, T. M., and Pelliccia, A. (1982). Economic analysis of water main breaks. J. AWWA, 74(3), pp. 140-147.

Wanas N., Auda G., Kamel S. K., and Karray F. (1998). On the optimal number of hidden nodes in a neural network. Electrical and Computer Engineering Conference, IEEE, Vol. 2, pp. 918-921.

Wylie, E.B., and Streeter, V.L. (1993). Fluid Transients in Systems. Prentice-Hall Inc., Englewood Cliffs, New Jersey, USA.

Vanrenterghem-Raven A., Eisenbeis P., Juran I. and Christodoulou S. (2003). Statistical modeling of the structural degradation of an urban water distribution system: Case study of New York city. World Water & Environmental Resources Congress and Related Symposia. Philadelphia, Pennsylvania, USA.

Vanrenterghem-Raven A. (2007). Risk Factors of Structural Degradation of an Urban Water Distribution System. Journal of infrastructure systems. Vol. 13, No 1, pp. 55-64.

Zhou F., Hicks F., Steffler P. (2004). Analysis of Effects of Air Pocket on Hydraulic Failure of Urban Drainage Infrastructure, Canadian Journal of Civil Engineering, 31: pp. 86-94.

M.Sc. Thesis:

Agar, F. (1999). Data reclamation for data mining. M.Sc. thesis, School of computing, Staffordshire University, UK.

Aslani, P. (2003). Hazard Rate Modeling and Risk Analysis of Water Mains, Project report. Polytechnic University, New York, USA.

Anita J. (2004). Fragility Analysis of Water Supply Systems .Cornell University, New York, USA.

Garry Doyle, (2000). The role of soil in the external corrosion of cast-iron water pipelines in University of Toronto, TORONTO, Canada.

Gayatri Nadimpalli, (2003). Estimating Leaks in Water Distribution Systems by Sequential Statistical Analysis of Continuous Flow Readings. Master of Science project Report in University of Cincinnati.

Guan, X. (1995). Condition and Replacement of Regina's Water Distribution System, M.Sc. Theses, University of Regina, Regina, SK, Canada.

Sacluti Fernando R. (1999). Modeling water distribution pipe failures using artificial neural networks. University of Alberta.

Siska, P. and Hung, K. (2001). Assessment of Krigging Accuracy in the GIS Environment. Presented at The 21st Annual ESRI International Conference, San Diego, CA, 2001.

Wanas N., and Kamel M. (2001). Combining Neural Network Ensembles". International Joint Conference on Neural Networks (UCNNOI), Washington, D.C., Jul 15 - 19. pp. 2952-2957.

Zhang T. (2006). Application of GIS and CARE-W Systems on Water Distribution Networks, Skärholmen, Stockholm, Sweden. Royal Institute of Technology, Sweden.

PhD Dissertation:

Agbenowosi Newland Komla, (2001). A Mechanistic Analysis Based Decision Support System for Scheduling Optimal Pipeline Replacement. Virginia Polytechnic Institute and State University, USA.

Andreou, S. (1986). Predictive models for pipe break failures and their implications on maintenance planning strategies for deteriorating water distribution systems. PhD thesis, MIT, Cambridge, MA.

Annie Vanrenterghem, (2003). Modeling of the structural degradation of an urban water distribution system. Polytechnic University, USA.

Eisenbeis, P. (1994). Modélisation statistique de la prévision des défaillances sur les conduites d'eau potable. Thèse de doctorat, Université Louis Pasteur, Strasbourg.

Kaara, A. F. (1984). A decision support model for the investment planning of the reconstruction and rehabilitation of mature water distribution systems. PhD thesis, MIT, Cambridge, MA.

Kleiner K. (1997). Water Distribution Network Rehabilitation: Selection and Scheduling of Pipe Rehabilitation Alternatives. Department of Civil Engineering University of Toronto.

Michael D. Royer, (2005). White paper on improvement of structural integrity monitoring for drinking water mains, EPA/600/R-05/038.

Misiunas Dalius, (2005). Failure Monitoring and Asset Condition Assessment in Water Supply Systems. Lund University. Lund - SWEDEN.

Mohammed S. EILA, (2005). Développement durable: une approche efficace pour la gestion de l'utilisation des sols Application à la ville de Gaza. Université de Lille 1, Francec.

Raed Jafar, (2006). Modélisation de la dégradation des réseaux d'eau en vue d'une gestion prévisionnelle. Université de Lille 1, France.

Røstum J. (2000). Statistical modeling of pipe failures in water networks. Norwegian University of Science and Technology (NTNU).

Smith, E.P. (1994). An Optimal replacement- Design Model for a Reliable Water Distribution Network System. PhD thesis, Virginia Polytechnic Institute and State University.

Book and Reports:

Abso Engineering Consulting Company (2000). Sanandaj's UFW report, Tehran, Iran.

Alberta Energy and Utilities Board, (1998). Pipeline Performance in Alberta 1980-1997. Report G.

American Water Works Service Company, AWWSC (2002). Deteriorating buried infrastructure management challenges and strategies.

AWWAFR, (2000). Investigation of grey cast iron water pipelines to develop a methodology for estimating service life. American water work association research foundation. ISBN 1-58321-063-6.

AWWA, (2001). Dawn of the Replacement Era-Reinvesting in Drinking Water Infrastructure. AWWA Water Industry Technical Action Fund, Denver.

Best Practices, (2003a). Best practices for utility-based data. Best Practice by the National Guide to Sustainable Municipal Infrastructure, Issue No. 1.0, Ottawa.

Best Practices, (2003b). Deterioration and inspection of water distribution systems. Best Practice by the National Guide to Sustainable Municipal Infrastructure, Issue No. 1.1, Ottawa.

Canadian National Guide to Sustainable Municipal Infrastructure (InfraGuide), 2002a. Environmental Protocols, Strategic Commitment to the Environment by Municipal Corporations, Ottawa, Ontario. ISBN 1–897094–50–7.

Cressie, N. (1991). Statistics for Spatial Data, Wiley, New York.

Cox, D. R., and Oakes, D. (1984). Analysis of Survival Data, Chapman and Hall, London, New York.

Diggle, P. J. (1983). Statistical Analysis of Spatial Point Patterns, Academic Press, London.

Gregory B. Baecher and John T. Christian, (2003) . Reliability and Statistics in Geotechnical Engineering. John Wiley & Sons , Ltd. ISBN: 0-471-49833-5.

Kilmeny J. Stephens, Earth Tech, and Janet Jackson, (2003). Main Rehabilitation Prioritization - Getting the Data together for OWASA.

Lim E. L. , Pratti R. (1996). Pipe Evaluation System. Internal report from consultants to Seattle Public Utilities.

Macmillan, (1986). Mechanical reliability, Great Britain.

Marks, D.H., S. Andreou, L. Jeffrey, C. Park, A. Zaslasky, 1987. "Statistical Models For Water Main Failures", U.S. EPA Technical Report, Cooperative Agreement No. CR810558

Mather, J. Blackmore, C. Petrie, and others. WS Atkins Consultants Ltd for HSE (2001). An assessment of measures in use for gas pipelines to mitigate against damage caused by third party activity .

Mavin, K. (1996). Predicting Pipe Failure Performance of Individual Water Mains. UWRAA, Research Report No. 114, 189.

Mays W., Sukru Ozger and Larry, (2001). Optimal Location of Isolation Valves in Water Distribution Systems : A Reliability / Optimization Approach.

Melina G., Kalles D. (2000). Water Network Maintenance Models. Technical Report at the University of Patras.

Mordak, J. and Wheeler, J. (1988). Deterioration of asbestos cement water mains, Final Report to the Department of the Environment, Water Research Center, Wiltshire, UK.

Morid S., Smakhtinb V. and Bagherzadehc K. (2006). International Journal of Climatology Drought forecasting using artificial neural networks and time series of drought indices.

Najjar, Y. (1999). Quick Manual for the Use of ANN program TR-SEQ1. Department of Civil Engineering, Kansas State University, Manhattan, Kansas, USA.

O'Day, Kelly, Richard Weiss, Suzanne Chiavari, and Dennis Blair, (1982). Water Main Evaluation for Rehabilitation/Replacement. Guidance Manual: Distribution Systems. Prepared for

American Water Works Association Research Foundation and US EPA Water Engineering Laboratory, Cincinnati, Ohio. (ISBN 0-915295-10-5).

Organization for Economic Cooperation and Development (2003). Calculation of Operational and Financial Performance Indicators for Azervodokanal Water and Sewerage Utilities. Baku.

Rajani, B. & McDonald S. (1995). Water pipelines break data on different pipe materials for 1992 and 1993. National Research Council of Canada, Ottawa. Report No. A-7019.1.

Report No WSAA 145, (1998). Prediction of Pipeline Failures from Incomplete Data.

Rossman, L. A., EPANET User Manual, (1993). Drinking Water Research Division, Risk Reduction Engineering Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Cincinnati, OH., July, 1993.

Roy F. Weston, Inc., and TerraStat Consulting Group, (1996). Pipe Evaluation System: Deterioration Model Statistical Analysis. Internal report from consultants to Seattle Public Utilities.

Pentecost Allan, (1999). Analysing environmental data Allan Pentecost. Harlow Longman.

SAS (1994). User's Guide. Version 6, Cary, NC, USA, SAS Institute Inc.

Stein Ing.& Partner Gmb H. (2005). European study of the performance of various pipe systems, respectively pipe materials for municipal sewage systems under special consideration of the ecological range of effects during the service life.

Stone S., Dzuray E.J., Meisegeier D., Dahlborg A. and Erickson M. (2000). Decision-support tools for predicting the performance of water distribution and wastewater collection systems. Logistics Management Institute. McLean VA 22102-7805.

The Sewerage Rehabiltation Manual (4th edition); WRc; 2001.

Vic Barnett, (2004). Environmental Statistics Methods and Applications, John Wily & Sons, Ltd.

Weimer, D. (2001). Water loss management and techniques, Technical report, DVGW, The German Technical and Scientific Association for Gas and Water.

WHO, (2001) . Leakage Management and Control - A Best Practice Training Manual.

WSSA Facts, (1998), The Australian Urban Water Industry, Water Services Association of Australia.

WS Atkins Consultants Ltd. Report, (2001). An assessment of measures in use for gas pipelines to mitigate against damage caused by third party activity. Research report which prepared by WS Atkins Consultants Ltd for the Health and Safety Executive.

Web site:

NWWEC, National Water and Wastewater Engineering Company, (2006). "Annually failure reports on water distribution system. "<http://www.nww.co.ir>.

Résumé

La défaillance des réseaux d'eau constitue un problème majeur en Iran, qui nécessite des investissements importants et l'élaboration d'une stratégie optimale pour la réhabilition des réseaux d'eau. Ce travail constitue une contribution à cet objectif. Il vise le développement des outils pour améliorer la gestion et la maintenance des réseaux d'eau. Il comporte la détermination des principaux facteurs affectant la défaillance des réseaux d'eau, l'élaboration d'un modèle de prévision fondé sur les Réseaux de Neurones Artificiels (RNAs), et le développement d'un modèle de survie. Ces approches ont été appliquées sur le réseau d'eau de la ville de Sanandaj en Iran.
Le travail de thèse a comporté différents parties, notamment : la collecte de données sur le réseau de la ville de Sanandaj (Iran), l'analyse spatiale et statistique de ces données, le développement d'un modèle basé sur les Réseaux de Neurones Artificiels et l'application de l'approche de survie.
L'analyse des données a permis la détermination de principaux facteurs à l'origine de la défaillance des réseaux d'eau. Deux modèles de régression (Multiple et Poisson) ont été employés pour la prévision du nombre de défaillances du réseau d'eau. Ces modèles ont été comparés à l'approche des Réseaux de Neurones Artificiels. La comparaison a montré tout l'intérêt d'utiliser cette dernière approche pour la prévision de la défaillance des réseaux d'eau. L'approche de survie a été utilisée pour étudier la durée de vie et étudier l'impact d'une intervention sur le réseau d'eau.

Mots clefs: Réseaux d'eau potable, Défaillance, Modélisation statistique, SIG, Réseaux de neurones, Analyse de survie, Prévision, Réhabilitation.

Abstract

A major challenge to Iranian water industry concerns the minimization of failures in water distribution system. This thesis constitutes a contribution for this objective. It includes a) assessment of the main indicators through statistical analysis; b) development of Artificial Neural Networks (ANNs) models for predicting pipes failure number; c) elaboration of a survival models for quantification avoided failure from network based on various rate of renewal. The use of theses approaches generates a quantitative picture of the condition and performance of mains network towards the optimization of the maintenance and rehabilitation programs. All neural networks and survival models were trained and tested on field data in Sanandaj city (Iran).
The methodology followed in this research includes field data collection, descriptive spatial and statistical analysis besides predictive modeling which incorporate Regression, ANNs and Survival models. Descriptive analysis of historical failure data based on statistical methods allowed the determination of factors affecting the evolution of water pipelines failure. Indeed, geostatistical analysis and spatial interpolation provide scientific bases for depicting spatial relationships and the strength of dependencies between failure incidents and environmental, hydraulic and other geographic covariates. Review of univariate statistical inferences, indices of bivariate relationship and multivariate data analysis assess the correlation between the affecting factors and identify the important variables for the occurrence of failures on the water mains. Two regression models (Multiple and Poisson) were used for the prediction of the number of failures in water mains. Artificial Neural Networks (ANNs) models were also developed to predict the number of failures in water mains. Comparison of ANNs and regression approaches reveals that the use of ANNs model in pipeline failure studies provides better prediction. Finally, four survival models were developed to simulate time to failure in water mains, and 3 stratified failure dataset.

Keyword: Water network, Pipeline, Failure, Statistical analysis, GIS, Neural Network, Survival analysis, Prediction, Rehabilitation.