Iron and Manganese Accumulation Potential in Water Distribution Networks

Ebenezer Danso-Amoako

School of Computing, Science and Engineering

University of Salford

Salford, UK

Submitted in Partial Fulfilment of the Requirements of the

Degree of Doctor of Philosophy in Civil Engineering

January 2016

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full name
ADALINE	ADAptive LInear NEuron
ART	Adaptive resonance theory
AI	Artificial intelligence
ANN	Artificial Neural Network
AC	Asbestos Cement
ARIMA	Auto regressive integrated moving average
ARNN	Auto-regression neural network
CI	Cast Iron
CoG	Centre of Gravity
CA	Classification accuracy
CTM	Cohesive Transport Model
DSR	Demand Satisfaction Ratio
DPM	Discolouration Propensity Model
DRAT	Discolouration Risk Analysis Tool
DRM	Discolouration Risk Modelling approach
DO	Dissolved Oxygen
DMA	District Metered Area
DWI	Drinking Water Inspectorate
DI	Ductile Iron
ES	Evolution Strategy
EP	Evolutionary Programming
FCR	Free chlorine residual
FIS	Fuzzy Inference System
GP	Genetic Programming
HDPE	High Density Polyethylene
IRC	Intersection Rule Configuration
JM	Jacobian matrix
KPI	Key Performance Indicator
LOM	Largest Of Maximum
LCS	Learning Classifier System
LM	Levenberg–Marquardt
MCL	Maximum Concentration Level
MSE	Mean Square Error
MOM	Middle Of Maximum
MAR	Multivariate auto-regression
NSF	National Sanitation Foundation

Abbreviation	Full name
NTU	Nephelometric Turbidity Units
NPT	Node Probability Table
NPDMA	number of properties in a DMA
PSM	Particle Sediment Model
PWG	Pennine Water Group
PE	Polyethylene
PVC	Polyvinyl Chloride
PODDS	Prediction of Discolouration events in Distribution Systems
PDA	Pressure-dependent Analysis
QCC	quarterly customer complaints
RBF	Radial Basis Function
RNN	Recurrent Neural Network
RPM	Re-suspension Potential Method
RMSE	Root Mean Square Error
SOFM	Self-organising feature maps
SOM	Self-Organising Map
SIM	Service Incentive Mechanism
SOMax	Smallest Of Maximum
ST	Steel
SQL	Structured query language
SSE	Sum of Square Error
TOC	Total Organic Carbon
THM	Trihalomethane
URC	Union Rule Configuration
WDN	Water Distribution Network
Ofwat	Water Services Regulation Authority in England and Wales
WSZ	Water Supply Zone
WHO	World Health Organisation

List of Chemical Symbols

Symbol	Full name
α-FeOOH	Goethite
Al	Aluminium
$Al(OH)_3$	Aluminium Hydroxide
$Al_2(SO_4)_3$	Alum, Aluminium Sulphate
As	Arsenic
Ca	Calcium
$CaCO_3$	Calcium Carbonate
Cu	Copper
DO	Dissolved Oxygen
FCR	Free chlorine residual
Fe	Iron
Fe_3O_4	Magnetite
$FeCO_3$	Siderite
HA	Humic Acid
Mg	Magnesium
$Mg(OH)_2$	Magnesium Hydroxide
Mn	Manganese
Na	Sodium
Ni	Nickel
P	Phosphorus
Pb	Lead
pH	Hydrogen ion (pH)
Sb	Antimony
THM	Trihalomethane
TOC	Total Organic Carbon

List of Symbols

Symbol	Full name
α_c	decay coefficient
β	regression coefficient
γ_k	impact coefficient for travel paths k
ΔN	change in turbidity
ε	error term in regression model
ϵ	parameter in Epanet-PDX model
η	learning rate in ANN
θ	parameter in Epanet-PDX model
μ	momentum value in ANN
ρ_w	density of water
σ	standard deviation
σ_l	width of the radial centre
τ	pipe boundary shear stress
τ^a	mean daily shear stress in a pipe
τ_{max}	maximum daily shear stress at the node
τ_s	current layer strength
$\bar{\tau}$	maximum daily shear stress at node
$\bar{\tau}^s$	variation of daily shear stress at node
$\varphi(.)$	Gaussian activation function for hidden neuron
ΔE	global energy in Boltzmann machine
$\mu A(x)$	membership function in a fuzzy set A , where x $\in$ universe of discourse
μ_k	a positive integer known as combination coefficient
A_s	pipe surface area
b	bias term in ANN
b’	power term to set for first order relationship in PODDS model
B_c	wall mass coefficient
c	water quality output concentration at a node
C_∞	final steady state concentration of particles in suspension
C_{int}	intercept term in linear equation
c_k	water quality source input
cl	centre of cluster in radial basis function network
C_{Ratio}	crossover ratio parameter
C_s	concentration of particles in suspension
C_T	stored turbidity volume of layer
C_{Tmax}	maximum turbidity potential

Symbol	Full name
C_w	mass of particles attached to the wall per unit weight of water
DF	the difference between ranks of values in a pair
d_p	diameter of pipe
ds	downstream node
E	Euclidean distance in self-organising feature map
e	training error
$E(w)$	global error function that depends on all the weights and biases
$E'(w)$	the gradient of the global error function
E_{cl}	total number of clusters in radial basis function network
Eh	redox
f'	power term in PODDS Model
g	acceleration due to gravity
grad	gradient
h	hidden node in ANN
H	head loss in a pipe
Hn_{nod}	head at node
I_m	identity matrix
itr	Iteration
k_p	gradient of layer strength in PODDS model
L_p	length of pipe
m_s	slope of the line
Mu	training gain parameter used in ANN
Mu_dec	Mu decrease factor
Mu_inc	Mu increase factor
M_w	mass of particles attached to the wall
NC	number of cycles in self-organising feature map
NI	number of input nodes in the ANN
NI	the number of input neurons
N_{itr}	number of iterations
NJ	number of hidden nodes in the ANN
NJ	number of hidden neurons in radial basis function network
NK	number of output nodes in the ANN
NL	number of hidden layers
NP	number of pipes connected to the node
n_s	the number of pairs of values in the sample

Symbol	Full name
NT	number of time intervals
N_{tp}	number of travel paths between a given input and output node
p_1	initial search direction in scaled conjugate gradient algorithm
PCC	percentage customer complaints per number of properties
P_s	gradient term in PODDS Model
Q^{req}	required supply at node
Q_t	flow rate
R	Pearson's correlation coefficient
r^*	neighbourhood radius in in self-organising feature map
r_1	initial steepest descent direction
R_h	hydraulic radius
r_s	Spearman's correlation coefficient
R_t	rate of release of sediment by the excess shear stress
Scale	Parameter in Gaussian mutation that represents variance of mutation during the first generation
Shrink	amount of shrink in the mutation in successive generations
S_o	hydraulic gradient
sp	sample size
T	output time
T_{Bolt}	temperature in Boltzmann machine
t_k	time delay for travel path k
us	upstream node
w	weight vector
X	set of input variables (sample data)
Xobs	observed value
Y	dependent variables

List of Publications

Danso-Amoako, E., & Prasad, T. D. (2013). ANN model to predict the influence of chemical and biological parameters on iron and manganese accumulation. *Proceedings of the 12th International Conference CCWI 2013: "Informatics for water systems and smart cities"*, (pp. 409-418). Perugia, Italy.

Danso-Amoako, E., & Prasad, T. D. (In press). Using fuzzy inference system to predict iron and manganese accumulation potential in water distribution networks. *Proceedings of the 13th International Conference CCWI2015 "Sharing the best practice in the water industry"*. Leicester, UK.

Prasad, T. D., & Danso-Amoako, E. (2013). Influence of chemical and biological parameters on iron and manganese accumulation In water distribution networks. *Proceedings of the 12th International Conference CCWI 2013: "Informatics for water systems and smart cities"*, *70*, pp. 1353-1361. Perugia, Italy.

Prasad, T. D., & Danso-Amoako, E. (2014). Predicting iron and manganese accumulation potential in water distribution networks using artificial neural network. *Proceedings of the 11th International Conference on Hydroinformatics HIC 2014, "Informatics and the environment: Data and Model integration in a heterogeneous hydro world".* New York, USA.

Acknowledgments

I would like to show my deep appreciation to my supervisor, Dr. Prasad Devi Tumula, for his directions, time, encouragement and helpful discussions throughout this research. Without his continual support, I would not have been able to complete this research. I also wish to express my heartfelt gratitude and sincere appreciation to my assistant supervisor, Dr. Saad Yousif, for his encouragement, advice, and suggestions.

Many thanks to United Utilities for funding this research. I would also like to thank all employees of United Utilities who helped in this research, especially Derek Clucas, Anna Provost, Adam Lechmere and Neil Croxton.

I would also like to thank my loving mother, Grace Mante, for her prayers, support and encouragement. Without her help, I would not have come this far in life.

Above all, I would like to thank Almighty God for giving me the strength and wisdom to complete this research. To Him be the praise, glory, and honour forever. Amen!

Declaration

I the undersigned declare that this work is my own origin and has not been produced anywhere. Appropriate due referencing has been acknowledged.

Signature ……  …….

Date ……19/01/2016……...

Abstract

The occurrence of discoloured drinking water at customers' taps, which is mainly caused by the deposition and release of iron (Fe) and manganese (Mn) in water distribution networks (WDNs), is a major concern for both customers and water companies. Increased concentrations of Fe and Mn in WDNs can lead to penalisation by the Drinking Water Inspectorate (DWI) and Water Services Regulation Authority in England and Wales (Ofwat). These high concentration levels can cause aesthetic problems such as giving water an unpleasant metallic taste and staining of laundry. It has also been found that increased Mn concentrations in drinking water can reduce intellectual function of children.

Despite efforts by water companies to comply with standards for drinking water, they continue to receive customer complaints related to water discolouration. Currently, most water companies identify high-discolouration-risk regions in WDNs by either selecting areas in the network with high concentrations of Fe and Mn from their routine sampling, or using data obtained from customer complaints related to discolouration. However, these risk assessment methods are imprecise, because only few selected nodes are sampled and not all customers who experience water discolouration complain. Moreover, considering that the water mains in England and Wales span approximately 315,000 km, monitoring Fe and Mn concentrations will always be a difficult and expensive task. It is therefore imperative for water companies to gain a practical understanding of the processes and mechanisms that lead to water discolouration, and to develop a model to identify the high-risk areas in WDNs so that remedial measures can be effectively implemented.

The factors that influence Fe and Mn accumulation from post-treatment to customers' taps through WDNs can be categorised into physical, chemical and biological. However, to date, researchers have only studied these factors partially or separately, but never in combination. None of the current models are able to predict discolouration/Fe and Mn accumulation potential for every node in WSZs using chemical, biological, and hydraulic/physical variables. This study took a holistic approach in investigating these factors. A five-year data set comprising of 36 water quality, hydraulic, and pipe-related variables covering 176 different district metered areas (DMAs) were analysed to identify relevant variables that influence Fe and Mn accumulation potential. Customer complaint data were also investigated for seasonal trends. Majority of the DMAs (67.44%) showed

significant peaks in customer complaints during summer. These spikes may be attributed to increased water consumption and warmer water temperatures during this period. An artificial neural network (ANN) model was developed using relevant variables identified through the data analysis. The model could predict Fe and Mn accumulation potential values for every node in a given water supply zone (WSZ). From the risk maps generated by the ANN model, it was observed that most of the regions in the network with high Fe and Mn accumulation potential also had high levels of customer complaints related to discolouration. Although the ANN model could predict Fe and Mn accumulation potential failures in WSZs, its black-box nature made it difficult to explain the causes of the failures, unless they were manually investigated.

To overcome the limitation in the ANN model, a fuzzy inference system (FIS) was developed to predict Fe and Mn accumulation potential for every node in WDNs and also capture the chemical, biological and physical processes as water travels through the network. The rules and weights of the rules for the FIS were calibrated using a genetic algorithm. The FIS is also able to determine the causes of the Fe and Mn accumulation potential failures. The ability of the developed models in this research to predict and indicate the causes of high Fe and Mn accumulation potential at the node level make them a unique and practical tool to detect high risk nodes in all regions in WDNs, including regions which have not been sampled. Both models could be of great benefit to water resource engineers and drinking water supply companies in managing water discolouration. They could also be used to investigate variables that influence physical, chemical and biological processes in WDNs.

CHAPTER 1: Introduction

1.1 Overview

Safe drinking water is essential for sustaining human life. An adequate, safe, and accessible supply of water should be available to everyone for both domestic and commercial use. According to the World Health Organization (WHO), safe drinking water is water that does not present any significant risk to human health over a lifetime of consumption, including different sensitivities that may occur during various life stages (WHO, 2006). In addition, safe drinking water should be aesthetically pleasing with respect to appearance, taste, and odour. Furthermore, it should not contain harmful concentrations of chemicals or pathogenic microorganisms (Australian National Health and Medical Research Council & Australian National Resource Management Ministerial Council [ANHMRC & ANRMMC], 2004). Although access to safe drinking water is considered a basic human right, more than one-sixth of the world's population lack reliable access to such water, with this problem being predominant in developing countries (WHO & UNICEF, 2006).

Although drinking water in developed countries is relatively safe, there are a number of issues that need to be addressed. Two main types of contaminant are considered to make drinking water unsafe. The first type, which is known as 'primary contaminants', include contaminants such as lead, copper (Cu), and nitrates which may have adverse health effects. The second type, known as 'secondary contaminants', include contaminants such as iron (Fe), manganese (Mn), and aluminium (Al) which can cause drinking water aesthetic problems such as unpleasant odour, taste, water discolouration, and staining of laundry. Increased Fe and Mn in drinking water, which is the main cause of discolouration, have long been considered to be an aesthetic problem. However, research by Wasserman et al. (2006) on increased Mn in drinking water indicated it can cause more than the traditional aesthetic issues. Increased concentrations of Fe and Mn can also lead to compliance failures, customer complaints, and loss of customers' confidence in drinking water companies. Unlike discolouration, compliance failures may not be visible to the eyes, but are assessed using analytical methods.

Several researchers have tried to develop models to predict discolouration in water distribution networks (WDNs). However, most of these models use only hydraulic or physical variables that influence discolouration in making predictions. Current drinking water discolouration risk models found in literature include the use of chemical, biological, or hydraulic/physical variables in making predictions; but not used in combination. As a result, these models are unable to capture all of the processes and mechanisms that influence discolouration/Fe and Mn accumulation. There is therefore the need for water companies to have a model that will use all the relevant variables to identify high-risk regions in WDNs, indicate the causes of failures in these regions, and if possible, find solutions to these problems. In this research, a comprehensive study on the processes that influence the accumulation of Fe and Mn particles in WDNs was conducted. Using relevant chemical, biological, and hydraulic/physical variables, models were developed to help drinking water companies to predict areas in WDNs with high-risk of Fe and Mn accumulation potential.

1.2 General problem statement

Fe is a naturally occurring element that is found in certain rocks and soils, and it constitutes approximately 5% by weight of the earth's crust. It is the fifth most abundant element in the earth's crust (Gschneidner, 1996). Unsurprisingly, a study conducted by Boxall, Skipworth and Saul (2003) on flushing samples collected in the UK identified Fe and Mn as the first and second most common water contaminants, respectively, irrespective of the pipe material in WDNs. A related study by Slaats (2002) showed that gradual accumulation of Fe and Mn particles in WDNs is the most common cause of water discolouration. Although elements such as silicon, calcium (Ca), Al, and Cu as well as organic compounds, can also cause water discolouration, they are not as prevalent as Fe and Mn (Teasdale, O'Halloran, Doolan, & Hamilton, 2007).

1.2.1 Health and aesthetic problems

Fe and Mn in drinking water have long been considered to cause only aesthetic problems; that is, they are secondary contaminants that have little or no adverse health effects. In fact, low concentrations of Mn and Fe are known to be essential for human health (Swistock, Sharpe, & Robillard, 2001). Although only low concentrations of Fe and Mn enter WDNs

after the treatment process, years of accumulation in distribution systems, as well as periodic re-release in significant quantities, and other adsorbed compounds associated with the deposits can result in more than the traditional aesthetic issues. For example, Wasserman et al. (2006) investigated the relationship between increased concentrations of Mn in drinking water and reduced intellectual functions of children.

High concentrations of Mn and Fe in WDNs can also give water an unpleasant medicinal or metallic taste (Swistock et al., 2001). Studies have attributed red-brown, yellow, yellow-brown, and brown colours of drinking water to corrosion of Fe (Yarra Valley Water, 1998). Black water has been attributed to excess concentrations of Mn and biofilms stripping (Sly, Hodgkinson, & Arunpairojana, 1990; Yarra Valley Water, 1998). This discoloured water could also lead to similarly coloured stains on laundry and porcelain, thereby prompting numerous customer complaints. Vegetables cooked with Fe-contaminated water become dark and look unappetizing, and Fe or Mn bacteria can cause black-brown slimy masses inside toilet tanks (Herman, 1996).

In response to these known issues, water companies have adopted expensive and sophisticated risk-based management systems for monitoring discolouration in WDNs. However, despite their efforts to comply with drinking water standards, they continue to receive customer complaints related to water quality. In this research, a customer complaint is defined as a record of a customer complaining directly to the water company regarding incidents such as discolouration, metallic taste, or slime. These complaints significantly undermine customers' confidence in water companies. An analysis of customer complaints related to the quality of water supplied by a UK water company over a five-year period showed that 34% and 7% of the complaints are related to discolouration and other aesthetic problems, respectively (Cook, Boxall, Hall, & Styan, 2005). Customers evaluate water quality by taste, sight, and smell. However, most substances that can be evaluated by human senses are secondary contaminants, and are often harmless (Department of Human Services & Department of Natural Resources and Environment, 2000). In fact, some of the highest health risks of water are attributed to substances that cannot be perceived by human senses (for example, bacteria and dissolved organic compounds). Although customer complaints are a good indicator of water quality, using them alone can be misleading, as not all customers complain. Nevertheless, they can be

very useful for predicting discolouration/Fe and Mn accumulation potential in WDNs when used in combination with other chemical, biological, and physical variables.

1.2.2 Compliance problems

High concentrations of Fe and Mn in WDNs can lead to compliance failures. The DWI has set the maximum concentration levels (MCLs) of Fe and Mn in drinking water to 200 and 50 µg/L, respectively. In general, water companies set post-treatment targets of Fe and Mn to approximately 3% of their respective MCLs. They do so to reduce the concentrations of Fe and Mn entering WDNs, thereby leading to reduced deposition. However, irrespective of how effective water is treated, very low concentrations of Fe and Mn may still enter the network from water treatment plants and gradually accumulate on the inner surface of pipe walls within WDNs. During hydraulic events, such as high flows created by bursts in water mains or high diurnal consumption of drinking water, these accumulated particles may be dislodged from the pipe walls, cause discolouration, and subsequently end up at customers' taps.

1.2.3 Financial losses

In April 2010, the Water Services Regulation Authority in England and Wales (Ofwat) introduced the Service Incentive Mechanism (SIM). This mechanism rates the performance of water companies based on customer satisfaction, and either rewards or penalises them. In view of this, it has become extremely important for water companies to reduce the number of customer complaints caused by drinking water discolouration. Water companies also receive fines from the DWI if the concentrations of Fe and Mn exceed their respective MCLs.

The deposits of Fe and Mn in WDNs can clog pipelines and decrease water pressure, thereby requiring more energy to pump water through the network. Furthermore, these deposits can increase pumping and rehabilitation costs (Vreeburg & Boxall, 2007). Moreover, the corrosion of iron pipes is an important chemical process in water discolouration. For this reason, several water companies have spent substantial amount of money replacing iron pipes with polyvinyl chloride (PVC) pipes, with the aim of decreasing discolouration in WDNs. However, customers still experience some discolouration in areas that are entirely networked with PVC pipes, although they do not corrode over time as they do not react with air and water (Vreeburg, 2007). A study by

Cerrato, Reyes, Alvarado and Dietrich (2006) indicated that this observation could be attributed to the deposition dynamics in PVC pipes. They observed that the Mn deposits on the walls of PVC pipes were loose because of their exceptionally smooth walls; as a result, were subjected to sloughing and discolouration under smaller shear forces than in iron pipes.

In a related study, Cook (2007) investigated plastic, asbestos cement, cast iron, epoxy lined, and cement and bitumen lined pipes, and observed no correlation between customer complaints related to water discolouration and these pipe types. Boxall et al. (2003) reported that, irrespective of the pipes used in WDNs, Fe and Mn were the first and second most common water contaminants, respectively. This result indicates that there are other factors in addition to pipe material that cause Fe and Mn particles to accumulate in WDNs.

1.2.4 Modelling difficulties

The processes influencing the accumulation and release of Fe and Mn in WDNs are highly complex, unpredictable, not fully understood, and difficult to model mathematically. The concentrations of Fe and Mn frequently change with time and space as water moves from the treatment plant to customers. The variability of source materials, hydraulics, biological and chemical reactions that occur within a network contribute towards creating a very complex environment that is difficult to understand.

Moreover, increased treatment costs, increased pumping and rehabilitation costs, fines, and sophisticated risk-based management systems are costing water companies significant amount of money. There is therefore an urgent need for water companies to not only gain a practical understanding of the processes and mechanisms that lead to compliance failures and discolouration, but also devise a comprehensive strategy to deal with such events. Water companies worldwide are urgently looking for solutions to prevent the above-described problems. Furthermore, there is also a strong need for a model that can predict the risk of Fe and Mn accumulation potential based on not only physical and hydraulic variables, but in combination with chemical variables and variables that influence biological processes.

1.3 Knowledge gaps

Water companies have to deal with two main problems: regulatory compliance failures and discolouration events, both of which can lead to penalisation and loss of confidence by customers, as well as concerns regarding potential health impacts. While discolouration can be detected by human eyes and prompts customers to complain, Fe and Mn compliance failures are assessed using analytical methods, because they cannot be detected by sight. Currently, many drinking water companies identify regions with high-risk of discolouration/Fe and Mn failures by either selecting areas in the network with high Fe and Mn concentrations from their random sampling, or by using customer complaints data due to water discolouration. These methods can be ineffective for two reasons. First, the large sizes of water supply zones (WSZs) make it difficult and expensive to monitor Fe and Mn concentrations. With about 315,000 km of water mains across England and Wales, it will be impossible to sample every node in large WSZs. Hence, regions which have high Fe and Mn concentrations that are not sampled will not be detected. Secondly, studies in the United Kingdom have shown that only 30% of customers that experienced discoloured water actually complained (Ewan & Williams, 1986). Similarly, a study conducted by Roseth and Rock (2003) in Melbourne, Australia, indicated that only 15% of customers who experienced water discolouration complained. These studies show that certain regions in WSZs with high discolouration risk/Fe and Mn accumulation potential will go undetected because complaints are not reported. Using Fe and Mn concentrations and customer complaints to identify regions with high-risk of discolouration or failures of Fe and Mn is desirable. However, a model that can predict the risk of Fe and Mn accumulation potential in every region in WDNs and indicate the causes of this risk will be more beneficial.

Hydraulic distance from source of water supply is a very important variable that influences Fe and Mn accumulation which has not been investigated thus far. It is the distance taken for water to travel from a source of water supply to a given node within a WDN. In general, the further water travels through WDNs, the more chlorine dissipates. This increases microbial growth, which increases biological oxidation of Fe and Mn, and subsequently increases Fe and Mn accumulation potential.

A number of discolouration risk assessment tools have been developed by researchers. They include the Particle Sediment Model (PSM) (Wu et al., 2003), Discolouration Risk Analysis Tool (DRAT) (Boxall & Husband, 2005), Resuspension Potential Method (RPM) (Vreeburg, Schaap, & van Dijk, 2004a), Discolouration Risk Modelling approach (DRM) (Dewis & Randall-Smith, 2005), Discolouration Propensity Model (DPM) (McClymont et al., 2011), and Pressure-dependent Analysis (PDA) model (Seyoum & Tanyimboh, 2014). Most of these models deal exclusively with the risk of discolouration based on physical/hydraulic variables such as water velocity, turbidity, shear stress, water age, turbopherosis, and pipe material. Very little research has been conducted on Fe and Mn accumulation on the inner surface of pipe walls. More importantly, none of the developed models are able to predict discolouration/Fe and Mn accumulation potential for every node in WSZs using chemical, biological, and hydraulic/physical variables.

Fe and Mn may be present in different complex species and compounds as well as in solution or particulate forms. They may also be loosely present within the water mains, or adhered to the inner surface of pipe walls. In this research, Fe and Mn were studied from the point of view of accumulation rather than from discolouration. The accumulation of Fe and Mn on the inner surface of pipe walls is influenced by the following factors:

(a) chemical reactions
(b) microbiological activity, and
(c) physical or hydraulic processes within the network.

Studies in this area have thus far mainly focused on the following issues:

1. Discolouration, with little research carried out to understand what factors affect accumulation and compliance problems.
2. The above mentioned factors have been investigated by researchers independently, rather than in combination.

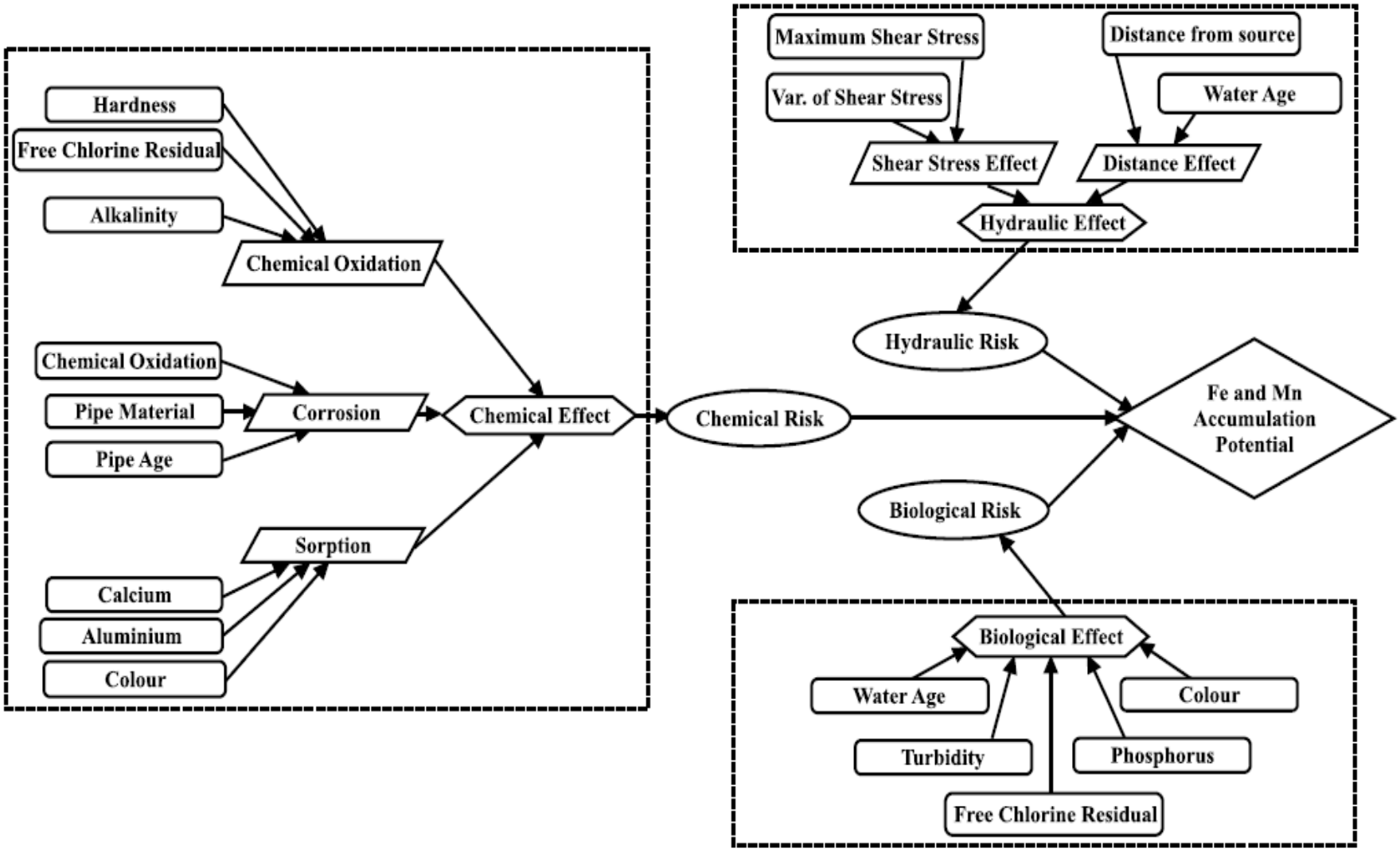


Figure 1.1 The Fe and Mn accumulation potential model

To date, researchers have only studied these factors partially or separately, but never in combination. Clearly, there is a limitation in their attempt to unravel the complex process of accumulation. In this research, the focus was on studying Fe and Mn accumulation potential holistically rather than independently. However, in order to study the factors that influence Fe and Mn accumulation potential holistically and estimate the combined effect on deposition dynamics, it is imperative to understand their influence both individually and in association. Here, the challenge is to correlate Fe and Mn accumulation potential with the relevant chemical, biological, and physical/hydraulic variables. A diagram of the developed model showing how all the variables were correlated with Fe and Mn accumulation potential is presented in Fig. 1.1. Chapters 2, 4, 5, and 6 discuss the effect of each of the variables on the accumulation of Fe and Mn in detail.

As mentioned above, this study took a holistic approach in investigating the factors that influence the accumulation of Fe and Mn particles from post-treatment to customers' taps through WDNs. This approach is important because continuous deposition due to these factors will lead to compliance failures, and eventually result in discolouration during hydraulic events such as opening of fire hydrants during fire extinguishing exercises and increased flow caused by increased water consumption. Mitigating the former problem will clearly mitigate the latter, as the two problems are directly linked. The ability of the developed models in this research to predict and indicate the causes of high Fe and Mn accumulation potential at the node level make them a unique and practical tool to guide drinking water companies in managing discolouration. These models can help in the maintenance of water mains, the development of cleaning protocols, and the development of operational and management strategies for water distribution at the national and international levels.

1.4 Aim and objectives

The main aim of this study is to develop cost-effective models to predict Fe and Mn accumulation potential using relevant chemical, biological, and hydraulic/physical factors that aid accumulation process in WDNs. These models will take a holistic approach to correlate Fe and Mn accumulation potential with relevant variables such as aluminium, alkalinity, free chlorine residual, pipe material, and maximum daily shear stress at nodes. The specific objectives are as follows:

1 To prepare a comprehensive literature review to find gaps in knowledge and identify variables other researchers have used that potentially influence Fe and Mn accumulation in WDNs.
2 To develop a model to extract/compute required physical/hydraulic variables such as hydraulic distance from source of water supply, maximum daily shear stress at node, water age, pipe age, and pipe material.
3 To analyse the hydraulic/physical and post treatment water quality data with the objective of identifying relevant variables that influence Fe and Mn accumulation in WDNs. Each of the variables will be studied in depth and as a complete system in order to understand the processes that lead to the accumulation of Fe and Mn in WDNs.
4 To develop risk assessment models using artificial intelligence techniques to predict Fe and Mn accumulation potential in WDNs. The developed risk assessment tools should be able to generate risk maps to help water resource engineers and drinking water companies to identify high-risk regions in WDNs.
5 To develop risk assessment models that would be able to indicate the causes of high-risk of Fe and Mn accumulation potential in order for water companies to find possible solutions to reduce it, since a reduction in Fe and Mn accumulation potential will reduce discolouration as the two are directly linked.

1.5 Thesis organisation

Chapter 1 presents the general overview of the research topic, general problems statement, knowledge gaps, and the aim and objectives of this research.

Chapter 2 presents a comprehensive literature review that identifies relevant variables that influence Fe and Mn accumulation in WDNs. In this chapter, the chemical, biological and hydraulic/physical processes that lead to Fe and Mn accumulation were discussed, and different models for predicting drinking water discolouration in WDNs were reviewed.

Chapter 3 presents a critical review on artificial intelligence based methods of modelling. This chapter reviews some applications of artificial neural network (ANN) models and fuzzy inference systems (FISs) in water resources.

Chapter 4 presents how a five-year customer complaint data were collated to identify suitable WSZs with low, medium and high customer complaints for analysis. The customer complaints data were investigated for seasonal trends. Also, a five-year post-treated water quality data set from the selected WSZs were analysed to identify relevant variables that influence Fe and Mn accumulation. The EPANET software was extended to extract/compute relevant hydraulic/physical variables. Each of the variables was studied in depth and as a complete system to understand the processes that lead to the accumulation of Fe and Mn in WDNs.

Chapter 5 shows how the ANN models were developed using the identified relevant variables to predict Fe and Mn accumulation potential in WDNs. The risk maps generated by the models were compared with maps of customer complaints due to water discolouration to investigate whether there were any correlations.

Chapter 6 presents two FISs developed using relevant variables that influence Fe and Mn accumulation potential in WDNs. The first FIS developed, the hierarchical rule-based expert FIS, uses expert knowledge to formulate rules and assigned weights to them in making its predictions. While the second FIS, the hierarchical data-driven FIS, uses genetic algorithm to optimise the rules and weights of the rules in making its predictions. The developed FISs are able to indicate the causes of high-risk of Fe and Mn accumulation potential.

Finally, the conclusion from this research and recommendations for future work are presented in Chapter 7.

CHAPTER 2: Literature Review

2.1 Introduction

The MCLs of Fe and Mn set by the DWI were discussed in Chapter 1. The need for drinking water companies to understand the processes and mechanisms that lead to Fe and Mn accumulation will also be discussed. The variables that influence accumulation can be grouped into three categories: (a) chemical variables such as alkalinity and chlorine that represents the chemical reactions within a WDN; (b) variables that influence biological processes such as phosphorus (P) and organic carbon that aid the accumulation; and (c) physical variables such as the age of pipes, shear stress, and water age. Because a significant percentage of customer complaints are related to water discolouration, more studies need to be conducted to tackle this problem. However, as mentioned in Chapter 1, water companies are striving for a better and practical understanding of the processes and mechanisms that lead to accumulation of Fe and Mn in WDNs, which thus far, have only attracted limited research.

In this chapter, a comprehensive literature review would be conducted to identify relevant variables that influence Fe and Mn accumulation from post-treatment, through the WDNs to customers' taps. The chemical, biological, and physical processes that lead to Fe and Mn accumulation would also be discussed. The following sections present a critical review of published literature in this field. Section 2.2 discusses studies on sediment accumulation in WDNs. A review of four main theories on the formation of discoloured water is presented in Section 2.3. Section 2.4 focuses on discolouration risk models. Finally, the summary of this chapter is presented in Section 2.5

2.2 Factors that influence sediment accumulation in water distribution networks

The accumulation of sediments in WDNs can be attributed to several factors, most of which are interrelated. These include complex physical, chemical and/or biological processes. Linking the particles found in a WDN to a particular discolouration source can be a very difficult task because of the numerous potential sources and the complex layout

of pipe networks (Gauthier, Gérard, Portal, Block, & Gatel, 1999). In an attempt to link particles of iron oxide deposits found in several PVC pipes, Gauthier et al. (1996) suggested they are likely to have travelled a considerable distance through the pipe network because they could not have originated in the PVC pipes themselves. The following sections review the factors that influence sediment accumulation in WDNs.

2.2.1 Influence of chemical variables on sediment accumulation

2.2.1.1 Iron

Fe can exist in aquatic systems (natural waters and their sediments) in several oxidation states: metallic iron (Fe), ferrous iron (Fe^{2+}), and ferric iron (Fe^{3+}). The oxidation state in which Fe exists in a particular aquatic system and the redox reactions (chemical oxidation-reduction reactions) in which it participates depend on the presence or absence of oxidising agents such as dissolved oxygen (DO) and chlorine. Fe can also be oxidised by some microorganisms (Sly, Hodgkinson, & Arunpairojana, 1988). Fe^{3+} is stable in oxygenated water but is usually insoluble in the particle and colloidal forms. On the other hand, Fe^{2+} is thermodynamically unstable in oxygenated water but is generally soluble (Teasdale et al., 2007). The source of Fe in drinking water is either from the ferrous pipes in WDNs or from the source of water supply after treatment.

2.2.1.2 Manganese

The chemistry of Mn is complex. It exists in several species with different oxidation states (Kohl & Medlar, 2006). Mn causes household problems only when it occurs in its particulate or oxidised form. In its soluble form, Mn is not visible to the human eye (United States Environmental Protection Agency [USEPA], 1994). Chemically, Mn occurs in all oxidation states between 0 and +7, with +2, +4, and +7 being more environmentally and biologically important (USEPA, 1994). Mn salts in the +2 and +7 states are chemically the most stable. Mn^{2+} and Mn^{7+} are soluble, whereas Mn^{4+} is insoluble (oxidised form, manganic dioxide). Mn is most stable in its +2 oxidation state, hence, most naturally occurring Mn is in the form of dissolved Mn^{2+} (American Water Works Association, 1999; USEPA, 2009). The next most common species is the particulate state; Mn^{4+}. At concentrations as low as 0.02 mg/L, Mn^{2+} compounds in solution form undergo oxidation in the presence of chlorine, DO, or bacteria to form black precipitates that get encrusted on pipe walls in WDNs (Bean, 1974). Mn has complicated redox kinetics, hence it is very difficult to chemically oxidise in pH environments typical of raw water (pH 6–8). It often

persists in soluble forms despite unfavourable thermodynamics. It has been reported that redox (Eh) and pH conditions do not completely explain the Mn distribution in groundwater (Homoncik, MacDonald, Heal, Ó Dochartaigh, & Ngwenya, 2010). Mn can also exist in the +3 state; however, this state is very unstable and usually reverts to the +2 state. Mn compounds in +5 states are not very common (American Water Works Association, 1999; Kohl & Medlar, 2006; USEPA, 2009).

2.2.1.3 Aluminium

Aluminium (Al) is the most abundant metallic element on the earth, comprising approximately 8% of the earth's crust (Gschneidner, 1996). Al salts such as alum are often used as coagulants during water treatment to reduce organic matter, colour, turbidity, and microorganism levels (WHO, 2006). Al is insoluble in water under neutral conditions (pH 6–9), except when it is in a complex organic form (Molot & Dillon, 2003). In the solution form, Al can exist as either an inorganic (Al^{3+}, $Al(OH)^{2+}$, AlF^{2+}) or complex dissolved organic carbon compound, with the latter being dominant in WDNs (Schintu, Meloni, & Contu, 2000). Increase in Al causes the formation of amorphous $Al(OH)_3$, which adsorbs Mn particles in WDNs (Wang et al., 2012).

2.2.1.4 Copper

Copper (Cu) can also cause water discolouration. Excessive concentration of Cu in drinking water can cause green or blue stains on household fittings. Cu mainly enters WDNs as a result of Cu salts which are used in reservoirs for algae control. They also enter WDNs from corrosion of Cu pipes (Cruse, 1971). Cu usually exists in the +1 and +2 oxidation states in solution form. Cu^{+} and Cu^{2+} species may form mineral precipitates such as carbonates, hydroxides, oxides, and phosphates. High pH usually limits the solubility of these species (American Water Works Association, 1999).

2.2.1.5 Organic matter

A study conducted by Gauthier, Barbeau, Milette, Block and Prevost (2001) showed that organic matter constitutes 40–76% of the total suspended solids in WDNs. In general, because such solids are of low density, they can be transported over large distances in WDNs if they are not deposited on the pipe walls.

2.2.1.6 pH of water

pH is a numeric scale ranged between 0 and 14 used to specify the acidity or alkalinity of an aqueous solution. Lower values of pH are more acidic, while higher values are more alkaline. Generally, increase in pH from 7 to 9 in WDNs has been found to increase pipe weight loss and corrosion rate (Stumm, 1960). However, the release of corrosion by-products decreases at higher pH (Hidmi, Gladwell, & Edwards, 1994). A contrasting study by Kashinkunti, Metz, DeMarco and Hartmann (1999) indicated that pipe weight loss and iron concentration decrease as pH increases from 8.5 to 9.2.

2.2.1.7 Alkalinity

Alkalinity is the ability of a solution to neutralize acids to the equivalence point of carbonate or bicarbonate. An increase in alkalinity helps to increase buffer capacity, thus keeping the pH of drinking water stable. Increase in alkalinity of water generally reduces pipe weight loss and corrosion rates (Kashinkunti et al., 1999). A study conducted by Kashinkunti et al. (1999) showed that customer complaints due to water discolouration were reduced when the alkalinity was maintained at 60 mg $CaCO_3$/L. In a related study, Naylor, Nicholas, Murry and Roddy (1993) investigated the effects of alkalinity and pH on the corrosivity of water and found that when alkalinity was higher than 50 mg $CaCO_3$/L, corrosion reduced within pH range of 7.5–8. Research by Gray (1994) also indicated that soft waters with alkalinity less than 50 mg $CaCO_3$/L are more likely to cause corrosion. A comprehensive research by Benjamin, Sontheimer and Leroy (1996) on the corrosion of iron and steel pipes and iron scale formation showed that in a low alkaline environment, the iron scales formed were thick (~2–3 cm), loose, and dark orange-brown in colour. These scales could easily be cracked or scraped off. On the other hand, the scales formed in a high alkaline environment were thin (≤1 mm), fairly uniform, hard, and tightly bound to the metal surface.

2.2.1.8 Dissolved oxygen

DO refers to the oxygen present in water. Corrosion rates generally increase with increasing DO concentration (Gedge, 1992). A study by Seo, Jung, Lee and Gee (1998) showed that the deterioration of drinking water is mainly caused by the corrosion of pipes in WDNs, and that DO concentration is the main factor that caused increased corrosion. During corrosion, DO serves as an electron acceptor (see Eqn. 2.1), and it oxidises ferrous iron (Fe^{2+}) (see Eqn. 2.2) or iron scales (see Eqns. 2.3 & 2.4), (McNeill, 2000).

$$Fe + 0.5O_2 + H_2O \rightarrow Fe^{2+} + 2OH^- \quad (2.1)$$

$$Fe^{2+} + 0.25O_2 + 0.5H_2O + 2OH^- \rightarrow Fe(OH)_3 \quad (2.2)$$

$$3FeCO_3 + 0.5O_2 \rightarrow Fe_3O_4 + 3CO_2 \quad (2.3)$$

$$4Fe_3O_4 + O_2 \rightarrow 6Fe_2O_3 \quad (2.4)$$

2.2.2 Influence of chemical processes on sediment accumulation

Raw water comprises of natural organic matter and various elements in different states, such as soluble ions, colloids, and particulates, which can contribute to water discolouration in WDNs if they are not removed during the treatment process. Coagulant chemicals such as aluminium sulphate (alum), ferric sulphate and ferric chloride can pass some residual amounts of contaminants into the distribution system (Teasdale et al., 2007). The changes in water chemistry caused by the treatment process and constant interaction of chemical variables such as DO, pH, alkalinity, and chlorine can influence Fe and Mn accumulation within pipes. There are a number of chemical processes enumerated by various researchers that aid Fe and Mn accumulation in pipes. Out of these, three major chemical processes, namely, chemical oxidation, corrosion, and sorption contribute significantly to the accumulation process. The following sections describe the chemical processes that influence the accumulation of sediments in WDNs.

2.2.2.1 Chemical oxidation

As water travels through WDNs, soluble Fe and Mn in the WDNs from the source of water supply undergo chemical oxidation. Chemical oxidation of Fe and Mn occurs when soluble Fe^{2+} and Mn^{2+} are converted to insoluble Fe^{3+} and Mn^{4+}, respectively, in the presence of an oxidising agent (Sly et al., 1990; Teasdale et al., 2007). Wallace and Campbell (1991) listed some oxidising agents that can oxidise Fe and Mn in order of effectiveness (from weakest to strongest) as: hypochlorite ion, chlorine dioxide, chlorine, hypochlorous acid, permanganate ion, hydrogen peroxide, ozone, and hydroxyl free radical.

Water utilities often add gaseous chlorine (Cl_2), chlorine oxidise, or hypochlorite to drinking water to protect it from harmful organisms or pathogens. Chlorination is the most commonly used method of disinfecting drinking water. Cl_2 undergo hydrolysis in drinking water by the reaction in Eqn. 2.5a. The hypochlorous acid (*HOCl*) formed from Eqn. 2.5a

is a weak acid which subsequently dissociates aqueous solution by the reaction in Eqn. 2.5b. Chlorine mainly exists at low, medium, and high pH as Cl_2, *HOCl*, and ClO^- (hypochlorite), respectively (Deborde & Von Gunten, 2008).

$$Cl_2 + H_2O \rightleftharpoons HOCl + Cl^- + H^+ \quad (2.5a)$$

$$HOCl \rightleftharpoons ClO^- + H^+ \quad (2.5b)$$

In general, chlorine oxidises faster and over a wider range of pH with Fe than with Mn (Odell, Cyr, & Prather, 1998). Chlorine dioxide is sometimes also used by water companies as a disinfectant because it is effective in the reduction of trihalomethanes (Wallace & Campbell, 1991). Equations 2.6a and 2.6b show how chlorine dioxide oxidises Fe^{2+} and Mn^{2+} to insoluble Fe^{3+} and Mn^{4+}, respectively. Usually, the reaction takes place within two to three seconds (Knocke et al., 1990). A detailed review on disinfection of drinking water is presented in Section 2.2.3.3.

$$Fe^{2+} + ClO_2 + 3H_2O \rightarrow Fe(OH)_3 + ClO_2^- + 3H^+ \quad (2.6a)$$

$$Mn^{2+} + 2ClO_2 + 2H_2O \rightarrow MnO_2 + ClO_2^- + 4H^+ \quad (2.6b)$$

2.2.2.2 Corrosion processes

Corrosion is a natural process that cannot be prevented but can be controlled. Corrosion of cast iron pipes is the most common cause of drinking water discolouration (DWI, 2007). It causes three main problems: (1) pipe mass is lost in the form of iron-bearing scales or soluble iron, (2) accumulated scales in pipes decrease the water capacity and increase the head loss, and (3) the release of soluble or particulate iron causes water discolouration and other aesthetic problems (McNeill, 2000).

For corrosion to occur, an anode, cathode, electrolyte, and metallic path are required. Oxidation and reduction reactions occur at the anode and cathode, respectively. Corrosion primarily occurs on the pipe wall with the anodic release of ferrous iron from iron metal (see Eqn. 2.7).

$$Fe \rightarrow Fe^{2+} + 2e^- \quad (2.7)$$

If the water has a higher pH, then the anodic reaction produces a surface film of ferric oxide (see Eqn. 2.8).

$$2Fe + 3H_2O \rightarrow Fe_2O_3 + 6H^+ + 6e^- \quad (2.8)$$

If DO is present in the system, reduction occurs at the cathode (see Eqn. 2.9). The hydroxyl ions cause the pH to increase, thus increasing the corrosion rate.

$$O_2 + 2H_2O + 4e^- \rightarrow 4OH^- \quad (2.9)$$

2.2.2.3 Formation of scales

Fe^{2+} is usually soluble in drinking water; however, it can also form small amounts of siderite ($FeCO_3$), which is deposited on pipe walls (Peng et al., 2010). Fe^{3+} forms insoluble oxides and hydroxides such as goethite (α-FeOOH), magnetite (Fe_3O_4), and hematite (α-Fe_2O_3) that can also precipitate and deposit on pipe walls (Gerke, Maynard, Schock, & Lytle, 2008; Sarin et al., 2004).

2.2.2.4 Influence of sorption variables on sediment accumulation

Sorption is a physical and chemical process by which adsorption and absorption take place simultaneously. A research study conducted by (Wang et al., 2012) on the adsorption of Mn^{2+} with amorphous $Al(OH)_3$ showed that adsorption mainly took place when the pH of drinking water was above 7.5. They observed that adsorption of Mn^{2+} with amorphous $Al(OH)_3$ was enhanced by co-existing of high concentrations of cations such as Ca^{2+} and Mg^{2+} due to the effects of co-precipitation contributed by newly formed $CaCO_3$ and $Mg(OH)_2$ on other solids. They also found that dissolved organic matter, especially humic acid (HA), enhanced adsorption of Mn^{2+}. They observed that the adsorptive capacity of Mn^{2+} with amorphous $Al(OH)_3$ were enhanced by the following co-existing substances listed in order of strength, from strong to weak as: HA, Mg^{2+}, and Ca^{2+}.

2.2.2.5 Pipe age

The age of pipes in WDNs has a significant effect on corrosion. The accumulation of corrosion by-products and suspended particles over years can reduce pipe diameter, increase roughness, and cause water discolouration. A contrasting research by (McNeill,

2000) however showed that corrosion rates are higher in newer pipes but soon stabilises as scales build up on the pipe walls.

2.2.3 Influence of biological processes on sediment accumulation

2.2.3.1 Biofilms

The biological processes that lead to water discolouration are very complicated. Microbial growth may lead to the formation of biofilms in WDNs. Biofilms are microorganisms that get attached to pipe walls and then multiply to form slime layers. Every WDN is susceptible to microbial growth and the resultant formation of biofilms, irrespective of the purity of water, type of pipe material, or presence of a disinfectant (National Research Council, 2005). Several definitions for biofilms have been published. USEPA (2002) defines biofilms as a complex mixture of microbes and organic and inorganic material accumulated amidst a microbial-produced organic polymer matrix attached to the pipe wall. Decho (2000) defines biofilms as aggregates of microorganisms such as mixed populations of bacteria, fungi, protozoa, algae, and higher organisms in the food chain such as nematodes and larvae. Biofilms that have Fe- and Mn-oxidising bacteria may contain high concentrations of inorganic content such as sediments, scales, and corrosion deposits (Cooperative Research Centre for Water Quality and Treatment [CRCWQT], 2005).

Biofilms can cause many problems. They can damage industrial equipment such as heat exchangers and cooling towers, and this can lead to inefficient energy transfer, energy loss, increased fluid friction, and corrosion (Xiong & Liu, 2010). Bacterial growth may contribute to pipe corrosion, increased demand for disinfectants, and nitrification reactions. They may also cause aesthetic problems such as giving water an unpleasant taste and odour (Servais, Laurent, & Randon, 1995; USEPA, 2009). Studies have shown that Fe and Mn deposition increases with microbial activity. As a result of cell death and flow dynamics, biofilms may release entrapped Fe and Mn into the bulk flow (Deines et al., 2010; Ginige, Wylie, & Plumb, 2011). The demand of biofilms for oxygen means that they may release, for example, Fe^{2+} during periods of extended anoxic conditions (Sarin et al., 2004).

Generally, increased flow velocity exhibits a negative correlation with the biofilms attached to a pipe wall (Donlan & Pipes, 1986). However, studies conducted by Becker (1998) showed that biofilms formed under higher flow velocities are often thinner but

firmer. Turbulent flow may cause shearing of biofilms from the pipe wall, causing bacteria to enter the water flow. Studies have also shown that suspended bacteria found in WDNs are produced by the detachment of biofilms from pipe walls, and not by the growth of organisms (Haudidier et al., 1988; van der Wende, Characklis, & Smith, 1989). A study conducted by Sly et al. (1998) showed that water velocity strongly influenced the nature of the biofilm in the early stages of microorganism development. They found that biofilms formed at higher velocities are more likely to accumulate Fe and Mn particles, and therefore, such biofilms are more likely to cause water discolouration.

Over 90% of the biomass in WDNs is present as biofilms on pipe walls (Deines et al., 2010; Flemming, 1998). The presence of organic carbon in water or on the pipe wall enhances biofilm production (van der Kooij, 2002). Biofilms have been found to enhance the accumulation of Fe and Mn particles, as well as calcium carbonate and other inorganic debris from the bulk flow (Geldreich, 1996). Furthermore, some microbes may oxidise Fe and Mn and increase their retention time in the network, whereas others may enhance the abiotic release of Fe from corrosion scales in the pipe (Cerrato et al., 2010). Recently, Ginige et al. (2011) showed that seasonal influence may affect biofilm production: production increased during the summer and autumn, whereas, during cooler periods, dead cells detached from pipe walls, together with the flow dynamics, increased water discolouration through the release of Fe and Mn particles into the bulk flow.

2.2.3.2 Biological oxidation of iron and manganese

Biofilms in WDNs contain a variety of microorganisms; however, only a few that oxidise Mn and Fe contribute to water discolouration. Such bacterial species have been identified by Sly et al. (1998) and LeChevallier, Babcock, & Lee (1987), and are listed in Table 2.1. Certain microorganisms such as *Crenothrix*, *Flavobactium*, *Pseudomonas*, *Leptothrix discophora*, and *Clonothrix* have been found to oxidise soluble Fe^{2+} and Mn^{2+} to insoluble Fe^{3+} and Mn^{4+}, respectively, thereby increasing the deposition in WDNs (LeChevallier et al., 1987; Sly et al., 1988; Vigliotta et al., 2007). *Bacillus spp*. have been found to reduce Mn to +4 or +2 state, whereas *Clostridium sp*., *Escherichia coli*, and *Enterobacter aerogenes* reduce insoluble Fe^{3+} to soluble Fe^{2+} (Cerrato et al., 2006; Emde, Smith, & Facey, 1992).

Table 2.1 Bacterial species found in WDNs that oxidise manganese and/or iron

Bacterial Species	Manganese	Iron
Arthrobacter[1]	Yes	No
Bacillus[1]	Yes	No
Enterobacter[1]	No	Yes
Flavobacterium[1]	Yes	Yes
Hyphomicrobium[2]	Yes	Yes
Metallogenium[1]	Yes	No
Micrococcus[1]	Yes	No
Pedomicrobium[2]	Yes	Yes
Pseudomonas[1]	Yes	Yes

[1] LeChevallier et al. (1987), [2] Sly et al. (1988)

2.2.3.3 Factors that influence biofilm formation

The factors that influence biofilm formation differ slightly in every network. The spatiotemporal non-uniformity of biofilms and several interrelated factors that lead to their growth make it very difficult to determine the dominant factor. Some potential factors that can influence the formation of biofilm are discussed below.

Availability of nutrients:

Biofilm bacteria need bioavailable forms of nutrients for growth in order to remain in WDNs. They require carbon, nitrogen, and phosphorus, with carbon being required in the greatest proportion. Some researchers have suggested a carbon:nitrogen:phosphorus ratio of 100:10:1 as being suitable for bacterial growth (CRCWQT, 2005). Trace amounts of some other nutrients are also required for the growth of biofilms, but these have not been investigated (LeChevallier, 1990). As bacteria mainly consume organic carbons, reducing the concentration of this nutrient can limit biofilm growth.

Disinfectants:

Chlorine, chlorine dioxide, and ozone are the three main primary disinfectants used by water companies to kill or prevent the growth of microorganisms in drinking water. Although drinking water is treated with these primary disinfectants before it is pumped through WDNs to consumers, its quality degrades with time as it travels through the network. To mitigate this problem, water in WDNs is treated with secondary disinfectants. The main secondary disinfectant used by water companies to treat drinking water is chloramine. Chloramine is obtained by adding ammonia to chlorine. The different types of

chloramine include monochloramine, dichloramine, trichloramine, and organic chloramines. The reactions that occur during the formation of chloramines are shown in Eqns. 2.10 – 2.12. At pH greater than 7.5, the chemical reaction that leads to the formation of monochloramine is dominant. When pH is between 4.5 and 5.0, the chemical reaction that leads to the formation of dichloramine occurs. Trichloramine is formed when pH is below 5 (Teasdale et al., 2007).

$$NH_4 + HOCl \rightarrow NH_2Cl + H_2O + H^+ \quad (2.10)$$

Monochloramine

$$NH_2Cl + HOCl \rightarrow NHCl_2 + H_2O \quad (2.11)$$

Dichloramine

$$NHCl_2 + HOCl \rightarrow NCl_3 + H_2O \quad (2.12)$$

Trichloramine

The type and concentration of disinfectant used can affect bacteria growth and biofilm formation in WDNs. It has been reported that increasing chlorine levels reduces biofilm formation (Gauthier et al., 1996; LeChevallier et al., 1987). Monochloramine has several advantages when used as a secondary disinfectant. It is the most common secondary disinfectant used for treating drinking water (USEPA, 2009). It is very effective because it does not dissipate quickly and provides longer-lasting protection (Teasdale et al., 2007). It is considered to be a less effective biocide for free cells, but it remains stable over a long duration. In addition, it better penetrates thick residuals and is less reactive. However, the effect of monochloramine on attached cells is very difficult to measure (USEPA, 2009; Zhang & DiGiano, 2001).

Temperature:

Although it is difficult to control the temperature of water in WDNs, it is a very important variable that influences bacterial growth rates (Ginige et al., 2011). Temperatures above 15°C promote bacterial growth. In addition, high temperatures can strongly influence the treatment plant efficiency, disinfection efficiency, and corrosion rates (LeChevallier, 1990; CRCWQT, 2005).

Water age:

The age of water in a WDN is the time taken for the treated water to travel from the source of water supply to a given node in the WDN. It may range from a few seconds to several days. The water age in a WDN depends on its mode of operation as well as physical variables such as the flow rate, pipe size, configuration, and amount of storage. WDNs with high flow rates and small pipe sizes will have a lower water age than those with low flow rates in large pipe sizes (National Research Council, 2005).

2.2.4 Influence of physical and hydraulic variables on sediment accumulation

Many studies have focused on the effects of physical variables on sediment accumulation (Vreeburg & Boxall, 2007). These variables can be grouped into (i) pipe-related variables such as the pipe material, pipe age, and pipe cleaning process, and (ii) hydraulic variables such as the water velocity, shear stress, diurnal variation, and turbophoresis (Vreeburg & Boxall, 2007). Boxall, Skipworth and Saul (2001) indicated that shear stress is the primary cause of sediment conditioning and re-release in pipe networks. Similarly, the pipe cleaning process employed may remove corroded pipe wall material in iron pipes, which may lead to further corrosion and scaling. Some of these variables and processes that influence sediment accumulation are discussed in the following sections.

2.2.4.1 Flow velocity

Studies have shown that flow velocity influences the accumulation process in pipes. In the Netherlands, a flow velocity of 1.5 m/s has been mandated to clean water mains. For self-cleaning, the velocities should be at least 0.4 m/s (Blokker & Vreeburg, 2005; van Boomen & Vreeburg, 1999). In the UK, the values range from 0.7 m/s for a 50-mm pipe to 1.3 m/s for a 200-mm pipe. Numerous laboratory and field studies have revealed that the generation of material layers is influenced by the range of daily flow patterns, with greater variability reducing material accumulation (Husband, Boxall, & Saul, 2008).

2.2.4.2 Shear stress

Ackers, Brandt and Powell (2001) recognised the importance of shear stress in the mobilisation of sediment, and recommended a value of 2.5 N/m^2 for complete flushing of material from pipe walls. Later, Boxall et al. (2001) developed the Prediction of Discolouration events in Distribution Systems (PODDS) model based on effective shear stress criteria. Boxall and Saul (2005) conducted extensive field studies on discolouration,

and concluded that deposition occurs in cohesive sediment layers and that conditioning shear stress is a function of the peak daily shear stress. Sly et al. (1988) studied biofilm development in WDNs, and observed that biofilms developed at a velocity of 0.5 m/s actively oxidised and deposited Mn, but those developed at 0.01 m/s did not affect Mn. Prince, Ryan and Goulter (2003) conducted continuous monitoring of turbidity and flow in WDNs, and their analysis of this data, along with customer complaint data and operations data, revealed that the largest proportion of turbidity spikes occurred during events that created abnormally high water velocities. Boxall and Prince (2006) analysed a large-diameter asbestos cement main, and proposed a minimum shear stress value of 1.12 N/m^2 for effective flushing. Husband and Boxall (2011) suggested that an ultimate shear stress of 1.2–1.8 N/m^2 is sufficient for the complete removal of sediment layers from plastic pipes. For rough iron pipes, no ultimate layer bonding strength was found for the flushing forces attained.

2.2.4.3 Turbophoresis

Turbophoresis is the tendency of a particle to move from a more turbid region to a less turbid one. In pipe flows, this means that particles move from the bulk flow toward the pipe wall, where they attach to cohesive layers. Vreeburg and Boxall (2007) carried out experimental investigations which indicated that, at low velocities, sediment accumulation occurs in the lower half of pipes (i.e. gravity settlement). At higher velocities, they observed that turbulence forces dominated the gravitational forces and influenced the accumulation process. Their experimental observations showed that turbophoresis forces exceeded gravitational force at velocities greater than 0.14 m/s.

2.2.4.4 Pipe material

Pipe material is an important variable that can influence sediment accumulation. Based on the material, pipes can be divided into ferrous and non-ferrous pipes. Studies have shown that networks that mainly use ferrous pipes are more prone to discolouration events because of corrosion (Cook et al., 2005). Contrary to the common belief that networks with plastic pipes are less prone to discolouration, recent studies have shown that deposits on plastic pipe walls are looser because of the smooth pipe surface, and are thus subject to sloughing under smaller shear forces compared to iron pipes (Cerrato et al., 2006; Husband & Boxall, 2011). This means that discolouration can also occur in regions of WDNs that are entirely networked with PVC pipes.

2.2.4.5 Pipe condition

Pipe condition strongly influences sediment accumulation, and, in turn, the risk of discolouration. Studies on flushing by Boxall et al. (2003) suggested that high turbidity levels occur during flushing operations in pipes that are in poor condition, because these pipes have not been cleaned or rehabilitated. Cook et al. (2005) studied the structural integrity of WDNs and water quality using pipe properties such as the burst frequency, diameter, age, and daily conditioning shear stress. They observed that DMAs with higher burst rates had fewer customer complaints related to discolouration; however, the inverse was not necessarily true. They attributed this to the cleaning effect of the bursts and leakages in removing deposits from pipes.

2.2.4.6 Pipe cleaning

Pipes can be cleaned by using various methods to flush accumulated sediment. The three most commonly used methods are water flushing, water/air scouring, and swabbing/pigging. The complete removal of all materials from pipes is only possible by using more abrasive methods such as swabbing and pigging. However, these methods can increase corrosion in ferrous pipes. Where established corrosion layers can effectively protect the underlying ferrous material aggressive cleaning methods can expose this underlying surface, which will then start to corrode more rapidly, thus generating material on the pipe wall and releasing ferrous ions into the bulk fluid (Slaats, 2002).

2.2.4.7 Water pressure

When water pressure in pipes with leakages is very low or negative, there is a high-risk that contaminants will be introduced into WDNs. Contaminants can enter the system through broken or cracked pipes, and also during or after maintenance and repair (Kirmeyer, Friedman, Martel, & Howie, 2001). Pipes are normally depressurised during repair or when hydrants are being used for extinguishing fires (Sadiq, Kleiner, & Rajani, 2007). At such times, backflow often occurs, causing water discolouration. During backflow, contaminated water (organic or inorganic particles) from factories, hospitals, and domestic water tanks flows back into WDNs (Prince, 2008). A study by Seyoum and Tanyimboh (2014) showed that a reduction in water pressure increases water age, and subsequently decreases water quality.

2.3 Water discolouration theories

There are several published theories on the formation of discoloured water. However, there are four main theories that researchers have used to explain how discoloured water is formed. These are the deposition and re-suspension theory, mobilisation theory, cohesion and erosion theory, and adhesion and striping theory. The following sections review these four theories.

2.3.1 Deposition and re-suspension theory

The deposition and re-suspension theory is based on the movement of non-cohesive, discrete, relatively large particles in river systems. The diameter of particles in river systems ranges from about 60 mm (gravels) to 0.2 mm (sand). The relatively large sizes of the particles make gravitational force dominant in the deposition and re-suspension process. This process also depends on physical properties of the particles, such as shape, specific gravity, and concentration (Boxall et al., 2001; Gauthier et al., 1996; Walski, 1991). If the force due to the flow of water acting on the deposited particles exceeds the gravitational and frictional forces, the particles will become entrained. The deposition and re-suspension process is cyclic in nature. It starts from the transportation of particles stage, goes through the deposition stage, erosion stage, and then again to the transportation stage. Figure 2.1 illustrates the deposition and re-suspension process in discoloured water formation.

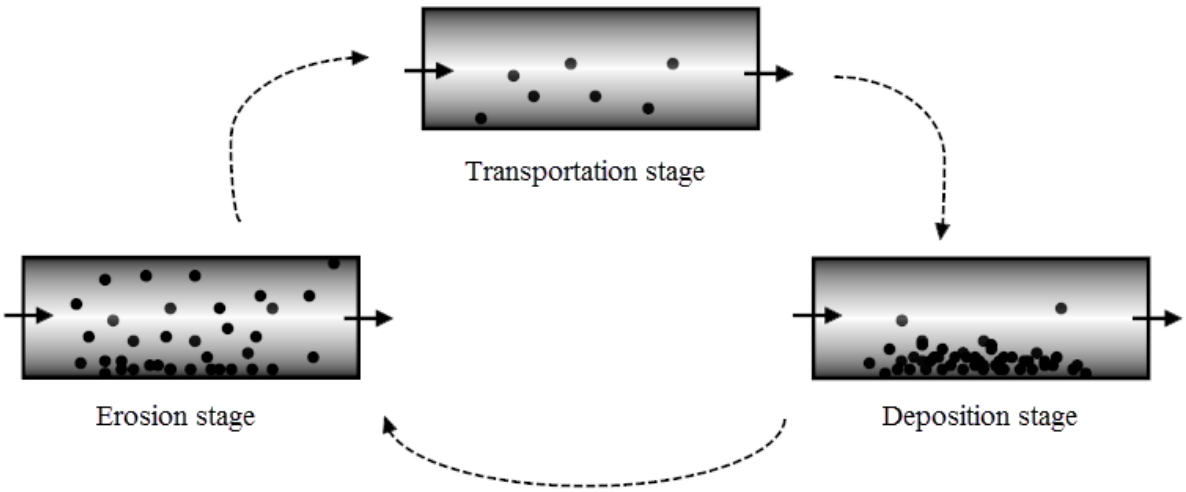


Figure 2.1 Deposition and re-suspension process in discoloured water formation

Unlike particles in river systems, particles in WDNs are very small. A large proportion of the particles found in WDNs are known to be less than 50 μm in diameter (Prince, 2008).

This means that gravitational force acting on particles will not be dominant in WDNs. Therefore, the deposition and re-suspension theory may not properly explain the formation of discoloured water in WDNs. Deposited particles in WDNs may be as a result of some other forces or mechanisms other than gravitational force.

2.3.2 Generation and mobilisation theory

Smith, Bisset, Colbourn, Hold and Lloyd (1997) and Boxall et al. (2001) used the generation and mobilisation theory to explain the formation of discoloured water in unlined cast iron pipes. In this theory, deposition occurs from corrosion of the unlined cast iron pipes. They observed that, in certain instances, the rate of Fe particles formed through corrosion was greater than its deposition rate. The deposited fine Fe particles were described by Boxall et al. (2001) as cohesive layers. When flow velocity or shear stress increases, deposited particles become mobilised and are transported together with suspended colloids to cause discolouration. Unlike the deposition and re-suspension process, the generation and mobilisation process is not cyclic. It starts with the generation of particles stage, goes through the mobilisation stage, and ends with the transportation of particles stage. It is assumed that the mobilised particles are not re-suspended. This assumption is inaccurate because mobilised particles do not remain in suspension forever. Figure 2.2 illustrates the generation and mobilisation process in discoloured water formation.

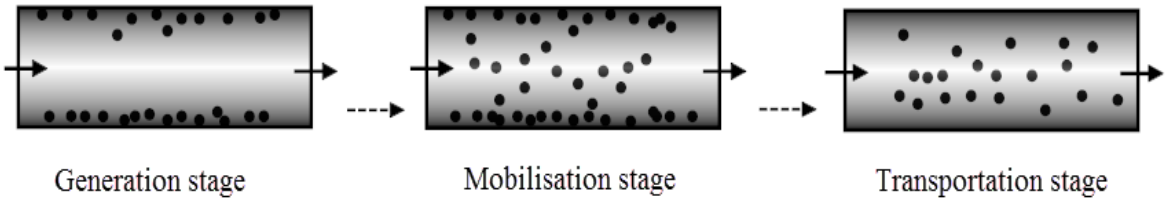


Figure 2.2 Generation and mobilisation theory in discoloured water formation

2.3.3 Cohesion and erosion theory

In the cohesion and erosion theory, particles cohere in layers and are deposited on the inner surface of pipe walls. Cohesion occurs as a result of biological,electro-chemical, and/or van der Waals forces that exist within the particles. If the shear stress applied on the surface walls exceeds these forces, the particles in the cohesive layers are eroded, causing discolouration to occur (Ackers et al., 2001; Boxall & Saul, 2005). The cohesion and

erosion process in discoloured water formation is illustrated in Fig. 2.3. Although water particles in WDNs are likely to undergo cohesion, the cohesion and erosion theory does not explain how particles adhere to the inner walls of pipes. Sorption, which was adopted in this study to explain this process, better explains particle deposition on pipe walls because it takes into account both the adsorption and absorption of particles.

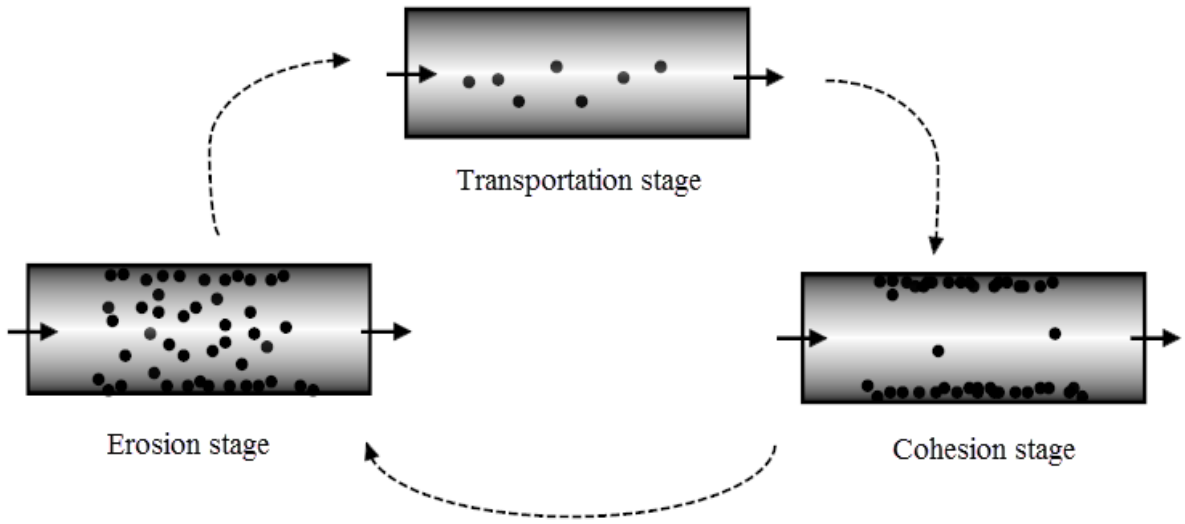


Figure 2.3 Cohesion and erosion process in discoloured water formation

2.3.4 Adhesion and stripping theory

Areas in the WDNs with very low shear stress, such as dead ends and redundant loops, are more susceptible to the formation of Fe and Mn oxide coatings, amorphous $Al(OH)_3$, biofilms, or scales from corrosion on the inner surface of pipe walls (Sly et al., 1990; Smith et al., 1997; Wang et al., 2012). The loose particles adhere to the scales or biofilms. During high flow events such as increased water consumption or opening of fire hydrants, the adhered particles on the pipe walls are stripped and mobilised to cause water discolouration. Figure 2.4 illustrates the adhesion and stripping theory in discoloured water formation. The adhesion and stripping theory may not properly explain the formation of discoloured water because it does not take into account the cohesion of particles in WDNs.

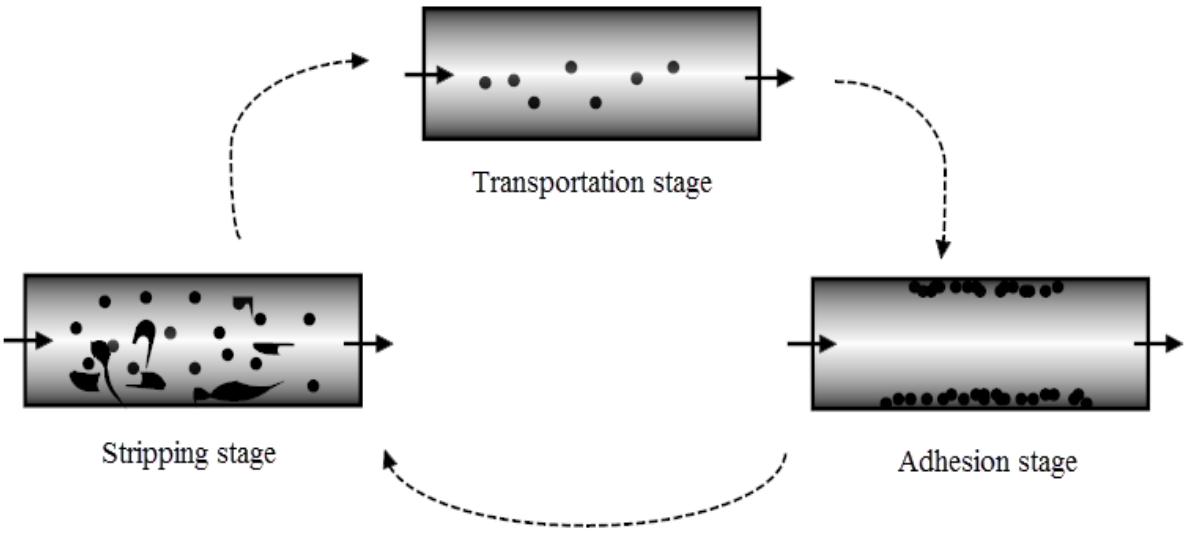


Figure 2.4 Adhesion and stripping theory in discoloured water formation

2.4 Discolouration risk models

Discolouration can be described as the mobilisation of sediment particles accumulated on pipe walls in WDNs. The characteristics of these sediment particles, such as their size, density, origin, and composition, vary greatly from one network to another. Sediment particles include organic and inorganic materials contained in the source water (Ellison, 2003; Gauthier et al., 2001; Lin & Coller, 1997; Slaats, 2002; Vreeburg, Schaap, & van Dijk, 2004b) or chemicals such as carbon, coagulants, and bio-particles from filters that are added to the water at the treatment plant (Boxall et al., 2003; Gauthier et al., 1999). To a certain extent, pipe corrosion, lining erosion, biofilm growth, and chemical reactions in WDNs may also produce sediments (Huck & Gagnon, 2004; LeChevallier et al., 1987; Sly et al., 1990; Walski, 1991). The external intrusion of contaminants during pipe rehabilitation and repair may also be a contributing factor (Prince, Goulter, & Ryan, 2001; Slaats, 2002). Sediments from the above-described sources are deposited on the inner surface of pipe walls in WDNs. Once deposited, these materials can be dislodged by excessive hydraulic forces produced by hydraulic events such as pipe burst, pipe flushing, and valve operations.

Early studies on discolouration were mainly based on collecting samples at different locations in WDNs. However, advanced measurements, loggings, and communication technologies over the past two decades mean that instruments are now available for

monitoring pressure, flow, and turbidity continuously and simultaneously at multiple locations. Variations in these variables can therefore be recorded to identify factors that influence discolouration events. The recorded data have been used to develop techniques to assist water companies identify and quantify discolouration risks in WDNs (Vreeburg, 1996; Vreeburg & Boxall, 2007). As mentioned in Chapter 1, many studies have been undertaken on the influence of various processes and mechanisms on water discolouration. However, they have been studied independently, rather than in combination. A number of researchers have published analytical tools to predict discolouration in the water industry. Some of these tools are reviewed in the following sections.

2.4.1 Prediction of Discolouration events in Distribution Systems model

The Prediction of Discolouration events in Distribution Systems (PODDS) model, also known as the cohesive transport model (CTM), is a computer-based discolouration model developed by Boxall et al. (2001) with the Pennine Water Group (PWG) at the University of Sheffield, UK. This model can predict the discolouration (turbidity) response to hydraulic changes in WDNs. PODDS is an extension of EPANET; a graphical software developed by the United States Environmental Protection Agency (USEPA). Boxall et al. hypothesised that sediment accumulation in WDNs occurs in layers, where each layer is conditioned according to the daily shear stress applied on it. The shear stress in a pipe is the force acting on an area of pipe wall perpendicular to the direction of flow. It can be mathematically expressed as Eqn. 2.13. The model uses the fundamental principle that discolouration is caused by mobilisation of layers of cohesive material attached to pipe walls. With this premise, the authors developed the PODDS model by incorporating the concepts of cohesive transport theory, which was developed to characterise the erosion of cohesive estuarine sediment.

$$\tau = \rho_w g R_h S_o \tag{2.13}$$

where R_h = hydraulic radius; S_o = hydraulic gradient; τ = boundary shear stress;
ρ_w = density of water; and g= gravity.

The PODDS model considers each cohesive layer to have a discolouration potential that corresponds to the strength of that layer. According to this model, the discolouration potential of layers away from the pipe boundary increases as their layer strength decreases.

This implies that a lower force is required to dislodge the top layer than the layers below it. The ultimate strength of the layers is theorised as being equal to the daily peak shear stress experienced within each pipe.

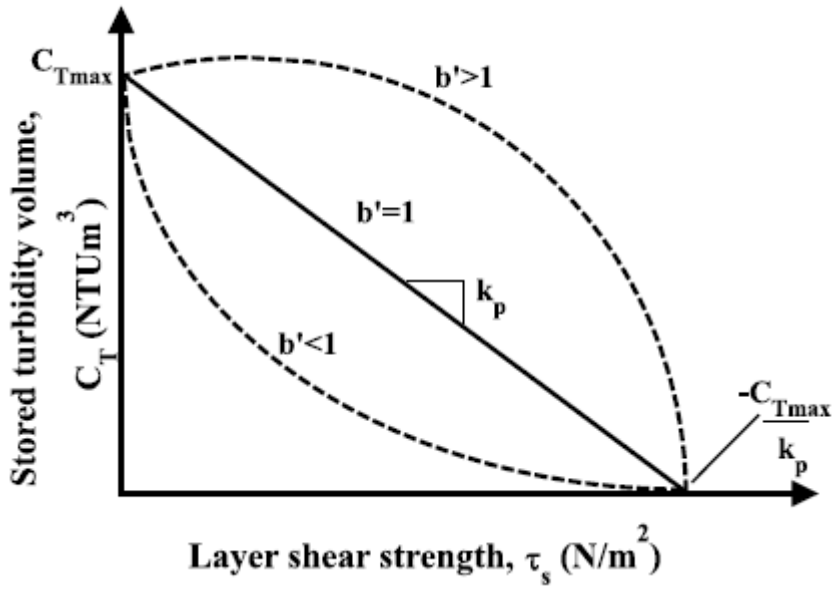


Figure 2.5 Representation of layer strength versus stored turbidity volume

The relationship between the strength of the layer and the stored turbidity volume (layer thickness) is expressed mathematically as Eqn. 2.14. A graphical representation of the relationship is presented in Fig. 2.5. From the graphs it can be observed that layers that have increased stored turbidity volume have reduced shear stress. The strength of the cohesive layer is determined by the shear stress applied within each pipe at constant peak daily flow. This means that the layer state is dependent on the daily shear stress generated by network hydraulic conditions. Therefore, sections of the pipe network that are subject to low daily peak shear stress, such as dead end pipes, redundant loops, oversized pipes, and zone boundaries, will have higher discolouration potential because low hydraulic forces can dislodge the attached material layers.

$$\tau_s = \frac{C_T^{b'} C^{b'} - C_{Tmax}}{k_p} \tag{2.14}$$

where k_p = gradient of layer strength in PODDS model; τ_s = current layer strength;
C_T = stored turbidity volume of layer; C_{Tmax} = Maximum turbidity potential;

and b′ = power term to set for first order relationship.

In an undisturbed system without any unusual flow, the developed cohesive layers are at their maximum discolouration potential as conditioned by the maximum daily peak flow rate. However, if there is an increase in the network demand (for example, due to operational activities or hydraulic events such as pipe bursts), this may disturb the prevailing equilibrium conditions and exert shear stress that exceed the conditioned shear stress on these cohesive layers. This may cause mobilisation of the cohesive sediment, which may subsequently lead to discolouration. Equation 2.15 is used to describe the mobilisation of cohesive sediment when exposed to a disturbing hydraulic force.

$$R_t = P_s(\tau - \tau_s)^{f'} \tag{2.15}$$

where τ = applied shear stress; f′ = power term; P_s = gradient term; and
R_t = rate of release of sediment by the excess shear stress.

The incremental change turbidity resulting from this erosion can be evaluated as:

$$\Delta N = \frac{R_t A_s}{Q} \tag{2.16}$$

where A_s = pipe surface area affected; Q = flow rate; and ΔN = change in turbidity.

The model is calibrated by measuring the flow rate and turbidity response of a system to a predicted flow rate and turbidity response. The variables used to describe the relationship between the strength of the cohesive layer and its discolouration potential are then optimised to achieve this calibration (Boxall et al., 2001).

The PODDS model does not take into account the decrease in the concentration of particles as they re-accumulate on the pipe wall during regeneration. The model's inability to address re-accumulation of mobilised particles makes it inaccurate, because particles do not remain suspended forever after they are entrained (Prince, 2008). Furthermore, because the model uses the assumptions of quasi-steady state modelling within EPANET, it does not use dynamic shear stress in its computations (Prince, 2008). The PODDS model does

not explicitly consider either the source of the material that is deposited or the mechanisms and processes that contribute to sediment accumulation. Furthermore, the empirical variables in the model need to be established through calibration before applying the model to a different hydraulic event in that section of the pipe. Because the mechanisms and processes that lead to discolouration are highly complex, the transfer of calibrated parameters from one system to another is questionable. Therefore, to use the model effectively, a table of these parameters needs to be established by conducting field trials covering a wide range of hydraulic and network conditions (Prince, 2008). Since the PODDS model only uses physical/hydraulic variables to make predictions, it may not properly explain the formation of discoloured water in WDNs.

2.4.2 Particle Sediment Model

The Particle Sediment Model (PSM), developed by Wu et al. (2003) at the Cooperative Research Centre, can be used to predict sediment accumulation in water distribution systems. PSM uses flow distribution in networks and inlet sediment concentrations to predict the mass of sediment deposited on pipe walls. It takes into account the gravitational settlement of sediment particles and the affinity between pipe walls and the sediment particles. The calculation of sediment accumulation in the distribution system is based on two mechanisms, namely, settling of particles under gravity and deposition of particles on pipe walls due to particle-wall surface interaction.

Under settling of particles under gravity mechanism, particles in the pipe are considered to be in one of three states:

(i) when the flow velocity exceeds the particle critical velocity, particles are subject to re-suspension,

(ii) when the flow velocity is below a certain limit, particles can settle, and

(iii) when the flow velocity is between the limits in (i) and (ii), particles will move upstream without settlement or re-suspension of material from pipe walls.

Experimental studies using a single loop pipe network were performed to demonstrate the deposition of particles on pipe walls due to particle-wall surface interaction. The results indicated that particles began to settle at velocities as high as 0.3 m/s, which contradicts the observations made by Boxall et al. (2003). This phenomenon was attributed to the attachment of sediment to pipe walls due to van der Waals forces. A set of equations was

proposed to predict the sediment concentration and model the process of sediment deposition on pipe walls (see Eqns. 2.17–2.19). These semi-empirical equations were obtained based on experimental data collected using a laboratory pipe network. Different pipe materials (mainly PVC and cast iron cement lined), sediment types, sediment concentrations, and flow regimes were used to generate the data.

$$\frac{\partial C_s}{\partial t} = -\propto_c (C_s - C_\infty) \tag{2.17}$$

$$C_w = \beta_c . C_\infty \tag{2.18}$$

$$\frac{M_w}{L_p} = C_w \frac{1}{4}\pi d_p^2 = B_c C_\infty \frac{1}{4}\pi d_p^2 \tag{2.19}$$

where C_s = concentration of particles in suspension; α_c = decay coefficient;
C_w = mass of particles attached to the wall per unit weight of water;
B_c = wall mass coefficient; d_p = pipe diameter;
C_∞ = final steady state concentration of particles in suspension; and
M_w/L_p = mass of particles attached to the wall per unit of pipe length.

PSM is an extension of investigations reported by Grainger et al. (2002), and is therefore subject to the limitations of their study. Grainger et al (2002) observed that because PSM was developed under laboratory conditions, the diurnal flows from the experimental setup did not replicate diurnal flows in real live WDNs. The model is yet to be validated with field data. Furthermore, PSM does not include chemical and biological variables. Therefore, the model may not properly explain the formation of discoloured water in WDNs.

2.4.3 Discolouration Risk Analysis Tool

Boxall and Husband (2005) developed a DMA-level tool, Discolouration Risk Analysis Tool (DRAT), for ranking pipes in order of their discolouration risk. This methodology of calculating risk is based on the PODDS theory. The tool was developed to help operational managers plan cleaning programmes. In this risk-based approach, a series of automated events are simulated and pipes ranked according to their simulated discolouration risk. The tool was then used to identify networks or specific pipes that present a potential discolouration risk. The predictions by the tool were only partially successful. The

localised variability in hydraulics and the difficulties in determining actual conditioning and mobilising shear stress could not all be incorporated into the hydraulic model. As a result, there were discrepancies between the measured and the modelled risk assessment. Because DRAT uses PODDS theory to compute the risk, it also has the limitations of PODDS model discussed in Section 2.4.1.

2.4.4 Re-suspension Potential Method

Re-suspension Potential Method (RPM) was developed by Vreeburg et al. (2004a) with Kiwi Water Research in the Netherlands. It is based on measurements of the capability of sediment within the distribution system to re-suspend using a standard flushing procedure. The pipes for which discolouration risks were assessed had minimum lengths of 315 m. To create hydraulic disturbance in a given pipe, a fire hydrant was opened to increase the velocity by an additional 0.35 m/s above the actual velocity and maintain for 15 minutes. Turbidity was continuously monitored and measured until it returns to its original level.

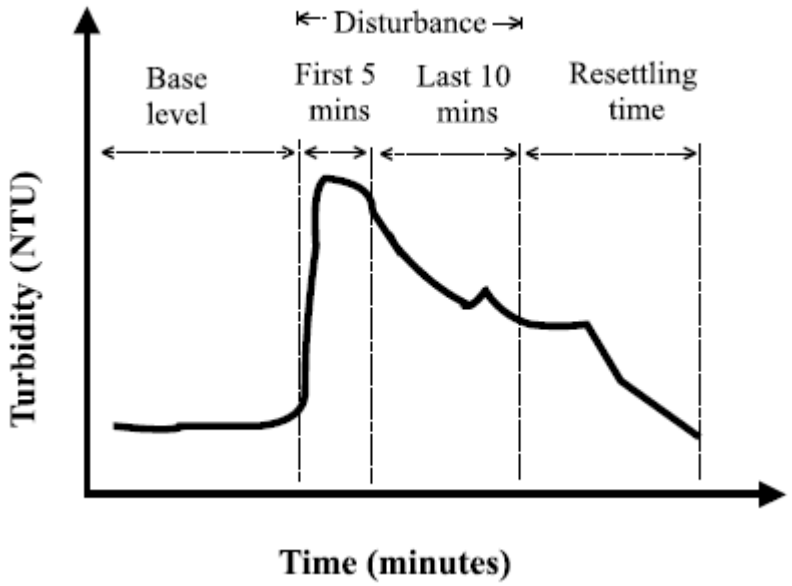


Figure 2.6 Turbidity trace results from the RPM test

A typical result of RPM is shown in the graph in Fig. 2.6. It consists of four regions that are used to rank discolouration risk. The first region, which comprises of the base level turbidity, is the level preceding the hydraulic disturbance. It is used to estimate the time for the turbidity to return to its initial level. The second region corresponds to the initial increasing turbidity during the first 5 minutes after the fire hydrant is opened. The third

region corresponds to the development of turbidity during the last 10 minutes after the fire hydrant is opened, whereas the last region corresponds to the resettling time. The resettling time is the time taken for turbidity to return to its initial level after the fire hydrant is closed. The discolouration risk ranking was done based on the maximum and average turbidity of the first 5 and last 10 minutes of the disturbance, and during the resettling time. (Vreeburg et al., 2004a, 2004b; Wricke et al., 2007).

A limitation of the RPM is that because it is mainly applied to pipes with large diameters between 100 and 150 mm, the difference in shear stress caused by increase in uniform velocity is insignificant. Consequently, changes in the hydraulic regime only cause a small increase in shear stress, which will be insufficient to raise turbidity levels for customers to complain (Vreeburg et al., 2004a). RPM uses only turbidity and water velocity as variables. Therefore, it may not be able to make good predictions in real live WDNs, where the formation of discoloured water is influenced by other hydraulic, chemical, and biological variables.

2.4.5 Discolouration Risk Modelling Approach

Discolouration Risk Modelling approach (DRM) is a risk assessment tool developed by Dewis and Randall-Smith (2005) in conjunction with Yorkshire Water and the Ewan Group. It generates a discolouration performance score for each pipe in WDNs, thus enabling operations managers and asset planners to make informed decisions. The propensity of each pipe to give rise to discolouration is expressed as a combination of the likelihood and consequence of either the pipe's failure in the entire network or a failure elsewhere that causes discolouration. The likelihood is assessed based on the pipes' tendency to burst, potential to cause discolouration, and sensitivity to flow changes that could cause discolouration. On the other hand, the consequence is assessed by the number of customers who could, potentially, be affected. The variables used in predicting the discolouration risk in DRM include pipe material, pipe age, rehabilitation history, and the history of Fe or Mn discharge. These variables are arranged in the form of a risk tree (based on standard fault tree analysis methodologies), with some modifications to suit the application. The hierarchical structure of the tree describes the dependencies between the variables and weights allocated to each node within the tree (Bhagwan, 2009; Dewis & Randall-Smith, 2005). A typical structure of the tree is shown in Fig. 2.7.

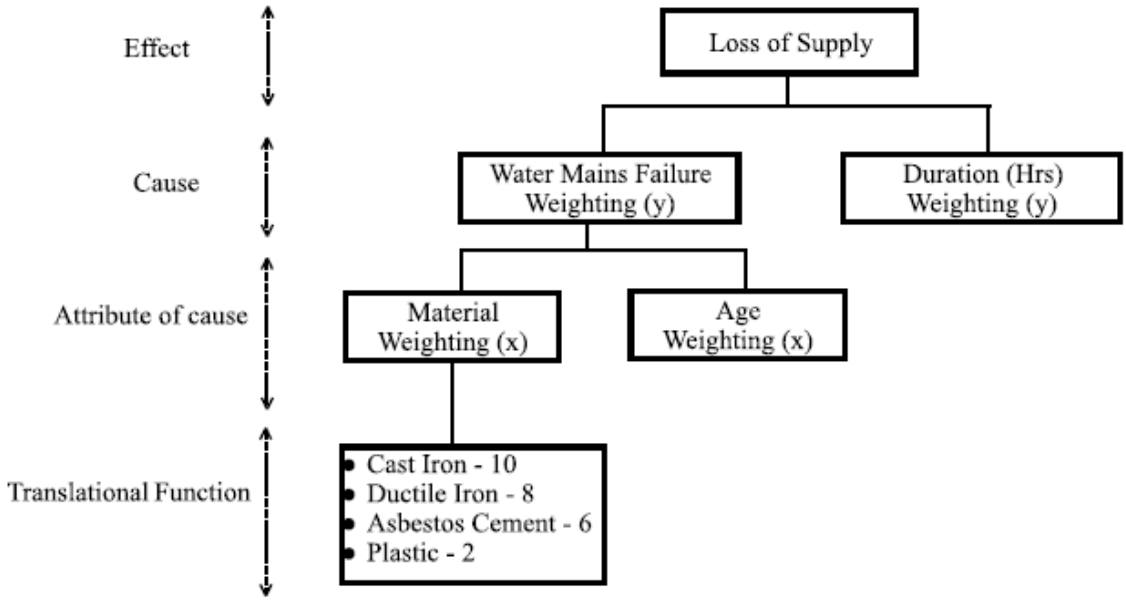


Figure 2.7 A hierarchical structure of the tree of the DRM

A limitation of DRM is that it assumes chemical and biological factors are insignificant in the formation of water discolouration, and therefore uses only hydraulic and pipe related variables for its predictions. However, as discussed in Sections 2.2.2 and 2.2.3, many researchers have found that chemical and biological factors significantly contribute to the formation of discoloured water.

2.4.6 Discolouration Propensity Model

The Discolouration Propensity Model (DPM), which is a replacement to the DRM, uses the CTM proposed by Boxall et al. (2001) to predict discolouration risk. It was developed by McClymont, Keedwell, Savic and Randall-Smith (2010) in association with Mouchel. DPM uses CTM to calculate the volume of deposited material, which is measured as turbidity (expressed in Nephelometric Turbidity Units, NTU). Furthermore, the model can also be used to calculate the volume of material mobilised in a given specific hydraulic event in the network such as valve closure and pipe bursts. It then ranks DMAs based on their ability to store discolouration-causing material. DPM also uses the EPANET software to calculate the dynamic hydraulic conditions of each network. A shear stress equation is then used to calculate each pipe's daily conditioning shear stress, which is subsequently used to assess the relative risks of pipes within a DMA. The daily conditioning shear stress values are obtained based on the hydraulic gradient values calculated by EPANET (McClymont et al., 2010, 2011). The implementation of DPM is based upon the shear

stress equations outlined in the PODDS model (see Eqns. 2.13-2.16). The results obtained from DPM in a case study network are shown in Table 2.2. The score for each DMA is calculated by summing the total potentially stored material of all pipes in each DMA and normalised by the total length of pipe in the DMA. The DPM is subject to the limitations of the PODDS model indicated in Section 2.4.1 because it uses CTM in its computations.

Table 2.2 Results from DPM case study ordered by DPM score from best to worse

DMA	Length Normalised Potential (DPM Score)	NTU Rank (Relative)
548	3.429×10^5	1
551	5.276×10^5	2
550	19.608×10^5	3
549	100.908×10^5	4
547	261.094×10^5	5

2.4.7 Pressure-dependent Analysis (PDA) model

Seyoum and Tanyimboh (2014) developed a Pressure-dependent Analysis (PDA) model which is integrated with a water quality model at the University of Strathclyde, UK. PDA, which is an extension to EPANET 2 known as EPANET-PDX, was used to investigate the effect of a range of hydraulic pressure conditions on water quality variables such as water age, concentration of chlorine and concentration of the disinfection by-product; trihalomethanes (THMs). The Epanet-PDX was developed by integrating a demand function given in Eqn. 2.20 into a global gradient algorithm. For detailed information on the PDA model, refer to Seyoum and Tanyimboh (2014).

$$Q_i(H_i) = Q_i^{req} \frac{\exp(\epsilon_i + \theta_i H_i)}{1 + \exp(\epsilon_i + \theta_i H_i)} \qquad (2.20)$$

where Q_i = demand at node i; Q_i^{req} = required supply at node i; H_i = head at node *i*; and ϵ and θ = parameters to be calibrated with relevant field data.

Simulations were performed using data from two WSZs obtained from a drinking water company in the UK. Graphs of water quality variables were plotted at various pressure conditions. At normal pressure, the Demand Satisfaction Ratio (DSR) was 100%. DSR is

the ratio of the flow available to the flow required. At pressure-deficient condition, DSR was less than or equal to 75%. They observed that under pressure-deficient conditions, water age and THM concentration increased, whereas the chlorine concentration decreased (see Fig. 2.8). Conversely, at normal pressure conditions, lower water age and THM levels, and higher chlorine concentrations were observed. This is because, generally, low pressure in WDNs reduces flow velocity, which causes water age and chlorine depletion to increase, and subsequently increases the formation of THMs. The PDA model is a very useful tool for predicting water quality in WSZs. However, the model could be improved by adding pipe-related variables such as pipe material and pipe age, and variables that influence biological processes such as phosphorus and carbon.

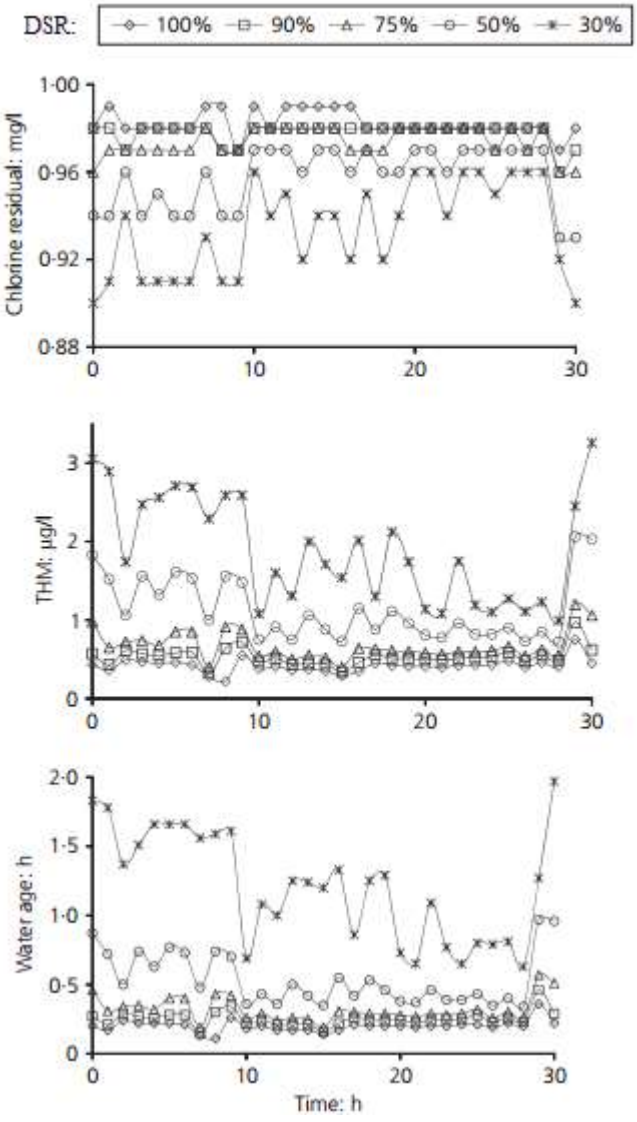


Figure 2.8 Graphs of water quality variables plotted at various pressure conditions Seyoum and Tanyimboh (2014)

2.4.8 Other discolouration risk models

Walski and Draus (1996) proposed a method for modelling turbidity in water during mains flushing operations. They observed that the amount of turbidity generated during mains flushing is proportional to the velocity generated in the water main during flushing. Samples were collected at 10 minutes intervals during the flushing operation. They observed that the measured turbidity was high during the initial period, but gradually reduced. Using these measured values, Walski and Draus (1996) proposed an empirical relationship between the turbidity and velocity. They concluded that the amount of turbidity generated during the flushing operation is influenced by the amount of material deposited in the mains since the previous discolouration event.

Ackers et al. (2001) proposed equations for sediment transport in closed conduits by extending their work on sediment transport in rivers. The Ackers–White equations were derived for uniform sediments transport in rivers. However, when particles of different sizes and densities are present, their settling behaviour cannot be explained by uniform sediment flow theories. Therefore, applying these equations, which are characterised by sediments having varying particle size distributions to WDNs will not yield desirable results.

Based on the cohesive transport theory, Boxall et al. (2001) proposed a mathematical model for predicting discolouration due to hydraulic events in WDNs. From a series of experiments and field studies, they observed that the sizes of particles present in sediment samples were not sufficiently large for gravity settlement. This implies that the hydraulic forces and mechanisms exert sufficiently large forces to keep the sediment particles in suspension and inhibit gravity settlement. From these observations, they concluded that mechanisms other than gravity settling forces caused particle discolouration in WDNs.

Husband et al. (2008) performed laboratory experiments to investigate drinking water discolouration in WDNs. Their model is based on the hypothesis that discolouration is caused by the erosion and transport of fine particles, mainly Fe and Mn, which are attached to the pipe walls. The model is also based on the hypothesis that particles are arranged in cohesive layers which gradually build up over a period of time and are conditioned by the daily flow pattern within the system. Erosion takes place when there are changes in shear stress on the walls. Graphs of turbidity against shear stress for dynamic and steady state

flow indicated an increase in turbidity response and an increase in boundary shear stress. They also observed a positive correlation between Fe concentration and shear stress. From these relationships, they suggested that the mobilisation of accumulated materials is influenced by daily shear stress, with greater variability reducing material accumulation.

2.5 Summary

A comprehensive literature review on drinking water discolouration models and the factors that influence the formation of drinking water discolouration showed that researchers have only studied each of the factors that influence Fe and Mn accumulation either partially or separately, but never in combination. Some of the variables identified to influence discolouration/Fe and Mn accumulation include pH, DO, shear stress, water age, carbon, nitrogen, pipe material, alkalinity, temperature, and Al. Most of the water discolouration models reviewed only used physical/hydraulic variables mainly in predicting drinking water discolouration. Hence, they do not capture some relevant factors that influence the formation of discoloured water in WDNs, and therefore may not properly explain the processes and mechanisms that lead to Fe and Mn accumulation.

CHAPTER 3: Artificial Intelligence Based Methods

3.1 Introduction

In Chapter 2, a comprehensive literature review was prepared to identify relevant variables that influence Fe and Mn accumulation in WDNs. In addition, various drinking water discolouration models were reviewed. In this chapter, a literature review on artificial intelligence based methods of modelling will be conducted. The remaining sections of this chapter are organised as follows: Section 3.2 presents a historical background of ANNs, explains what they are, and compares them with conventional models. This section also discusses how ANNs are currently being applied in water resources and in other disciplines. In addition, the section also gives a critical review on the various types of ANNs. Section 3.3 presents a critical review on FISs. It reviews fuzzy set concepts and fuzzy inference process. The section also gives some benefits and limitations of FIS, and reviews some applications of FISs in water resources. Section 3.4 gives a review on genetic algorithm. Finally, the summary of this chapter is presented in Section 3.5.

3.2 Artificial neural networks

There are some problems that human brains can solve, that mathematical formulae or computer algorithms cannot solve. This category of problems can be solved by learning from previous examples. The ability of the human brain to adapt and learn from a given set of data makes it possible to solve such problems. Artificial Neural Network (ANN) is an artificial intelligence (AI) based method of modelling that attempts to mimic the learning processes of the human brain by developing complex mathematical relationships from a given set of data (Smith, 1993). In view of this, ANNs do not need a detailed formulation of the underlying processes in their computations. Even though ANNs are very powerful tools, and are sometimes regarded as the future of computing, some researchers are reluctant to apply then because of their black-box nature. While conventional (traditional) models try to explain the underlying modelling processes, ANN models on the other hand are more data-driven and rely heavily on the data that describes the dependent and independent variables.

3.2.1 Historical background of artificial neural networks

ANN was first introduced in the 1940s by Warren McCulloch, a neurophysiologist, and Walter Pitts, a young mathematician (McCulloch & Pitts, 1943). Since then, it has been through a long period of development. Pitts and McCulloch (1947) indicated how ANN could perform abstraction by learning generalised rules from specific instances. They applied this concept on recognition of spatial patterns and classification tasks. Hebb (1949) proposed the Hebbian rule which explained the adaptation of neuron in the brain during the learning process. However, he could not verify this rule because of lack of neurological research.

In the 1950s, researchers started building computer models of ANNs by combining psychological and biological insight. The world's first neurocomputer (the Snark) was designed and built by Marvin Minsky in 1951 as part of his PhD research. Although the Snark operated well from a technical point of view, it was never implemented (Minsky, 1954). In 1958, Frank Rosenblatt, a neurobiologist who is referred by many researchers as the father of neurocomputing, developed the first successful neurocomputer (the Mark I perceptron) at Cornell University. This neurocomputer was capable of using a perceptron learning algorithm to recognise characters by means of a 20 × 20 pixel image sensor (Rosenblatt, 1958). Nevertheless, this was limited by the perceptron's inability to classify patterns that are not linearly separable in the input space.

After the successful invention of the neurocomputer and digital computers, there was an explosion of research on neural networks in the 1960s. Researchers started to move from logic circuits to machine learning. Widrow and Hoff (1960) introduced the ADAptive LInear Neuron (ADALINE), which has an advantage over perceptron learning because of its ability to learn and adapt to new data. ADALINE was the first neural network to solve a real-world problem. It was then used in most analogue telephones to adaptively filter and eliminate echo in real time. The activation function of ADALINE is linear, which limits its applicability to only linearly separable problems. ADALINE is currently used commercially as adaptive equalizers in telephone lines (Rogers & Kabrinsky, 1991).

Although Werbos (1974) developed a learning procedure to train ANNs known as back-propagation of errors, it was not used until after a decade. Little research was conducted on neural networks until the mid-1980s because of lack of high-performance computing

systems. The PDP Research Group (1986) used the back-propagation algorithm with multi-layer perceptrons to solve non-linear separable problems. From that time, the application of ANNs in water resource engineering has almost been explosive. ANNs have moved from being a mere research tool to a powerful tool for solving real-world problems. In 1987, the first Institute of Electrical and Electronic Engineering (IEEE) international conference on neural networks attracted more than 1800 attendees. The following section gives some applications of ANNs in water resources.

3.2.2 Applications of artificial neural network in water resources

In recent years, ANNs have been successfully applied in a broad range of areas including science, engineering, telecommunication, technology, and business (Widrow, Rumelhard, & Lehr, 1994). Some specific applications include remote sensing, stock trading, speech and handwriting recognition, face recognition, e-mail spam filtering, credit scoring, fraud detection, and medical diagnosis. ANNs have also been used in environmental engineering to develop hydrological models, taking advantage of their ability to capture and learn both linear and complex non-linear relationships from modelling data, especially in situations where the underlying physical relationships are not fully understood (Lingireddy & Brion, 2005).

Maier and Dandy (1996) used a back-propagation ANN model to obtain a 14-day forecast of the salinity of the Murray River in South Australia. The K-fold cross-validation method was used to validate the model. The average absolute percentage errors for the model ranged from 5.3–7.0%, indicating the model gave good prediction. This model could help save money, because high salinity levels in the Murray River cost consumers in Adelaide approximately $US22 million per year in damages (Maier & Dandy, 1996). The model could be further improved by using an optimisation technique to fine-tune the number of hidden neurons and layers to obtain appropriate numbers to use during the training process.

Aafjes, Verberne, Hendrix and Vingerhoeds (1997) used a combination of expert systems and ANN to predict water consumption at Friesland, Netherlands. They used a two-year data set which included independent variables such as hourly precipitation, global radiation, temperature, and air pressure. The day of the week and past holidays were also used as independent variables. The dependent variable was hourly water consumption. They also used a traditional statistical-based model, auto regressive integrated moving

average (ARIMA) to make water consumption predictions from the same data. They observed that the ANN model gave better predictions than the ARIMA model.

A number of factors can contribute to drinking water discolouration. One of them is the dissipation of residual chlorine (a chemical that kills microorganisms or prevents their growth), which eventually leads to increased biological oxidation. Rodriguez, West, Powell and Serodes (1997) used ANN and traditional first-order modelling approach to predict residual chlorine in WDNs. They observed that the ANN model gave better predictions than the first-order model.

Many researchers have also used ANNs to predict raw water quality (DeSilets, Golden, Wang, & Kumar, 1992; Zhang & Stanley, 1997). Knowledge of the concentrations of incoming raw water quality variables such as turbidity, Fe, Mn, water colour, and coliform bacteria in advance is very important in drinking water treatment process, because it enables water utilities to optimise the treatment process to prevent inadequate or over-treatment of the raw water. For instance, insufficient chlorination in the treatment process can increase microbial re-growth, and subsequently cause waterborne diseases like typhoid fever, cholera, and hepatitis A. Over-chlorination can lead to an increase in customer complaints due to the taste and smell of chlorine. Zhang and Stanley (1997) developed a back-propagation ANN model that uses a five-year data set consisting of variables such as river flow rate, precipitation, and turbidity to forecast raw water colour. The ability of this ANN model to deal with multiple complex nonlinear input variables makes it an improvement on other conventional models. The model was able to reasonably predict all the peaks and recognise 355 out of 365 patterns. ANNs have also been used to forecast turbidity and colour removal through enhanced coagulation (Stanley, Baxter, Zhang, & Shariff, 2000), predict source water salinity (DeSilets et al., 1992), and forecast the dose of alum and polymers required for coagulation (Mirsepassi, Cathers, & Dharmappa, 1995).

Gautam (1999) used an auto-regression neural network (ARNN) model to predict the level of Lake Ijsselmeer at North Holland. The input variables for the model were wind speed, discharge of the lake, daily low tide water level, and water level of the sea. Data from two seasons were divided into two and used to train and verify the ARNN model. Figure 3.1 shows a graph of the observed (target) and predicted water levels by the ARNN model.

From the graph it was observed that the model was able to predict most of the peak levels of the lake.

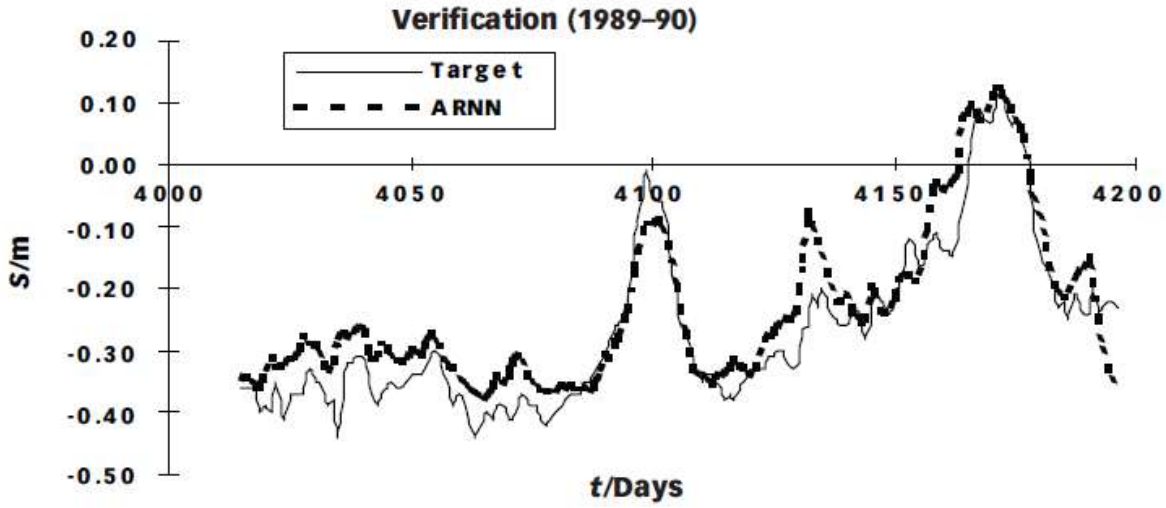


Figure 3.1 Graph of the observed (target) and predicted water levels by the ARNN model Gautam (1999)

Lint and Vonk (1999) developed an ANN model to predict water levels in reservoirs for South Holland Province Water Authority because the expert system they had previously used gave inaccurate results. The water authority needed to know the water levels of the reservoirs 24 hours in advance in order to optimise the pumping of water from high-level reservoirs to low-level reservoirs during night, when energy costs are cheaper. The input variables used for the model were pump status, precipitation, water level, and temperature at hourly intervals for the preceding 12 hours. Other input variables required by the model were one-hour-in-advance predicted temperature and precipitation. The output variable was 24-hours-in-advance water level of reservoirs at time steps of one hour. The model was able to predict water level with a coefficient of determination value of 0.71. In a related study, Raman and Sunilkumar (1995) used multivariate auto-regression (MAR) and an ANN model to predict monthly reservoir inflows at two sites in Kerala, India, namely, Mangalam and Pothundy reservoirs. A data set from the two reservoirs over a 14-year period was used to train and test the model. The four input variables used were two consecutive normalised monthly inflow values for each of the reservoirs. Table 3.1 shows the mean of the historic and generated inflow series by the ANN and MAR models. Comparing the two models, they observed that the ANN model generated better results than the MAR model.

Table 3.1 Mean of the historic and generated inflow series by the ANN and MAR models

	Mangalam reservoir			Pothimdy reservoir		
Month	Mean Historic data	ANN model	MAR model	Mean Historic data	ANN model	MAR model
January	1.254	1.397	1.412	4.57	4.123	5.464
February	0.641	0.643	0.962	0.91	0.808	1.088
March	0.2	0.21	0.178	0.503	0.443	0.578
April	0.413	0.392	0.641	0.531	0.489	0.901
May	0.898	0.744	1.199	0.836	0.848	0.752
June	12.528	12.507	13.19	8.519	8.594	9.668
July	24.275	24.416	23.131	15.543	15.616	15.691
August	24.452	24.091	24.432	15.615	15.148	14.457
September	10.63	10.324	11.323	5.14	5.134	5.124
October	9.644	9.993	7.998	4.18	4.029	4.072
November	5.633	5.787	4.929	4.683	4.268	4.627
December	1.477	1.778	1.88	4.479	4.633	4.486

3.2.3 How artificial intelligence models differ from traditional models

Despite the fact that models play an important role in water resource engineering, it is often very difficult to simulate the behaviour of natural systems. This is partly because of the complex non-linear nature of their data, and also the interactions that occur within many natural systems are poorly understood. Furthermore, water resource data often have skewed distributions, inter-related independent variables, and discontinuous functions, making traditional modelling methods difficult or impossible (Lingireddy & Brion, 2005). Therefore, AI computing approaches such as ANNs, Bayesian networks, and fuzzy logic have been used in recent year in simulations, forecasting, and predictions, especially in cases where traditional models fail. For example, image and speech recognition problems can be solved using AI models. However, they are beyond the scope of traditional models

Although AI models are increasingly being applied in water resource engineering, they are still viewed with scepticism by some researchers that use conventional statistical or mathematical models, due to their black-box nature (Kingston, 2006). While traditional (‘white-box’) models are seen as pre-supposed mechanism which are derived from prior understanding, some researchers have argued that ANNs should no more be viewed as black-box. Recently, some researchers have proposed various rule extraction algorithms from ANNs to explain or understand how the networks solve problems (Augasta &

Kathirvalavakumar, 2012; Bhalla, Bansal, & Gupta, 2012; Setiono, Baesens, & Mues, 2008). These rules can be grouped into three main methods: decompositional, pedagogical, and eclectic. In the decompositional methodology, rules are generated by inspecting the weights of each neuron in each layer. In the pedagogical technique, rules are formulated from the empirical analysis of the input-output pattern. Finally, in the eclectic methodology, rules are formulated from a combination of some elements of both decompositional and pedagogical methodologies (Bologna, 2001).

Traditional models make their computations using rules, formulae, and concepts. However, the underlying processes of many real-world problems are too complex to explain. Traditional models are ideal for solving numeric and analytic problems such as computing head losses due to friction in pipes using Hazen–William and Chezy–Manning equations, simulating water flows and water age (residence time) in pipes, and using hydrological transport models in flood forecasting. AI models make their predictions by learning relationships and patterns from the modelling data. In view of this, it is important to train every neural network with modelling data before it can be used to make predictions.

3.2.4 Structure of Artificial Neural Network

Because ANNs try to emulate the learning process of the human brain, it is a good idea to first look at the structure and mechanisms of the brain. The full mechanism of the brain remains a mystery, however some aspects of its functions are known. As neuroscience research advances and provides a better understanding of how the human brain works, researchers will engineer better solutions to problems that traditional methods of modelling cannot solve. The human brain has the ability to learn from examples, retain knowledge, and adapt to different conditions. This is normally referred to as experience. Humans acquire knowledge and build experience over time. For example, the brain is able to recognise familiar faces. Similarly, olfactory sensory neurons send messages to the brain to identify different types of smell. Although computers are very good at performing complex computations, they are very poor at recognising even simple patterns. Many animals have better pattern recognition capabilities than current computers.

The human brain is made up of many cells, of which neuron cells are the main functional units. Unlike other cells in the human body, neurons do not die and cannot be replaced, although they increase in size until they are around 18 years of age. However, neurons in

the brain can divide and form new cells during foetal development and early infancy (Williams & Herrup, 1988). It is estimated that the average human brain is made up of approximately 100 billion neurons and 100 trillion synapses of connections (Williams & Herrup, 1988). A lower estimate of 69 billion neurons has been published by Herculano-Horzel (2009). Each neuron is able to connect to about 200,000 other neurons. Neurons constitute three main components: dendrites (branched, arm-like filaments attached to the cell body which carry electrical signals to the cell body), cell body (containing the nucleus), and axons (very long slender projections which carry information away from the cell body). Figure 3.2 shows a schematic representation of a neuron cell. Basically, information is transmitted to and from the brain through a complex electro-chemical process. A schematic diagram of the flow of information from three neighbouring neurons is shown in Fig. 3.3. The arrows in the diagram show the flow of information from one neuron to another. The arrows pointing towards the dendrites indicate the signals received from sensory organs which are then transmitted to the neurons. Signals are also transmitted between neurons via the axons to the dendrites of other neurons through a special membrane called the synapse (highlighted by circles).

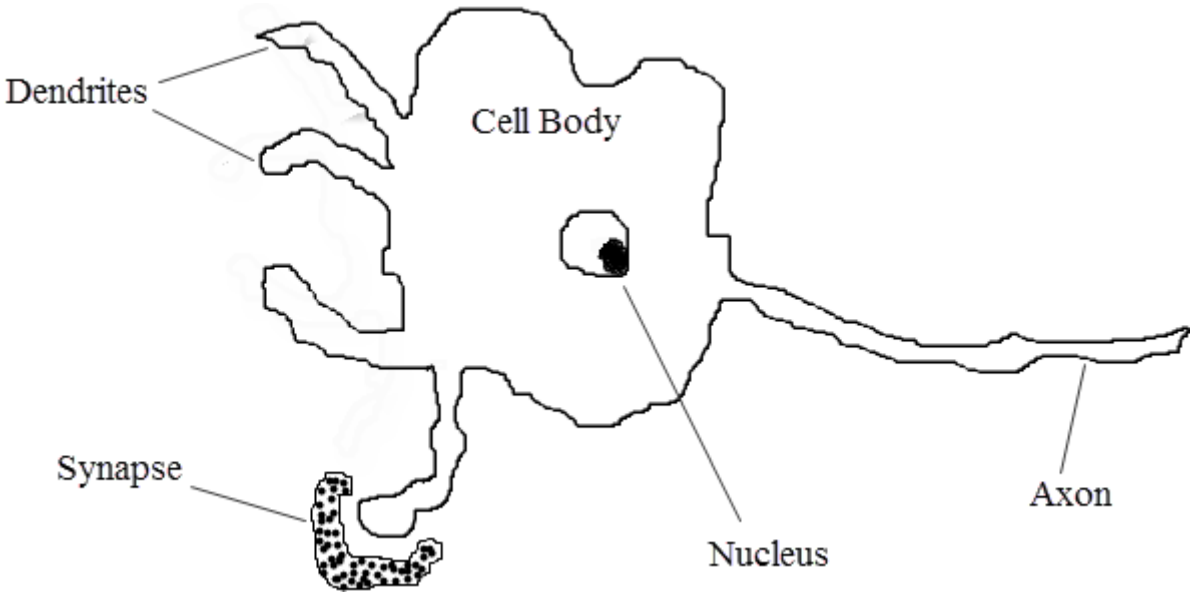


Figure 3.2 Schematic diagram of a neuron cell

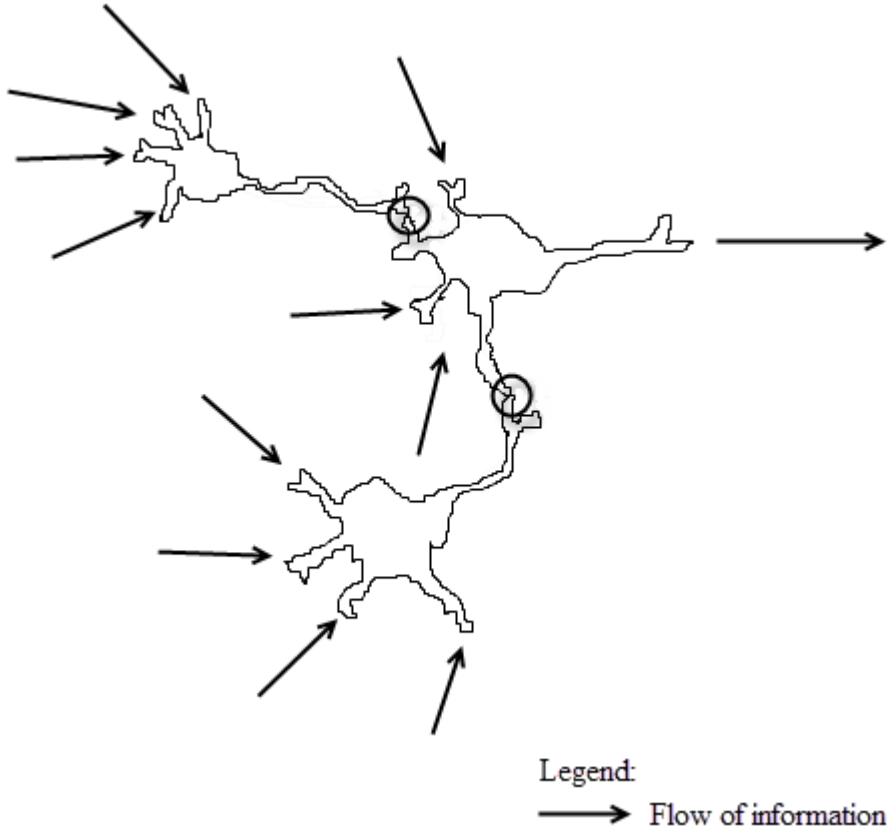


Figure 3.3 Schematic diagram of the flow of information from one neuron to another

3.2.5 Classification of Artificial Neural Networks

Researchers have classified ANNs in several ways. However, there are two common types of classifications that are often seen in literature. In the first type of classification, ANNs are categorised according to the arrangement of neurons and connection patterns. Neural networks in this type of classification can be further grouped into three categories, namely: feed-forward neural networks, feedback neural networks, and self-organised maps (Lobbrecht, Dibike, & Solomatine, 2002). Figure 3.4 shows a tree diagram of classification by arrangement of neurons and connection patterns. In the second classification, neural networks are classified according to their learning algorithm. They can be further categorised as ANNs with supervised learning algorithms (where networks learn from known input data to fit known output data) and ANNs with unsupervised learning algorithms (where networks organise known input data without any desired output data). Typical examples of supervised learning algorithms include back-propagation, ADALINE, and Boltzmann machines; whereas examples of unsupervised learning algorithms include counter-propagation, Hopfield networks, and adaptive resonance theory (Lobbrecht et al.,

2002). In the following sections, Back-propagation algorithm is discussed in detail as this method will be used in this research. Whereas other ANNs such as Radial basis, Boltzmann machine and Self-organising map ANNs are discussed briefly. For detailed information on these methods, the reader may refer to the book by Patterson (1996).

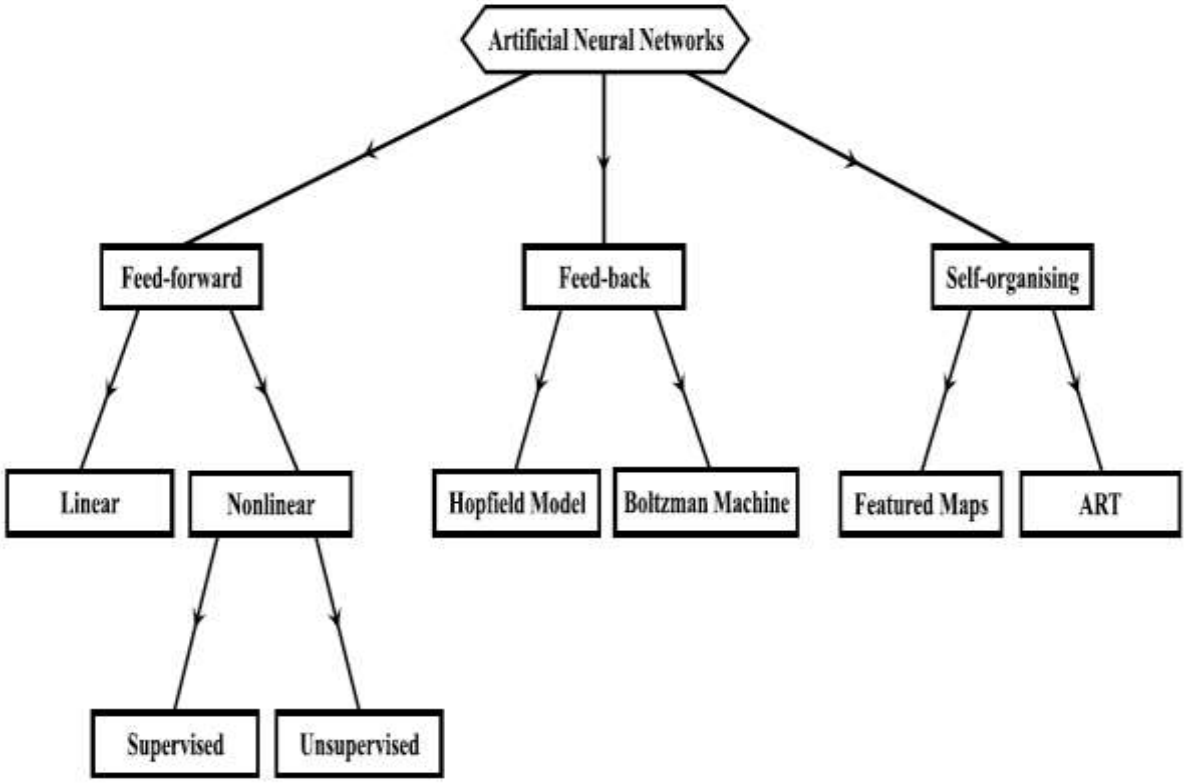


Figure 3.4 Classification of ANNs by arrangement of neurons and connection patterns

3.2.5.1 Back-propagation neural networks

Back-propagation neural network is a feed-forward neural network first introduced by Bryson and Ho (1969) for optimisation. However, it gained recognition after Rumelhart, Hinton and Williams (1986) used it in their research on learning procedures. Back-propagation neural networks are currently the most commonly used algorithm for training ANNs. The robust nature of back-propagation neural networks make them able to solve a broad range of problems. Some applications include weather prediction, speech recognition, water quality prediction, and credit scoring. Unlike the early single-layered ANNs (for example, ADALINE), which had a limitation of being able to solve only linear separable problems, back-propagation neural networks are able to solve both linear and non-linear separable problems.

A three-layered back-propagation neural network consisting of an input, hidden, and output layer will be used to explain the computations of the back-propagation algorithm (see Fig. 3.5). There is no limit to the number of hidden layers of a neural network. However, most problems require only one or two hidden layers. It is important to use an appropriate number of hidden neurons in back-propagation neural networks. Too many hidden neurons will cause overfitting; thus memorising the training data set instead of learning to generalise the trends within the data. Overfitted models have poor prediction performance. In contrast, if the hidden nodes are not enough, it can result in underfitting. Underfitted models have reduced learning capabilities and are too simple to solve problems.

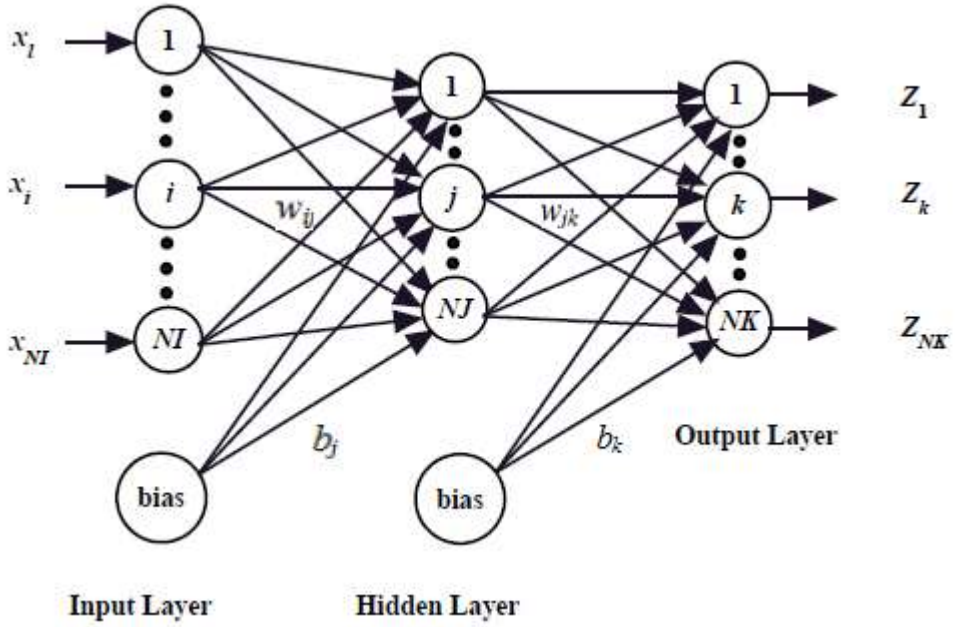


Figure 3.5 A three layered back-propagation neural network

Back-propagation neural networks requires known data for both the input and output nodes to make its computations. In view of this, it is classified as a supervised learning ANN. It also requires a differentiable activation function. Back-propagation neural networks operate in two modes, namely, the mapping and learning mode. During the mapping mode, input and target vectors are presented to the network, whereas weights and biases are randomly or specifically assigned to each of the connections during the learning mode. A bias is a constant that allows the activation function to be transformed either to the left or right during the learning process in order to bring the predicted vector close to the target

vector. The objective of back-propagation during the training process is to capture the underlining functional relationship between the input vector and the target vector, and reduce the error between the predicted and target vectors.

To explain the back-propagation process, let us use the i^{th}, j^{th}, and k^{th} node in the input, hidden, and output layers of the neural network, respectively for the computations. NI, NJ, and NK are the number of nodes in the input, hidden and output layers, respectively The input node passes a value (x_i) from the input vector to all the nodes in the hidden layer. The j^{th} hidden node then computes the weighted sum of the input values based on its weight (w_{ij}) (see Eqn. 3.1). The net output of the hidden node (h_j) is computed using a sigmoidal activation function (see Eqn. 3.2). Similarly, each of the output nodes receives inputs from the hidden nodes. The weighted sum of the k^{th} output node from the hidden nodes based on its weight (w_{jk}) is given in Eqn. 3.3. The net output of the output node (z_k) is computed using Eqn. 3.4. The error is calculated using Eqn. 3.5 to test whether predicted values are close to the target (measured) values. If the computed error is greater than a tolerance value, the error is back-propagated through the network. New weights and biases are then re-assigned and updated using a chain rule and an optimisation method known as gradient descent. The whole process is done iteratively to minimise the error and subsequently move the predicted values closer to the target values. The back-propagation process can be summarised in the following steps:

Step 1: Calculation of the net input of the j^{th} hidden node

If an input vector $X = (x_1, x_2, x_3, \ldots x_{NI})$ is applied to the input nodes, then the net input of the j^{th} hidden node (sum_j) is defined as the weighted sum of the connection from i^{th} input node to the j^{th} hidden node plus a bias term (b_j) applied to the hidden layer. It is mathematically expressed as:

$$sum_j = \sum_{i=1}^{NI} w_{ij} x_i + b_j \tag{3.1}$$

where w = connection weight; NI = the number of input nodes ($i = 1, 2, \ldots, I$); and
w_{ij} = connection between the i^{th} and j^{th} node

Step 2: Calculation of the net output of the j^{th} hidden node

The net output of the j^{th} hidden node (h_j) is computed using a sigmoidal activation function as follows:

$$h_j = \frac{1}{1 + e^{-sum_j}} \tag{3.2}$$

Step 3: Calculation of the weighted sum of the j^{th} hidden node to the k^{th} output node

The weighted sum of the k^{th} output node from the hidden nodes based on its weight (w_{jk}) is expressed as the weighted sum of the connection from j^{th} hidden node to the k^{th} output node plus a bias term. This can be calculated as follows:

$$sum_k = \sum_{j=1}^{J} w_{jk} h_j + b_k \tag{3.3}$$

where NJ = the number of hidden nodes ($j = 1, 2, ..., NJ$);

Step 4: Calculation of the net output of the k^{th} output node

The net output of the k^{th} output node (z_k) is computed using a sigmoidal activation function as:

$$h_k = \frac{1}{1 + e^{-sum_k}} \tag{3.4}$$

where b_k = bias term applied to the output layer; z_k = the net output of the k^{th} node; and w_{jk} = connection between the j^{th} and k^{th} node.

Step 5: Evaluation of the performance

The calculated error (E) is used to test the performance of the model. This can be expressed mathematically as:

$$E = \frac{\sum_{n=1}^{N}\sum_{k=1}^{NK} Y_{nk} - Y_{nk}^{\text{Pred}}}{2NK} \tag{3.5}$$

where N = the number of records (examples or data points) in the data set;

NK = the number of output nodes ($k = 1, 2, \ldots, NK$); and

Y_{nk}^{Pred} = predicted output of k^{th} node (output node) for n^{th} data sample.

Step 6: Weight adjustment

If the computed error is greater than a tolerance value, the error is back-propagated through the network and new weights are assigned. Using any of the gradient based optimisation methods (such as Levenberg–Marquardt method or Scaled conjugate gradient; see Section 3.2.6) the weights can be adjusted as follows

$$\Delta w(itr) = -\eta \frac{\partial E}{\partial w} + \mu \nabla w(itr - 1) \tag{3.6}$$

where η = learning rate. This is a parameter that ranges between zero and one that controls how fast a neural network learns. ANNs with low learning rates takes longer to train than those with high learning rates; Δw= change in weight;

μ = momentum value. This is a parameter that ranges between zero and one which is used to speed up convergence and maintain generalisation performance. An ANN without a momentum has a high risk of getting stuck in a local minimum; and

itr = iteration number. Iteration is the act of repeating a process with the aim of approaching a desired target.

3.2.5.2 Radial basis function neural networks

The application of radial basis function (RBF) in ANNs was initially introduced by Broomhead and Lowe (1988) at the Royal Signals and Radar Establishment. RBF neural networks are a type of feed-forward network which usually have three layers (single input, hidden, and output layers). Unlike other feed-forward neural networks, they cannot have more than one hidden layer. Even though RBF neural networks can use functions such as multiquadric and inverse multiquadric function as activation functions, the Gaussian

function is commonly used in this network. Each of the neurons in the hidden layer contains a Gaussian function which stores a prototype of one of the input vectors that is being classified. The ability of these functions to interpolate multi-dimensional scattered data makes them very useful. RBF networks can be trained faster than other multi-layered neural networks if there are not too many input variables (Hwang & Bang, 1997). They are able to efficiently approximate complex functions or data with a single hidden layer; a task that requires multiple hidden layers for back-propagation neural networks (Poggio & Girosi, 1990). However, RBF neural networks cannot model networks with many input variables (Lobbrecht et al., 2002).

Conceptually, RBF neural networks make predictions quite similar to how K-Nearest Neighbour models make prediction. They perform classification by measuring the input's similarity to examples from the training set. As an example, let us consider a RBF neural network which stores class A prototype as a class after training the network. When the network is presented with a new input, each neuron computes the Euclidean distance between the input and its prototype, and output a value between zero and one. If the input is equal to class A prototype, the output of that RBF neuron will return a value one and therefore classified as class A, since they are similar. As the Euclidean distance between the input and prototype grows, the output of that RBF neuron reduces exponentially towards zero. A diagram showing the centre and width of a RBF is presented in Fig. 3.6.

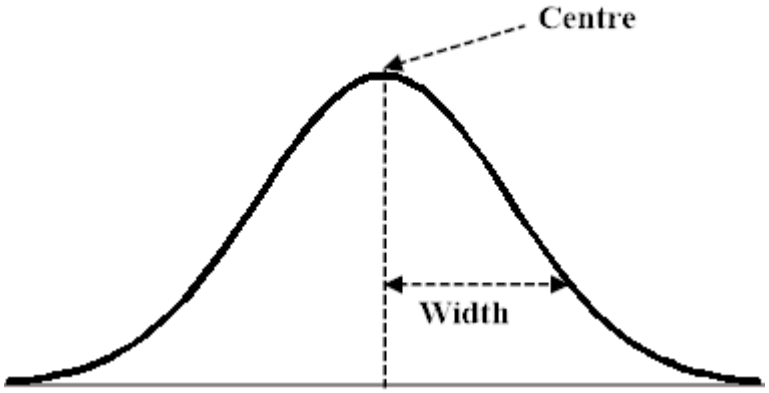


Figure 3.6 Radial basis function

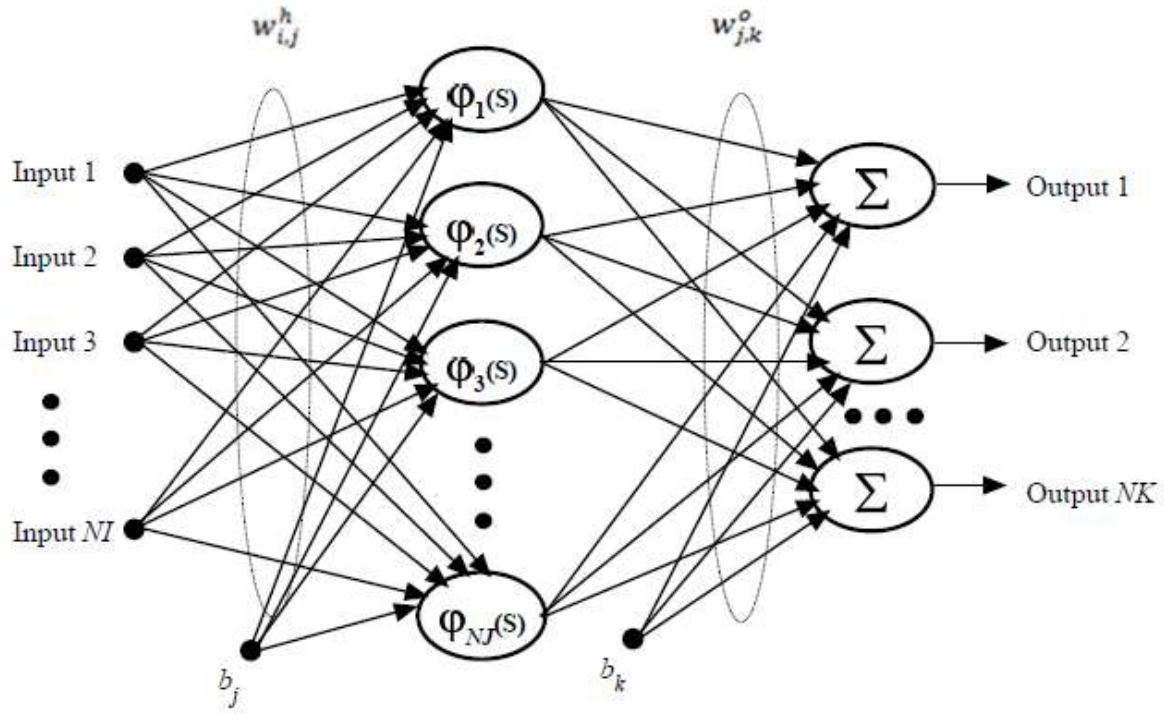


Figure 3.7 RBF network with NI inputs, NJ hidden neurons, and NK outputs

RBF neural networks are sometimes referred to as hybrid neural networks, because they use both supervised and unsupervised learning algorithms in their computations. The unsupervised algorithm determines the centres and width of the RBF, while the supervised algorithm computes the weights. A diagram showing a RBF network with *NI* inputs, *NJ* hidden neurons, and *NK* outputs is presented in Fig 3.7. The algorithm for training these types of networks can be summarised as follows:

i. The net input (s_j) of the hidden neuron (j) from the input vector $X = (x_1, x_2, x_3, \ldots x_{NI})$ is calculated as:

$$s_j = [x_1 w^h_{1,j}, x_2 w^h_{2,j} \ldots x_i w^h_{i,j} \ldots . x_{NI} w^h_{NI,j}] \quad (3.7)$$

where *NI* = number of input neurons; j = index of hidden neurons;

$w^h_{i,j}$ = input weight between input neuron i and hidden neuron j in at the hidden layer h; and $x_i = i^{th}$ input neuron.

ii. Define the number of clusters E_{cl}, where each cluster is represented a hidden neuron. Determine the centres of the clusters using an unsupervised learning algorithm (clustering technique). This is calculated as:

$$c_j = \frac{1}{E_{cl}} \sum_{e=1}^{E_{cl}} x_e \tag{3.8}$$

where cl = centre of cluster; E_{cl} = total number of clusters; and x_e = eth cluster.

iii. Calculate the width of the radial centre of each of the hidden neurons.

This is calculated as:

$$\sigma = \frac{1}{N} \sum_{i=1}^{N} \| x_i - cl \| \tag{3.9}$$

where N = the number of training samples in the cluster;

x_i = the ith training samples in the cluster; and σ = width of the radial centre.

iv. Compute the output from the hidden neurons. The output of the hidden neuron *j* is calculated as:

$$\varphi_j(s_j) = exp\left(- \frac{\| s_j - cl_j \|^2}{\sigma_j} \right) \tag{3.10}$$

where $\varphi_l(.)$ = the Gaussian activation function for hidden neuron *j*.

v. Compute the output from the output neurons. The output neuron *k* is calculated as:

$$o_k = \sum_{j=1}^{NJ} \varphi_j(s_j)\, w_{j,k} + w_k \tag{3.11}$$

where *m* = index that denotes the output neuron *m*;

o_k = output from the output neuron; w_k = the bias weight of output unit *k*;

$w_{j,k}$= the weight between hidden neuron *j*; and output neuron *k*.

vi. Calculate the error between the computed output and the target.

vii. If the computed error is greater than a tolerance limit, adjust the variables in the RBF neural network and repeat the above steps until the stopping criterion is met.

3.2.5.3 Recurrent neural networks

Recurrent neural networks (RNNs) are a type of feedback ANN in which the units contain at least one connection to form a directed cycle. The units in RNNs can either be fully connected (Fig. 3.8(a)) or partially connected (Fig. 3.8(b)). In partially connected RNNs, only some of the units are connected concurrently. Whereas in fully connected RNNs, all the units are connected concurrently making it impossible to apply a back-propagation

algorithm. The feedback connection feature in RNNs makes them dynamic, thus enabling it to solve both continuous and discrete time-dependent problems. For a partially connected RNN, selecting an appropriate number of layers and hidden nodes is very important prior to training the network. Partially connected RNNs can be trained using population-based methods such as evolutionary algorithms and particle swarm optimisation. These methods have the ability to perform both the parameter learning and structure learning simultaneously during the training process (Shin & Xu, 2009).

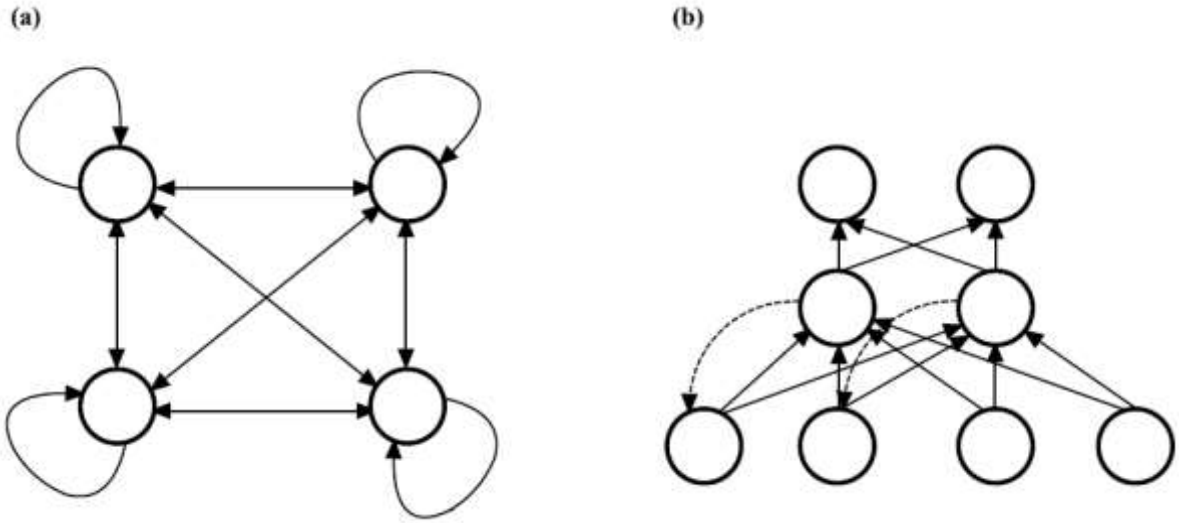


Figure 3.8 The architecture of a: (a) fully connected and (b) partially connected RNN

Some applications of RNN include learning formal grammar, music composition and speech recognition. Despite their versatility, they have some limitations: it is sometimes very difficult to determine a suitable architecture, number of units, and number of time lags. These difficulties in choosing optimal properties can result in poor predictions from RNNs (Lobbrecht et al., 2002). For detailed information on RNN, the reader may refer to the book by Patterson (1996).

3.2.5.4 Boltzmann machine

Boltzmann machine is a stochastic RNN that was first developed by Ackley, Hinton and Sejnowski (1985). Its mode of operation is quite similar to that of Hopfield network. They both have fully connected networks that are trained by minimising their energy state. However, the network topology of Boltzmann machine differs from that of Hopfield network. The neurons in Boltzmann machine have hidden neurons, whereas there are no hidden neurons in Hopfield network. In a Boltzmann machine, neurons are grouped into

input, hidden, and output neurons. Figure 3.9 shows a schematic diagram of Boltzmann machine with 3 hidden nodes (grey circle) and 4 visible nodes (white circle). To train a Boltzmann machine, the network is run from a high temperature and decreased gradually to a low temperature using Eqn. 3.12. The network repeatedly cycles through the states until it reaches a steady state. This process is known as simulated annealing. The Boltzmann machine has a relatively slower learning rate compared to back-propagation ANN. It is also sometimes difficult to adjust the temperature and determining the equilibrium state during the simulated annealing process.

$$p_i = \frac{1}{\left[1 + exp\left(\frac{\Delta E_i}{T_{Bolt}}\right)\right]} \qquad (3.12)$$

where p_i = probability of the i^{th} neuron; T_{Bolt} = temperature; and ΔE = global energy.

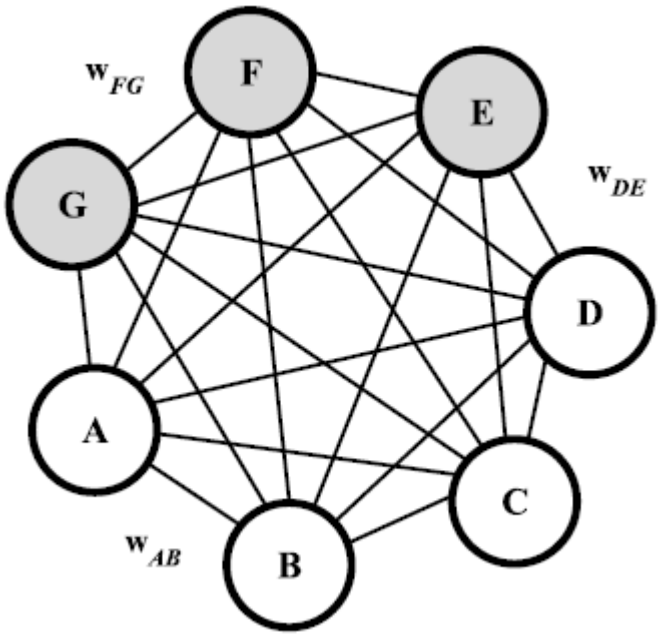


Figure 3.9 Graphical representation of Boltzmann machine

3.2.5.5 Self-organising maps

Self-organising maps (SOMs) were developed by Kohonen (1990), and are therefore sometimes referred to as Kohonen networks. SOMs can be divided into two, namely, adaptive resonance theory (ART) and self-organising feature maps (SOFM). The ART uses an unsupervised learning algorithm that takes an input vector which consists of one-

dimensional array of values and transfers it to its best match in a recognition field. It has the ability to recognise previously learnt categories or create new categories if it is presented with new analogue or binary input vectors. SOFMs use unsupervised learning algorithm to project high-dimensional data to a one-, two-, or three-dimensional data space (known as maps) while preserving key features of the input space. SOFMs consist of an input layer, connection weights, and an output layer. Figure 3.10 shows the structure of a SOFM. The generated maps have reduced dimensions of groupings of similar data items. SOFMs are instrumental in solving problems with high-dimensional data. Such problems cannot be solved by humans, as we cannot visualise data with high dimensions. Some applications of SOMs include pattern recognition, clustering, speech recognition, and market segmentation (for example, grouping customers according to their buying criteria). An advantage of SOMs is that the output results can be easily understood and interpreted, but a major drawback is that they require many data to develop meaningful clusters (Sonali, 2014).

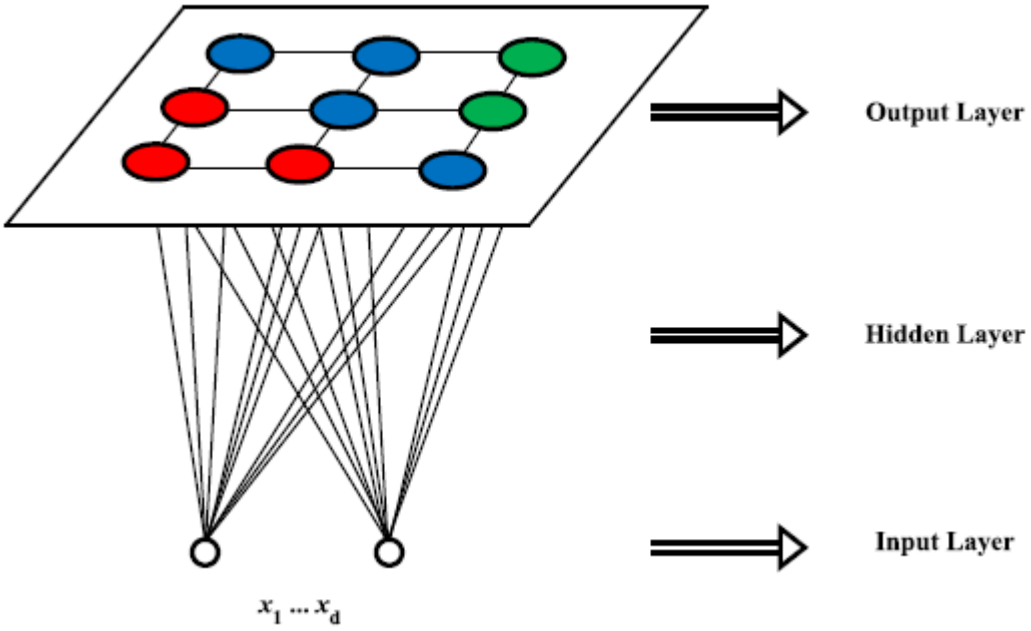


Figure 3.10 Graphical representation of a self-organising feature map

Unlike other ANNs, SOFMs use a neighbourhood function which does not change the topological features of the input space. They transform data in two main modes: through vector quantisation to train the input data, and a mapping for the classification. The learning algorithm for SOFM is listed below:

i. Assign random weights to the network. The connection weight between the input neuron i at the input layer and neuron k at the output layer can be expressed as:

$$W_i = \{w_{i,k}: i = 1, \dots, NI;\ k = 1, \dots, NK\} \tag{3.13}$$

where NI = number of neurons in input layer; and

NK = total number of neurons output layer

ii. Assign the input vectors (patterns) to the network. The input vector X is given as:

$$X = \{x_i:\ i = 1, \dots, NI\} \tag{3.14}$$

iii. Select the best matching (winning) node. This is done by calculating the Euclidean distance (E) between the input vector X and the weight vector W_i for each neuron i using Eqn. 3.15. The node with the shortest Euclidean distance is selected as the winning node. This node exhibits the greatest similarity with the input vector.

$$E(X) = \sqrt{\sum_{i=1}^{NI} (x_i - w_{ik})^2} \tag{3.15}$$

iv. The winning node and its neighbouring nodes are updated using Eqn. 3.16. This equation adjusts the weight of the winning node and its neighbours towards the input vectors in order to preserve the topology of the map.

$$w_{ik}(NT + 1) = w_{ik}(NT) + \eta(NT)F(NC, r^*)[x_i - w_{ik}(NT)] \tag{3.16}$$

where NT = number of iterations; $F(NC, r^*)$ = neighbourhood function;

NC = number of cycles; η = learning rate with a value between 0 and 1 that controls how fast the SOFM learns; r^* = neighbourhood radius;

$w_{ik}(NT + 1)$ = new connection weight between the input node i and output node k; and

$w_{ik}(NT)$ = old connection weight between the input neuron i and output node k.

v. Repeat steps ii.–iv. until a stable network configuration is attained.

3.2.6 Training of artificial neural networks

ANNs have three important components; namely, connection weights, summation function (aggregate function), and transfer function. The ANN receives one or more inputs, and multiplies each input by its connection weight. These weights are numerical values that are initially randomly assigned to each node, and are then adjusted during the training process. In ANN modelling, training is the process whereby the connection weights are adjusted until the predicted values are close to the measured values. This adjustment makes some of the independent variables more significant than others in the prediction of the dependent variable. The products of each input and connection weight are aggregated and passed through an activation (transfer) function. The activation function of a node defines the output node for a given set of input data.

The three most commonly used activation functions during the training process in ANNs are the sigmoid (logistic), hyperbolic (tanh), and linear activation functions (Zeng, 1999). Figure 3.11 shows graphs of these functions. The sigmoid activation function is the most widely used activation function (Ozkan & Erbek, 2003). It takes any input value and returns an output value between 0 and 1. If an ANN has more than one hidden layer, the output values serve as input values to the output layer, which is transformed again using a transfer function into output values between 0 and 1. A mathematical representation of the sigmoid activation function is shown in Eqn. 3.17. The hyperbolic activation function is the second most popular activation function (Ozkan & Erbek, 2003). This function takes the input values and returns output values in the range [-1, +1]. Equation 3.18 shows a mathematical representation of the hyperbolic activation function. A mathematical representation of the linear activation function is presented in Eqn. 3.19. This function gives an output that is linearly proportional to the input. As a result, it is unable to solve non-linear problems. This was one of the limitations of the early ANN models.

$$y_o = \frac{1}{1 + e^{-x_{in}}} \tag{3.17}$$

$$y_o = \frac{e^{x_{in}} - e^{-x_{in}}}{e^{x_{in}} + e^{-x_{in}}} \tag{3.18}$$

$$y_o = m_s x_{in} \tag{3.19}$$

where m_s = slope of the line; x_{in} = input neuron; and y_o = output neuron.

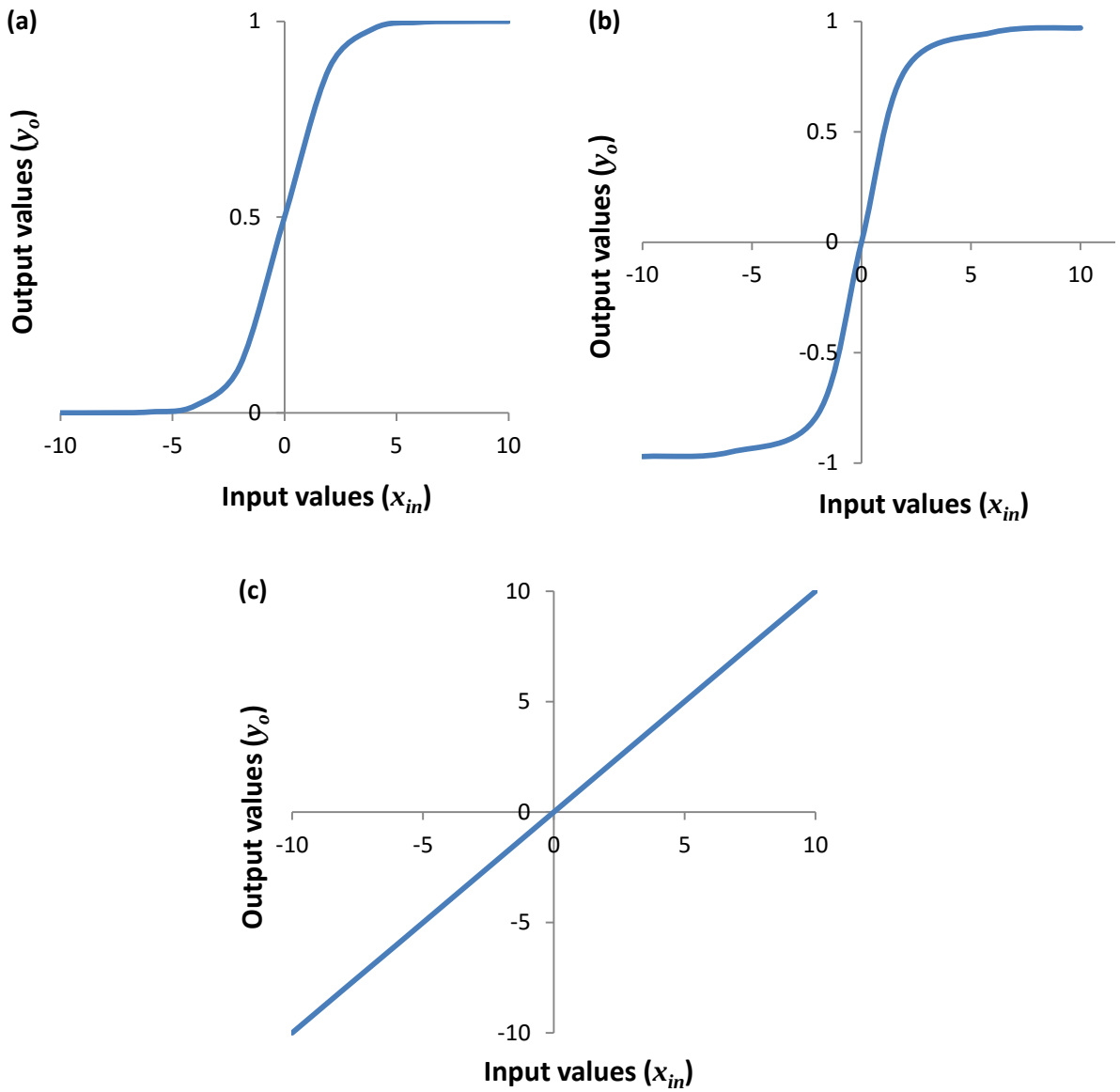


Figure 3.11 (a) Sigmoid (b) Hyperbolic and (c) Linear activation function

ANNs use optimisation algorithms to optimise connection weights and biases. Optimisation help to minimise the error in the objective function. Some common objective functions are mean square error (MSE) and sum of square error (SSE). There are several optimisation algorithms that can be used to optimise ANNs. They include the Levenberg–Marquardt (LM) algorithm, which updates weights and biases according to LM non-linear optimisation; scaled conjugate gradient back-propagation, which updates weights and bias values based on the scaled conjugate gradient method; and Bayesian regulation back-propagation, which also uses LM optimisation to update weights and biases. Other customised optimisation algorithms have been developed to improve the performance of ANNs. These include the artificial bee colony algorithm (Shah, Ghazali, Nawi, & Deris,

2012), hybrid particle swarm optimization–back-propagation (Zhang, Lok, & Lyu, 2007), and particle swarm optimisation (Mendes, Cortez, Rocha, & Neves, 2002).

3.2.6.1 Levenberg–Marquardt algorithm

The LM algorithm is used to solve non-linear least squares problems in non-linear and ANN models. It has the ability to find solutions to problems even if it starts very far off the final minimum. LM algorithm can be presented mathematically as shown in Eqn. 3.20. It is able to switch between two algorithms; the Gauss–Newton algorithm and the gradient descent algorithm during the training process. If the combination coefficient μ_k is very small (approaching zero), Eqn. 3.20 approximates to Gauss–Newton algorithm which is represented by Eqn. 3.21. On the other hand, if μ_k is very large, it can be interpreted as the learning coefficient in the gradient descent method. Equation 3.20 then approximates to Eqn. 3.22 and uses the gradient descent method.

$$w_{k+1} = w_k - (JM_k^T JM_k + \mu_k I_m)^{-1} JM_k e_k \tag{3.20}$$

$$w_{k+1} = w_k - (JM_k^T JM_k)^{-1} JM_k e_k \tag{3.21}$$

$$w_{k+1} = w_k - \frac{1}{\mu_k} grad_k \tag{3.22}$$

where μ_k = a positive integer known as combination coefficient; I_m = identity matrix; *grad* = gradient, which is the first-order derivative of the total error function;
JM = Jacobian matrix; *e* = training error; *k* = the index of iteration; w = weight vector; and ^{T} = transposition.

3.2.6.2 Scaled conjugate gradient algorithm

The scaled conjugate gradient algorithm was developed by (Moller, 1993) to avoid the time-consuming line search methods of optimisation. This algorithm chooses the search direction and the step size by using information from second order approximation. An algorithm for scaled conjugate gradient algorithm is presented in Algorithm 3.1

Algorithm 3.1 Scaled conjugate gradient algorithm

a. Choose the initial weight vector w_1 and set $k = 1$

$$p_1 = r_1 = -E'(w_1) \tag{3.23}$$

where k = index of the iteration; r_1 = initial steepest descent direction

p_1 = initial search direction; w_1 = initial weight vector;

$E(w)$ = global error function that depends on all the weights and biases; and

$E'(w)$ = the gradient of the global error function.

b. Calculation of the second-order derivative δ_k during the k^{th} iteration is given by:

$$\delta_k = p_k^T E''(w_k) p_k \tag{3.24}$$

where ^{T} = transposition; p_k = search direction during the k^{th} iteration; and

w_k = weight vector during the k^{th} iteration.

c. Calculate the step size α_k:

$$\mu_k = p_k^T r_k \tag{3.25a}$$

$$\alpha_k = \frac{\mu_k}{\delta_k} \tag{3.25b}$$

where r_k = steepest descent direction during the k^{th} iteration.

d. Update the weight vector

$$w_{k+1} = w_k + \propto_k p_k \tag{3.26}$$

$$r_{k+1} = -E' w_{k+1} \tag{3.27}$$

e. If mod $N_{itr} = 0$, then

restart algorithm

$$p_{k+1} = r_{k+1} \tag{3.28}$$

else

$$\beta_k = \frac{|r_{k+1}|^2 - r_{k+1} r_k}{\mu_k} \tag{3.29}$$

where $1 \leq k \leq N_{itr}$; and N_{itr} = number of iterations.

f. If $r_k \neq 0$ then

Set $k = k + 1$ and go to step b

Else

Terminate and return w_{k+1} as desired minimum.

(Moller, 1993)

3.2.6.3 K-fold cross-validation

There are several types of ANN models that can be used to solve different types of problems. Each has its own advantages and disadvantages. For instance, RBF neural networks are not able to model networks with many input variables (Lobbrecht et al., 2002). Despite the versatility of recurrent neural networks, it is sometimes very difficult to choose optimal parameters, which may result in poor predictions (Lobbrecht et al., 2002). The Hopfield networks also have the limitation of not being able to converge if too many patterns are stored (Lippmann, 1987). Back-propagation neural networks are one of the most commonly used ANNs. They are very flexible to use. Some advantages of back-propagation neural networks listed by Priddy and Paul (2005) are:

- They are easy to use and implement.
- They can solve a wide range of problems.
- They have the ability to solve complex non-linear problems.

In ANN modelling, the data are generally partitioned into training, validation, and testing sets. During the training process, ANNs usually require large amounts of training data to capture variations in the entire search space. This helps to prevent the model from overfitting, i.e. memorising the training data instead of learning to generalise the trends in them. There are two main reasons why overfitting occurs in ANNs: either the training data set is too small, or the input and hidden nodes are too many.

Fields such as aviation, marine benthic ecology, and disease diagnosis often have limited data sets. Under such circumstances, the scarcity of data often results in biased partitions. When this occurs, the training data set may not have enough data to make a generalisation. Since the data set used for the model development in this research was relatively small, it was very important to have a suitable approach for dividing the data into training, validation, and testing sets while still maintaining a high level of confidence in the results. A number of researchers have used different ANN methods to solve problems with small data sets. Li, Chen and Lin (2003) combined the functional virtual population method with an ANN model to solve dynamic manufacturing problems with small data sets. Mao, Zhu, Zhang and Chen (2006) also proposed a posterior probability technique that estimates missing data in small data sets used for ANN modelling. K-fold cross-validation is the most common method used to form unbiased partitions of small-sized data sets. This

method has been successfully applied by researchers in a broad range of disciplines, including biological sciences, chemistry, and water resource engineering, in instances where the modelling data set is small (Shahriari & Shahriari, 2014; Singh & Gupta, 2012; Zhang, Wang, Ji, & Phillips, 2014).

In the K-fold cross-validation, the data set used in the modelling is randomly divided into K distinct equal subsets. K-1 subsets are used for the training, while the remaining subset is used to validate the performance of the ANN model. The process is repeated K times, each time using a different K subset to validate the performance of the model, while using the remaining K-1 subsets to train it. The model performance indicator (objective function), root mean square error (RMSE) is averaged across the K trials for both the training and validation data sets as suggested by Hanrahan (2011). A diagram to illustrate how the training and validation subsets were divided and applied in the model is presented in Fig. 3.12.

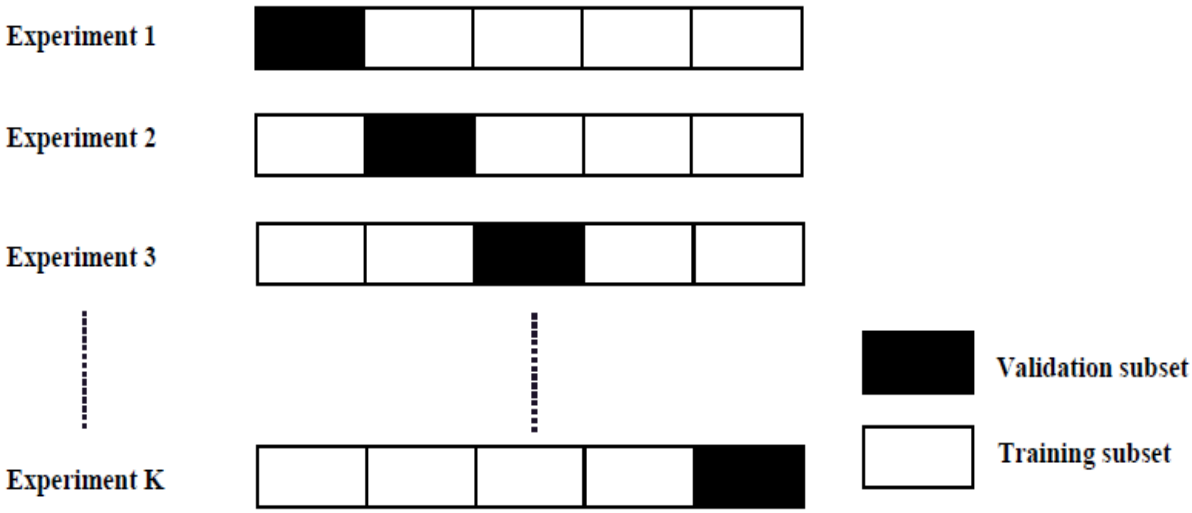


Figure 3.12 Schematic depiction of the K-fold cross-validation method

Just like any other methodology, K-fold cross-validation has its own benefits and limitations. An advantage of this method is its ability to reduce lucky and unlucky splits (biased partitions) by using each of the randomly partitioned subsets for training and validation exactly once. Since the model has to be run K times, it can be slow and requires considerably high computational resources to train the network.

3.2.7 Advantages and disadvantages of artificial neural networks

Researchers have created the impression that ANNs can solve every problem. This misconception has caused disappointments if users fail to obtain good results from ANNs. Just like any other model, ANNs offer a number of advantages and limitations. These include the following:

3.2.7.1 Advantages of artificial neural networks

- They can detect linear or complex non-linear relationships between independent and dependent variables (Hinton, 1992). This is a distinct advantage over traditional statistical methods.
- They can learn from observed examples by adjusting their internal weights to reduce the error between the desired output and the actual output. The three major algorithms used by ANNs for learning are supervised learning, unsupervised learning and reinforcement learning.

3.2.7.2 Disadvantages of artificial neural networks

- They are frequently referred to as black-boxes because it is difficult to understand their internal operation.
- They require considerable computational resources to simulate.
- They are prone to overfitting, which may lead to poor predictive performance. Overfitting occurs when a model memorises training data instead of learning to generalise from trends.
- They require a large amount of data to train, validate and test a model.

3.3 Fuzzy inference system

Fuzzy logic is a kind of logical system that deals with reasoning that is approximate rather than crisp or precise. It can be used to translate sophisticated statements from natural language (qualitative knowledge) into mathematical formalisms (numerical reasoning). Fuzzy logic is useful for finding precise solutions from vague, ambiguous, or uncertain data (McNeill & Thro, 1994). Instead of mathematically modelling complex data, fuzzy logic incorporates rules that include words such as ‘IF’, ‘AND’, and ‘THEN’.

3.3.1 Historical background of fuzzy inference system

It is believed that the first person to use fuzzy logic was Gautama Buddha. He was born in about 563 BC, and is the founder of Buddhism. His philosophy was full of grey statements (what the Western world would describe as contradictions). He believed that a statement could be both true and false at the same time. In other words, something can be X and not-X simultaneously. Some 200 years later, Buddha's philosophy was sharply refuted by a Greek scholar, Aristotle. Aristotle believed there were no grey areas in things pertaining to the world. He believed that they were either true or false, black or white, hot or cold, X or not-X. These two philosophies spread independently. Buddhism was accepted by Indians, whereas Aristotle's philosophy was accepted by Greek scholars, and later by the Western world. Aristotle's philosophy was later proofed by logic and accepted by scientists. He is credited with the development of formal logic (Degnan, 1994), and it is therefore sometimes referred to as Aristotelian logic or binary logic.

Aristotelian logic ruled the Western world for over 2,000 years. During this time, the scientific community found it very difficult to embrace the concept of uncertainty or fuzzy reasoning. They believed uncertainties were detrimental, and should be avoided as much as possible. In the late 19th century, scientists began to realise that some problems could not be addressed by crisp theories and laws that did not consider uncertainty. For example, they found that Newtonian mechanics could not solve problems at the molecular level. As a result, they started replacing Newtonian mechanics with statistical mechanics, which could be explained by probability theory; a theory that captures some level of uncertainty (Ross, 2010). In 1923, Jan Lukasiewicz, a Polish philosopher and logician who many regard as one of the most important historians of logic, introduced the theorem of multi-valued logic. He established a relationship between his theorem and the traditional Aristotelian logic (Lukasiewicz, 1963).

During the early 20th century, probability theory was the leading concept for describing uncertainty. However, this theory was challenged by Max Black, a British-American philosopher who lectured at Cornell University. In 1937, he published a paper on vagueness in which he proposed a logic for vague terms (Black, 1937). In 1965, Lotfi Askar Zadeh, a professor in electrical engineering, developed a continuous-valued logic which he termed fuzzy set theory (Zadeh, 1965). Unlike Aristotelian logic and probability theory, fuzzy logic was able to address several uncertainties. This was a significant step

towards solving problems with uncertainties. Zadeh is credited with the invention of fuzzy logic. In fuzzy set theory, the sets consist of members that have various levels of belonging, known as memberships. These memberships are defined over a universe of discourse called membership functions (Zadeh, 1965).

After Zadeh's introduction of fuzzy set theory, a number of researchers have made significant contributions to the development of fuzzy logic. One of the significant contributions is the industrial application of fuzzy logic in engineering systems control. Ebrahim H. Mamdani, a British professor who lectured at Imperial College, was the first to apply fuzzy logic in this area. He showed how fuzzy logic can be used to control dynamic plants (Mamdani, 1974). Sugeno (1985) introduced a similar fuzzy model to Mamdani's model to control systems. Sugeno's model is still in use today, and is regarded as one of the suitable tools for modelling non-linear systems. The main difference between these two models is that the output of Mamdani's model is a fuzzy set, whereas the output membership function of Sugeno's model is linear.

3.3.2 Fuzzy set concepts

In mathematics, a fuzzy set is a set with elements that have varying degrees of membership. This concept directly contradicts with crisp sets, where the elements in their set have full membership. A crisp set is defined by a bivalent truth function which assigns a value of either 0 or 1 to each element of the universe of discourse. This means that an element is either a full member of a set or is not at all. Some examples of crisp sets are: a set of odd numbers or set of even numbers. In such sets, the boundaries are precisely defined. Hence, it is easy to determine which set a given element belongs to.

In real life, there are many instances where judgement or evaluation of information cannot be precise. In such instances, there are ambiguity and uncertainty in the evaluation. For example, it will be difficult to precisely define the boundaries of a class of intelligent students. This is because the word intelligent is relative; what someone may classify as intelligent, another may classify as not intelligent. Fuzzy sets provide a mathematical way of representing this uncertainty by allowing partial memberships with intermediate values between 0 and 1. The following definitions explain the basis of fuzzy set operations.

Membership function

A membership function $\mu_A(x)$ in a fuzzy set A is a curve that defines how each element x in the universe of discourse U is mapped to a membership value between 0 and 1 for all $x \in U$. The membership function is maximum when $\mu_A(x) = 1$ and minimum when $\mu_A(x) = 0$. At the maximum or minimum value, the fuzzy set becomes a crisp set.

Intersection

The intersection of two fuzzy sets A and B denoted by $A \cap B$ for all elements of x in the universe of discourse U is defined by Eqn. 3.30.

$$A \cap B = \mu_{A \cap B}(x) = min[\mu_A(x), \mu_B(x)] \qquad for\ all\ x \in U \tag{3.30}$$

Union

The union of two fuzzy sets A and B denoted by $A \cup B$ for all elements of x in the universe of discourse U is defined by Eqn. 3.31.

$$A \cup B = \mu_{A \cup B}(x) = max[\mu_A(x), \mu_B(x)] \qquad for\ all\ x \in U \tag{3.31}$$

Compliment

The complement of the fuzzy set A is denoted by $\bar{A}$ for all elements of x in the universe of discourse U is defined by Eqn. 3.32.

$$\bar{A} = \mu_{\bar{A}}(x) = 1 - \mu_{\bar{A}}(x) \qquad for\ all\ x \in U \tag{3.32}$$

Equal sets

Two fuzzy sets A and B are said to be equal if $\mu_A(x) = \mu_B(x)$ for all elements of x in the universe of discourse U.

Subset

A fuzzy set A is a subset of a fuzzy set B if $\mu_A(x) \leq \mu_B(x)$ for all elements of x in the universe of discourse U.

Empty set

A fuzzy set *A* is said to be empty if $\mu_{\mathrm{A}}(x) = 0$ for all elements of *x* in the universe of discourse *U*.

De Morgan's laws

De Morgan's laws states that the negation of a conjunction of two fuzzy sets *A* and *B* is the disjunction of the negations of the sets, and the negation of a disjunction in the sets is a conjunction of the negations of the sets. The laws can be expressed as:

$$\overline{A \cap B} = \overline{A} \cup \overline{B} \tag{3.33}$$

$$\overline{A \cup B} = \overline{A} \cap \overline{B} \tag{3.34}$$

Commutative laws

Two fuzzy sets *A* and *B* are said to be commutative if:

$$A \cup B = B \cup A \tag{3.35}$$

$$A \cap B = B \cap A \tag{3.36}$$

Distributive laws

Three fuzzy sets *A*, *B* and *C* are said to be distributive if:

$$(A \cup B) \cup C = B \cup (A \cup C) \tag{3.37}$$

$$(A \cap B) \cap C = B \cap (A \cap C) \tag{3.38}$$

3.3.3 Types of fuzzy membership functions

A membership function in a fuzzy set defines the degree to which an element in the universe of discourse maps to a membership value between 0 and 1. The horizontal axis defines an input variable or element, and the vertical axis represents the value of the membership function. Membership functions are used to map non-fuzzy inputs to fuzzy outputs and vice versa. The following sections discuss some types of fuzzy membership functions that can be represented in fuzzy sets.

3.3.3.1 Triangular membership function

The elements *x* in a triangular membership function $\mu(x)$ is specified by three parameters (*A*, *B*, and *C*), where (A < B <C). Figure 3.13 shows a triangular membership function with parameters A = 2, B = 5, and C = 8. It can be mathematically expressed as:

$$\mu(x) = triangle(x; A, B, C) = max\left(min\left(\frac{x-A}{B-A}, \frac{C-x}{C-B}\right), 0\right) \qquad (3.39)$$

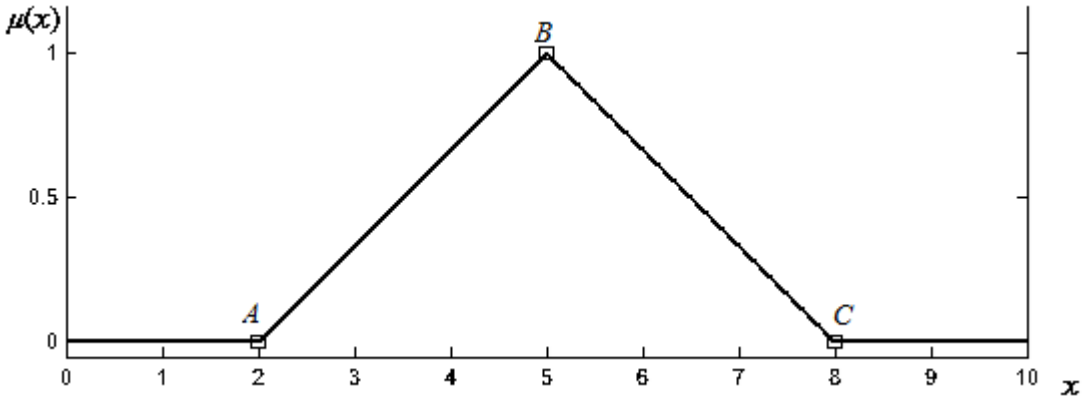


Figure 3.13 Triangular membership function

3.3.3.2 Trapezoidal membership function

The elements *x* in a trapezoidal membership function $\mu(x)$ can be specified by four parameters (*A*, *B*, *C*, and *D*), where ($A \leq B \leq C \leq D$). Figure 3.14 shows a trapezoidal membership function with parameters $A = 2$, $B = 3$, $C = 7$, and $D = 8$. The membership function can be represented mathematically as:

$$\mu(x) = trapezoid(x; A, B, C, D) = max\left(min\left(\frac{x-A}{B-A}, 1, \frac{D-x}{D-C}\right), 0\right) \qquad (3.40)$$

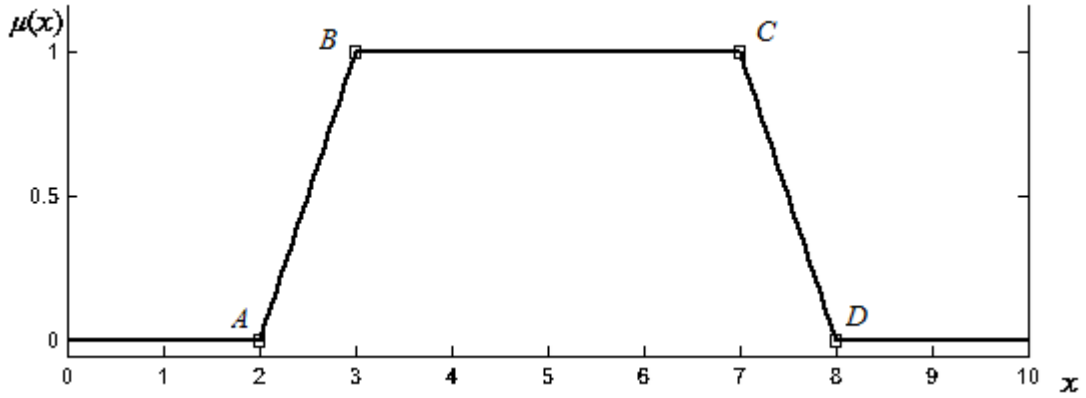


Figure 3.14 Trapezoidal membership function

3.3.3.3 Gaussian membership function

The elements x in a Gaussian membership function $\mu(x)$ is specified by two parameters (C and σ). The parameters σ_{gau} and c represent the centre and width of the function, respectively. Figure 3.15 shows a Gaussian membership function with parameters $\sigma = 1$ and $C = 5$. The function can be represented mathematically as:

$$\mu(x) = gaussian(x; \sigma, C) = e^{\left(\frac{-(x-C)^2}{2\sigma^2}\right)} \tag{3.41}$$

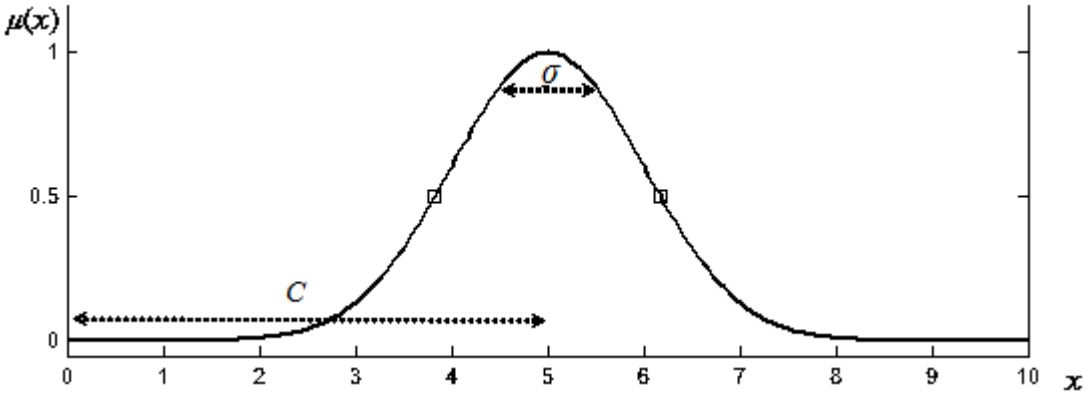


Figure 3.15 Gaussian membership function

3.3.3.4 Generalised bell membership function

The elements x in a generalised bell membership function $\mu(x)$ is specified by three parameters (*A, B, and C*). The parameter *C* represents the centre and width of the function.

Parameters *a* and *c* are used to adjust the width of the curve, whereas *B* is used to control the slope. Figure 3.16 shows generalised bell membership function with parameters *A* = 2, *B* = 3, and *C* = 4. The function can be represented mathematically as:

$$\mu(x) = bell(x; A, B, C) = \frac{1}{1 + \left|\frac{x - C}{A}\right|^{2B}} \tag{3.42}$$

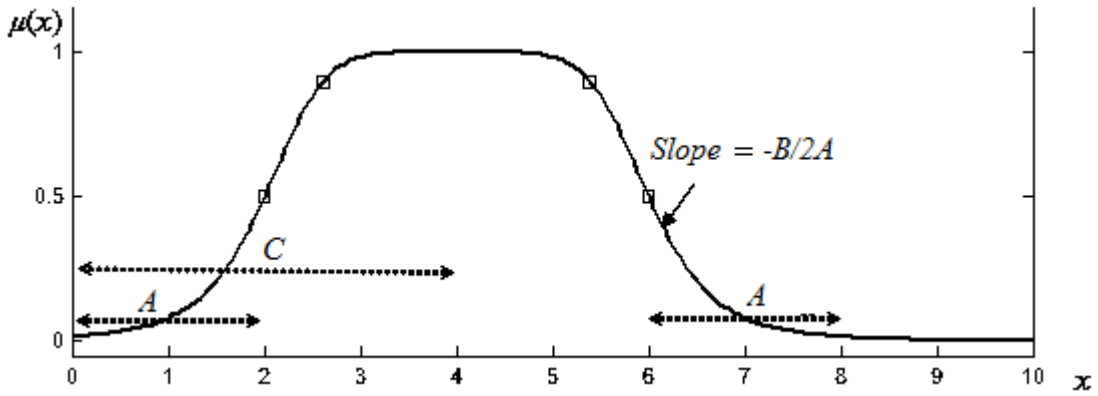


Figure 3.16 Generalised bell membership function

3.3.3.5 Sigmoidal membership function

The elements *x* in a sigmoidal membership function $\mu(x)$ is specified by two parameters (*A* and *C*). The parameter *C* represents the centre of the function. The parameter *A* controls the slope of the membership function. Figure 3.17 shows a sigmoidal membership function with parameters *A* = 2 and C = 5. The function can be represented mathematically as:

$$\mu(x) = sig(x; A, C) = \frac{1}{1 + e^{[-A(x-C)]}} \tag{3.43}$$

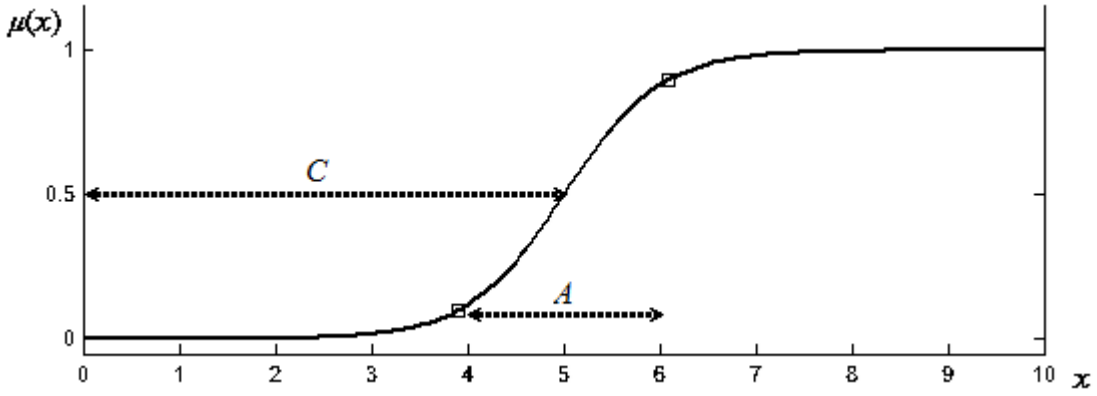


Figure 3.17 Sigmoidal membership function

3.3.4 Fuzzy inference process

Fuzzy inference is the process where fuzzy logic is used to map a given input to an output. The mapping is done based on expert knowledge of the system from which decisions can be made. The three main types of fuzzy inference methods are the Sugeno fuzzy inference, Mamdani fuzzy inference, and Tsukamoto fuzzy inference. Mamdani fuzzy inference is the most common fuzzy model used today. Its process can be performed in six main steps:

Step 1: Formation of fuzzy rules

Step 2: Fuzzification of input variables

Step 3: Application of fuzzy operator

Step 4: Implication

Step 5: Aggregation

Step 6: Defuzzification

3.3.4.1 Formation of fuzzy rules

Fuzzy logic tries to mimic human control logic by using descriptive language in its operations just like human operators. Rule-based expert systems translate expert knowledge written in natural language into fuzzy rules. Instead of mathematically modelling complex data, they use rules with simple statements made up of words such as 'IF', 'AND' and 'THEN'. These rules are expressed in syntax such as:

IF 'x is A' AND 'y is B' THEN 'z is C'

The expressions 'x is A' and 'y is B' are known as the antecedents, whereas the expression 'z is C' is known as the consequent. The input variables are x and y, and the output variable is z. A, B and C are the linguistic values. Each linguistic value is defined by a membership function in the universe of discourse. Some examples of fuzzy rules are:

If temperature is COLD then turn heating on HIGH.

If temperature is LOW then turn heating on OFF.

3.3.4.2 Fuzzification of input variables

Fuzziffication is the process of transforming crisp numeric input values into fuzzy values of appropriate fuzzy sets through membership functions. An output fuzzified value between 0 and 1 is returned irrespective of the value of the crisp input variable. Figure 3.18 shows a membership function curve, $\mu(x)$, which describes the fuzzy set 'temperature is cosy'. After fuzzyfying the crisp value of temperature at 27 ºC, a value of 0.66 was obtained.

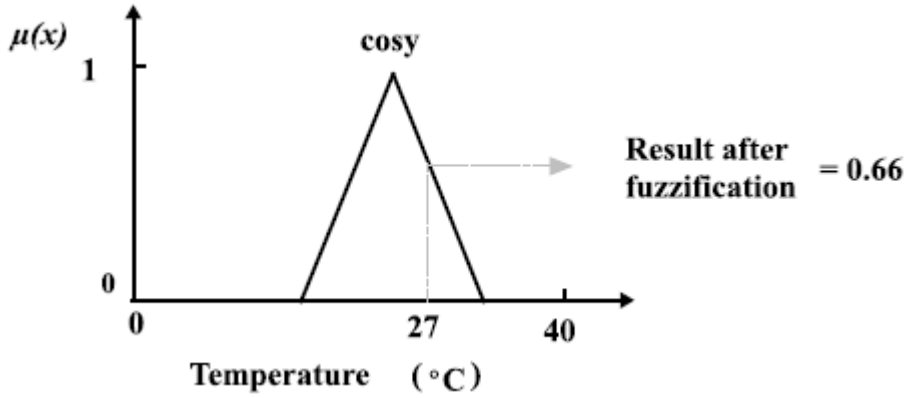


Figure 3.18 Fuzzification of input variable temperature

3.3.4.3 Application of fuzzy operator

If the antecedent of the fuzzy rule has more than one linguistic set, the fuzzy operator AND or OR is used to combine the fuzzy membership values. The fuzzy linguistic sets, "temperature is cosy" and "humidity is high" will be used to demonstrate the application of the fuzzy operator AND. Figure 3.19 shows membership function curves, $\mu(x)$ and $\mu(y)$, which describes the input variables temperature and humidity, respectively. The linguistic sets "temperature is cosy" and "humidity is high" gave results of 0.66 and 0.34, respectively after fuzzification. Applying the AND operator defined in Eqn.3.30, the value 0.34 is selected as the antecedents of the fuzzy rule.

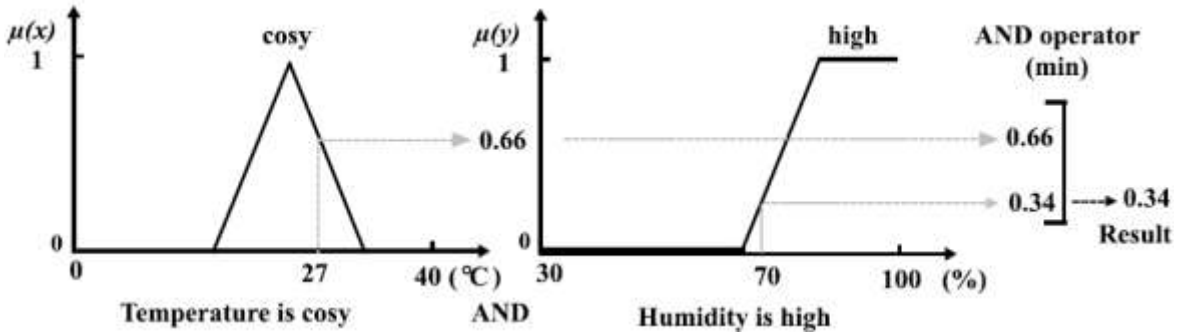


Figure 3.19 Illustration of the fuzzy operator AND

3.3.4.4 Application of the Mamdani minimum implication method

The Mamdani minimum implication method specifies how the membership function of the output linguistic variable is truncated. The AND or OR operator can be used in the implication process depending on the logic required to solve the problem. Figure 3.20 illustrates how the AND operator was used to truncate the output linguistic variable 'climate'. Using the AND operator defined in Eqn.3.30, the output linguistic variable is truncated where the fuzzified value is minimum ($\mu(z)$=0.34).

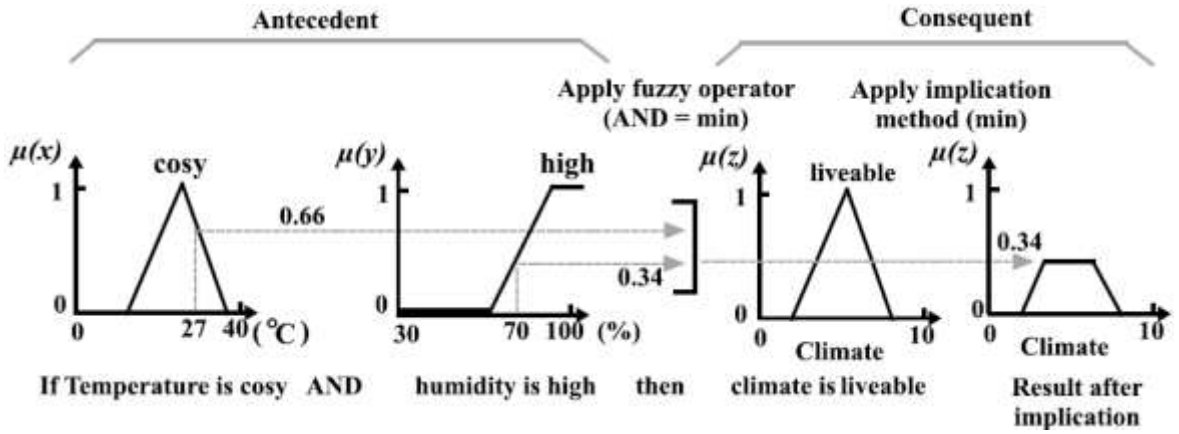


Figure 3.20 Illustration of the implication method

3.3.4.5 Aggregation

Figure 3.21 shows a diagram to illustrate the Mamdani inference system process. From the figure, membership function curves $\mu(x)$, $\mu(y)$, and $\mu(z)$ are used to describe the input variables temperature, humidity, and climate, respectively. The fuzzy rules, "If temperature is not cosy then climate is harsh", "If temperature is cosy AND humidity is high then climate is liveable", and "If temperature is cosy AND humidity is LOW then climate is comfortable" will be used to demonstrate the aggregation method. The process

of combining the output sets of each of the rules into the single fuzzy set is known as aggregation. Depending on the logic required to solve a given problem, the ‘max’, ‘sum’, or ‘probabilistic or’ function can be applied in the aggregation process. In this example, the ‘max’ function was used for illustration. The aggregated output set is the union (sum) of the output sets of each of the rules. The aggregated output set was obtained using Eqn. 3.31.

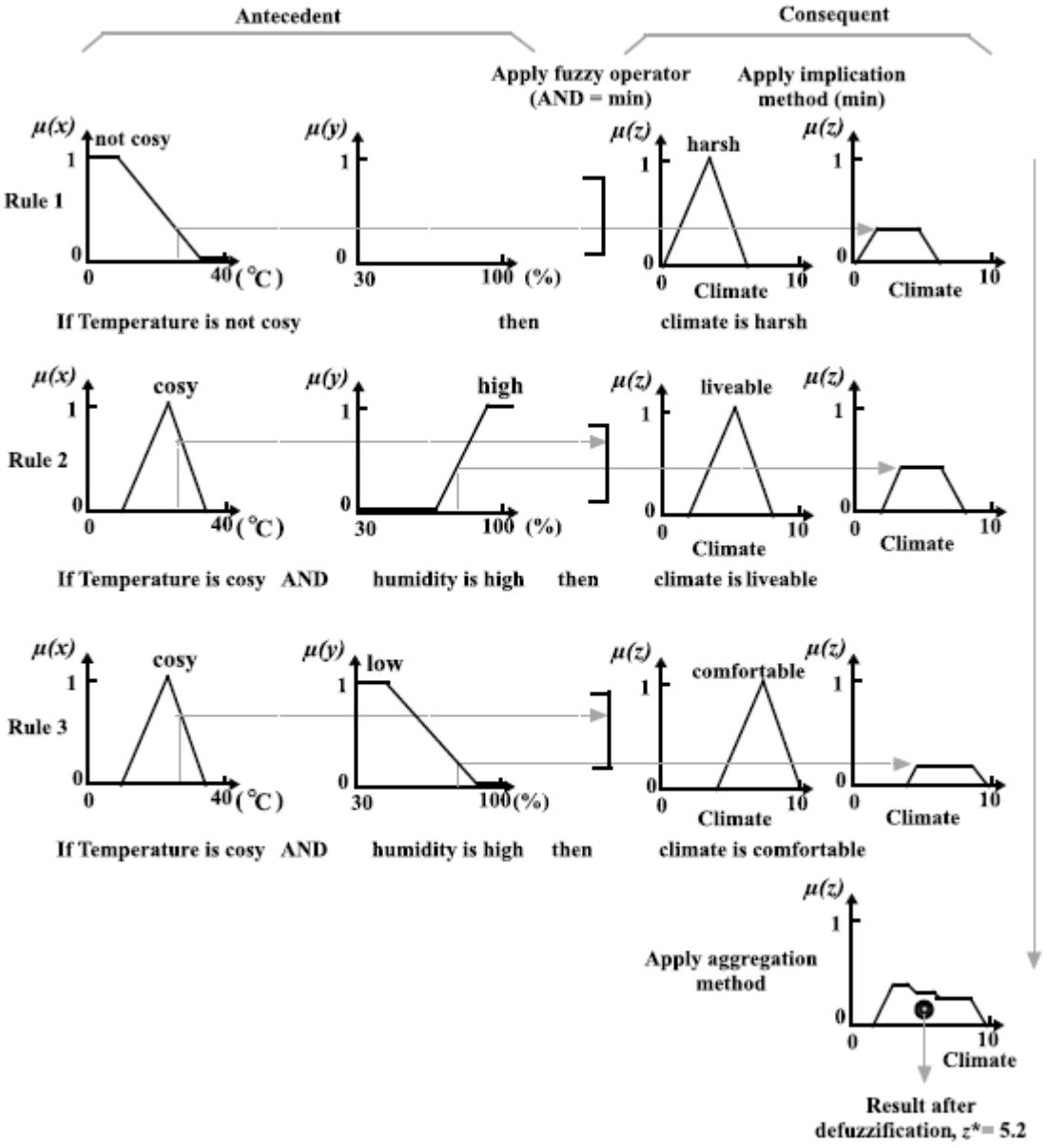


Figure 3.21 Illustration of the Mamdani inference system process

3.3.4.6 Defuzzification

Defuzzification is the final step of the FIS process. It is the process of converting the aggregated fuzzy output sets into a crisp value. There are several ways a fuzzy output can be defuzzified to a crisp value. The appropriate method of difuzzification to choose depends on a number of factors, including computational efficiency, shape of membership functions and the type of model being developed. For example, the centroid method will be more appropriate to use in quantitative models (van Leekwijck & Kerre, 1999), whereas the middle of maximum (MOM) will be more appropriate in qualitative models (Saletic, Velasevic, & Mastorakis, 2002). Other methods of defuzzification include the weighted average, bisector, smallest of maximum (SOMax), and largest of maximum (LOM) method.

In the weighted average method of defuzzification, each membership function in the output is weighted by its respective membership values. This method is frequently used in fuzzy applications because it is computationally fast. However, it has the disadvantage of not being able to process asymmetrical membership functions. This method can be mathematically expressed as:

$$z^* = \frac{\sum \mu_C(\bar{z}).\bar{z}}{\sum \mu_C(\bar{z})} \tag{3.44}$$

where $\bar{z}$ = the centroid of each of the symmetrical membership functions;
Σ = the algebraic summation; and z^* = defuzzified crisp value.

The bisector method of defuzzification divides the output aggregated membership function into two equal areas. The vertical line that divides the membership function into two equal areas corresponds to the deffuzzified crisp value. The bisector method sometimes gives the same results as the centroid method. This method of defuzzification is computationally fast and gives good results in fuzzy sets with symmetrical membership functions. However, it gives inaccurate results in fuzzy sets with asymmetrical membership functions (Ginart, Sanchez, Links, & Back, 2002).

MOM method of defuzzification, also known as the mean of maximum, takes the means of the points where the membership functions are at their maximum. This method is

computationally efficient. The SOMax method of defuzzification takes the smallest of the points (i.e. the leftmost point) where the membership functions are at maximum. The LOM method of defuzzification takes the largest of the points (i.e. the rightmost point) where the membership functions are at maximum.

The centroid method, which is also known as the centre of gravity (CoG) method, which is the most common technique of defuzzification, was developed by (Sugeno, 1985). The CoG method was used to defuzzify the aggregated fuzzy output sets as shown in Fig. 3.21. It computes a crisp value representing the centre of gravity of the aggregated fuzzy output sets. It can be mathematically expressed as Eqn. 3.45. A crisp defuzzified crisp value of 5.52 was obtained after using the CoG method. This indicates that when temperature is 27 ºC and humidity is 70 %, then climate is liveable.

$$z^* = \frac{\int \mu_C(z).z\mathrm{d}z}{\int \mu_C(z)\mathrm{d}z} \tag{3.45}$$

where z^* = the defuzzified crisp value which is the vertical line through the centre of gravity.

3.3.5 Application of fuzzy inference system in water resources

Ever since the emergence of fuzzy logic, it has been criticised for several reasons. Some researchers have opposed its application because of a natural reluctance to embrace new technology, especially when the change is seen as revolutionary (Bouslama & Ichikawa, 1992). Others have been sceptical about fuzzy logic applications because they believe probability theory is able to solve problems that have all kinds of uncertainties. Some even believe that fuzzy logic is probability theory in disguise, while others think probability theory is the only sensible method to solve problems with uncertainty (Ibrahim, 2004). It would have been ideal for the critics of fuzzy logic to first check whether its objectives were being achieved before criticising it. The objectives are to let computers reason like humans, and to enable linguistic computing (computing with words). Although these objectives have not been fully achieved, some level of success has clearly been attained. It is important for critics to realise that fuzzy logic models, like any other model, cannot solve every problem. They have their own limitations, therefore, alternate methods of modelling should be used if they are unable to solve a given problem (Ibrahim, 2004).

In recent years, fuzzy logic has been increasingly applied in water resource engineering. Sadiq, Kleiner and Rajani (2004) used a combination of fuzzy techniques and an analytic hierarchy process (AHP) to develop a tree-based structure that predicts the risk of water quality failure in WDNs. They defined the risk of each item that contributes to the failure using fuzzy numbers to capture fuzziness in the qualitative linguistic definitions. However, their hierarchical model did not include sorption; an important process that contributes to water discolouration. Moreover, the fuzzy rules for their model were formulated using expert knowledge. Consequently, it may not give accurate predictions on new WDNs. In a related study, Kord and Ashgari Moghaddam (2014) applied both fuzzy logic and kriging models in the evaluation of ground water quality using variables such as pH, iron, total dissolved solids, and electrical conductivity. The outputs of the predicted drinking water quality were categorised as ‘not acceptable’, ‘desirable’, and ‘acceptable’. They observed that the fuzzy model gave better predictions than the kriging model.

A hierarchic fuzzy logic model with 550 formulated rules was used by Gharibi et al. (2012) to develop a water index that measures the quality of drinking water supplied to dairy cattle from Karun River, Iran. Their model was trained on a four-year database consisting of 20 relevant input variables from 2007–2010. The relevant input variables included biochemical oxygen demand, temperature, turbidity, faecal coliform, dissolved oxygen, total dissolved solids, alkalinity, arsenic, and lead. Biochemical oxygen demand is very important variable in water quality modelling because it measures the quantity of oxygen used by microorganisms. The results from the model indicated that the water from Karun River had a low to medium water quality score. This fuzzy model could be very useful for assessing the quality of drinking water supplied to dairy cattle. The model’s prediction accuracy could have been improved if hydraulic variables were incorporated.

Islam, Sadiq, Rodriguez and Francisque (2013) used a fuzzy-based model to assess raw water quality at Clayburn watershed in British Columbia, Canada. Their proposed model estimated pollutants loads discharged from various land uses such as highways/roads, agriculture, livestock, forests, and pasture land. After carrying out monthly and yearly analyses from the predicted results, they observed that highways/roads, and agriculture had an adverse impact on water quality, whereas forest gave the best water quality. The model can help water resource engineers to make informed land use decisions to improve water quality.

Chang et al. (2014) developed a quick and reliable model using neuro-fuzzy logic to estimate arsenic concentrations in Huang Gang Creek, Taiwan, which is a mix of stream water and hot springs. Input variables used in developing the model included one-month antecedent rainfall, nitrite nitrogen, temperature, DO, and Pb. Observations from their model indicated that lower temperatures, higher nitrite-nitrogen concentrations, and higher one-month antecedent rainfall resulted in high arsenic concentrations. This model could be used in the management of arsenic pollution in rivers and streams.

3.3.6 Benefits and limitations of fuzzy inference system

FIS is a very important tool in fuzzy set theory since it has the ability to automate system control and make decision analysis. However, like every good tool it may not be able to solve every problem. Hence, it has its own advantages and disadvantages. The following section summarises some advantages and disadvantages of FIS.

3.3.6.1 Benefits of fuzzy inference system

- It is cheaper to use than most traditional methods of modelling. Even though some traditional methods of modelling can give more precise results than FIS, they may be too costly or time intensive to model. Precision is expensive, but may not be always necessary. In other words, it may not be necessary to obtain an exact result when an estimated result from FIS is sufficient.
- It can express natural language as fuzzy logic rules. By so doing, complex problems can be converted into simpler problems using these rules.
- It can use expert knowledge in formulating fuzzy rules to perform tasks such as target tracking, systems control and water quality prediction.
- It can be used to solve highly complex problems which analytic or numeric formulations cannot solve. Some problems are too complex to use conventional functions to define their causes and effects.
- Fuzzy logic is easy to learn and use.
- Its tolerance nature enables modelling with imprecise and inaccurate data.

3.3.6.2 Limitations of fuzzy inference system

- As the system's complexity increases, more rules become necessary. As a result, combining these rules to obtain a good solution becomes increasingly difficult.

- A substantial amount of time is required to correctly tune the membership functions and adjust the rules in order to obtain accurate predictions.
- FIS may not be suitable for solving problems that require high level precision solutions. For instance, fuzzy logic cannot be used to determine the ballistic trajectory of a missile intended to hit a target from a long range.

3.4 Genetic algorithm

3.4.1 Overview genetic algorithm

Over the years, humans have gradually developed artificial intelligence that enables us to predict natural phenomena such as rainfall, temperature, snow, and the causes of diseases. Since 1980s, they have been increased research in artificial intelligence in the area of neural networks (imitates the human brain), fuzzy inference system (emulates human imprecise reasoning), and evolutionary algorithm (mimics evolution). Genetic algorithm is the most common evolutionary algorithm used by researchers. It is a global optimisation algorithm that was introduced by Holland (1975) at the University of Michigan. As the name indicates, genetic algorithm uses the natural phenomenon of evolution to find a solution to a problem by iteratively selecting fit candidates from a population of solution candidates to create offspring. This is repeated for several generations, each time creating offspring that are fitter than their parents. This algorithm is sometimes referred to as "survival of the fittest" because for each generation, fitter candidates are selected from a population for crossover (mating).

Traditional methods of optimisation are slower in finding solutions to problems that have complex search space. Conversely, genetic algorithm is suitable for solving computational problems that usually have a number of possible solutions. Its computational parallelism functionality makes it able to simultaneously search for different solutions to a problem in an efficient way. It requires little information to effectively search through poorly understood search space.

3.4.2 Mechanism of genetic algorithm

Figure 3.22 shows the six main steps in the genetic algorithm process. These steps include the encoding of chromosomes, initialisation of the population, evaluation of the objective

function, selection of chromosomes, crossover, and mutation. The last three steps are referred to as the genetic operators. The following sections present in detail the steps in the genetic algorithm process.

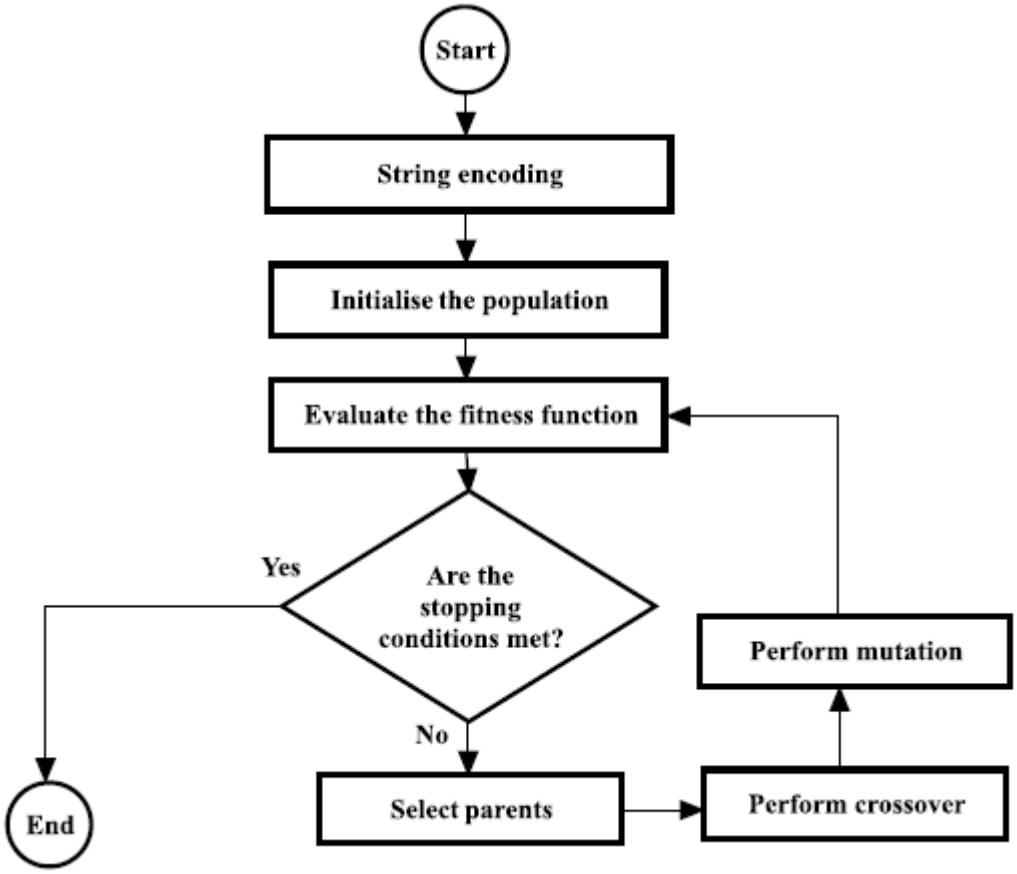


Figure 3.22 The genetic algorithm process

3.4.2.1 Encoding of chromosomes

Living organisms are made up of cells. Each of these cells consists of one or more chromosomes (strings of DNA). A chromosome consists of many identifiable subunits known as genes. In genetic algorithm, chromosome refers to a candidate solution encoded as a bit string. The bit strings encoded in a particular parameter are referred to as genes. A collection of chromosome is known as population. There are two methods of encoding in genetic algorithm. In the first method, binary encoding, the encoded bit strings in chromosomes are presented in binary strings of 0s and 1s. This method is often used for solving function optimisation problems. Figure 3.23 (a) shows some example of chromosomes with binary encoding. The second method, permutation encoding, is often

used for solving sequencing or ordering problems. The strings in the chromosomes of this method consist of numbers ranging from 1 to *n* numbers (see Fig. 3.23 (b)).

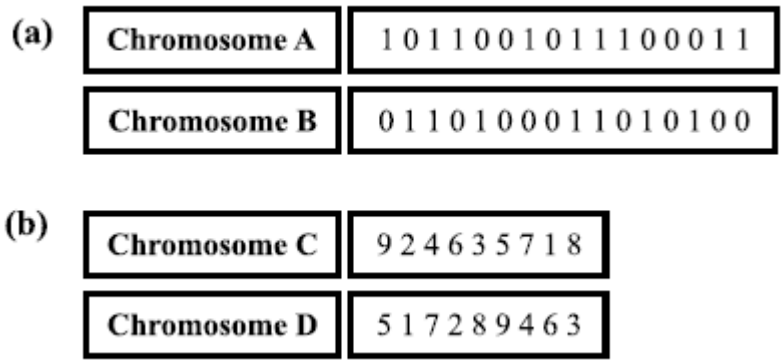


Figure 3.23 Examples of (a) chromosomes with binary encoding and (b) chromosomes with permutation encoding

3.4.2.2 Population initialisation

Population initialisation is a very important step in the genetic algorithm process. Using bias initial population can reduce convergence speed or the performance of the model (Rahnamayan, Tzihoosh, & Salama, 2007). If there is no information about a solution to the problem, the parent population before the first generation is usually generated randomly. Randomly generated techniques such as pseudo-random number generator or chaotic number generator can be used to generate random numbers within the range defined in the chromosomes.

3.4.2.3 Objective function evaluation

Objective function is a function used to evaluate the performance of individuals in the population. In minimisation problems, an individual chromosome in the problem domain that has the lowest numerical value is the fittest, whereas in maximisation problems an individual with the highest numerical value is the fittest. An example of an objective function *f*(.) is given in Eqn. 3.46. To evaluate the objective function, the bit string in the chromosome is translated to a real number y_{real} and substituted into the objective function. The value returned by the objective function is known as the fitness value of the candidate solution.

$$f(y_{real}) = y_{real} + |\sin(32 * y_{real})|, \qquad 0 \leq y_{real} \leq \pi \tag{3.46}$$

3.4.2.4 Selection

After the evaluation of the objective function, the selection operator is used to eliminate the worst chromosomes due to the low fitness value and select two healthy parents for the generating of new offspring. The most common selection method in genetic algorithm is the rank-based selection scheme. This method ranks fitness value for each chromosome and select the healthy parent chromosomes for crossover.

3.4.2.5 Crossover

The crossover operator is applied to the selected healthy parent chromosomes to produce offspring. Crossover operators for binary strings differ from permutation strings. In permutation strings, it is a requirement that each element appears only once in the string, whereas elements in binary strings can repeat itself. The three main crossover operators used in genetic algorithm are the standard one-point crossover, one-point order crossover, and heuristic crossover. Figure 3.24 illustrates how the standard one-point crossover is applied to binary strings. In this figure, two new offspring are formed by swapping elements in the tail parts of the strings (seventh to ninth position in the string).

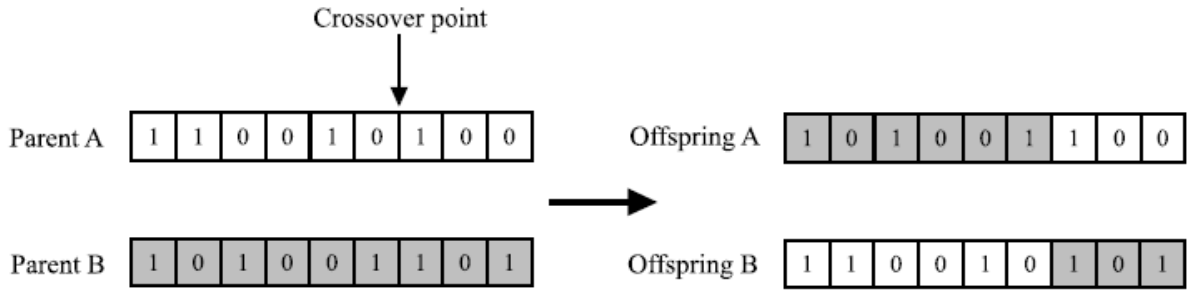


Figure 3.24 Illustration of standard one-point crossover for binary strings

Figure 3.25 shows how the one-point order crossover is applied to permutation strings. From the figure, a crossover point is selected to divide the parent strings. After the crossover, the head part of Parent B becomes the head part of Offspring A and the head part of Parent A becomes the head part of Offspring B. The strings in the tail part of Parent A are reordered in the order of the appearance in Parent B and become the tail part of Offspring B. Similarly, the strings in the tail part of Parent B are reordered in the order of the appearance in Parent A and become the tail part of Offspring A.

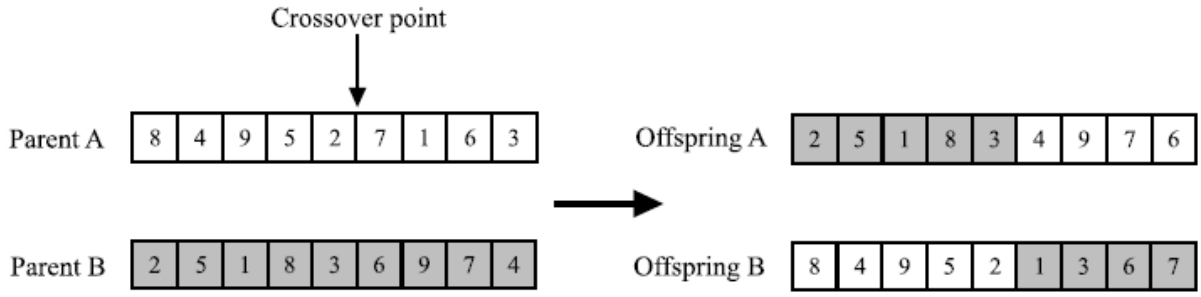


Figure 3.25 Illustration of standard one-point crossover for permutation strings

Heuristic crossover method uses the fitness values of the two parent chromosomes (*Parent 1* and *Parent 2*) with numerical representation to determine the direction of the search. The offspring generated is placed in a line drawn between the two parents, nearer the parent with the better fitness value. If *Parent 1* has the better fitness value than *Parent 2*, then the offspring is generated by the following equation:

$$Offspring = Parent\ 2 + C_{Ratio} * (Parent\ 1 - Parent\ 2) \quad (3.47)$$

where C_{Ratio} = the crossover ratio (usually a uniform random number) which specify how far the offspring is from the parent with a better fitness value.

3.4.2.6 Mutation

Mutation is a genetic operator that is used to maintain genetic diversity from one generation to another generation and prevents premature convergence of the genetic algorithm. It enables the genetic algorithm to search a broader space. Bitwise mutation and Gaussian mutation are mutation methods often used in genetic algorithm. In the bitwise mutation process, some elements in the solution string are randomly selected and inverted. Figure 3.26 illustrates how elements in the solution binary String A at positions 3, 6, and are inverted to form a new binary String B.

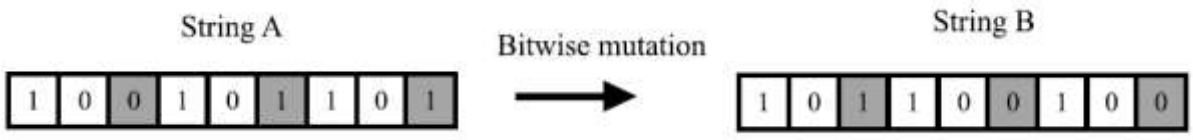


Figure 3.26 Illustration of bitwise mutation for binary strings

In the Gaussian mutation method, a random number ($RAND_{Gaussian}$) is taken from a Gaussian distribution with mean 0 to each entry of the parent vector. The standard deviation of the Gaussian distribution is determined by the parameters '*Scale*' and 'Shrink'. The parameter '*Scale*' represents variance of mutation during the first generation, whereas the parameter '*Shrink*' represents Amount of shrink in the mutation in successive generations. The equation for calculating '*Scale*' and mutation is given by Eqns. 3.48 and 3.49, respectively.

$$Scale = Scale\ -\ Shrink * Scale * \left(\frac{Currnt\ generation}{Total\ generation}\right) \tag{3.48}$$

$$Mutated\ chromosome = Chrosome + Scale * RAND_{Gaussian} \tag{3.49}$$

3.5 Summary

A critical review on AI-based methods of modelling in this chapter showed that these methods have the capability of learning from data and also able to cope well with uncertainties in data. They are also able to model data which have complex non-linear relationships between the independent and dependent variables. Given the complex nature of the processes that lead to the formation of Fe and Mn accumulation/water discolouration in WDNs, AI-based methods of modelling such as ANNs and FISs may be more appropriate to solve these types of problems.

CHAPTER 4: Data Acquisition and Exploratory Data Analysis

4.1 Introduction

In order to understand the processes and mechanisms that lead to Fe and Mn compliance failures, it is necessary to identify the relevant variables that influence Fe and Mn deposition in WDNs. In Chapter 2, studies by other researchers were reviewed to identify relevant variables that influence Fe and Mn accumulation. However, there are other important variables such as hydraulic distance from source of water supply and variation of daily shear stress which also influence Fe and Mn accumulation that have not been investigated thus far. In this chapter, a five-year customer complaint data set was collated with the objective of identifying WSZs with low, medium, and high levels of customer complaints for further analysis. Fourteen WSZs were selected for this analysis. A five-year data set comprising 37 chemical and biological water quality variables from the selected WSZs was analysed to identify relevant variables that influence Fe and Mn accumulation. EPANET was extended to extract relevant hydraulic and pipe-related variables such as maximum shear stress, average water age, pipe type, and pipe age from the network files of the WSZs. The computed hydraulic variables were also analysed to determine their effect on Fe and Mn accumulation. In subsequent chapters, the relevant variables identified will be used to develop models for predicting Fe and Mn accumulation potential. The remaining sections of this chapter are arranged as follows. The data collection methods for this research are presented in Section 4.2. Section 4.3 explains how the data were prepared for this study. The methodology for identifying relevant variables is presented in Section 4.4. The results are presented and discussed in Sections 4.5–4.8. Finally, the summary of this chapter is presented in Section 4.9.

4.2 Data collection

In this study, a five-year data set comprising 37 water quality variables covering 14 water supply zones (WSZs), consisting of 176 different DMAs, provided by an industrial partner was analysed. Customer complaints data for the WSZs covering the same period and study area were also provided by the drinking water company. The 37 water quality variables are listed in Table 4.2. In the context of this research, the sites of interest included WSZs with high, medium, and low levels of customer complaints in order for the models to capture all levels of discolouration and to remove any form of bias. The 14 WSZs used for the research had the following customer complaints levels: WSZ10, WSZ1, and WSZ8 had low customer complaints; WSZ7, WSZ5, WSZ6, WSZ11, and WSZ3 had medium customer complaints; and WSZ2, WSZ12, WSZ14, WSZ13, WSZ4, and WSZ9 had high customer complaints. The WSZs were selected from regions around Blackburn, Bolton, and Liverpool.

4.3 Data preparation

The existence of outliers in data sets is likely to have deleterious effects on models. Outliers can cause significant misinterpretations in statistical estimates in parametric or nonparametric tests (Zimmerman, 1998). They can also decrease normality, which eventually leads to Type I error (incorrect rejection of a true null hypothesis) and Type II error (failure to reject a false null hypothesis) (Zimmerman, 1994). In view of this, both transformed/normalised and untransformed/unnormalised data were used to develop the models in this research. It is important to remove all outliers because they can skew the distribution of data, which can lead to inaccurate model predictions. However, some researchers believe that outliers should not be removed, because extreme values do exist and are not always errors. It has been argued that removing outliers which are not caused by errors is a way of manipulating data to obtain better results, and can prevent models from predicting extreme values (Deyo, 2010; Osborne & Overbay, 2004).

4.3.1 Customer complaints data

Water companies worldwide currently use analysis of customer complaints data as the primary method of identifying areas in WDNs with high-risk of discolouration (Prince et al., 2003). A simple ‘rule of thumb’ used by Sly et al. (1990) is that, if Mn and/or Fe levels rise(s) above their respective MCLs of 50 µg/L and 200 µg/L, customer complaints

increase. For the purpose of this study, only customer complaints data relating to discoloured water and slime were reviewed. This is because increased Fe and Mn concentrations are known to cause drinking water discolouration and leave slimy masses inside sinks and toilet tanks (Boxall et al., 2003; Herman, 1996; Slaats, 2002). Using ArcGIS, the addresses where customers complained were geocoded to X, Y hydraulic file coordinates and assigned to their nearest respective nodes in the WSZs.

The raw water quality data received from the drinking water company included other customer complaints that were not relevant for this research. Since increased Fe and Mn concentrations are associated with discoloured water and black-brown slimy masses inside toilet tanks and sinks, only customer complaints data relating to discoloured water and slime were reviewed. The customer complaints data were exported into a Microsoft Access database. The Structured query language (SQL) code to retrieve customer complaints data relating to discoloured water and slime from the Microsoft Access database is given in Appendix K.

Since highly populated DMAs have a higher propensity for more customers to complain, there is the need to remove any population bias. For this purpose, the customer complaints data were normalised by dividing it with the number of properties (service connections) in each DMA and multiplying it by 1000. This removed any form of bias caused by the differences in DMA populations. This procedure is used by many water utilities (Prince et al., 2003). The customer complaints per 1000 properties can be mathematically expressed as:

$$\text{QCC per } 1000 \ = \frac{\text{Quarterly customer complaints } \times 1000}{\text{NPDMA}} \qquad (4.1)$$

where QCC = quarterly customer complaints; and
NPDMA = number of properties in a DMA.

4.3.2 Water quality variables

The water quality data provided by the drinking water company were sampled at different frequencies. Water quality variables with low sampling frequencies could not be investigated, because there were insufficient data from these variables to make analysis.

The street/house addresses of the sampled water quality variables were recorded. Using ArcGIS, the addresses were geocoded to X, Y hydraulic file coordinates and assigned to their nearest respective nodes in the WSZs. The geocoding was done by the drinking water company.

The initial five-year post-treated water quality data and the five-year customer complaints data obtained from the drinking water company were provided in Microsoft Excel format. The data were exported into a Microsoft Access database to undergo data preparation. Microsoft Excel and Microsoft Access have their own strengths and weaknesses. Databases are more efficient for aggregating data (for example, calculating monthly customer complaint numbers and yearly averages of Fe concentrations). The relational database nature of Microsoft Access enables two or more tables to be joined. For instance, the customer complaints table can be joined to the water quality table to determine the relationships between water quality variables and customer complaints. SQL makes interacting with and retrieving data from Microsoft Access databases flexible and very easy. The SQL code to retrieve and merge the hydraulic data from the hydraulic table and the yearly averages of Fe and Mn from the water quality table for WSZ2 is given in Appendix L. Microsoft Excel has an advantage of being able to give different graphical representations of data for analysis.

Outliers, which are extreme data points that deviate significantly from other data points, were removed from the data set. The method of detecting outliers by Smith and Subandoro (2007) was used in this research. This method classifies data which are more than five standard deviations from the mean as outliers. Mathematically, a data point (X_{data}) is classified as an outlier if:

$$|X_n - \bar{X}| > 5\sigma \qquad (4.2)$$

where X_n = n^{th} data point; $\bar{X}$ = mean of the data points; and
σ = standard deviation of the data points.

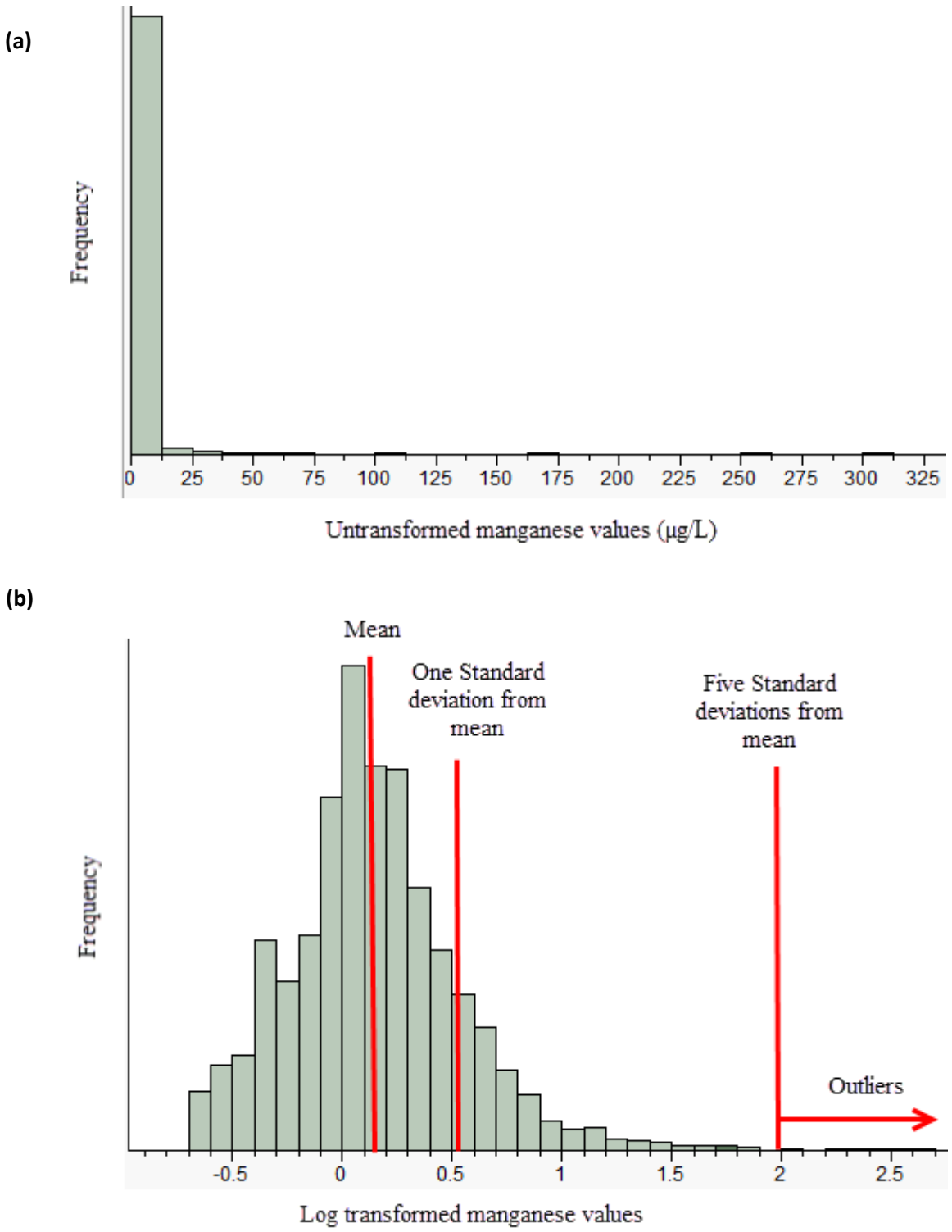


Figure 4.1 Distribution of (a) untransformed Mn data with outliers (b) logarithmic transformed Mn data with outliers

The data set used for model development consisted of approximately 0.5% outliers. Figure 4.1 shows the distribution of the untransformed and logarithmic transformed Mn data with outliers. The presence of these outliers in the data set could have been due to data entry errors, measurement errors, instrument failure, sampling errors, or contaminants introduced into WDNs from pipe bursts. They could also have come from deposited particulates after sudden increase in flow due to pipe bursts or the opening of fire hydrants during flushing operations and fire extinguishing exercises. All outliers were removed from the data set to prevent any disproportionate strong influence on the ANN models' predictions.

4.3.3 Hydraulic and pipe-related variables

4.3.3.1 Maximum daily shear stress at node

To investigate the influence of shear stress on Fe and Mn accumulation potential, the EPANET software was extended to extract all pipe and node variables. From the software, the shear stress was computed every 15 minutes for 24 hours for each pipe in the network, and the maximum daily shear stress for each pipe was recorded. The maximum daily shear stress in a pipe can be mathematically expressed as Eqn. 4.3. This equation for calculating hydraulic shear stress in pipes was adopted from PODDS model (see Section 2.4.1 for more details). Because shear stress has a pipe property, a methodology was devised to calculate the maximum daily shear stress at each node. The maximum daily shear stress at a given node was calculated by summing the maximum daily shear stress of the pipes connected to the node, and dividing it with the number of pipes connected to that node. This can be mathematically expressed as Eqn. 4.4.

$$\tau = \frac{\rho_w g d_p H}{4 L_p} \tag{4.3}$$

$$\bar{\tau} = \frac{\sum_{j=1}^{NP} \tau_j}{NP} \tag{4.4}$$

where τ_j = maximum daily shear stress of a pipe j connected at a node;
ρ_w = density of water; g = acceleration due to gravity; H = head loss; L_p = length of pipe;
$\bar{\tau}$ = maximum daily shear stress at the node; d_p = diameter of pipe; and
NP = number of pipes connected to the node.

4.3.3.2 Variation of daily shear stress at node

The diurnal variation in WDNs due to the continuous variation of drinking water demand causes the shear stress in pipes to vary from time to time. The shear stress acting on the walls at peak demand times is higher than that at off-peak times. The EPANET software was extended to compute the variation of daily shear stress in each pipe every 15 minutes for 24 hours. The variation of daily shear stress was calculated using the formula:

$$\tau^s = \sqrt{\frac{\sum_{k=1}^{NT} (\tau_k - \tau^a)^2}{NT - 1}} \tag{4.5}$$

where τ^a = mean daily shear stress in a pipe; *NT* = number of time intervals;
τ_k = daily shear stress at the k^{th} time interval; and
τ^s = variation of daily shear stress of a pipe.

The variation of daily shear stress in pipes was converted to variation of daily shear stress at nodes by summing the value in each pipe connected to the node and dividing it by the number of pipes connected to it. It is expressed mathematically as:

$$\bar{\tau}^s = \frac{\sum_{j=1}^{NP} \tau_j^s}{NP} \tag{4.6}$$

where $\bar{\tau}^s$ = variation of daily shear stress at node.

4.3.3.3 Water age

The age of water in WDNs, often referred to as residence time, is the time taken for treated water to travel from the treatment plant to a given node. This is a vital variable that can help to determine the extent of disinfectant loss in WDNs. It may range from a few seconds to several weeks. The EPANET software was used to compute the water age for all the nodes in the network after 72 hours of simulation.

4.3.3.4 Hydraulic distance from source of water supply

Hydraulic distance from source of water supply is the distance travelled by water from the source of water supply to a given node within a WDN. To measure this variable, the

EPANET software was extended to calculate the hydraulic distance of each node from the source of water supply. The program makes use of two main algorithms; namely, Particle Backtracking Algorithm (PBA) and Shortest Route Algorithm (SRA).

The PBA was adopted from that developed by Shang, Uber and Polycarpou (2002), and was modified to suit this model. PBA can track particle movements in water from any node in a WDN to the source(s) of the water supply. PBA uses Lagrangian time-driven method which runs in reverse time for its computations; that means it runs opposite to the hydraulic simulation time. Furthermore, it is able to trace all the flow paths and their corresponding time delays between any given node and the source(s) of water supply. However, for the purpose of this research, only the flow paths were needed to compute the hydraulic distance from the source of water supply. A detailed algorithm with and without multiple storage tanks has been published by (Shang *et al., 2002*).

Under the assumption of first-order chlorine decay reaction, the PBA developed by Shang et al. (2002) models output concentrations as a variable depending on input concentrations, network hydraulics, and physical characteristics of the pipe network (see Eqn. 4.7). For a single water quality source, PBA describes the output concentration (c) as a linear function of the input source strength (c_s) for the travel paths (hydraulic paths) between a given source node and output node as follows:

$$c(T) = \sum_{k=1}^{NTP} \gamma_k c_k (T - t_k) \tag{4.7}$$

where *NTP* = number of travel paths between a given input and output node;
T = output time; c_k = water quality source input;
γ_k= impact coefficient for travel paths *k*, which is the sensitivity of output concentration to path input concentration. This takes into account chlorine decay in pipes and storage tanks, and flow mixture at junctions; c = water quality output concentration; and
t_k = time delay for travel path *k*.

Figure 4.2 is a simple pipe with steady hydraulic conditions that illustrates how PBA works. Although PBA is able to trace flow paths and computes their corresponding time

delays in WDNs with multiple water sources (storage tanks), for simplicity, a pipe with no storage tank will be used to illustrate how it works. The same concept can be used in networks with multiple storage tanks. The following parameters need to be defined in order to explain the PBA process:

us = upstream node; *ds* = downstream node; L_p = length of pipe;
μ' = positive flow velocity directed from upstream to downstream;
x' = the position of the particle along the pipe; t' = particle travel time;
γ = unitless impact coefficient; k = composite first-order decay coefficient;
T' = initial algorithm time of the current hydraulic period.

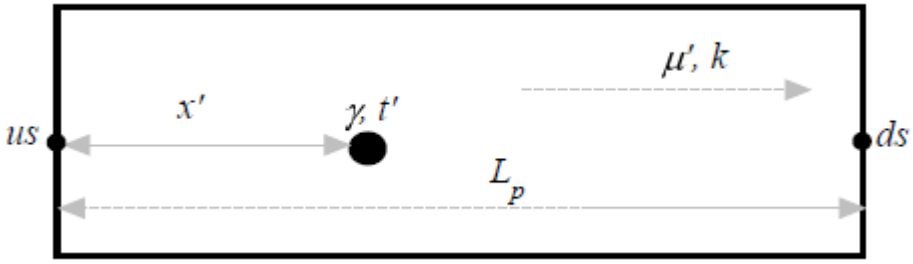


Figure 4.2 Illustration of particle backtracking algorithm in a single pipe without tank

At the beginning of the PBA, the algorithm time is initialised to zero in a single pipe (see Fig 4.2). The particle (water parcel) is transported (backtracked) in the network in reverse time. The parameters t and γ are then updated until the particle reaches the upstream node *us* or the algorithm time equals the beginning of the hydraulic period (i.e. equals T). If the particle reaches the upstream node, and the node at the upstream is not the source node, then it is split among all inflows. On the other hand, if the algorithm time equals the beginning of the hydraulic period, the flow conditions are updated. Algorithm 4.1 shows how the particle is backtracked in a single pipe during a single hydraulic period.

Algorithm 4.1 Algorithm to backtrack a particle in a single pipe during a hydraulic period

1. ***Compute*** *the time ($\delta t'$) the particle remains in the pipe and the current hydraulic condition.*

   If *$\mu(T'-t') < x$ (i.e. the particles is still in the pipe)* ***then***

   $\delta t' = T'-t'$

   Else if *the particles has reached upstream node d* ***then***

$\delta t' = x'/\mu'$

End if

2. ***Update*** *the parameters of the particle*
   - $x' = x' - \mu'\delta t'$
   - $t' = t' + \delta t'$
   - $\gamma = \gamma \exp(-k\, \delta t')$
3. ***If*** *the particle reaches upstream node before time T′* ***then***

   Split the particles into different path flows.

   End if

The SRA uses graphshortestpath, which is a Matlab in-built function, to find the shortest paths and all possible paths from each node to all the source of water supply (reservoirs/tanks). Whereas the PBA was used to find the hydraulic paths and determine which of the reservoirs/tanks supplied the nodes with water. If there is only one source of water supply and the PBA indicates that there is a hydraulic path between the source of water supply and the node, the distance between them is recorded as the hydraulic distance from source of water supply. However, if there is more than one source of water supply to a node, their average hydraulic distances to the node are computed and record as the hydraulic from source of water supply. The flow chat and source code for calculating the hydraulic distance from source of water supply to the nodes are presented in Fig 4.3 and Appendix D, respectively.

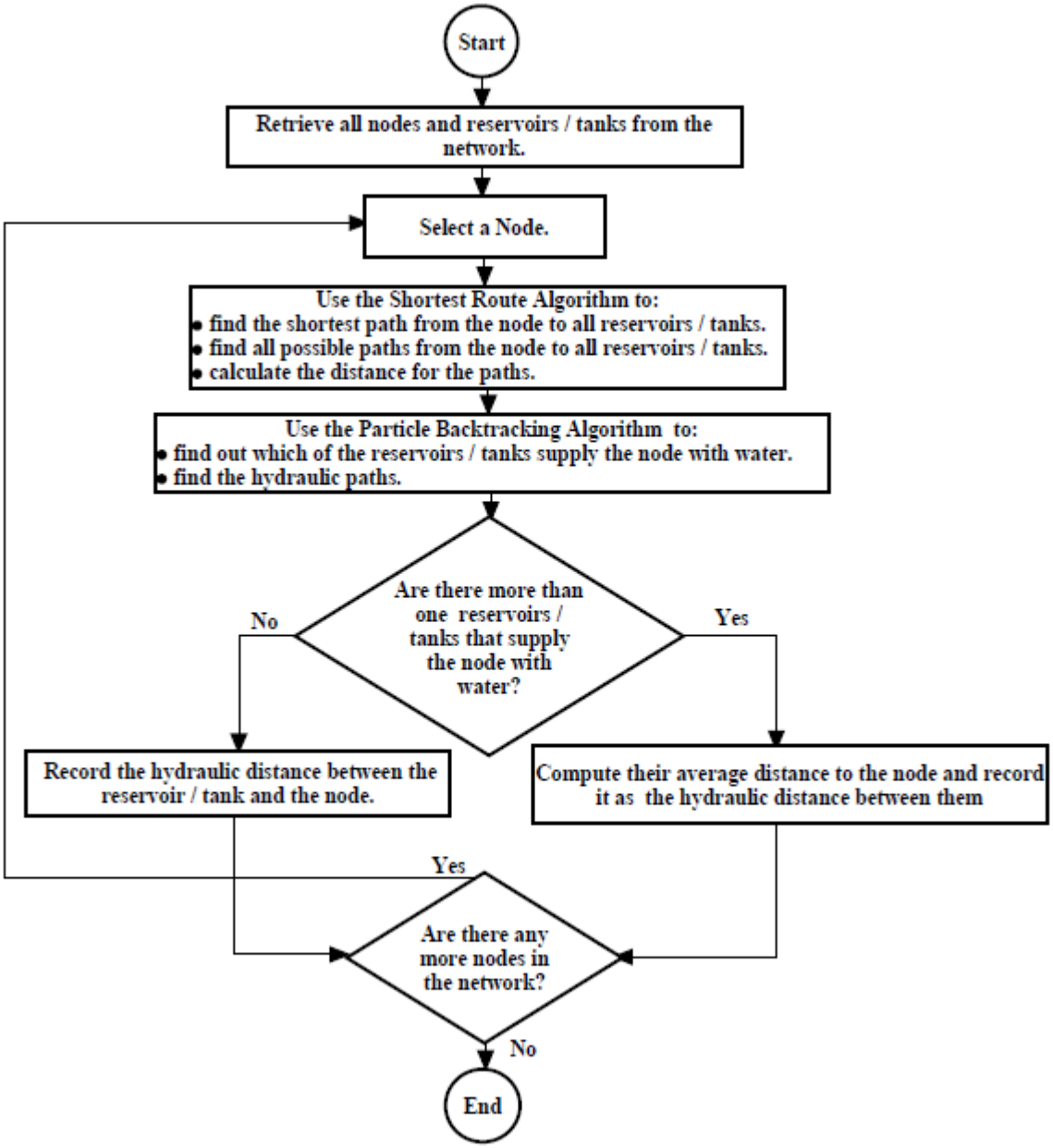


Figure 4.3 Flow chart for calculating the hydraulic distance

4.3.3.5 Pipe material

Each pipe material was assigned a value in the range between zero and one and termed pipe material index. Plastic materials were given values very close to zero, while ferrous materials were given values close to one. Pipe materials were arranged from low to high in the following order of susceptibility to corrosion: Polyethylene (PE) → Polyvinyl Chloride (PVC) → High Density Polyethylene (HDPE) → Asbestos Cement (AC) → Ductile Iron (DI) → Steel (ST) → Cast Iron (CI). There were a few missing pipe material data in the hydraulic files. The table compiled by Bhave (1991), which lists pipe materials and their

corresponding range of roughness values, was used to estimate these missing data (see Table 4.1). Pipe roughness has a strong correlation with pipe material (Bhave, 1991). For instance, pipe materials such as CI and DI are known to have higher roughness values than PE and PVC. Pipe roughness also depends on pipe age. As pipe age increases, the accumulation of sediments and corrosion by-products on the inside of the pipe walls also increases. This causes the pipe roughness to increase, which eventually leads to a reduction in the pipe's inner diameter (Christensen, 2009). From Table 4.1, pipe roughness values for uncoated CI ranged from 0.15–0.6 mm. Therefore, new uncoated CI pipes were given values close to 0.15 mm, whereas old uncoated CI pipes were given values close to 0.6 mm.

Table 4.1 List of pipe materials and their corresponding range of pipe roughness values

Pipe material	**Pipe roughness (mm)**
Asbestos cement	0.015 - 0.03
Bitumen/Cement lined	0.03
Wrought iron	0.03 - 0.15
Galvanised/Coated cast iron	0.06 - 0.3
Uncoated cast iron	0.15 - 0.6
Ductile iron	0.03 - 0.06
Uncoated steel	0.015 - 0.06
Coated steel	0.03 - 0.15
Concrete	0.06 - 1.5
Plastic, PVC, PE	0.02 - 0.05
Glass fibre	0.06
Brass, cooper, lead	0.003

(Bhave, 1991)

4.3.3.6 Pipe age

Pipe age data were provided by the water company. However, a few of them were missing in the network files. A Matlab program was written to extract the age of all pipes in the network. Where the pipe age data were missing, the pipe roughness was used to estimate the missing data. Pipe roughness was used because it is known to have a strong correlation with pipe age (Christensen, 2009). Linear regression models were developed to estimate the missing pipe age data for each pipe material using their respective pipe roughness values for each of the WSZs under investigation. Figure 4.4 (a) and (b) show sample graphs of the linear regression models used to estimate pipe age for CI and DI pipes, respectively, in WSZ1 and WSZ2. A Coefficient of determination (R^2) value of 0.69 and

0.63 observed for the models for estimating CI and DI indicates it predicts reasonably well. Substituting the values of pipe roughness, slope of the regression line, and intercept term into the equations in Fig. 4.4 (a) and (b), the pipe age can be estimated. The algorithm for estimating missing pipe age data is presented in Appendix M.

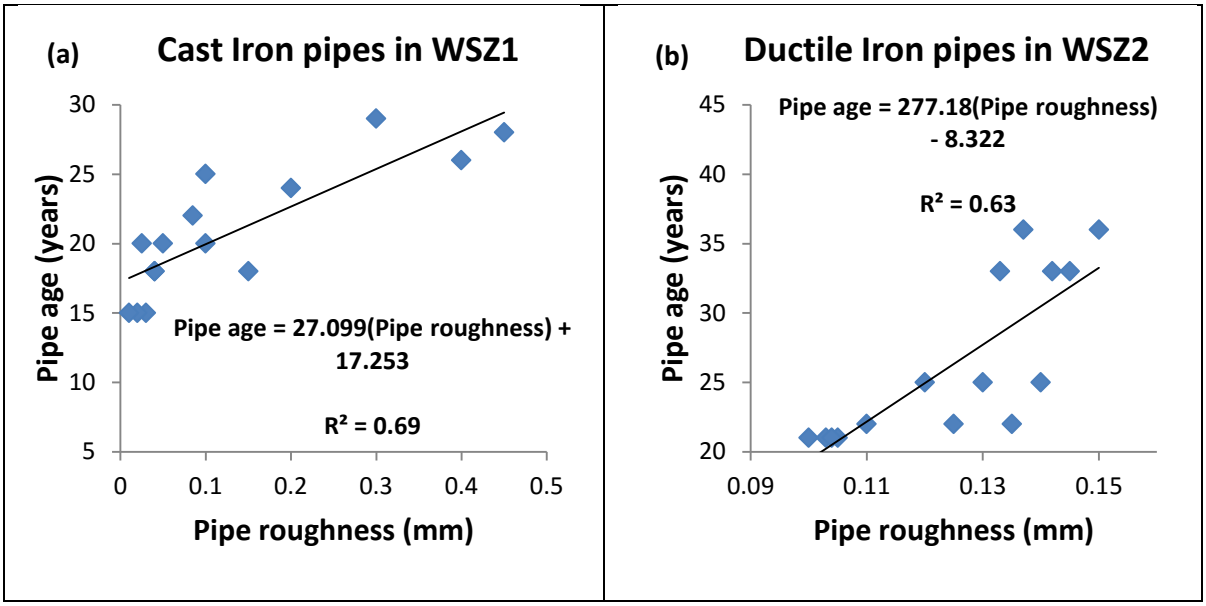


Figure 4.4 Linear regression models used to estimate pipe age for CI and DI pipes

Since the water quality variables were sampled at the nodes, pipe age, which has a pipe property, was converted to a node property in order to make analysis with other variables with node properties possible. This was done by summing the pipe ages of pipes connected to a given node and dividing it with the number of pipes connected to it (similar to Eqn. 4.4).

4.4 Analytical Methods

4.4.1 Spearman's rank correlation

Unlike the Pearson correlation, Spearman's rank correlation is a nonparametric measure that is used to determine correlation between two variables that may have linear or nonlinear (monotonic) relationship (Cohen, 1988; Puth, Neuhäuser, & Ruxton, 2015). Spearman's correlation coefficient, r_s, is a measure of how two variables correlate with each other. It can have a positive or negative value between 0 and 1. A positive value of r_s indicates a positive correlation between the two variables, whereas a negative value of r_s indicates an inverse relation between the variables. The classification of the strength of r_s

is subjective; thus, it depends on the type of data and the purpose of the study. While an r_s value of ±0.7 may be classified as high in clinical research, it may be regarded as medium or low in a research in aeronautics. For this research, because Fe and Mn depend on many variables, some of which are interrelated, high values of R were not expected. Therefore, the classification by Cohen (1988) where $|r_s| > 0.5$ was classified as strong correlation, $0.3 \leq |r_s| \leq 0.5$ as moderate correlation and $0 < |r_s| < 0.3$ as weak correlation was adopted. The equation for calculating r_s is given as:

$$r_s = 1 - \frac{6\sum_{i=1}^{n_s} DF_i^2}{n_s({n_s}^2 - 1)} \tag{4.8}$$

where n_s = the number of pairs of values in the sample; and
DF_i = the difference between ranks of values in i^{th} pair.

The aim of this study is not to predict any variable, but to understand the influence of chemical and biological processes on Fe and Mn accumulation in WDNs. In view of this, Spearman's rank correlation analysis was performed at the DMA level. Fe was used as the dependent variable, and was plotted against each of the 36 water quality variables (dependent variables) in turn. Similarly, Mn was also plotted against each of the 36 water quality variables. When computing r_s for a given pair of variables in a given DMA, it is important to compare its value with those from other DMAs for the same pair of variables. This will give an idea as to whether the two variables are significantly correlated or are correlated by chance. Figure 4.5 shows selected plots to illustrate strong, moderate, and weak correlations between Fe (and Mn) and some water quality variables.

The percentages of graphs at the DMA level with negative or positive correlations of Fe and Mn against the water quality variables were also determined. The knowledge of how an independent variable negatively or positively correlates with a dependent variable is very important because it helps in the formulation of fuzzy rules in FISs. Details of the formation of fuzzy rules are presented in Sections 3.3.4.1 and 6.3.4

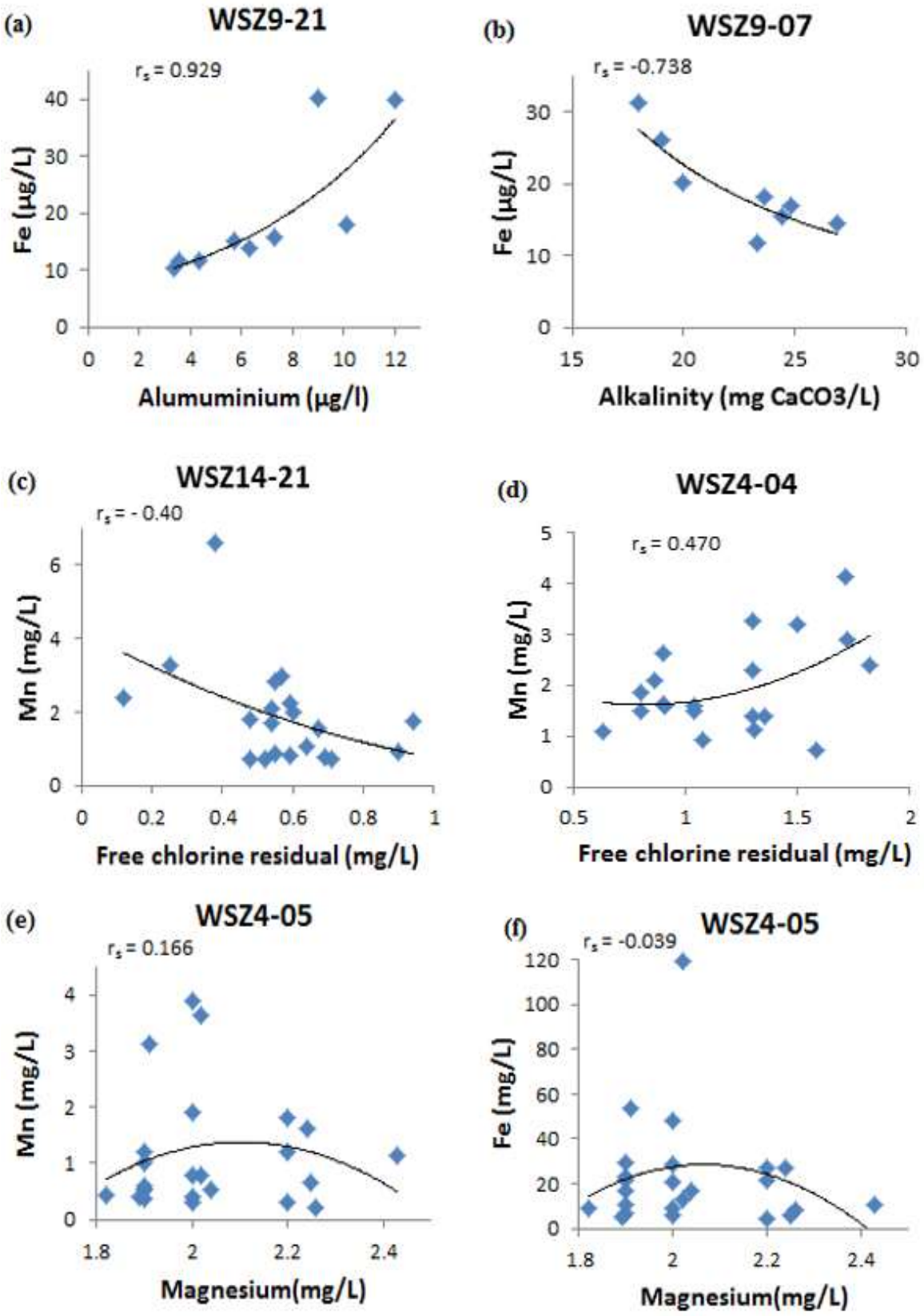


Figure 4.5 Plots showing ((a) and (b)) strong, ((c) and (d)) moderate, and ((e) and (f)) weak correlations between Fe (and Mn) and selected water quality variables

4.4.2 Linear regression

Regression models have been applied in almost every field of study, including economics, medicine, political science, sociology, and psychology. They have also been extensively used in water resource engineering. Some of the research done in water resource engineering includes the work of Murdoch and Shanley (2006), who used segmented regression analysis to assess water quality trends. Rajendra Prasad, Sadashivaiah and Ranganna (2011) used a regression model to predict total dissolved solids based on electrical conductivity values, while Christensen, Rasmussen and Ziegler (2002) developed a real-time water-quality monitoring model that uses regression analysis to estimate nutrient and bacteria concentrations in Kansas Streams, USA. Joarder, Raihan, Alam and Hasanuzzaman (2008) conducted research that used a linear regression equation to predict ground water quality with variables such as electrical conductivity, calcium, and dissolved solids.

Despite extensive use of regression models in the past few decades, they have been superseded by sophisticated models with strong learning capabilities, such as ANN and neuro-fuzzy logic models because of their learning capabilities. Also, the requirement that the variables of most regression models must be continuous and normally distributed makes them inappropriate to use on some data.

Pearson's correlation coefficient, *R*, was used to determine any existing correlations between customer complaints and selected water quality variables. The equation for calculating *R* is given as:

$$R = \frac{\sum_{i=1}^{sp}((X_i - X)(Y_i - \bar{Y}))}{\sqrt{\sum_{i=1}^{sp}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{4.9}$$

where R = Pearson's correlation coefficient; X = independent variable;
Y = dependent variable; $\bar{X}$ and $\bar{Y}$ are the mean of X and Y, respectively; and
sp = the number of observations.

R is a measure of how well a model is likely to make predictions of future outcomes. In a positive correlation, as the values of predictive variable increase, values of determinant

variable also increase. On the other hand, an inverse or negative correlation occurs when the values of predictive variable increase and the values of the determinant variable decrease. *R*, whether positive or negative, range in strength from strong to weak between 0 and 1. For this research, the classification of *R* given by Rodgers and Nicewander (1988) was adopted. In their research, they classified $|R| > 0.5$ as strong correlation, $0.3 \leq |R| \leq 0.5$ as moderate correlation and $0 < |R| < 0.3$ as weak correlation.

4.5 Analysis of chemical variables

Attributing the causes of Fe and Mn accumulation in WDNs to a particular factor can sometimes be very difficult, because the factors that contribute to the accumulation process are complex and interrelated. Some water companies use the following rule of thumb to explain the causes of Fe and Mn accumulation:

- If the Fe/Mn ratio is <10, then biological oxidation is the cause;
- If the Fe/Mn ratio is > 20, then corrosion is the cause;
- If the Fe/Mn ratio is between 10 and 20, then the interpretation is uncertain (Teasdale et al., 2007).

The following sections present the analysis of chemical variables that contribute to Fe and Mn accumulation.

4.5.1 Chemical oxidation analysis

Table 4.2 shows the percentages of Fe (and Mn) graphs for the 36 different water quality variables at the DMA level that exhibited positive or negative correlations. It was observed that 71.28% of graphs had positive correlations between Mn and temperature; whereas 65.00% of graphs had positive correlations between Fe and temperature. The positive correlations observed is due to the vital role temperature plays in the chemical oxidation of soluble Fe^{2+} and Mn^{2+} to insoluble Fe^{3+} and Mn^{4+}, respectively, in WDNs. This result echoes that of a study by Van Benschoten, Lin and Knocke (1992), who observed that increase in temperature increases corrosion rates and the rates of chemical oxidation of Fe and Mn. Table 4.3 shows the percentage of graphs with strong, moderate and weak correlations between Fe (and Mn) and 36 water quality variables. It was observed that 37.96 and 47.20% of the graphs exhibited strong correlations when temperature was plotted against Fe and Mn, respectively.

Many researchers have reported that an increase in pH increases chemical oxidation and corrosion rates in iron pipes (Hidmi et al., 1994; Kashinkunti et al., 1999; Stumm, 1960). However, it was observed from Table 4.3 that there were poor correlations between Fe (and Mn) and pH. This could be due to the low variation in pH observed in the data. The average and standard deviation values of pH observed in the data were 7.25 and 0.3, respectively. As the pH was fairly constant, it did not have any significant influence on chemical oxidation or corrosion rates in iron pipes. The low variation in pH was expected because the water under investigation is for drinking purposes.

From Table 4.2, it was observed that a very high percentage (94.59 %) of graphs had positive correlations when Fe was plotted against Mn. Similarly, it was observed from Table 4.3 that a very high percentage (72.51 %) of the graphs had strong correlations when Fe was plotted against Mn. It was also observed from the data set that, whenever there were Fe failures, Mn also failed. Unsurprisingly, researchers have found several similarities between Fe and Mn. Both Fe and Mn can be chemically oxidised by oxidising agents such as dissolved oxygen and free chlorine residual (FCR) (Knocke et al., 1990; Odell et al., 1998). Furthermore, they can be biologically oxidised by microorganisms such as *Crenothrix*, *Flavobactium*, and *Enterobacter aerogenes* (LeChevallier et al., 1987; Sly et al., 1988).

From Table 4.2, it was observed that 64.71% of the regression graphs of Mn against alkalinity (AKLA) showed negative correlation. Similarly, 70.65% of the graphs of Fe against alkalinity were negatively correlated. Increasing alkalinity causes Fe and Mn concentrations to decrease, because increasing alkalinity increases the buffer capacity of the water and also helps to form calcium or magnesium carbonate layers within the distribution network, thereby reducing corrosion rates. These results conform to a study by Kashinkunti et al. (1999) on alkalinity. They observed in their study that fewer customer complaints regarding water discolouration (which is mainly caused by increased Fe and Mn concentrations) were received when the alkalinity concentration was maintained at a high value of 60 mg $CaCO_3$/L. Studies by Naylor et al. (1993) also showed that corrosion reduces when alkalinity concentrations are higher than 50 mg $CaCO_3$/L.

Table 4.2 Percentages of graphs that exhibited positive or negative correlation when Fe (and Mn) was plotted against selected water quality variables

Variable	Mn		Fe	
	% Positive Correlation	% Negative Correlation	% Positive Correlation	% Negative Correlation
Alkalinity	35.29	64.71	29.35	70.65
Al*	79.61	20.39	77.90	22.10
Ammonia	55.68	44.32	60.05	39.95
Sb*	76.45	23.55	41.65	58.35
As*	45.92	54.08	55.40	44.60
Benzo(b)fluoranthene	59.61	40.39	62.07	37.93
Bromodichloromethane	81.24	18.76	72.95	27.05
Calcium hardness	74.82	25.18	40.87	59.13
Ca*	73.17	26.83	42.58	57.42
Chloride	69.95	30.05	33.59	66.41
Total residual chlorine	40.79	59.21	31.13	68.87
Colour	72.55	27.45	75.29	24.71
Conductivity	59.76	40.24	42.61	57.39
Cu*	43.69	56.31	42.04	57.96
Dibromochloromethane	61.08	38.92	36.06	63.94
FCR	38.64	61.36	33.08	66.92
Fe*	94.59	5.41	-	-
Pb*	52.18	47.82	55.01	44.99
Mg*	73.98	26.02	54.83	45.17
Magnesium hardness	72.51	27.49	45.15	54.85
Mn*	-	-	94.59	5.41
Ni*	43.27	56.73	44.06	55.94
Nitrite	37.43	62.57	36.39	63.61
Nitrate	38.01	61.99	41.26	58.74
Nitrite plus Nitrate	42.09	57.91	40.71	59.29
Total Oxidised Nitrogen	41.74	58.26	39.30	60.70
pH	53.31	46.69	48.04	51.96
P*	70.02	29.98	63.42	36.58
Na*	76.61	23.39	65.43	34.57
Tetrachloroethane	66.07	33.93	60.19	39.81
Temperature	71.28	28.72	65.00	35.00
Tribromomethane	54.55	45.45	45.75	54.25
THM	81.03	18.97	65.35	34.65
Trichloromethane	75.25	24.75	69.07	30.93
Trichloroethane	65.55	34.45	62.75	37.25
Turbidity	78.80	21.20	82.30	17.70

* These are measured totals; e.g. Fe* contains (Fe^{2+} and Fe^{3+}).

Table 4.3 Percentage of graphs with strong, moderate and weak correlations when Fe (and Mn) was plotted against selected water quality variables

	Correlation Strength of Fe (%)			Correlation Strength of Mn (%)		
Variable	**Strong**	**Moderate**	**Weak**	**Strong**	**Moderate**	**Weak**
Alkalinity	49.56	25.32	25.12	40.67	30.01	29.32
Al*	40.23	29.80	29.97	45.52	29.95	24.53
Ammonia	26.02	27.08	46.90	24.46	25.73	49.81
Sb*	25.23	25.00	49.77	36.17	45.82	18.01
As*	27.46	30.76	41.78	25.06	35.81	39.13
Benzo(b)fluoranthene	42.67	25.93	31.40	25.03	30.33	44.64
Bromodichloromethane	15.46	48.23	36.31	35.26	33.58	31.16
Calcium hardness	28.23	29.4	42.37	32.24	29.89	37.87
Ca	30.86	27.46	41.68	35.45	27.52	37.03
Chloride ions	26.19	20.46	53.35	38.55	41.84	19.61
Total residual chlorine	18.67	23.81	57.52	21.42	29.57	49.01
Colour	31.53	32.09	36.38	27.14	29.44	43.42
Conductivity	35.26	22.77	41.97	32.51	20.49	47.00
Cu*	15.91	29.16	54.93	13.99	27.41	58.6
Dibromochloromethane	19.84	21.48	58.68	21.37	35.06	43.57
FCR	35.29	39.07	25.64	31.26	29.47	39.27
Fe*	-	-	-	72.51	22.77	4.72
Hardness Total as $CaCO_3$	29.49	18.26	52.25	32.19	30.05	37.76
Pb*	17.91	18.01	64.08	16.59	24.90	58.51
Mg*	28.05	28.43	43.52	28.94	29.68	41.38
Magnesium hardness	26.55	27.28	46.17	28.2	25.19	46.61
Mn	72.51	22.77	4.72	-	-	-
Ni*	21.83	19.65	58.52	23.99	27.52	48.49
Nitrite	17.23	22.15	60.62	19.17	27.00	53.83
Nitrate	28.69	21.68	49.63	24.84	20.08	55.08
Nitrite plus Nitrate	22.38	19.53	58.09	28.87	20.04	51.09
Total Oxidised Nitrogen	33.93	19.57	46.5	31.82	28.00	40.18
pH	19.26	23.60	57.14	12.91	17.85	69.24
P*	23.56	30.12	46.32	28.88	35.95	35.17
Na*	24.91	31.85	43.24	21.00	30.40	48.60
Tetrachloroethane	35.79	25.03	39.18	41.28	25.67	33.05
Temperature	37.96	28.29	33.75	47.20	23.24	29.56
Tribromomethane	29.85	21.62	48.53	30.17	20.07	49.76
THM	58.65	25.46	15.89	64.89	21.36	13.75
Trichloromethane	62.40	20.68	16.92	70.65	15.63	13.72
Trichloroethane	53.45	21.89	24.66	28.90	25.85	45.25
Turbidity	42.40	23.56	34.04	38.22	26.91	34.87

* These are measured totals; e.g. Fe* contains (Fe^{2+} and Fe^{3+}).

4.5.2 Corrosion analysis

Corrosion is known to be one of the most common causes of drinking water discolouration (DWI, 2007). There are several similarities between corrosion and chemical oxidation. Increase in FCR, dissolved oxygen, and hardness cause both corrosion and chemical oxidation to increase. Also, increased corrosion and chemical oxidation levels causes Fe concentrations to increase. On the other hand, increase in alkalinity decreases both corrosion and chemical oxidation. The main difference between corrosion and chemical oxidation is that chemical oxidation occurs in all types of pipes, whereas corrosion mainly occurs in ferrous pipes. The source of Fe in ferrous pipe mainly comes from the inner surface of the ferrous pipe walls. The age of ferrous pipes in WDNs also has a significant effect on corrosion. The accumulation of corrosion by-products over many years can reduce pipe diameter and cause water to discolour. Knowing the causes of increased Fe concentrations in drinking water will determine the kind of solution drinking water companies will provide.

A graph of Fe (and Mn) concentrations against pipe material index for most of the WSZs showed that, generally, ferrous pipe materials had higher Fe and Mn concentrations than non-ferrous pipe materials (see Fig. 4.6 (a) and 4.6 (b)). Similarly, it was observed that most WSZs had high Fe and Mn concentrations with increasing pipe age (see Fig. 4.6 (c) and 4.6 (d)). These observations conform to studies by Cook et al. (2005) and Cerrato et al. (2006), who found that networks with mainly ferrous pipes are more prone to discolouration due to corrosion.

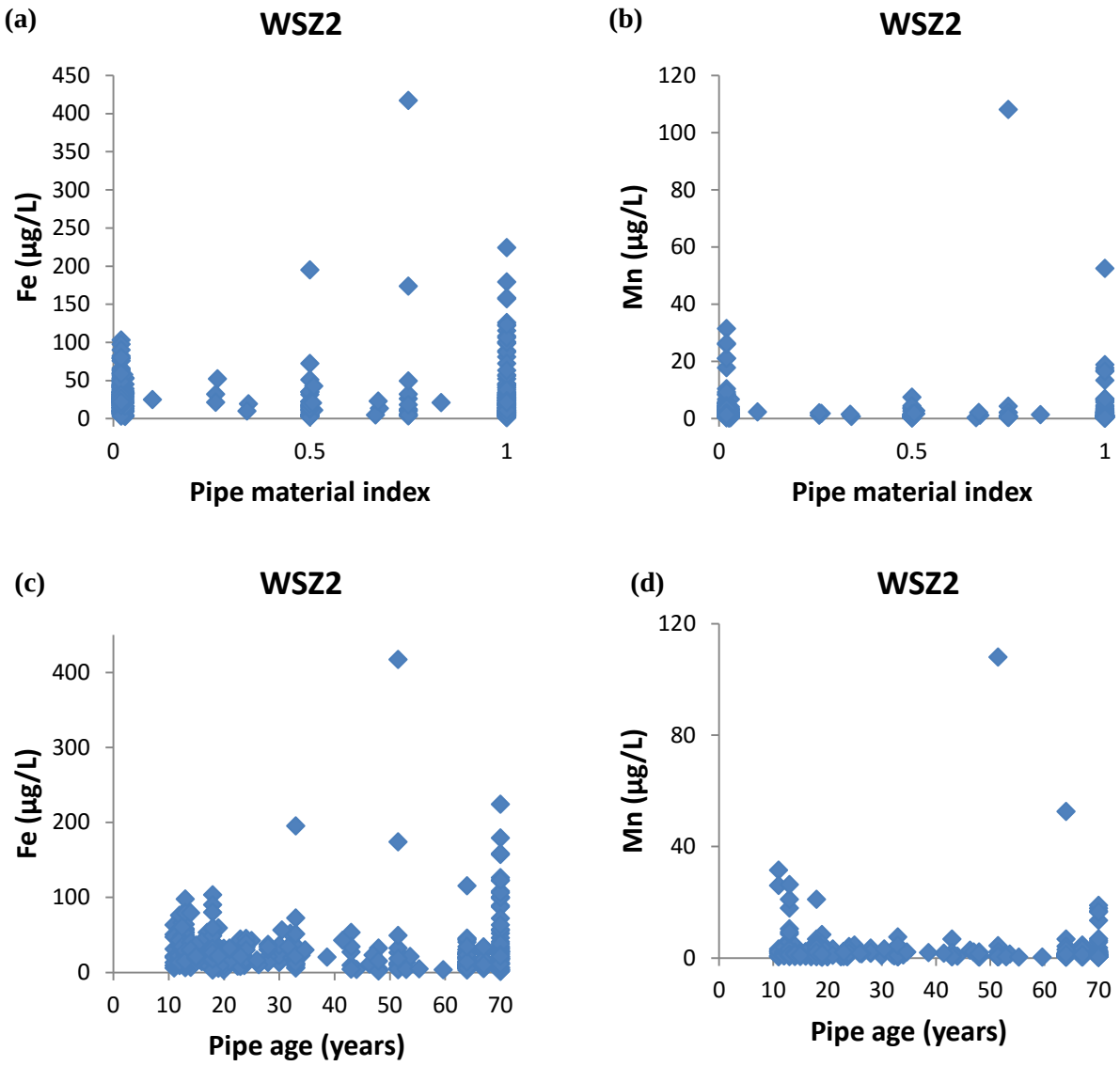


Figure 4.6 Variation of Fe (and Mn) with pipe material index and pipe age

4.5.3 Sorption variables analysis

From table 4.2, a high percentage of the graphs (77.90 and 79.61 %) exhibited positive correlations when Fe and Mn were plotted against Al, respectively. Residual amounts of Al enter the distribution system as a result of the coagulant, aluminium sulphate (Alum/$Al_2(SO4)_3$). Alum is added to raw water for the removal of particulates, dissolved substances, and colloids as part of the drinking water treatment process. Increased alum concentration may result in the formation of amorphous $Al(OH)_3$, a compound that numerous researchers have found to have sorption capabilities (Dayton & Basta, 2005; Wang et al., 2012). The high percentage of positive correlations could be due to the sorption of Fe and Mn on amorphous $Al(OH)_3$.

4.6 Analysis of variables affecting biological processes

It was observed that 65.00% of the graphs exhibited positive correlation when Fe was plotted against water temperature (see Table 4.2). Similarly, 71.28% of the graphs exhibited positive correlation when Mn was plotted against water temperature. The positive correlations can be attributed to the fact that increase in water temperature increases bacterial growth rates and biological oxidation of soluble Fe^{2+} and Mn^{2+} to insoluble Fe^{3+} and Mn^{4+}, respectively, in WDNs.

From Table 4.2, it was observed that significant percentages of graphs exhibited positive correlation when Fe (and Mn) was plotted against the organic variables bromodichloromethane, dibromochloromethane, and THM. This could be due to the reaction of FCR with natural organic matter (NOM) when chlorine dissipates. When chlorine decays, it reacts with NOM to form toxic disinfection by-products such as bromodichloromethane, dibromochloromethane, and THM. Furthermore, when chlorine decays, it increases biological oxidation (thus, increases Fe and Mn accumulation). This observation conforms to studies by Di Cristo, Esposito and Leopardi (2013) and Seyoum and Tanyimboh (2014), who observed that low FCR concentrations generally correspond to high THM levels in drinking water. Although these organic variables had significant positive correlations with Fe (and Mn), they could not be used in the ANN models and the FISs developed in Chapters 5 and 6, respectively, because they were not sampled frequently.

Although FCR is an oxidising agent and was expected to chemically oxidise Fe and Mn. However, it was observed that significant percentages (66.92 and 61.36 %) of graphs exhibited negative correlation when Fe (and Mn) was plotted against FCR, respectively (see Table 4.2). This could be due to the fact that high concentrations of FCR kill bacteria that help to biologically oxidise Fe and Mn in the distribution system. Numerous studies have shown that water colour is an indirect measure of total organic carbon (TOC), the main nutrient for bacteria (Effler, Schafran, & Driscoll, 1985; Evans, 1988; Gorham, Underwood, Martin, & Ogden, 1986). Since high levels of TOC influence biofilm formation, it explains why significant percentages of graphs exhibited positive correlation when Fe was plotted against colour (75.29 %) and Mn against colour (72.55 %).

Further investigations on variables that influence biological processes were carried out by plotting Fe and Mn against FCR for all 176 DMAs. The graph of Fe against FCR showed that there were a significant number of Fe failures (concentrations exceeded the MCL of 200 µg/l) when FCR was less than 0.8 mg/l (see Fig. 4.7). However, there were no Fe failures when FCR exceeded 0.8 mg/l. Likewise, a graph of Mn against FCR showed that there were a significant number of Mn failures (concentrations exceeded the MCL of 50 µg/l), but no failures when FCR was greater than 0.8 mg/l (see Fig. 4.8). This could be an indication that most of the oxidation that occur within the distribution system could be microbial induced and that free chlorine residual concentrations above 0.8mg/L were able to kill or reduce the growth of Fe- and Mn-oxidising bacteria and hence prevented them from oxidising soluble Fe^{2+} and Mn^{2+} to Fe^{3+} and Mn^{4+} precipitates, respectively. FCR is needed in the water distribution system to deactivate the growth of microorganisms and preserve water quality.

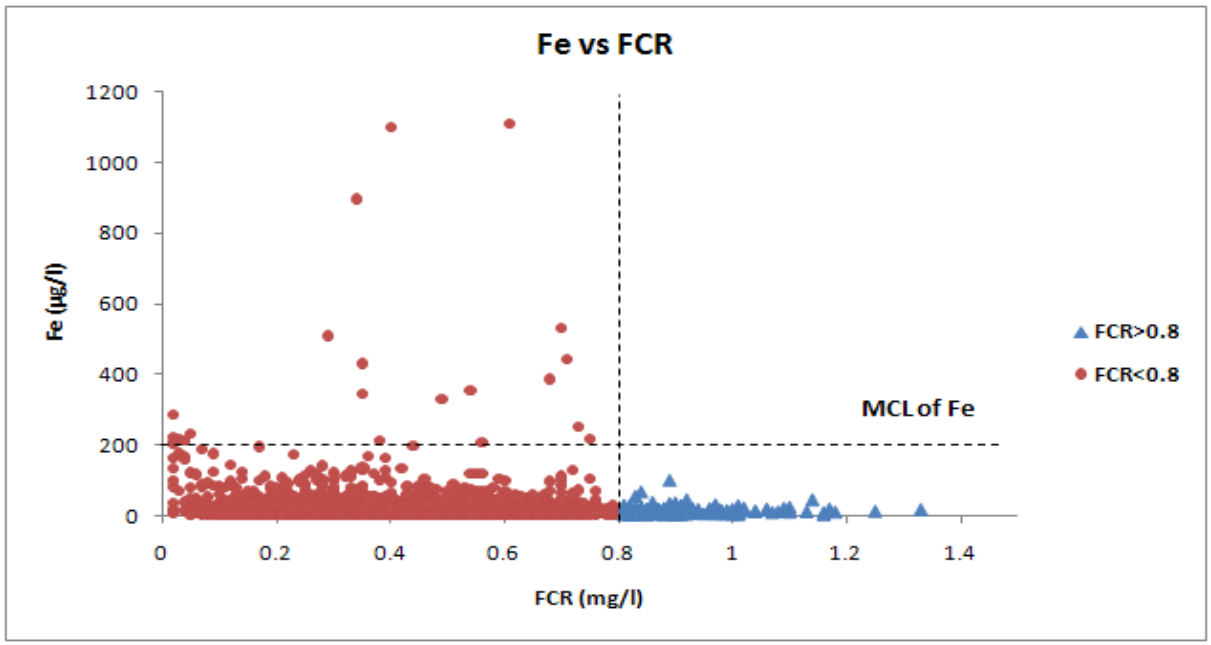


Figure 4.7 Variation of Fe with free chlorine residual for all 176 DMAs

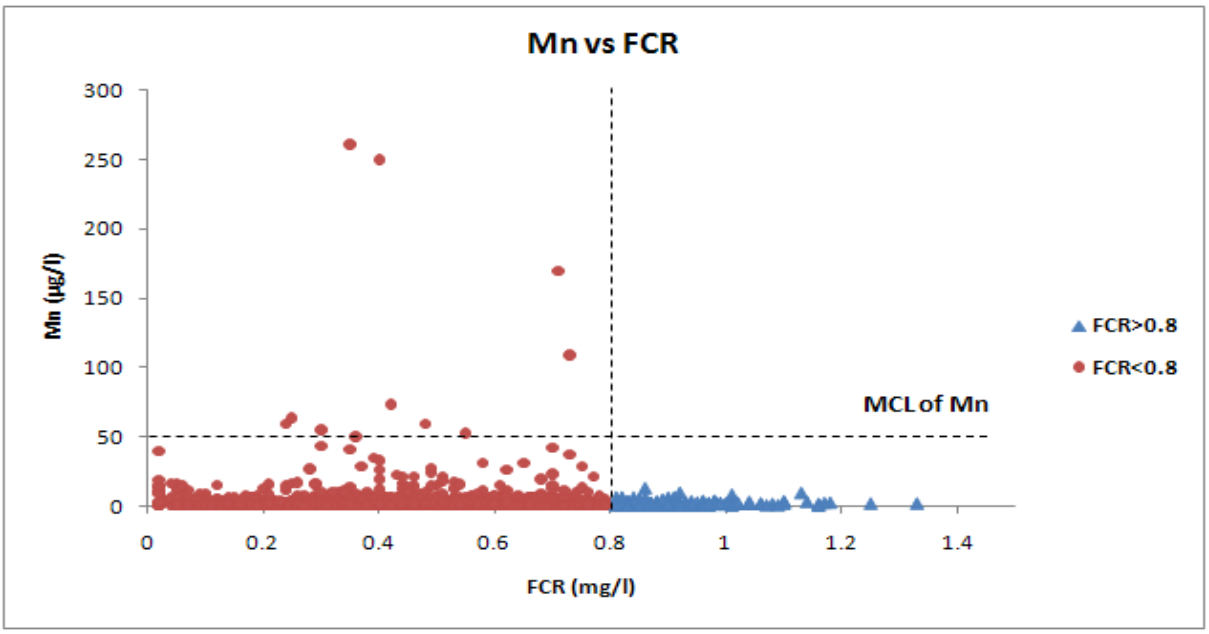


Figure 4.8 Variation of Mn with free chlorine residual for all 176 DMAs

4.7 Customer complaints analysis

Quarterly customer complaints were plotted against some selected water quality variables at the DMA level and their Pearson's correlation coefficient were determined. The percentages of graphs that exhibited positive correlation when quarterly customer complaints were plotted against quarterly averages of water quality variables are presented in Table 4.4. The top three percentages of graphs observed were Mn, Fe and Al were 65.67, 58.82, and 55.74%, respectively. This explains why these three variables have all been used as Key Performance Indicators (KPIs) in customer complaints studies (Bernal, Cardenoso, Fabrellas, Matia, & Salvatella, 1999; Ewan & Williams, 1986; Gauthier et al., 1999).

Table 4.4 Percentages of graphs that exhibited positive correlation when quarterly customer complaints were plotted against quarterly average water quality variables

Variable	%Positive Correlation	%Negative Correlation
Mn	65.67	34.33
Fe	58.82	41.18
pH	52.17	47.83
Turbidity	55.74	44.26
FCR	52.44	47.56
Total residual chlorine	54.88	45.12
Al	57.63	42.37

Table 4.5 Percentage of graphs with different levels of correlation when quarterly customer complaints were plotted against quarterly average water quality variables

Variable	%Strong Correlation	%Moderate Correlation	%Weak Correlation
Mn	38.81	20.90	40.30
Fe	33.82	19.12	47.06
pH	13.04	28.99	57.97
Turbidity	36.07	18.03	45.9
FCR	13.41	35.37	51.22
Total residual chlorine	14.63	36.59	48.78
Al	33.90	22.03	44.07

From Table 4.5, it was observed that the top four strongest correlations between quarterly customer complaints and quarterly average water quality variables were Mn, turbidity, Al, and Fe. Although these four variables have been used as KPIs in customer complaints studies, they had relatively low correlations. The weak correlation strength exhibited by the customer complaints data with some of the water quality variables could be due to a number of factors. First, it could be attributed to the numerous and interrelated water quality variables that causes water discolouration. Secondly, it could be due to the effect of physical and hydraulic variables that also contribute to water discolouration. Finally, some customers tend not to complain, even when they are dissatisfied with water quality. A survey conducted by Sydney Water, Australia, indicated that only 7% of customers that experienced aesthetically unpleasant water over a 12 month period complained (Roseth, 2002). In a related survey at South East Water in Melbourne, Australia, only 15% of customers who experienced water discolouration complained (Roseth & Rock, 2003). A study conducted by Ewan and Williams (1986) in the UK showed that only around 30% of customers that experienced discoloured water actually complained. Evins, Liebeschuetz

and Williams (1990) also pointed out in their research that if customers do not complain, it does not always imply water quality standards are good. In a related study, Chadderton, Christensen and Henry-Unrath (1992) noted that customer complaints could be misleading in that a single discolouration event at a DMA could lead to multiple customer complaints. This could consequently send incorrect signals of high-risk of water discolouration at that DMA. However, with all the above shortcomings, it is advisable for water companies to carry out some preliminary analysis of customer complaints data along with water quality data to gain insight into network deposition dynamics.

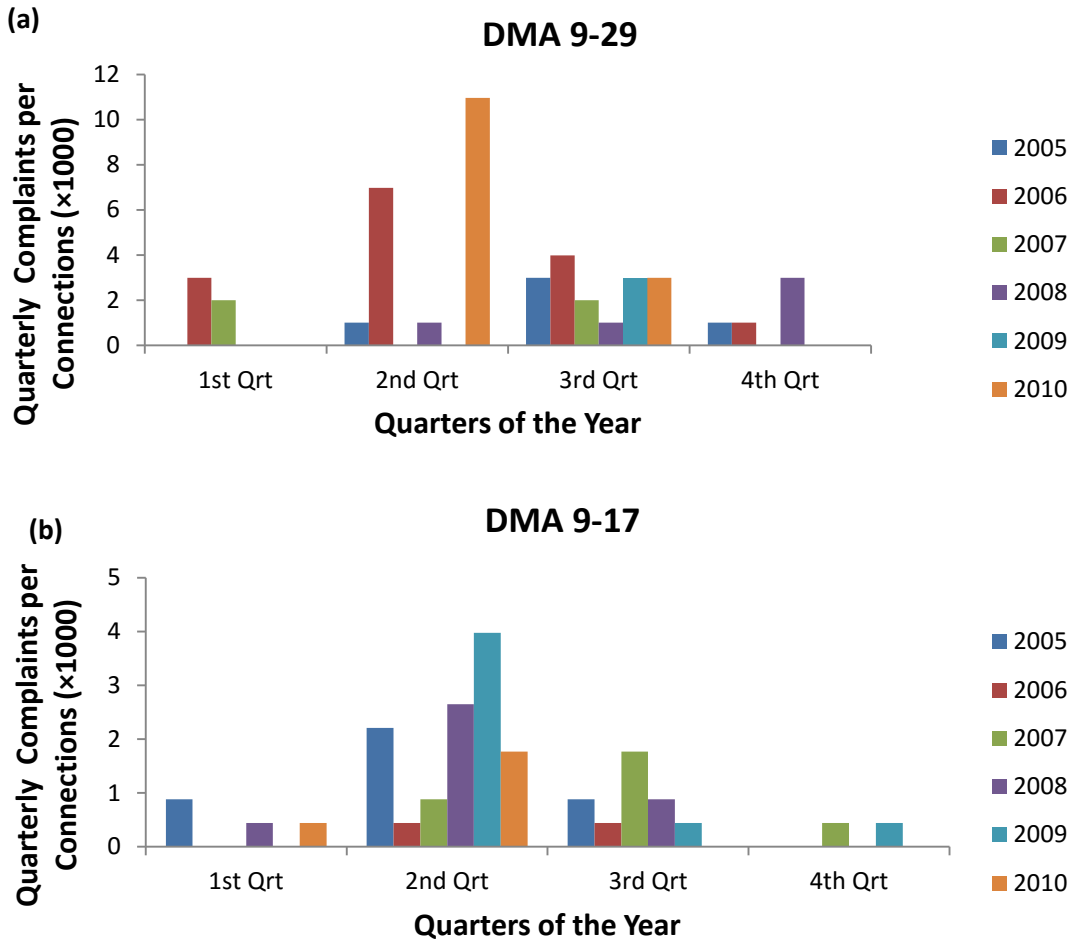


Figure 4.9 Seasonal variations of customer complaints in DMAs

In the UK, temperatures are relatively high during the second and third quarters of the year, whereas they are very low during the first and fourth quarters. High temperatures promote bacterial growth, which causes biological oxidation from soluble Fe and Mn to insoluble precipitates in WDNs. High temperatures also cause chemical oxidation of Fe and Mn. Unsurprisingly, analysis of customer complaints data showed that, in total, 116 out of the

176 DMAs exhibited clear seasonal variations of customer complaints, with peaks during the second and third quarters of the year. Figure 4.9 shows seasonal variation of customer complaints for two DMAs. The high number of customer complaints during this period could also be attributed to high water consumption. Water companies normally experience high demand for water during the second and third quarters of the year. This excess demand causes increased flow velocity and shear stress, which dislodges accumulated Fe and Mn from the pipe walls, causing water discolouration. Although customer complaints is a good KPI for water discolouration, the variable and subjective nature of the manner in which customers complain make using it alone for making predictions sometimes misleading.

4.8 Analysis of hydraulic variables

4.8.1 Analysis of maximum daily shear stress at node

Fe and Mn concentrations were plotted against the maximum daily shear stress at nodes for each of the WSZs (see Figs. 4.10 (a), (b) and 4.11 (a), (b)). Low maximum daily shear stress regions were found in sections of the pipe network that have dead ends and redundant loops, whereas high maximum daily shear stress regions were mainly found in trunk mains and regions with high water demand. From these graphs, it was observed that areas with high maximum daily shear stress had low Fe and Mn concentrations. This is because Fe and Mn precipitates are unable to accumulate on the pipe walls under these high hydraulic conditions. This observation echoes that of a study by Boxall et al. (2001). They observed that, in general, high shear stress regions are subject to low accumulation potential. These regions also have low water age; as a result, biological oxidation of soluble Fe^{2+} and Mn^{2+} to insoluble Fe^{3+} and Mn^{4+}, respectively, is minimal. In contrast, it was observed that regions with low daily maximum shear stress had high concentrations of Fe and Mn. This is because the shear stress exerted on the pipe walls in these regions are not high enough to dislodge any deposits of Fe, Mn, or biofilms. This condition creates a conducive environment for sorption to take place and also the accumulation of Fe and Mn particles. Low shear stress causes water age to be increased in these regions. These stagnant conditions promote the growth of bacteria, increase biological oxidation and result in the deterioration of water quality. Low shear stress regions are generally subjected to high Fe and Mn accumulation.

From Figs. 4.10 (c) and 4.11 (c), it was also observed that regions with high maximum daily shear stress had low turbidity, whereas regions with low maximum daily shear stress had high turbidity. This observation conforms to a research by Boxall et al. (2001, 2003), who suggested that discolouration materials are more likely to accumulate in networks subjected to low conditioning daily shear stress than networks with high conditioning daily shear stress. This means that pipes with low conditioning shear stress have higher discolouration potential than pipes with high conditioning shear stress.

Regions with high maximum daily shear stress had low FCR, whereas regions with low maximum daily shear stress had high FCR (see Figs. 4.10 (d) and 4.11 (d)). This is because low maximum daily shear stress regions mostly occur at dead ends, where water age is very high and chlorine dissipation is quite rapid. Since FCR helps to kill or reduce microbial growth, regions with low concentrations of FCR are more susceptible to biological oxidation of Fe and Mn than regions with high FCR. This results in high Fe and Mn concentrations in chlorine dissipated regions.

It can also be seen that most regions with low maximum daily shear stress also have high Al concentrations, whereas regions with high maximum daily shear stress have low Al concentrations (see Figs. 4.10 (e) and 4.11 (e)). Regions with high Al concentration tend to form amorphous $Al(OH)_3$ in WDNs, a compound that numerous researchers have identified to have sorption capabilities (Dayton & Basta, 2005; Wang et al., 2012). The amorphous $Al(OH)_3$ tends to adsorb and absorb Fe and Mn particles, and accumulates on pipe walls, resulting in high concentrations of Fe and Mn.

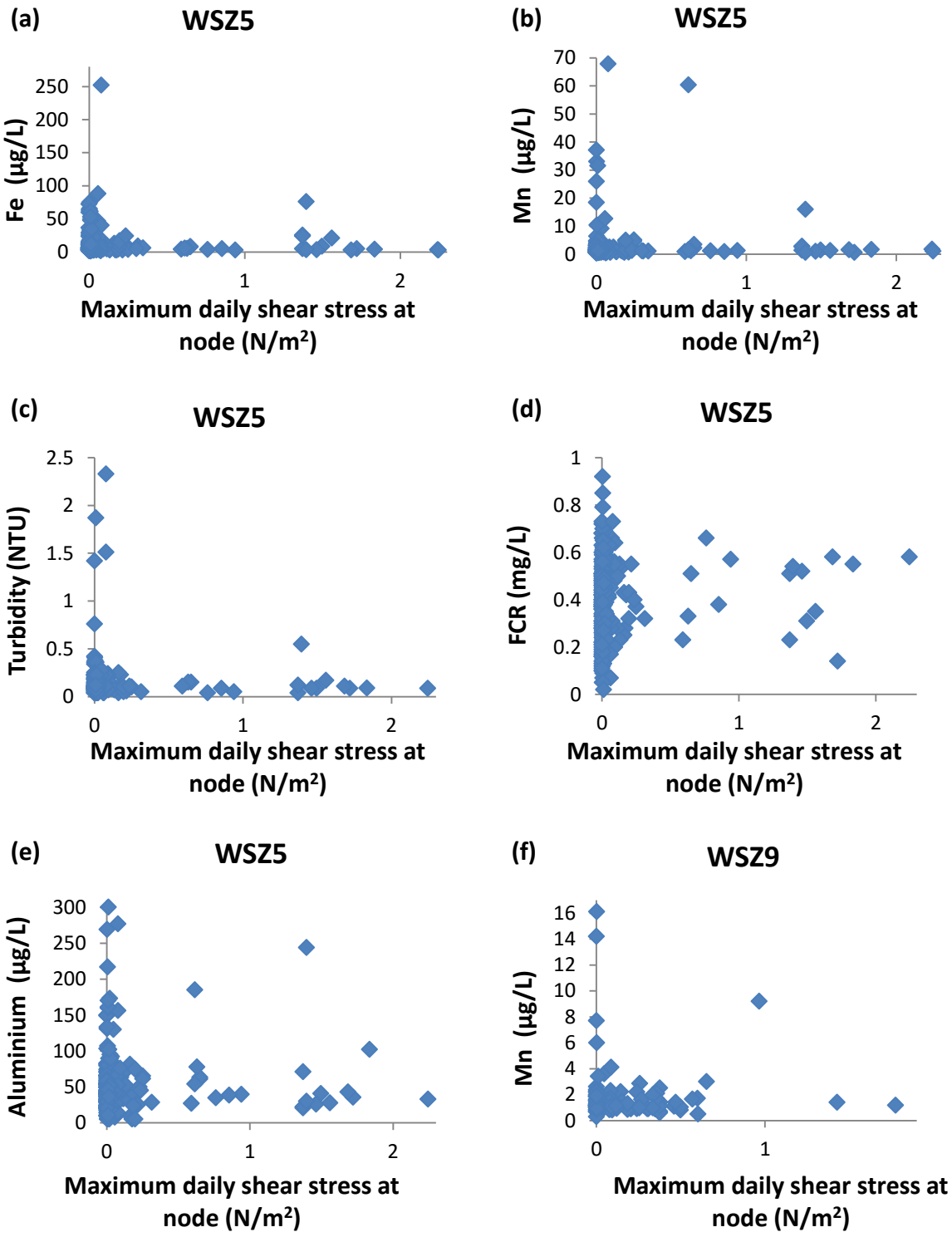


Figure 4.10 Variation of water quality variables with maximum daily shear stress

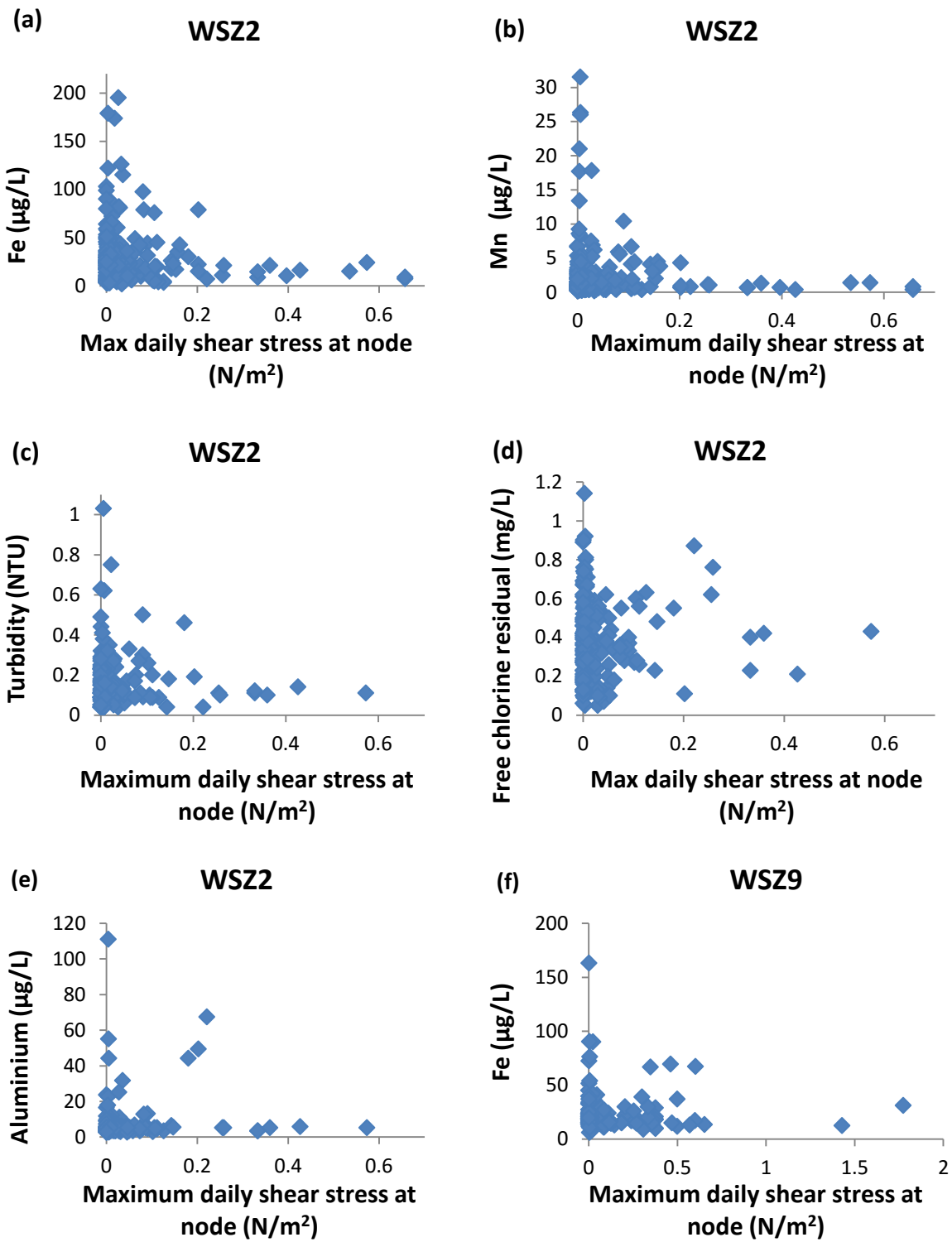


Figure 4.11 Variation of water quality variables with maximum daily shear stress

4.8.2 Analysis of variation of daily shear stress at node

From Fig. 4.12 (a), (b), and (d), it was observed that Fe, Mn, and Al all had high concentrations at low variation of daily shear stress and low concentrations at high variation of daily shear stress. Nodes with low variation of daily shear stress generally

have low disturbance in the WDNs. This condition makes Fe and Mn particles deposit or attach to the walls (sorption) easily without being dislodged. Hence, Fe, Mn, and Al have high concentrations in regions with low variation of daily shear stress. Conversely, nodes with high variation of daily shear stress generally have high disturbance in the WDNs. Hence, Fe, Mn, and Al particles in these regions are not able to accumulate on the pipe walls. This results in low concentrations of these water quality variables in these regions. In general, pipes with high variation of daily shear stress have low Fe and Mn accumulation potential, whereas pipes with low variation of daily shear stress have high Fe and Mn accumulation potential.

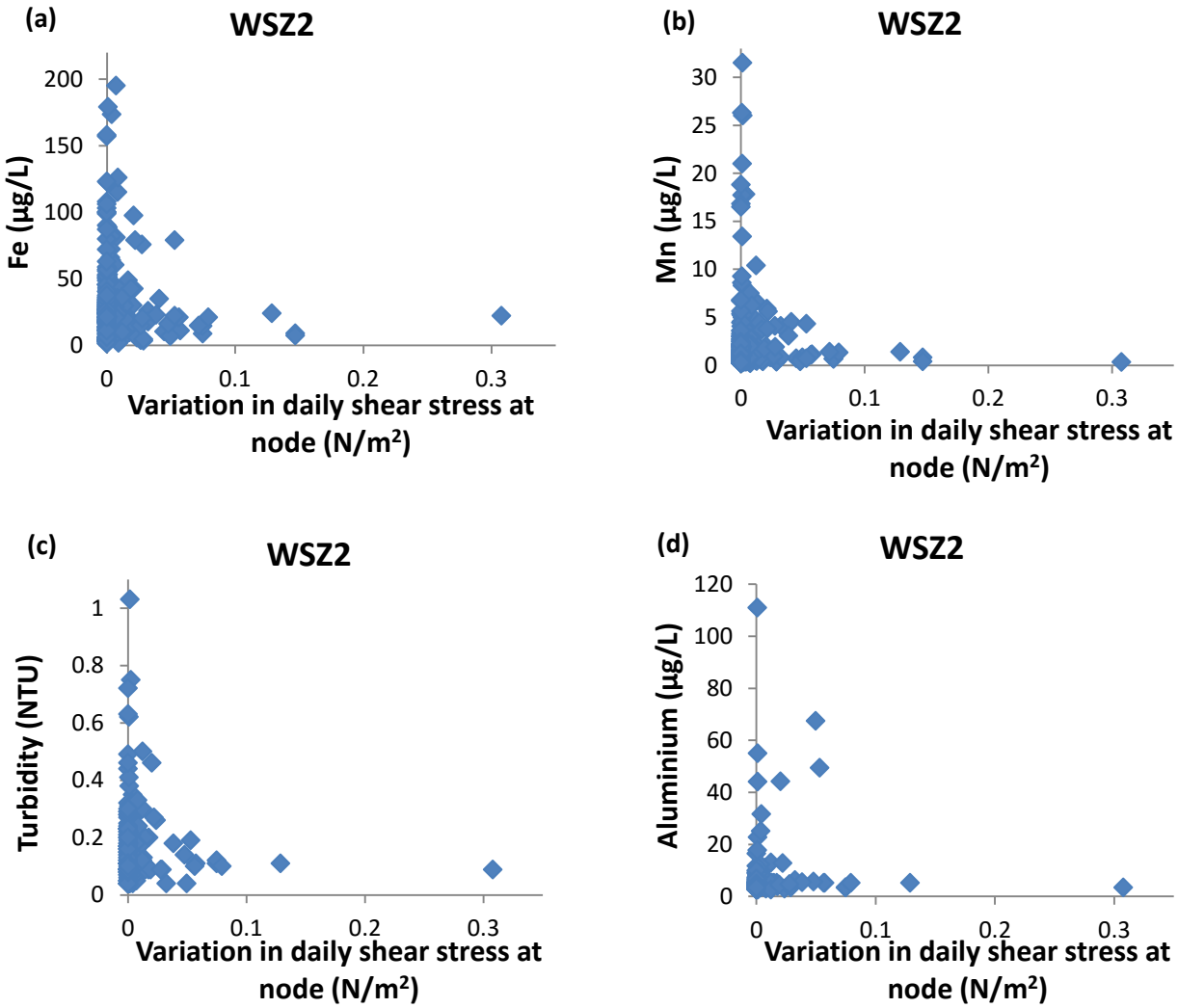


Figure 4.12 Variation of water quality variables with variation of daily shear stress

It was observed that regions with low variation of daily shear stress had high turbidity levels, whereas regions with high variation of daily shear stress had low turbidity levels

(see Fig. 4.12 (c)). The high turbidity levels in regions with low variation of daily shear stress are due to the high tendency of sediments to be deposited on the pipe walls under these conditions. In regions with low variation of daily shear stress, unexpected events such as high flows created by water mains bursts or opening of fire hydrants can increase variation of daily shear stress. This causes loose sediments to be re-suspended, and subsequently lead to increased turbidity levels. In contrast, sediments are unable to accumulate on pipe walls in regions with high variation of daily shear stress. This is because the high shear stress experienced in these regions does not allow the deposited or attached sediments to pile up before they are dislodged from the pipe walls.

4.8.3 Analysis of water age

Water age is a very important variable that influences water quality within the distribution system. Figure 4.13 shows the distribution of water age after 72 hours of simulation at WSZ2. It was observed that high water ages were predominantly found in regions with dead ends and redundant loops from the distribution of water age in all WSZs. The water age in a WDN also depends on its mode of operation and physical variables such as the flow rate, pipe size, configuration, water demand, system design, and amount of storage. WDNs with high flow rates and small pipe sizes will have a lower water age. Water quality problems associated with water age include poor taste, bad odour, increased microbial growth, discolouration, and increased water temperature.

As water travels through WDNs, it goes through a number of bio-chemical processes. When drinking water stagnates in the network, chlorine dissipates, water temperature increases, and its quality degrades. This means regions with high water age create a conducive environment for microbial growth. Researchers have found that microorganisms such as *Crenothrix, Flavobactium, Pseudomonas, Leptothrix discophora,* and *Clonothrix* are able to biologically oxidise soluble Fe and Mn to insoluble Fe and Mn (LeChevallier, 1987; Sly et al., 1988). These bacteria assist in the formation of biofilms on the pipe walls. The Fe and Mn precipitates formed are easily attached to these gelatinous biofilms. This causes regions with high water age to often have high Fe and Mn concentrations. The design of WDNs can help reduce water discolouration due to increased water age. It is recommended that dead ends be prevented or looped, reservoir turnover be increased, oversized mains be reduced, and stagnant zones be routinely flushed.

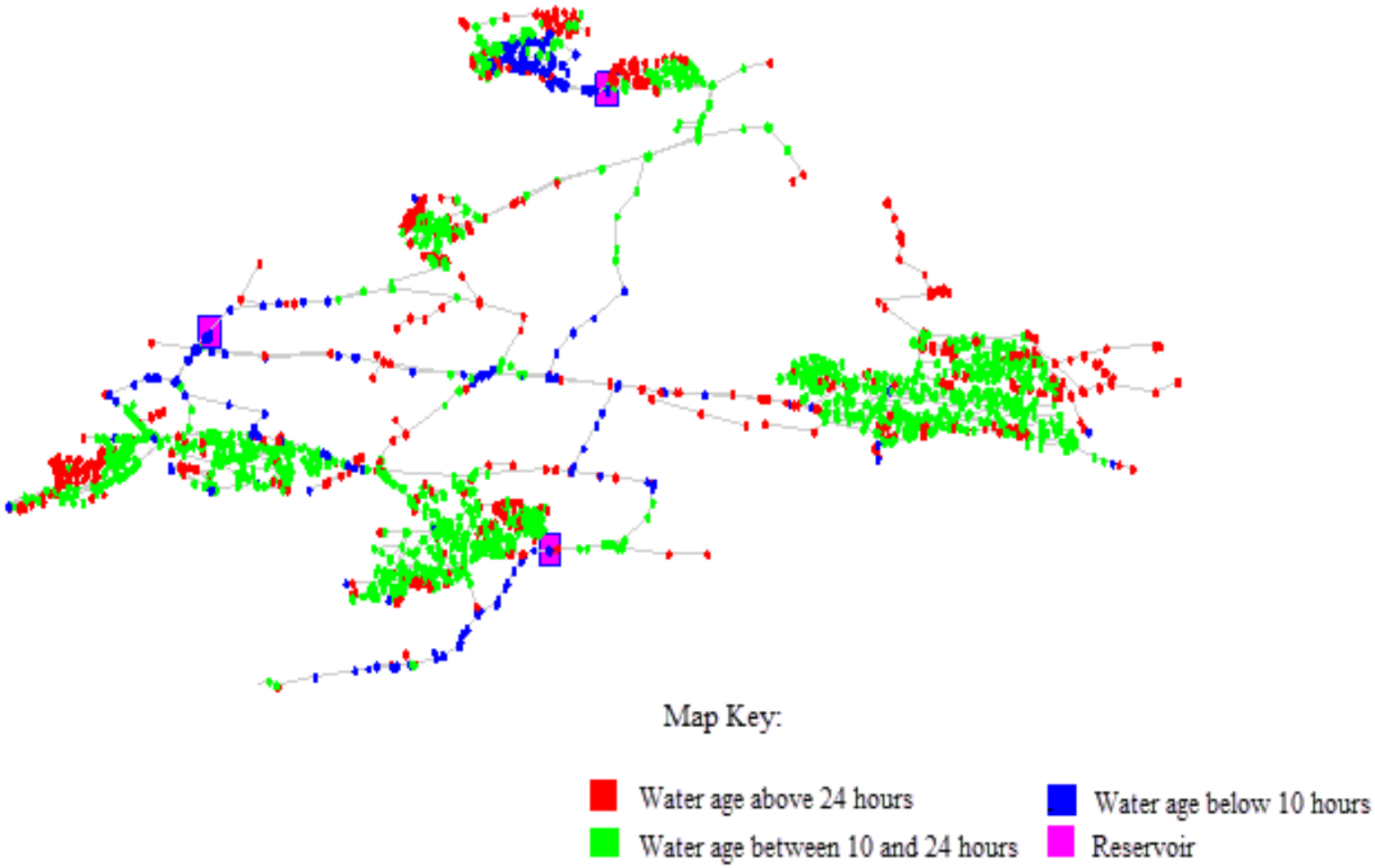


Figure 4.13 The distribution of water age after 72 hours of simulation at WSZ2

4.8.4 Analysis of hydraulic distance from source of water supply

Figure 4.14 shows a gradual increase in Fe and Mn concentrations as hydraulic distance from source of water supply increases. In general, the further water travels through WDNs, the higher the water age and the more chlorine is dissipated. Because chlorine is a disinfectant, it suppresses or kills Fe- and Mn-oxidising bacteria, preventing the biological oxidation of soluble Fe and Mn to insoluble Fe and Mn. Hence, regions with shorter hydraulic distances from source of water supply have lower Fe and Mn concentrations. In contrast, regions with longer hydraulic distances from source of water supply have lower concentrations of FCR. This increases microbial growth, which causes biological oxidation of Fe and Mn, and subsequently leads to increased Fe and Mn concentrations.

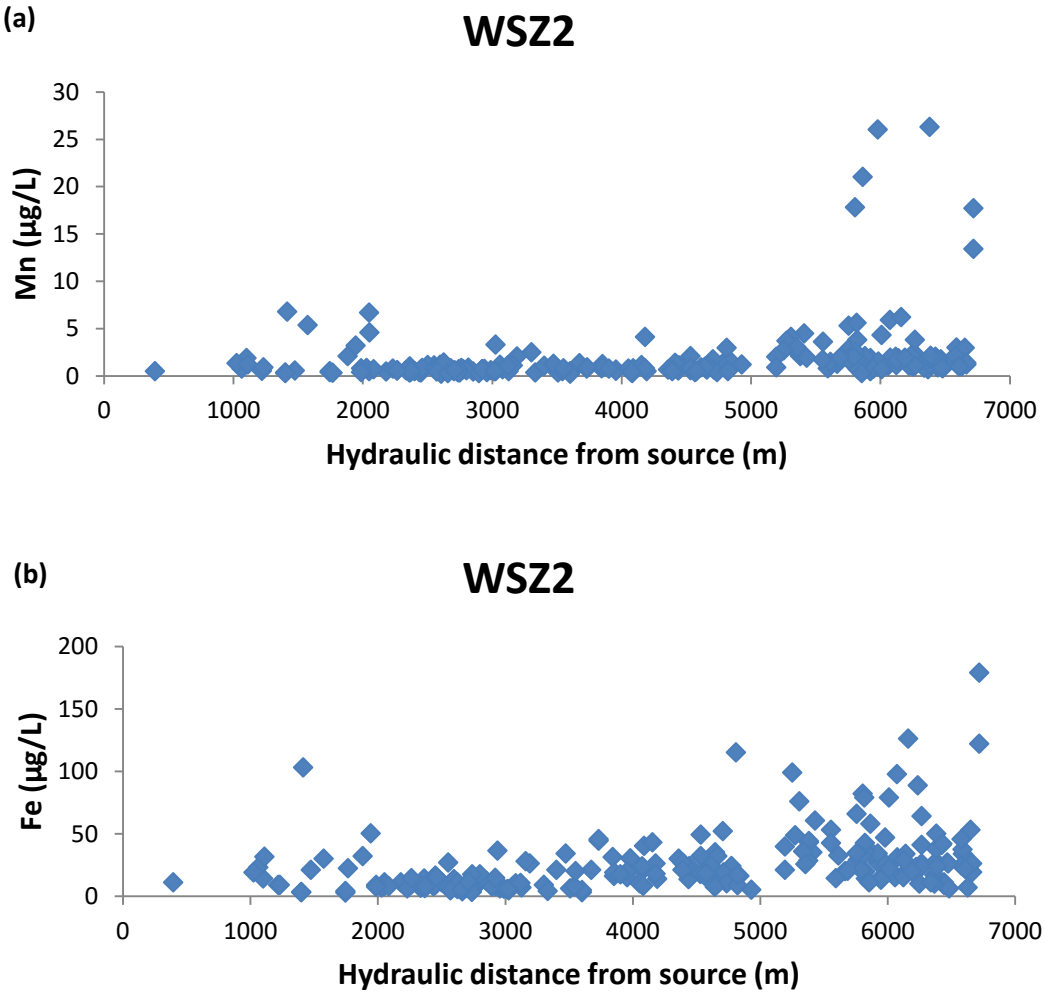


Figure 4.14 Variation of Fe (and Mn) with hydraulic distance in WSZ2

4.9 Summary

After analysing the five-year data set, the findings indicate that alkalinity is a very important variable in drinking water for reducing discolouration. It was observed that increase in alkalinity lowered both Fe and Mn concentrations. Alkalinity serves as a buffer that prevents large variations in pH; a variable when in excess can increase chemical oxidation. Furthermore, high concentrations of alkalinity help to form protective layers of calcium or magnesium carbonate within WDNs, which reduces corrosion rates.

Both Fe and Mn showed high positive correlations with Al. This finding is a probable consequence of residual amounts of Al entering the WDNs after the raw water has been treated with the coagulant $Al_2(SO4)_3$. When these residual amounts occur in high quantities, amorphous $Al(OH)_3$ can be formed, a compound that has been found to have sorption capabilities. This subsequently leads to the sorption of Fe and Mn particles on amorphous $Al(OH)_3$.

The seasonal trend of customer complaints observed with peaks during the second and third quarters of the year could be due to high temperatures. High temperatures are known to enhance biological and chemical oxidation of Fe and Mn. They could also be due to excess consumption of water during this period. Increased water consumption increases water velocity, dislodges accumulated Fe and Mn particles from the pipe walls and subsequently leads to water discolouration.

It was observed for all the DMAs in this study that when FCR concentrations were above 0.8mg/L, neither Fe nor Mn concentrations exceeded their respective MCLs. This indicates that most of the oxidation within the distribution system may be microbial induced, and that FCR concentrations above 0.8mg/L were able to kill or reduce the growth of Fe- and Mn-oxidising bacteria. The highly correlated variables observed in this study can be used to develop AI-based methods like ANNs, FIS, Bayesian networks and neuro-fuzzy models to estimate the risk of Fe and Mn compliance failures in WDNs.

CHAPTER 5: Artificial Neural Network Model for Predicting Accumulation Potential

5.1 Introduction

Understanding the processes that influence water discoloration and identifying the regions in WDNs that have high-risks of Fe and Mn accumulation are of paramount importance to all drinking water companies. Despite their efforts to comply with drinking water standards, water utilities continue to experience compliance failures and receive customer complaints related to water quality. Given the non-linear relation of water quality variables and the complex interactions that occur within them, AI-based methods may prove most effective for modelling water discoloration. This chapter describes the development of two different ANN models, based on the relevant variables that were identified in Chapters 2 and 4. The first model, ANN(t), uses relevant hydraulic, water quality, and pipe-related variables to make its predictions. The second model, ANN(t,ψ), uses relevant hydraulic, pipe-related, and yearly averaged water quality variables to make its predictions. The remaining sections of this chapter are arranged as follows: Section 5.2 describes how the data was transformed prior to the development of the models. These data processing steps were carried out for both the input and output data. Section 5.3 presents how the measured Fe and Mn accumulation potential was calculated. Section 5.4 describes the development of the two ANN models. Specifically, it discusses the tuning of the model parameters and the determination of relevant input variables. The ANN(t) model results and discussion are presented in Section 5.5. This model was used to investigate the relationship between Fe and Mn accumulation potential and input variables. It was also used as a sensitivity analysis tool to further reduce the number of relevant variables identified in Chapter 4. Section 5.6 presents the results and discussion for the ANN(t,ψ). This model was used to predict Fe and Mn accumulation potential for every node in a given WDN. In addition, the model generated risk maps using the predicted Fe and Mn accumulation potential values. These risk maps were compared with maps of customer complaints related to water discoloration, to determine if any correlations existed. Finally, the summary of this chapter are presented in Section 5.7. The developed models can be very useful in helping water resource engineers to identify WDN regions with high discoloration risk.

5.2 Data transformation

After removing outliers from the data set in Chapter 4, the data was transformed before using it to develop the ANN models. Data transformation is a very important step in ANN model development. The success of the ANN model is highly dependent on how effectively the input and output data are represented. Transforming or scaling the raw input data may also improve model prediction.

It has been suggested in the past that input data used in ANN modelling do not require transformation (Grissom, 2000). However, some researchers have found that transforming both the input and output data helps to improve the performance of ANN models (Bowden et al., 2003; Shi, 2000). The three main types of data transformation researchers use are linear transformation, mathematical functions, and statistical standardisation (Bowden, Dandy, & Maier, 2005). Linear transformation is the most commonly used transformation technique in ANN modelling. Usually, the data are transformed to values between 0 and 1 or -1 and 1. This is done to ensure that variables with small input ranges are not dwarfed by variables with large input ranges in the ANN training process. Mathematical functions such as logarithm and the square root can also be used to transform data with positive values (Bowden et al., 2005). In the statistical standardisation technique, the mean is subtracted from the measured value, and the result is divided by the standard deviation. These techniques of transforming data are sometimes referred to as normalisation.

Data transformation is very important in ANN development because it can improve models' predictions and reduce their computation time. However, some researchers are reluctant to accept this concept (Grissom, 2000; Rothery, 1988). They argue that transformed data do not always revert back to their original untransformed form when inverse transformation techniques are applied. Research on clinical trials conducted by Grissom (2000) showed that the means of transformed data can sporadically reverse the difference of the means of the original untransformed data. Though this can sometimes be disturbing, it should not dissuade researchers from normalising their data, as the benefits generally outweigh the disadvantages.

Some ANNs give better results if the data used for modelling is linearly transformed, whereas others tend to give improved results if the data is transformed by mathematical

functions such as logarithm or square root. In ANNs such as SOMs, it is important to linearly transform the modelling data between zero and one, because larger input data dwarf the contributions of smaller input data, resulting in improper classification (Anderson & McNeill, 1992). Transforming the data helps to convert skewed distributed data to normal distributed form in order to minimise the effect of extreme values. It also helps to provide a smooth mapping of input data to output data in the training process of ANNs.

In this research, the guidelines provided by Howell (2007) and Tabachnik and Fidell (2007) on how to select the best normalisation method for a given set of data were followed. The models were developed using both transformed and untransformed data. The two transformation methods used were linear transformation and logarithmic transformation. Using the linear transformation technique, all the data for each variable were transformed between zero and one by the following equation:

$$Y_i = \left(\frac{X_i - X_{min}}{X_{max} - X_{min}}\right) \tag{5.1}$$

where X_i and Y_i are raw value and transformed value of sample i, respectively; and X_{min} and X_{max} are minimum and maximum raw values of the variable, respectively.

Before using the logarithmic transformation technique, a normality test was performed on both the input and output data. Only variables with skewed data were normalised because it was unnecessary to transform data that were already normally distributed. The threshold value for skewness is subjective (Hair et al., 1998). However, as a rule of thumb, the skewed cut-off points suggested by Hair et al. (1998), whereby values greater than 1 or less than -1 are considered as skewed, were used to normalise the data. The formula used in the logarithmic transformation is given in Eqn. 5.3. Figure 5.1 shows graphs of the distribution of turbidity levels before and after normalisation. Before the normalisation, the distribution of turbidity levels was negatively skewed with a skewness value of 1.59. However, this was reduced to -0.09 after the normalisation. The skewness of a set of data can be computed as:

$$Skewness = \frac{1}{sp}\sum_{i=1}^{sp}\left(\frac{X_i - \bar{X}}{\sigma}\right)^3 \quad (5.2)$$

$$Y_i = log_{10}(X_i) \quad (5.3)$$

where σ = standard deviation; sp = sample size; X = sample data; and $\bar{X}$ = sample mean.

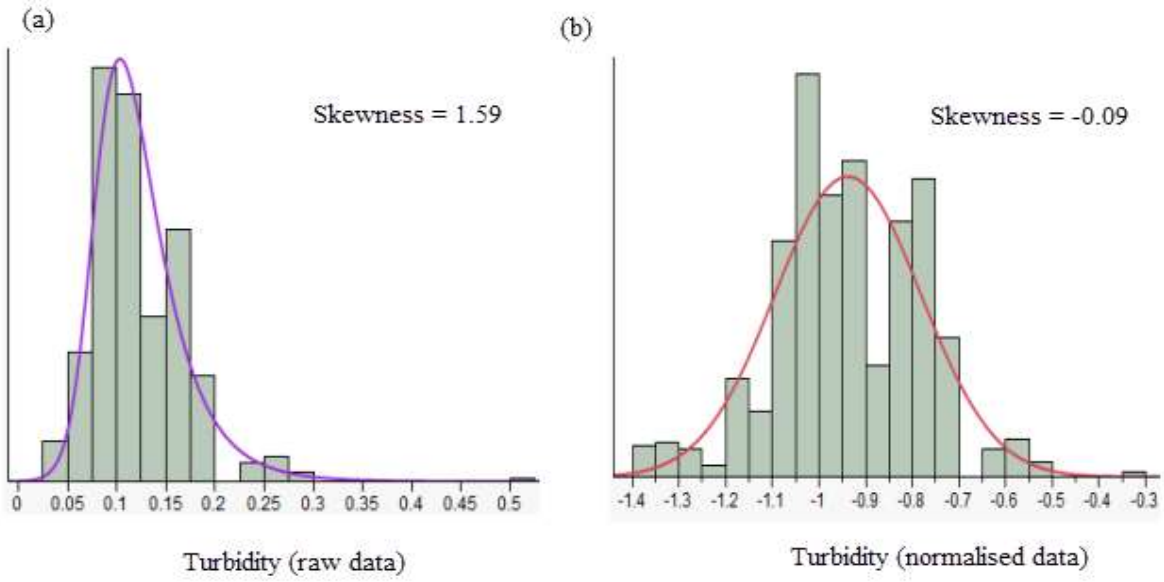


Figure 5.1 Distribution of turbidity data (a) before and (b) after normalisation

5.3 Calculation of measured Fe and Mn accumulation potential

Fe and Mn accumulation occurs in WDNs when they are chemically or biologically oxidised to form insoluble Fe and Mn. The higher the concentrations of Fe and Mn, the more accumulation will occur on the pipe walls. As discussed in Chapter 1, research study conducted by Boxall et al. (2003) on flushing samples collected in the UK identified Fe and Mn as the first and second most common water contaminants, irrespective of pipe material used in WDNs. Related studies conducted by Bowden, Dandy and Maier (2003) and Slaats (2002) showed that gradual accumulation or sudden increase of Fe and Mn particles in WDNs were the most common causes of water discolouration. It was also observed from the data used for the modelling that Fe had a strong positive correlation with Mn in all the WSZs. 94.59% of the graphs exhibited positive correlation when Mn was plotted against Fe at the DMA level (Table 4.2). Also, 72.51% of the graphs exhibited strong correlation when Mn was plotted against Fe at the DMA level (Table 4.3). Figure 5.2 shows sample plots of Fe against Mn at the DMA level. Similar correlations can be

observed in other DMAs. In most cases, when there are Fe failures, there are also Mn failures. In view of the above mentioned similarities between Fe and Mn, Fe and Mn accumulation potential was used as the output variable for the ANN models. When the concentrations of Fe and Mn exceed their respective MCLs, they are more likely to cause discolouration and reduce intellectual function of children (Boxall et al., 2003; Wasserman et al., 2006). It was therefore derived by normalising Fe and Mn with their respective MCLs of 200 and 50 µg/L permitted by the DWI and then aggregated. It was normalised to prevent Fe from dwarfing the contribution of Mn, as they have different magnitudes of concentrations that cause water to discolour. Equation 5.4 shows how it was calculated.

$$Fe\ and\ Mn\ accumulation\ potential = \frac{Fe}{200} + \frac{Mn}{50} \tag{5.4}$$

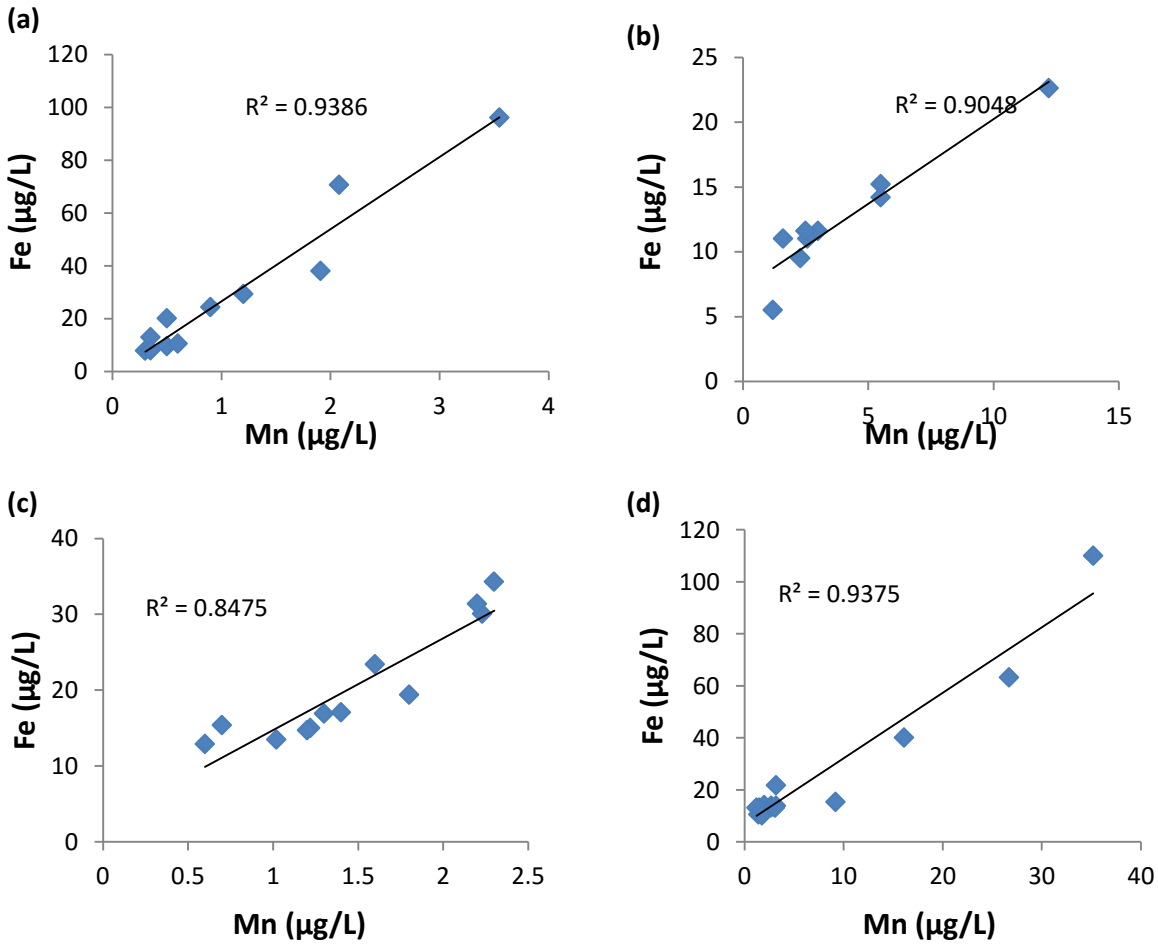


Figure 5.2 Correlation between Fe and Mn at district metered area (a) DMA4-08 (b) DMA7-03 (c) DMA8-03 (d) DMA9-17

A cumulative frequency curve of the measured Fe and Mn accumulation potential for all the WSZs is presented in Fig. 5.3. The measured data were obtained from WSZs with all levels of customer complaints to allow the model to capture all levels of discolouration and remove any form of bias. It was observed that the 90th percentile defines the inflection point in Fe and Mn accumulation potential (which corresponded to the value 0.25 of the measured Fe and Mn accumulation potential) and therefore values above the 90th percentile were subsequently classified as high-risk.

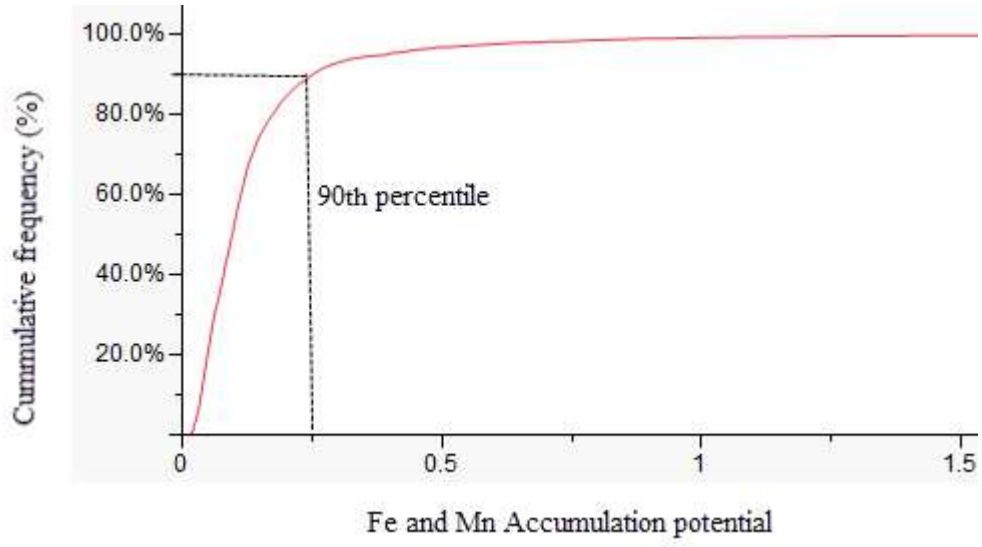


Figure 5.3 Cumulative frequency curve of the measured Fe and Mn accumulation potential

The percentage of customer complaints due to drinking water discolouration per property in the WSZ with the highest customer complaints was also used to further substantiate this upper limit. Customer complaints was used because Fe and Mn are known to be the main causes of drinking water discolouration as indicated in Chapter 1 (Boxall et al., 2003; Slaats, 2002).

$$PCC = \frac{Customer\ complaints\ in\ the\ WSZ}{Number\ of\ properties\ in\ the\ WSZ} \times 100 \qquad (5.5)$$

Where PCC = percentage of customer complaints per number of properties in the WSZ.

The above calculation of PCC ($\frac{543}{17602} \times 100 = 3.08\ \%$) in Eqn. 5.5 is performed for WSZ with the highest complaint. However, because only 30% of customers that experience drinking water discolouration in the UK actually complain, therefore, the percentage of customer complaints per property was multiplied by a factor of 3.333 (=100/30) to obtain the approximate number of complaints (3.08% * 3.333 $\approx$ 10%). In view of this, the top 10% of all the measured Fe and Mn accumulation potential values (values above 90th percentile) were classified as high-risk. Although there is no clear-cut explanation for using 70th percentile as the lower limit of the medium-risk classification, it was observed from a number of trial runs that the 70th percentile provided better classification results. Therefore, measured Fe and Mn accumulation potential values between 70th and 90th percentile were classified as medium-risk and below 70th percentile as low-risk, respectively.

It would have been ideal to use monthly or quarterly averages of water quality variables as inputs for the developed models because customer complaints and concentrations of Fe and Mn exhibit seasonal variations. However, because some of the water quality variables were not sampled frequently, yearly averages of water quality variables were computed at each node and used as input variables.

5.4 Model development

As indicated in Chapter 3, ANNs are a form of artificial intelligence method of modelling that attempt to emulate the learning process of the human brain. They have the ability to learn from past information, retain knowledge, and adapt to different conditions. In the learning process, the input and output data are trained by adjusting the connection weights between the neurons iteratively. ANN models are similar to multiple non-linear regression models, since they can both solve non-linear problems (Nakhaei & Irannajad, 2013). However, ANNs are more flexible and can solve more complex problems. Unlike conventional models, ANN models are data-driven and rely heavily on the quality and quantity of the data that describes the input and output variables. Although ANNs have many advantages (see Chapter 2), they require a large amount of data to train the network. This limitation can be overcome by carefully selecting the data for training to represent the entire population.

In this research, two different ANN models were developed using the relevant variables identified in Chapter 4 to predict the Fe and Mn accumulation potential. The first model, ANN(t), uses relevant hydraulic, water quality, and pipe-related variables to make its predictions. The input variables for this model include measured water quality variables such as alkalinity, P, turbidity, and hardness. They also include hydraulic variables such as the maximum daily shear stress at node, variation of daily shear stress at node after 24 hours of simulation, and average water age at node after 72 hours of simulation. The ANN(t) model can be used as a sensitivity analysis tool to determine the effect of the input variables on Fe and Mn accumulation potential.

The second model, ANN(t,ψ), uses pipe-related, hydraulic and yearly averaged water quality variables to make its predictions. This model can be used as a risk assessment tool to predict Fe and Mn accumulation potential for every node in a given WSZ. The same hydraulic and physical variables were used to develop both models. However, some assumptions were made in obtaining the input water quality variables for the ANN(t,ψ) model. In order to predict the Fe and Mn accumulation potential for every node, a base data set of the measured data for all the nodes is required. Although the hydraulic and pipe-related variables had base data for all the nodes, it was impossible to have water quality data for all nodes in the network. This is because the large sizes of WSZs make drinking water companies unable to sample every node. In fact, there are some parts of the WSZs that were seldom sampled.

From the five-year water quality data, it was observed that the standard deviations for the majority of variables within each DMA were small. This is because nodes in the DMAs are in close proximity to each other. Hence, their bio-chemical conditions were similar. With the exception of a few water quality variables in WSZ1 and WSZ3, which exhibited high standard deviations because of multiple water supply sources, the remaining water quality variables had small standard deviations. It was therefore assumed that, at any given time, the concentrations of chemical variables and variables that influence biological processes in a given DMA were approximately the same. With this assumption, the water quality variables at each node in a DMA were obtained by calculating yearly averages of these variables within that DMA. It would have been ideal to have monthly or quarterly averages as input water quality variables. However, it was impossible because they were not sampled frequently and there would have been many missing data if they were used.

A feed-forward back-propagation neural network was used to develop both the ANN(t) and ANN(t,ψ) models. They are regarded as feed-forward neural networks because they require both input and output data for prediction. The ANN(t) model consists of input, hidden, and output layers. The input and output layers comprise at least one input and output node, respectively, and the hidden layer(s) contains the hidden nodes. Although there is no limit to the number of hidden layers, it was observed in this research that the optimal number of hidden layers for the developed models was two. Individual models were developed for each WSZ using their respective measured data. Using the combined measured data for all five WSZs, another model was developed. Figure 5.4 shows a diagram of the developed multi-layered ANN model.

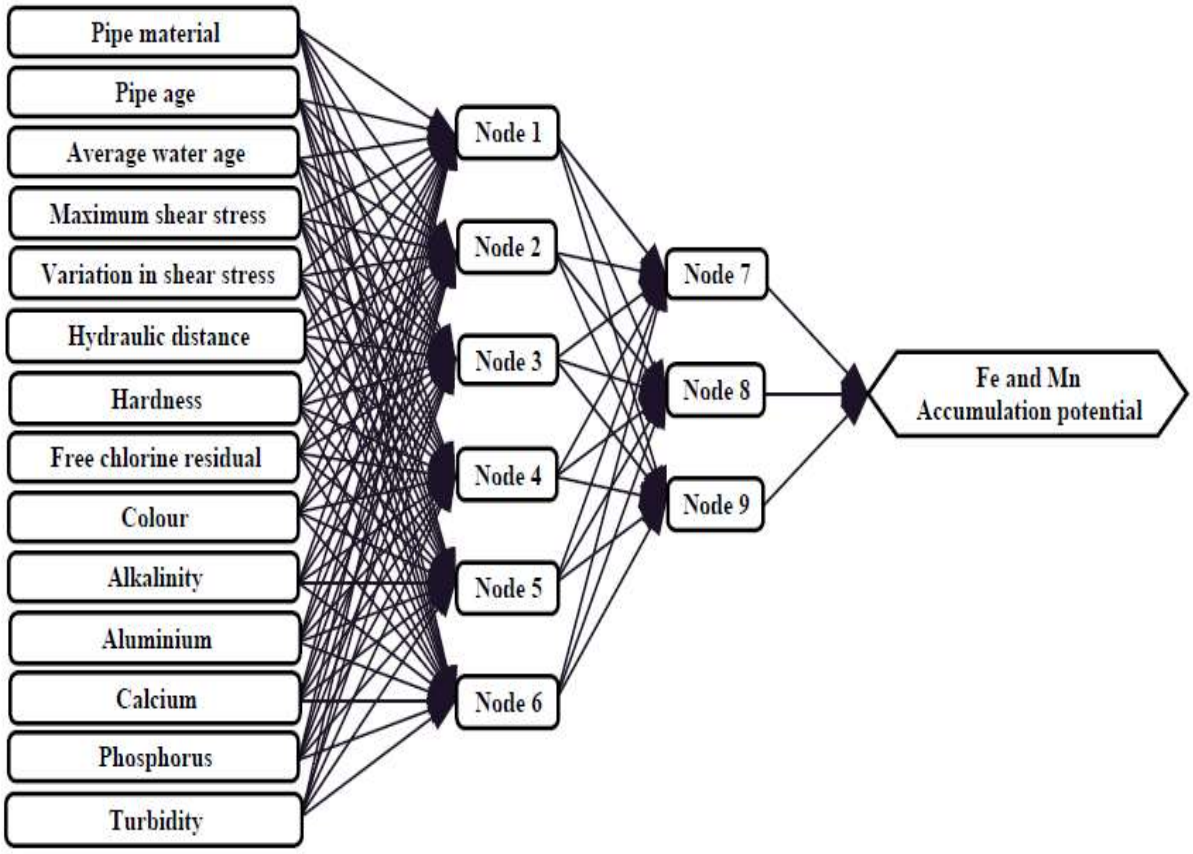


Figure 5.4 The developed ANN(t) model

In this study, K-fold cross-validation method described in Section 3.2.6 was used to develop the models. The data was divided into (K-folds) 5-folds. 15% of the data were randomly selected for testing the model before applying the 5-fold cross-validation method on the remaining 85% of the training and validation data. This means that 80% of the

remaining data was used to train the model and 20% for validation. The process was repeated five times, each time with different subsets to train and validate the model.

In this study, the LM back-propagation algorithm (trainlm) described in Section 3.2.6 was used to optimise the weights and biases of the ANNs. LM is the fastest back-propagation algorithm, and is usually the preferred choice of supervised algorithm for most researchers (Vinay, Vinay, & Ravindra, 2014). However, this procedure requires more computer memory than other back-propagation algorithms. The algorithm updates the connection weights using LM optimisation.

Back-propagation algorithm can be divided into two phases. The first phase involves the forward-propagation of the training patterns and the backward-propagation of the output activation through the network. The second phase adjusts the connection weights. At the beginning of the training process, the network is initialised by assigning random weights and biases to the connections. Using the summation function and the sigmoid activation function, the net output of the hidden nodes and output nodes is calculated. The LM algorithm is then applied to calculate new weights and biases to minimise the error between the predicted output values and the measured output values at the output layer. If the error computed is greater than a tolerance value or the other stopping criteria have not been met, the error is back-propagated through the network. The process is repeated by assigning different weights and biases, recalculating, and updating them until the stopping criteria are reached. The stopping criteria are constraint parameters that ensure the training process does not go on indefinitely. The algorithm for the ANN models is presented as a flow chart in Fig. 5.5.

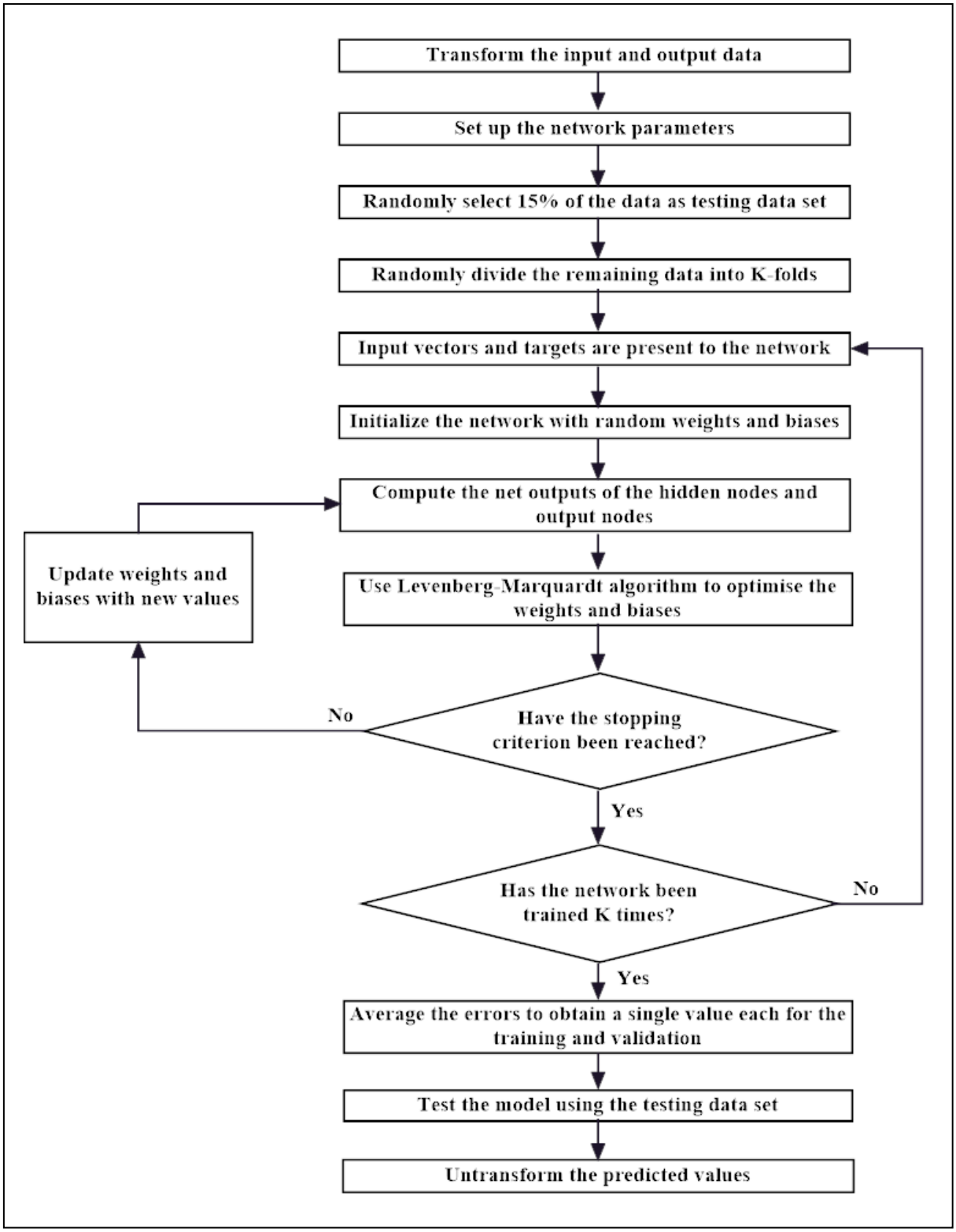


Figure 5.5 Flow chart showing the algorithm for the ANN(t) model

5.4.3 Tuning of the ANN(t) model parameters

Building a back-propagation ANN involves the specification of a number of parameters, including choosing the relevant input variables and selecting an appropriate number of hidden nodes and layers. It also involves choosing an appropriate learning algorithm and activation function, and the tuning of the network parameters. Selecting the best combination of these parameters can be a difficult and time consuming task. The following sections describe how the ANN(t) model was developed and the optimum network parameters were obtained during the training process.

5.4.3.1 Selection of relevant input variables

The selection of appropriate input variables is very important for the development of any type of model because it reduces the cost of collecting unwanted data and improves model performance. In ANNs, the inclusion of inappropriate input variables can mislead the training algorithm, which may result in sub-optimal solutions instead of global optimal solutions.

In Chapter 4, statistical analyses on water quality data were performed to identify the variables that influence Fe and Mn accumulation. Variables which exhibited high correlations with Fe and Mn were selected as relevant variables. Using these relevant variables, an ANN(t) model was developed. The ANN(t) model attempts to learn the complex process of Fe and Mn accumulation by taking a holistic approach to combine all the relevant parameters. This model also serves as a tool to further reduce the number of relevant input variables identified in Chapter 4, while maintaining acceptable performance in the prediction of Fe and Mn accumulation potential. Initially, all the relevant variables were used as independent variables. A windows-based user-friendly software was then developed using the ANN equations to help fine-tune the model by selecting the most significant variables that influenced Fe and Mn accumulation potential.

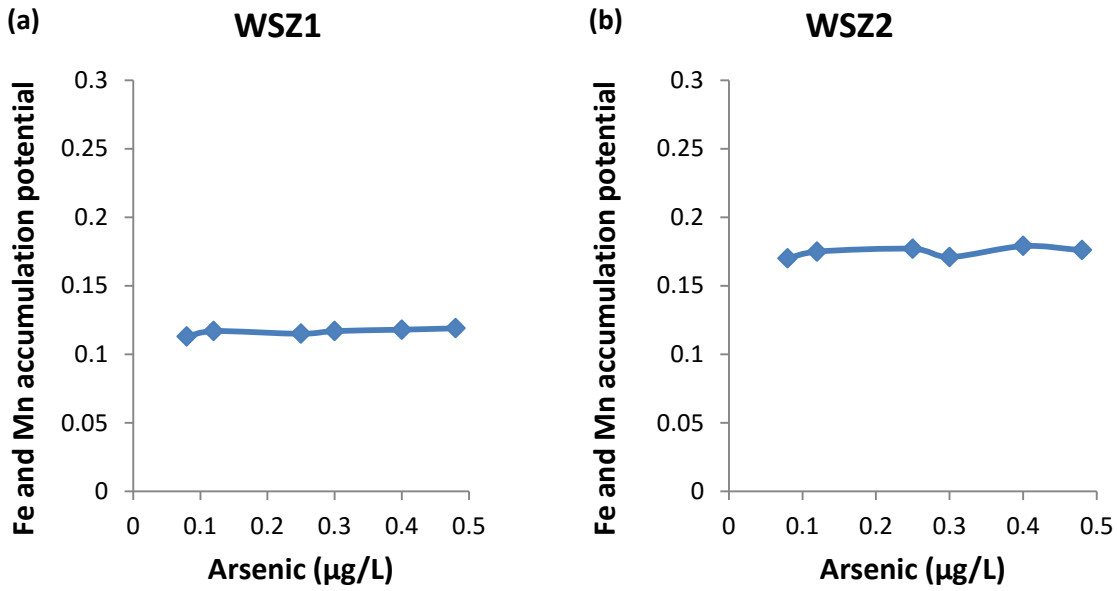


Figure 5.6 Relationship between Fe and Mn accumulation potential and arsenic

Five different ANN(t) models (software) were developed for each of the five WSZs using their respective data. A sixth software was developed using the combined data from all the WSZs. From the software, an input variable value was varied to find its effect on Fe and Mn accumulation potential, while all other variables were kept constant at their respective average values. Since the variables that contribute to the accumulation of Fe and Mn may vary slightly for every WSZ, the procedure was repeated for each WSZ. This procedure was also used for the software developed using the combined data from all the WSZs. The variation of variables values that did not have significant effect on the predicted Fe and Mn accumulation potential were removed from the network. For example, Fig. 5.6 shows the relationship between Fe and Mn accumulation potential and arsenic in WSZ1 and WSZ2. Constant relationships between arsenic and Fe and Mn accumulation potential were observed as shown in the prediction profiler graphs. This shows that arsenic did not have a significant relation with Fe and Mn accumulation potential. Hence, it was removed from the model. Similarly, prediction profiler graphs showing the relationship between Fe and Mn accumulation potential and lead (Pb) concentration in WSZ1 and WSZ2 is shown in Fig 5.7. From the graphs, it was also observed that Pb made insignificant contribution to Fe and Mn accumulation potential. Therefore, Pbwas also removed from the model. The network was rebuilt with a reduced number of input variables each time removing the insignificant variables until an acceptable performance was obtained.

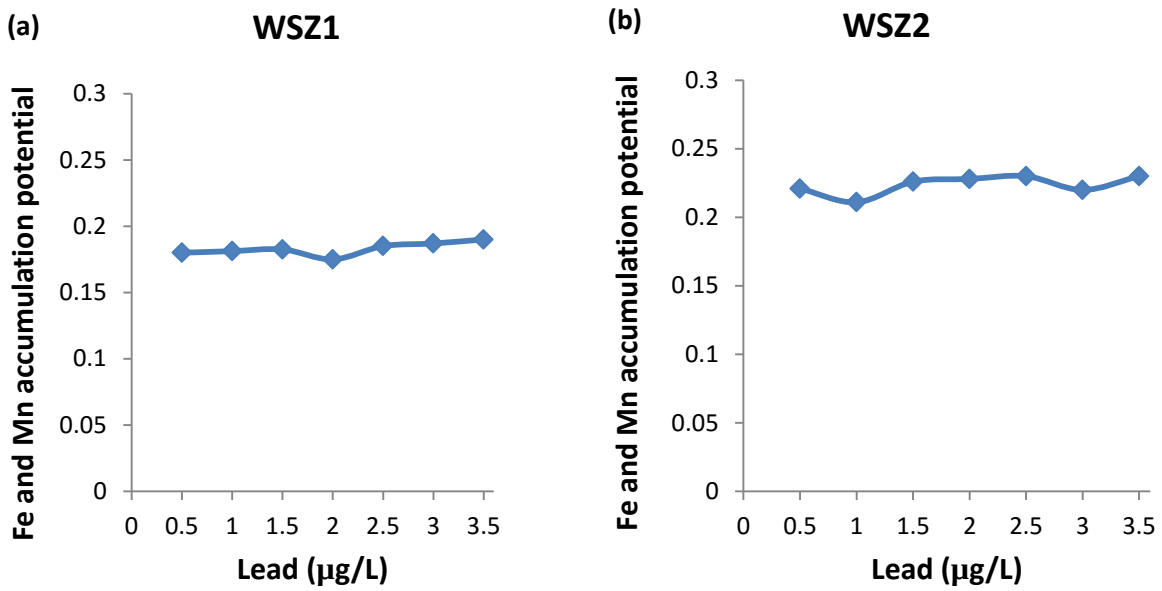


Figure 5.7 Relationship between Fe and Mn accumulation potential and lead

After training the network, the variables that significantly influenced Fe and Mn accumulation were grouped into three categories:

(a) Chemical variables contributing to the chemical reactions within a WDN; namely, Al, alkalinity, turbidity, hardness, calcium, FCR, and colour.

(b) Variables that influence biological processes; namely, P, colour, average water age, FCR, and turbidity.

(c) Physical/hydraulic variables; namely, maximum daily shear stress at a node, variation of daily shear stress at a node, hydraulic distance from source of water supply to a node, pipe age, and pipe material index.

Some of the variables were classified under more than one category. For instance, FCR is a chemical variable because chlorine is an oxidising agent. Hence, it has a significant influence on chemical oxidation. FCR can also be classified as a variable that influence biological process because it is a disinfectant. Therefore, an increase in its concentration can help reduce the formation of biofilms and biological oxidation (USEPA, 2009). A detailed analysis of the effect of each of the variables on Fe and Mn accumulation is presented in Section 5.5.3.

5.4.3.2 Choosing the appropriate number of hidden nodes and layers

Selecting an appropriate number of hidden nodes and layers is very important in the ANN development process. Too many neurons and hidden layers tend to increase computation/training time and cause the model to memorise the input data, whereas models with too few neurons and hidden layers may not give accurate predictions. Using an inappropriate number of hidden neurons and layers may lead to over- or under-fitting of the models (Sheela & Deepa, 2013). Many researchers have used trial-and-error approaches to select the optimum number of hidden neurons and layers in ANNs (Devi, Rani, & Prakash, 2012; Sheela & Deepa, 2013). A disadvantage of using trial-and-error methods is that they can be time consuming if a large number of trials are required.

Various heuristic methods for choosing the optimum number of hidden neurons and layers in ANNs have been proposed. Jinchuan and Xinzhe (2008) proposed a heuristic method in which the optimum number of hidden neurons is dependent on the number of inputs, outputs, sample size, and complexity of the network architecture. They derived a formula for calculating the optimum number of hidden neurons (*NJ*) as follows: $NJ = (NI + \sqrt{sp}\,)/NL$, where *NL* is the number of hidden layers, *NI* is the number of input neurons, and *sp* is the sample size. Shibata and Ikeda (2009) used the formula $NJ = \sqrt{NI * NK}$ to estimate the optimal number of hidden neurons. Hunter, Yu, Pukish III, Kolbusz and Wilamowski (2012) developed a heuristic method for calculating the optimum number of hidden nodes. They used the formulae $NJ = NI + 1$, $NJ = 2NI + 1$ and $NJ = 2^{NI} + 1$ to calculate hidden nodes of MLP networks, bridged MLP networks and fully connected cascade ANNs, respectively. Although these proposed methods are useful, there is no guarantee that they give the optimum number of neurons and hidden layers. Instead, they should be used as guidelines in choosing an appropriate number of hidden nodes and layers.

With these guidelines, an ANN was developed to predict Fe and Mn accumulation potential while simultaneously varying the number of hidden nodes and layers. The maximum number of training epochs was set to its default value of 1000. To ensure reliability of the ANN model predictions, it was run 30 times for each combination of number of hidden nodes and layers, and the performance of the model was then averaged. Initially, the program was run with a single layer using varying hidden nodes between 1

and 15. Thereafter, it was run by varying the number of hidden nodes in the first layer (between 1 and 15) and the second layer (between 3 and 8) simultaneously. The algorithm for choosing the appropriate number of hidden nodes and layers is presented in Appendix N. The performance indicators (RMSE and classification accuracy (CA)) were averaged after 30 runs. A small RMSE and large CA denote better model predictions. The combination of the number of hidden nodes and layers that resulted in the best model performance was selected as the optimum parameter value for the model. Table 5.1 presents the average performance of the ANN(t) model with various combinations of hidden nodes in the first layer and four nodes in the second layer using the test data set for WSZ2 after 30 runs. The best performance was obtained with five and four hidden nodes in the first and second hidden layers, respectively (in bold).

Table 5.1 Average performance of the ANN(t) model on the testing data set for WSZ2

Average RMSE on testing data set	**Average CA on testing data set (%)**	**Hidden nodes in 1st layer**	**Hidden nodes in 2nd layer**
0.1404	56.85	1	4
0.1416	58.47	2	4
0.1620	52.10	3	4
0.1395	58.47	4	4
0.1309	**59.23**	**5**	**4**
0.1492	55.06	6	4
0.1527	50.62	7	4
0.1389	59.19	8	4
0.1612	54.26	9	4
0.1510	57.00	10	4
0.1602	53.45	11	4
0.1524	58.09	12	4
0.1501	56.55	13	4
0.1469	57.90	14	4
0.1547	53.41	15	4

5.4.3.3 Choosing the appropriate activation function

The choice of activation function can have a significant effect on the performance of ANN models. Selecting an appropriate activation function in ANNs can be quite challenging for new users. The nature of the training data will influence the choice of activation function. For non-linear, separable data, using a linear activation function can result in poor model prediction. A sigmoid or hyperbolic function is more appropriate in such cases. The three

functions discussed in Section 3.2.6 were used to develop separate models. From the results, it was observed that the sigmoid activation function produced the best model performance since it had the least (0.1258) RMSE value and the highest CA value (61.54). Although more or less similar model performance was achieved using either hyperbolic or sigmoidal activation function, over all using sigmoidal function provided slightly better results. Therefore, the sigmoidal activation function was selected for further analysis. Table 5.2 presents the average performance of the ANN(t) model in modelling WSZ2 with the three different activation functions.

Table 5.2 Average performance of the ANN(t) model using three different activation functions for WSZ2

	Sigmoid activation function	**Linear activation function**	**Hyperbolic activation function**
Average RMSE on testing data set	**0.1258**	0.1746	0.1307
Average CA on testing data set (%)	**61.54**	51.89	61.41

5.4.3.4 Tuning of network parameters

The network was rebuilt with the sigmoid activation function and the optimum number of hidden nodes and layers. All the other parameters were kept constant at their default values except the parameter to be tuned. The model was then run 30 times for each varying range of values of the parameter being tuned. The model's performance indicators were then averaged. This was done to ensure consistency of the model's predictions. The first parameter tuned was the minimum gradient magnitude, which is a constraint parameter that causes the training process to stop if the performance gradient falls below a specified value. As the gradient approaches this value, the change in error will be insignificant and the network performance will stop improving. From Table 5.3, it was observed that the model gave the best average performance on the testing data set when the minimum gradient magnitude was 1×10^{-5} (in bold) for WSZ2. The network was then rebuilt with the optimum value of the minimum gradient. The remaining parameters were tuned in turns using the same procedure until the best possible performance of the model on the testing data set was obtained. The algorithm for tuning the network parameters is presented in Appendix O.

Table 5.3 Average performance of the ANN(t) model on the testing data set using different minimum gradient values for WSZ2

Average RMSE on testing data set	Average CA on testing data set (%)	Minimum gradient magnitude
0.1244	59.37	0.01
0.1290	62.43	0.001
0.1256	60.81	0.0001
0.1235	**62.06**	**1E-05**
0.1253	60.54	1E-06
0.1294	61.01	1E-07
0.1278	59.67	1E-08
0.1296	60.90	1E-09
0.1271	60.21	1E-10
0.1386	62.17	1E-11

After updating the network with the optimum minimum gradient magnitude, the learning rate parameter was tuned by running the model with varying values of it from 0.001 to 0.3. This parameter controls the speed at which the neural network learns (converges). If the learning rate is too high, the objective function diverges. As a result, the ANN will not be able to learn from the data. On the other hand, if the learning rate is too small the model takes a long time to converge to a solution. The best average performance on the testing data was observed when the learning rate value was 0.1 for WSZ2 (see Table 5.4). The network was then updated with the optimum values of the learning rate.

Table 5.4 Average performance of the ANN(t) model on the testing data set using different learning rate values for WSZ2

Average RMSE on testing data set	Average CA on testing data set (%)	Learning rate
0.1205	62.69	0.001
0.1318	58.95	0.008
0.1320	62.58	0.005
0.1264	63.13	0.01
0.1307	61.90	0.08
0.1267	62.92	0.05
0.1205	**63.19**	**0.1**
0.1326	61.95	0.15
0.1355	60.04	0.2
0.1239	61.67	0.3

The initial training gain (Mu) is used to increase or decrease the step-size of the training process. It is used to prevent the ANN from converging at a local minimum. A high initial Mu helps the ANN to converge faster. However, high values can lead to overshooting the local minimum. A very low initial Mu does not also guarantee the avoidance of the system being trapped in local minimum and can slow down the training process. The Mu factor was tuned by running the model with varying values of it between 1E-05 to 0.5. From Table 5.5, it was observed that the model gave the best average performance on the testing data set for WSZ2 when initial Mu was 0.001 (in bold).

Table 5.5 Average performance of the ANN(t) model on the testing data set using different initial Mu values for WSZ2

Average RMSE on testing data set	Average CA on testing data set (%)	Initial Mu
0.1195	64.60	1E-05
0.1249	60.01	5E-05
0.1204	63.72	0.0001
0.1173	**65.33**	**0.001**
0.1196	61.69	0.01
0.1250	58.26	0.04
0.1223	62.22	0.08
0.1208	58.40	0.1
0.1186	60.05	0.2
0.1260	57.72	0.5

The parameters Mu increase and Mu decrease factors were used to control the weights during the training process. After updating the ANN(t) model with values of already tuned parameters, the model was run with various values of Mu increase factor. Table 5.6 shows that the best average performance of the ANN(t) model on the testing data set for WSZ2 was attained when Mu increase factor was set to 10 (in bold). The network was rebuilt with the optimum Mu increase factor and run 30 times with various values of Mu decrease factor. From Table 5.7 it was observed that the best average performance of the ANN(t) model on the testing data set for WSZ2 was obtained when Mu decrease factor was 0.001 (in bold). The network was then updated with the optimum value.

Table 5.6 Average performance of the ANN(t) model on the testing data set using different Mu increase factor values for WSZ2

Average RMSE on testing data set	Average CA on testing data set (%)	Mu increase factor
0.1263	59.33	0.01
0.1185	64.72	0.1
0.1220	60.05	1
0.1213	63.20	3
0.1208	65.16	7
0.1176	**65.45**	**10**
0.1190	65.39	15
0.1305	60.01	20
0.1211	63.50	30
0.1260	61.56	50

Table 5.7 Average performance of the ANN(t) model on the testing data set using different Mu decrease factor values for WSZ2

Average RMSE on testing data set	Average CA on testing data set (%)	Mu decrease factor
0.1154	**65.82**	**0.001**
0.1170	65.10	0.01
0.1268	63.76	0.05
0.1264	62.02	0.08
0.1281	59.28	0.1
0.1181	63.86	0.12
0.1203	64.44	0.15
0.1228	61.79	0.2
0.1270	58.63	0.5
0.1235	59.50	1

LM back-propagation algorithm was used in the model until this point in the tuning process. Using all the tuned parameters, the network was rebuilt with a different optimisation algorithm, the scaled conjugate gradient backpropagation, to compare its performance with the LM back-propagation algorithm. This algorithm updates weights and biases based on the scaled conjugate gradient method (Hsieh, 2008). Tables 5.7 and 5.8 show that the average testing classification accuracy reduced from 65.82 to 59.37 when the scaled conjugate gradient back-propagation algorithm was used to train the network. It was also observed that the average RMSE increased from 0.1154 to 0.1268. These results indicate that the LM back-propagation algorithm gives better predictions than scaled

conjugate gradient back-propagation algorithm with the data used for the modelling. Table 5.9 shows all the tuned parameter values used in the ANN(t) model for WSZ2. The results of the tuned parameters for the remaining WSZs are presented in Appendix P.

Table 5.8 Average performance of the ANN(t) model on the testing data set using the scaled conjugate gradient backpropagation algorithm for WSZ2

Performance indicator	Scaled conjugate gradient backpropagation
Average RMSE on testing data set	0.1268
Average CA on testing data set (%)	59.37

Table 5.9 The tuned ANN(t) model parameter values for WSZ2

Name	Tuned value	Description of parameter
Show	5	The display of epochs within display
Epochs	1000	The maximum number of iteration
Goal	0	Performance goal
Min_grad	1.00E-05	Minimum gradient magnitude
Mu	0.001	Initial Mu
Mu_inc	10	Mu increase factor
Mu_dec	0.001	Mu decrease factor
η	0.1	Learning rate
1^{st} layer nodes	5	The number of nodes in 1^{st} layer
2^{nd} layer nodes	4	The number of nodes in 2^{nd} layer
Sigmoid activation function		The activation functions used in the model
Levenberg–Marquardt algorithm		Optimisation algorithm used in the model

5.4.4 Training of the artificial neural network models ANN(t,ψ)

The methodology used in developing the ANN(t,ψ) model, is similar to the ANN(t) model. The same method was used in training and tuning the network parameters. The main difference between them is how water quality variables were calculated. The ANN(t) model uses the actual water quality variables (not averaged) for modelling, whereas the ANN(t,ψ) model makes its prediction using yearly averages water quality variables as input variables for every node in the WSZ. In view of this, the ANN(t) model is unable to make predictions in regions where water quality variables were not sampled. However, base data were available at every node for all pipe-related variables and hydraulic variables

for the ANN(t,ψ) model. The pipe-related variables used were pipe age and pipe material index, whereas the hydraulic variables were maximum daily shear stress at node after 24 hours of simulation, average water age at node after 72 hours of simulation, variation of daily shear stress at node after 24 hours of simulation and hydraulic distance from source of water supply.

The ANN(t,ψ) model is able to predict yearly Fe and Mn accumulation potential for each node. A matlab program was written to plot the risk maps for the predicted Fe and Mn accumulation potential as well as customer complaints due to discolouration. The generated risk maps can visually show the distribution of Fe and Mn accumulation potential in WSZs. The source code for the program is given in Appendix F.

5.5 Results and discussion of the ANN(t) models

Six different ANN(t) models were developed to investigate the effect of individual input variables on Fe and Mn accumulation potential. The user-friendly software developed using the models made it easy to investigate the effect of the input variables by simply changing the values in the text box of the software. Prediction profiler graphs of Fe and Mn accumulation potential against individual input variables were plotted. The effect of combined model variables on Fe and Mn accumulation potential were also investigated.

5.5.1 Performance indicators for the ANN(t) models

Different evaluation measures can be adopted in evaluating the performance of ANN models. Coefficient of determination (R^2), CA, RMSE, mean square error (MSE), and sum of square error (SSE) have all been used as performance indicators in ANNs. It is a good research practice to use more than one performance measure to determine the accuracy of a model. In this research, CA and RMSE were used to evaluate the models' prediction accuracy. The RMSE is a measure of the difference between the predicted values of a model and its measured values. The equation for calculating RMSE is given in Eqn. 5.7. Since Fe and Mn accumulation potential has no units of measurement, RMSE also has no units of measurement. The smaller the RMSE value, the better the predictive power of the model.

$$RMSE = \sqrt{\frac{1}{sp}\sum_{i=1}^{sp}(Y_i - X_i)^2} \quad (5.7)$$

As indicated in Section 5.3, the output variable, Fe and Mn accumulation potential, was classified as high, medium and low. Confusion matrices (contingency tables) were used to represent the predicted results of these classifiers in a clean and unambiguous way. CA is widely used in confusion matrices to determine prediction accuracy of the various classifiers (Valverde-Albacete, 2014). The overall CA of a model is the percentage ratio of the number of samples correctly predicted by the model to the total number of samples. The formula for calculating this performance indicator is given in Eqn. 5.8. In general, higher classification accuracy signifies better model performance. However, this is not always the case because it is possible for a model to have a high CA even if it is unable to correctly predict a single value in a particular class. This problem is known as the accuracy paradox. To ensure that the developed models were not exhibiting accuracy paradox, the percentage of each correctly predicted class was computed. Equations (5.9)–(5.11) were used to calculate the percentage for the low, medium, and high classes, respectively.

$$Overall\ CA = \frac{Total\ samples\ corectly\ predicted}{Total\ number\ of\ samples} \times 100 \quad (5.8)$$

$$CA\ (low) = \frac{Low\ samples\ corectly\ predicted}{Total\ number\ of\ low\ samples} \times 100 \quad (5.9)$$

$$CA\ (medium) = \frac{Medium\ samples\ corectly\ predicted}{Total\ number\ of\ medium\ samples} \times 100 \quad (5.10)$$

$$CA\ (high) = \frac{High\ samples\ corectly\ predicted}{Total\ number\ of\ high\ samples} \times 100 \quad (5.11)$$

5.5.2 Performance of the ANN(t) models

Six ANN(t) models were developed using the untransformed data with and without outliers. Five of the models used their respective WSZ data for training, whereas the sixth model used combined data from all five WSZs for training. The best performance values

for each model are presented in Tables 5.10–5.12. For comparison purposes, the models were also developed using logarithmic transformed data and the linear transformed data. Table 5.10a shows the performance of ANN(t) model results when untransformed data with outliers was used for training. The low CAs and high RMSE values observed in the testing data sets indicate that the model does not predict well. In view of this, all models in this research were developed using data without outliers.

Table 5.10a Performance of the ANN(t) models using untransformed data with outliers

Performance indicator	WSZ1	WSZ2	WSZ3	WSZ4	WSZ5	WSZAll
Overall Training CA (%)	75.65	79.43	83.49	87.95	76.59	72.34
Overall Testing CA (%)	56.49	50.09	49.59	57.49	45.95	53.49
Training CA - low (%)	79.26	83.24	87.34	92.56	80.97	76.28
Training CA - medium (%)	63.27	80.64	76.52	88.55	78.52	70.18
Training CA - high (%)	60.94	70.52	80.06	75.88	69.39	65.33
Testing CA - low (%)	61.49	55.34	60.00	63.91	58.24	70.54
Testing CA - medium (%)	57.27	51.79	20.00	50.00	42.61	60.00
Testing CA - high (%)	50.86	42.83	40.00	55.07	45.00	0.00
Training RMSE	0.0458	0.0375	0.0354	0.0394	0.0421	0.0974
Validation RMSE	0.0648	0.8497	0.0708	0.0958	0.0819	0.1277
Testing RMSE	0.2491	0.1854	0.2084	0.2064	0.1954	0.2328
Training data points	153	142	69	168	160	692
Validation data points	38	36	17	42	40	173
Testing data points	33	30	16	35	34	148

Table 5.10b presents the performance of ANN(t) model results when untransformed data was used for training. The high CA and low RMSE values observed in the testing data sets when each of the individual WSZ data was used for training shows that the predicted values are similar to the measured values. This indicates that the model is likely to predict Fe and Mn accumulation potential reasonably well on new data sets.

Table 5.10b Performance of the ANN(t) models with untransformed data

Performance indicator	WSZ1	WSZ2	WSZ3	WSZ4	WSZ5	WSZAll
Overall Training CA (%)	80.54	77.51	96.43	81.82	97.42	74.52
Overall Testing CA (%)	75.76	73.33	75.00	65.71	85.29	74.32
Training CA - low (%)	89.34	83.16	97.22	91.34	98.90	90.35
Training CA - medium (%)	69.27	72.73	95.82	60.00	83.33	30.34
Training CA - high (%)	84.29	66.67	90.57	70.97	71.43	38.37
Testing CA - low (%)	76.00	83.33	85.71	76.92	93.55	88.18
Testing CA - medium (%)	60.00	55.56	0.00	25.00	0.00	35.71
Testing CA - high (%)	100.00	66.67	100.00	40.00	100.00	30.00
Training RMSE	0.0312	0.0211	0.0128	0.0362	0.0216	0.0841
Validation RMSE	0.0384	0.0516	0.0183	0.0425	0.0364	0.0818
Testing RMSE	0.1209	0.1012	0.1111	0.1452	0.0410	0.0792
Training data points	148	135	67	159	156	665
Validation data points	37	34	17	39	38	167
Testing data points	33	30	16	35	34	148

For better predictions, ANNs require large data sets that have been sampled adequately from the entire search space in order to have sufficient instances from which to make a generalisation. In other words, they require large data sets to improve their prediction capabilities. Contrary to this notion, it was observed that the ANN model that used the combined data from all five WSZs for prediction gave relatively poor results as shown by the testing CA and RMSE values in Table 5.10b. This could be due to not having enough instances of data to represent the entire search space from the combined data. It could also be due to the fact that Fe and Mn accumulation occur under slightly different conditions for each WSZ. Therefore, combining the data sets confused the training process, resulting in relatively poor predictions.

Figure 5.8 shows the confusion matrix for the untransformed testing data set after predictions from the ANN(t) model for WSZ1. The model correctly predicted 3 out of 3 (100%) high-risk values, 3 out of 5 (60%) medium-risk values, and 19 out of 25 (76%) low-risk values. The overall CA of 75.76% for the testing data set suggests the ANN(t) model for WSZ1 will make good predictions when applied on new data sets.

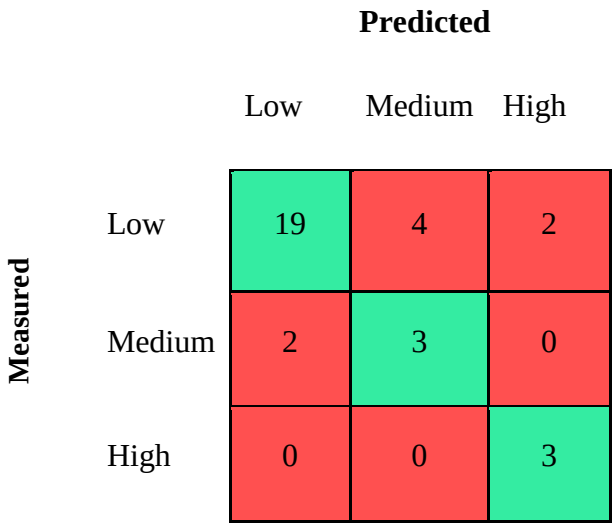


Figure 5.8 Testing data confusion matrix from the ANN(t) model for WSZ1 when untransformed data was used for training

Table 5.11 Performance of the ANN(t) models with logarithmic transformed data

Performance indicator	WSZ1	WSZ2	WSZ3	WSZ4	WSZ5	WSZAll
Overall Training CA (%)	75.68	79.88	95.24	85.86	97.42	77.88
Overall Testing CA (%)	72.73	63.33	87.5	65.71	88.24	70.27
Training CA - low (%)	91.80	85.26	97.22	96.06	99.45	92.35
Training CA - medium (%)	51.02	75.00	90.91	55.00	66.67	41.38
Training CA - high (%)	21.43	70.00	0.00	83.87	71.43	38.37
Testing CA - low (%)	92.00	61.11	100.00	76.92	96.77	88.18
Testing CA - medium (%)	20.00	66.67	0.00	25.00	0.00	21.43
Testing CA - High (%)	0.00	66.67	0.00	40.00	0.00	10.00
Training RMSE	0.0759	0.1195	0.0450	0.1162	0.0486	0.2139
Validation RMSE	0.1504	0.1675	0.0595	0.2395	0.0672	0.2195
Testing RMSE	0.2462	0.2747	0.0993	0.2606	0.2091	0.2522
Training data points	148	136	67	158	156	666
Validation data points	37	33	17	40	38	166
Testing data points	33	30	16	35	34	148

Table 5.11 presents the results from the ANN(t) models when logarithmic transformed data was used for training. It was observed that the ANN(t) models that used untransformed data for training gave better predictions than the ANN(t) models that used logarithmic transformed data. Although the ANN(t) model for WSZ1 gave a moderately high overall CA of 72.73% on the testing data set, however, the testing CA for high-risk was 0%. This means the model was unable to predict high-risk values of Fe and Mn accumulation potential. Similarly, the ANN(t) model using combined WSZ data for

training correctly classified only 10% of high-risk Fe and Mn accumulation potential. These poor performances could be due to some of the transformed data not reverting back to their original untransformed form when the logarithmic inverse was applied to the predicted results. Grissom (2000) and Rothery (1988) experienced similar results when they inverse-transformed the means of the transformed data to an untransformed scale. Because inverse-transformation of the transformed data does not always revert back to its original untransformed form, some researchers are reluctant to transform their data (Grissom, 2000).

Table 5.12 Performance of ANN(t) models with linear transformed data

Performance indicator	WSZ1	WSZ2	WSZ3	WSZ4	WSZ5	WSZAll
Overall Training CA (%)	85.96	84.02	97.62	83.84	90.72	77.40
Overall Testing CA (%)	72.73	73.33	75.00	82.86	85.29	75.68
Training CA - low (%)	90.98	89.58	98.61	96.85	95.58	93.84
Training CA - medium (%)	87.76	81.40	90.91	50.00	50.00	42.76
Training CA - high (%)	35.71	70.00	100.00	74.19	88.83	20.93
Testing CA - low (%)	76.00	77.78	85.71	96.15	93.55	91.82
Testing CA - medium (%)	60.00	66.67	0.00	25.00	0.00	32.14
Testing CA - high (%)	66.67	66.67	100.00	60.00	100.00	20.00
Training RMSE	0.0861	0.0697	0.0010	0.1037	0.0753	0.1413
Validation RMSE	0.1222	0.1194	0.0533	0.1486	0.1565	0.1751
Testing RMSE	0.2108	0.2025	0.2184	0.2660	0.2602	0.2319
Training data points	148	136	67	159	155	666
Validation data points	37	33	17	39	39	166
Testing data points	33	30	16	35	34	148

The results of the ANN(t) models when linear transformed data was used for training are presented in Table 5.12. It can be observed that the models using individual WSZs for prediction gave better results than the model using the combined data. The model for the combined data predicted only 20% of the high-risk values of Fe and Mn accumulation potential. It was observed that the ANN(t) models that used linear transformed data for prediction gave slightly better results than the ANN(t) models that used logarithmic transformed data. However, the models that used untransformed data for prediction gave the best results and were subsequently used to develop the software.

5.5.3 Effect of individual model variables on Fe and Mn accumulation potential

To investigate the effect of changes in input variables on the output variable, a user-friendly and cost-effective software was developed to predict Fe and Mn accumulation

potential. The software takes different input variable values and predicts Fe and Mn accumulation potential. It can also be used as a sensitivity tool to determine the relationship between input variables and Fe and Mn accumulation potential. It was developed using Microsoft Visual Basic. The source code for the software is given in in Appendix B.

To plot the prediction profiler graphs, all input variables were kept at their default (average) values, except the variable which its sensitivity was being determined. The values of the input variable which its sensitivity was being tested was varied a number of times and the predicted Fe and Mn accumulation potential values were recorded. Using these input variable and predicted values, prediction profiler graphs were plotted. These graphs show the correlation between each of the input and output variables.

As mentioned previously, measured values of Fe and Mn accumulation potential above the 90^{th} percentile of the combined data were classified as high-risk, those between the 70^{th} and 90^{th} percentile as medium-risk, and those below the 70^{th} percentile as low-risk. The 90^{th} and 70^{th} percentiles correspond to Fe and Mn accumulation potential values of 0.25 and 0.15, respectively. In this research, all WSZs were classified using these ranges.

Keeping all the input variables constant at their respective default values, Al concentration values were varied while their corresponding Fe and Mn accumulation potential values were computed. Profiler prediction graphs were then plotted from the values. From the graphs in Fig. 5.9, it can be observed that increasing Al concentration generally increases Fe and Mn accumulation potential. Al enters WDNs as a result of aluminium salt $Al_2(SO_4)_3$ (alum) which is used as a coagulant during the water treatment process to reduce organic matter, colour, turbidity, and microorganism levels (WHO, 2006). Increase in Al concentration would result in the formation of amorphous $Al(OH)_3$; a compound that has sorption capabilities (Dayton & Basta, 2005; Wang et al., 2012). The increase in Fe and Mn accumulation potential could be attributed to the sorption of Fe and Mn on amorphous $Al(OH)_3$. Although Al concentrations exhibited a positive correlation with Fe and Mn accumulation potential in both WSZ1 and WSZ3, it was observed that it was a more significant parameter in the latter WSZ.

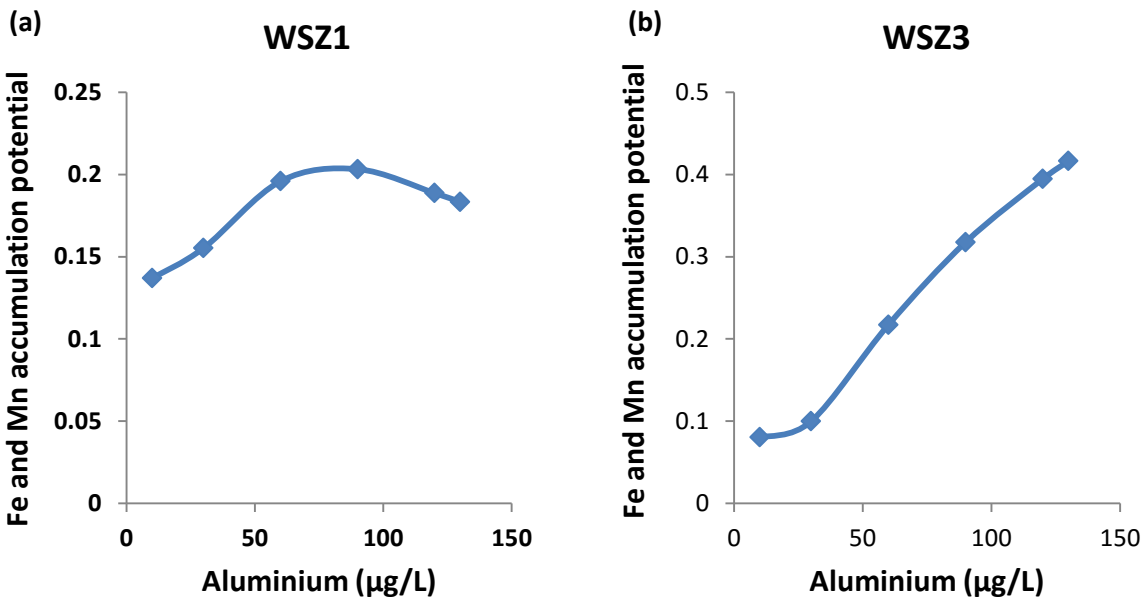


Figure 5.9 Relationship between Fe and Mn accumulation potential and aluminium

The prediction profiler graphs of the models for WSZ1 and WSZ5 in Fig. 5.10 show negative correlations between Fe and Mn accumulation potential and Ca concentration. This may be due to the formation of calcium carbonate ($CaCO_3$) in the presence of DO in the WDNs. Increase in $CaCO_3$ concentration increases alkalinity levels and subsequently causes Fe and Mn accumulation potential to reduce. This finding conforms to the research on alkalinity conducted by Naylor et al. (1993) and Kashinkunti et al., (1999). Kashinkunti et al., (1999) observed that fewer customer complaints regarding water discolouration were received when the alkalinity was maintained at 60 mg $CaCO_3$/L, whereas Naylor et al. (1993) found that alkalinity above 50 mg $CaCO_3$/L reduced corrosion within the a pH range of 7.5–8. Furthermore, $CaCO_3$ serves as a corrosion inhibitor by forming protective scales on the inner walls of ferrous pipes. These scales can form films thick enough to prevent drinking water from coming into direct contact with ferrous pipes in WDNs and subsequently reduce Fe failures.

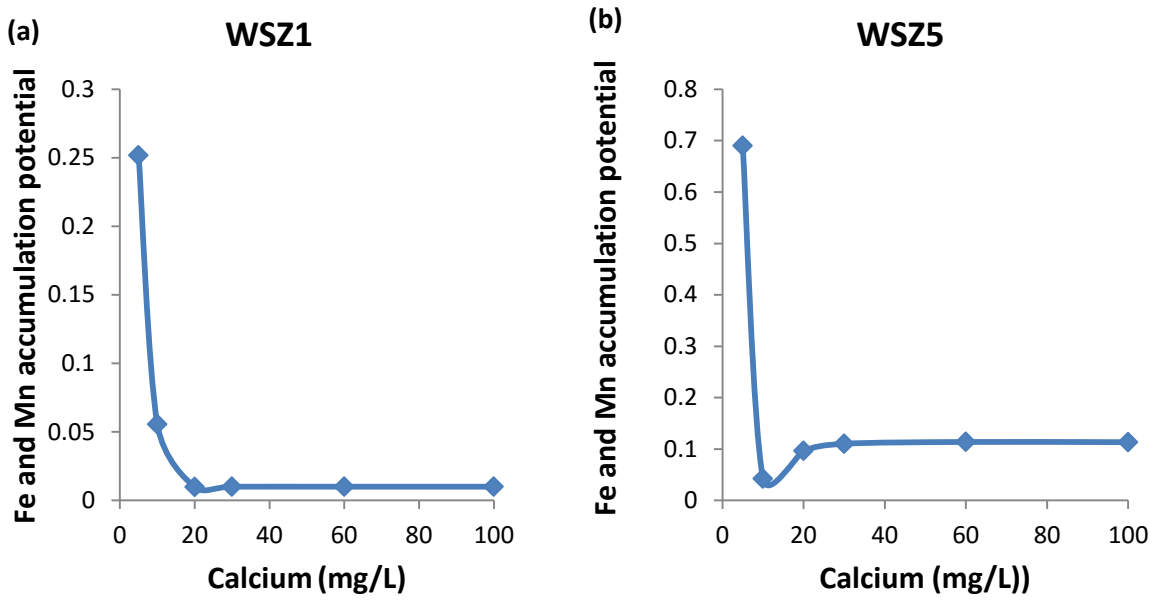


Figure 5.10 Relationship between Fe and Mn accumulation potential and calcium

FCR has a dual effect on Fe and Mn accumulation potential. Since FCR is an oxidising agent, moderately high concentrations can chemically oxidise soluble Fe^{2+} and Mn^{2+} to insoluble Fe^{3+} and Mn^{4+}. Also, because it is a disinfectant, a moderately high concentration of FCR helps to kill Fe- and Mn-oxidising bacteria. This subsequently prevents or reduces biological oxidation of Fe and Mn in WDNs. The prediction profiler graphs in Fig. 5.11 (a) and (b) have similar characteristics, but differ in values of Fe and Mn accumulation potential. Both graphs have negative correlation when biological oxidation of Fe and Mn is dominant and positive correlation when chemical oxidation of Fe and Mn is dominant. High values of Fe and Mn accumulation potential were observed when FCR concentrations were below 0.8 mg/L. This is because low concentrations of FCR promotes microbial growth and increases biological oxidation of Fe and Mn in that range. This finding conforms to the observations made in the analysis of Fe and Mn with FCR in Section 4.6, where it was observed that whenever FCR concentrations were below 0.8 mg/L, there were many Fe and Mn failures. The concentrations of FCR within this range are however not strong enough to cause chemical oxidation.

The gradual increase in Fe and Mn accumulation potential observed when FCR exceeds 0.8 and 0.6 mg/L in graphs in Fig 5.11 (a) and (b) respectively, is due to the dominant effect of chemical oxidation of Fe and Mn in the WDNs. Also, the relatively high concentrations of FCR help to reduce microbial growth and decrease biological oxidation.

Due to the dual effect of FCR, water resource engineers should be careful when choosing the optimal value of chlorine, since high concentration of it can give drinking water a very strong odour and unpleasant taste, whereas low concentration of it deteriorate water quality (Teasdale et al., 2007).

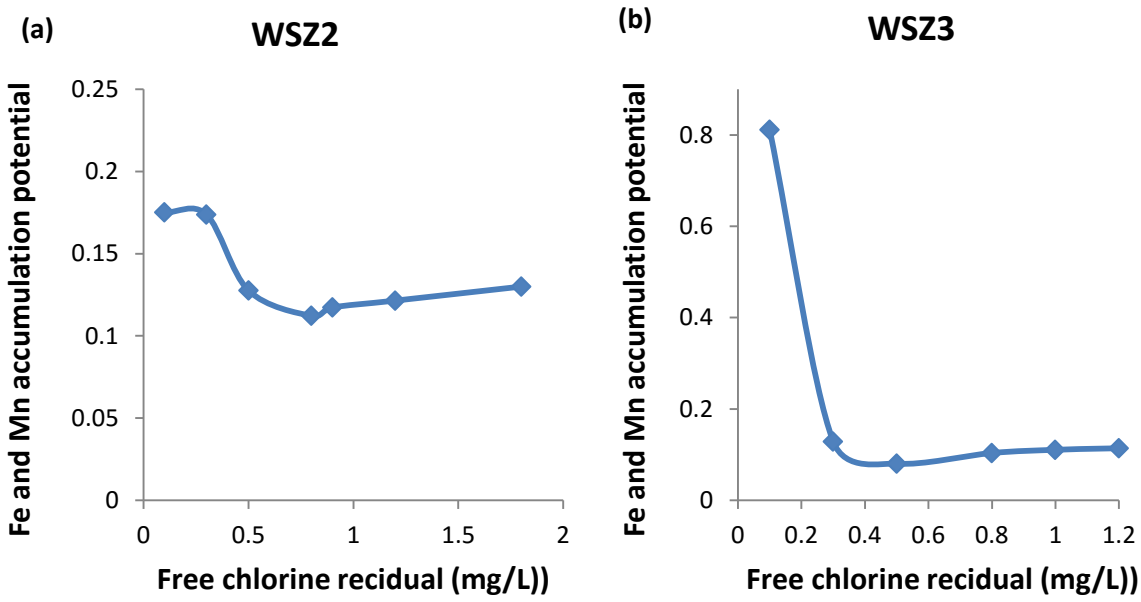


Figure 5.11 Relationship between Fe and Mn accumulation potential and FCR

Several studies have shown that total organic carbon (TOC) concentrations are strongly correlated with colour (Effler et al., 1985; Evans, 1988; Gorham et al., 1986). Colour was used as an indirect measure of TOC because there were no data available for it. TOC has often been used by researchers as a measure to appraise the potential formation of biofilms because increased levels of it in the drinking water enhance biofilm formation (van der Kooij, 2002). The strong positive correlation between TOC and biofilm formation is because the bacteria responsible for the formation of biofilms mainly use carbon as a bioavailable form of nutrient (CRCWQT, 2005). Unsurprisingly, the prediction profiler graphs in Fig 5.12 show a strong positive correlation between Fe and Mn accumulation potential and colour.

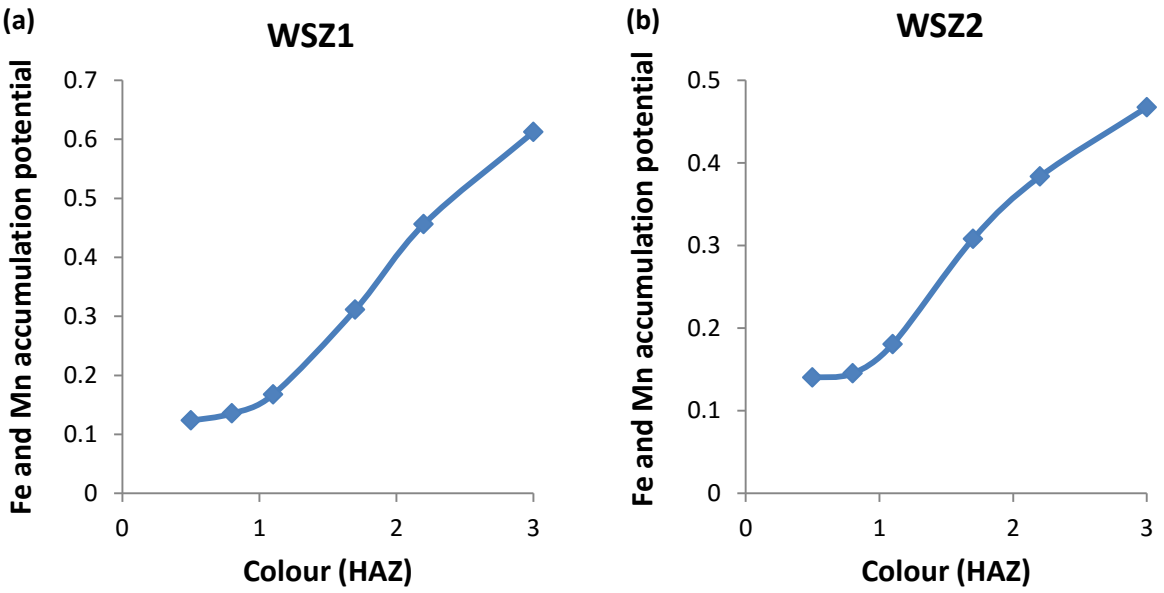


Figure 5.12 Relationship between Fe and Mn accumulation potential and colour

Figure 5.13 illustrates the highly positive correlation between hardness and Fe and Mn accumulation potential in the models for WSZ1 and WSZ5. Hardness is a measure of total dissolved minerals, mainly Ca and Mg. However, dissolved ions such as Fe, Mn, Al, and zinc may also contribute to hardness (WHO, 2011b). This explains the reason for the positive correlation observed in the prediction profiler graphs. Usually water with hardness in the range between 0–60 mg/L is classified as soft, 60–120 mg/L as moderately hard, and above 120 mg/L as hard (WHO, 2011b). Although hardness does not pose any harmful threats to human health, it can cause many domestic and industrial problems. They can cause the breakdown of boilers and cooling towers, leave deposits of lime scale in kettles and water heaters and difficulties in lathering soap because of the formation of the complex Ca^{2+} and Mn^{2+} compounds.

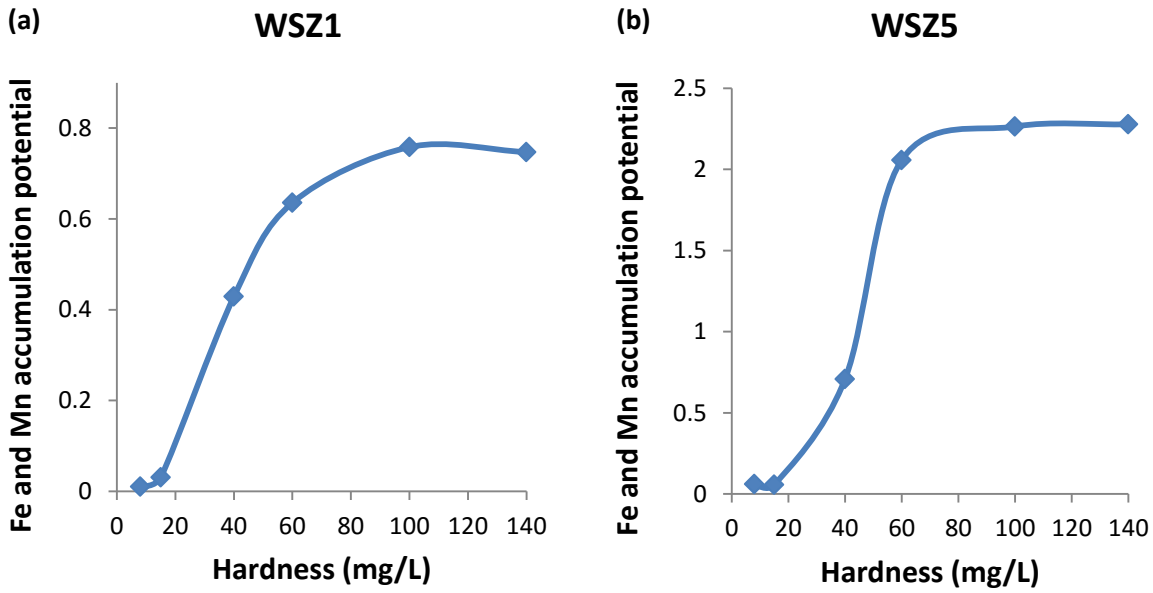


Figure 5.13 Relationship between Fe and Mn accumulation potential and hardness

A highly positive correlation was observed between Fe and Mn accumulation potential and turbidity, as shown in Fig. 5.14. Turbidity in WDNs is caused by suspended inorganic or organic matter, which tends to block the transmission of light through water. Microorganisms such as Fe- and Mn-oxidising bacteria attach themselves to these suspended particles. High turbidity levels therefore enhance microbial growth, increase biofilm formation, and subsequently cause the biological oxidation of Fe and Mn. Furthermore, high levels of turbidity enhance biological oxidation by serving as a shield to inhibit microorganisms from disinfection (WHO, 2011a). The high positive correlation could also be due to the condition of the pipes in WDNs, as suggested by Boxall et al. (2003). In their studies on flushing, they found that WDNs with low discolouration risks (turbidity levels) were regularly cleaned or rehabilitated, whereas WDNs with high turbidity levels were as a result of poor pipe conditions, i.e. lack of cleaning or rehabilitation. Years of sediment accumulation, including Fe and Mn precipitates, causes high Fe and Mn accumulation potential and turbidity levels.

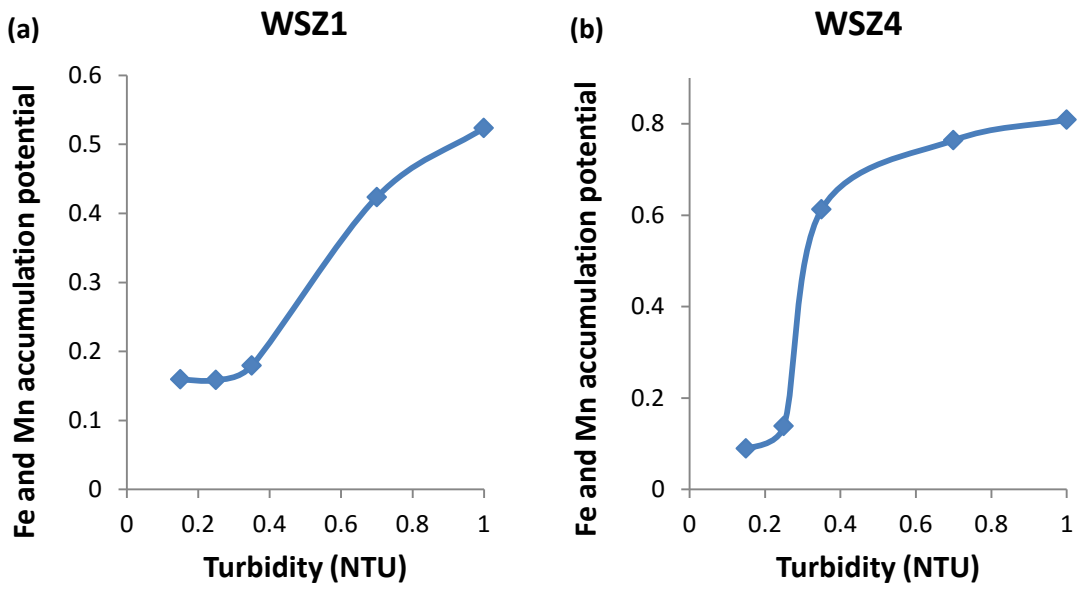


Figure 5.14 Relationship between Fe and Mn accumulation potential and turbidity

The prediction profiler graphs in Fig. 5.15 show that, in general, Fe and Mn accumulation potential increases with hydraulic distance from source of water supply. However, Fig 17 (a) shows a gradual decrease in Fe and Mn accumulation potential when hydraulic distance from source of water supply exceeds 4,000 m. This gradual decrease could be due to the effect of other contributing variables to Fe and Mn accumulation potential. In general, the further water travels through a WDN, the higher the water age and the more chlorine dissipates within the system. Since chlorine is a disinfectant, it suppresses the growth of Fe- and Mn-oxidising bacteria, preventing the biological oxidation of soluble Fe and Mn to insoluble Fe and Mn. Hence, regions with short hydraulic distance from source of water supply have low Fe and Mn concentrations. In contrast, regions with long hydraulic distances from source usually have low concentrations of FCR. This increases microbial growth, which causes biological oxidation of Fe and Mn and subsequently leads to increased Fe and Mn accumulation potential. Similar relationships between Fe and Mn accumulation potential and the input variables for the remaining WSZs are presented in Appendix Q.

Another reason why regions with long hydraulic distance from source of water supply also have high values of Fe and Mn accumulation potential could be due to the low flow rates and velocities in pipes experienced at the periphery of WDNs. In general, peripheries of WDNs have long hydraulic distance from source of water supply, low flows, low

velocities as well as low shear stress. As found by Boxall et al. (2001), lower conditioning daily shear stress will result in a higher accumulation of Fe, Mn, and other particles in pipes.

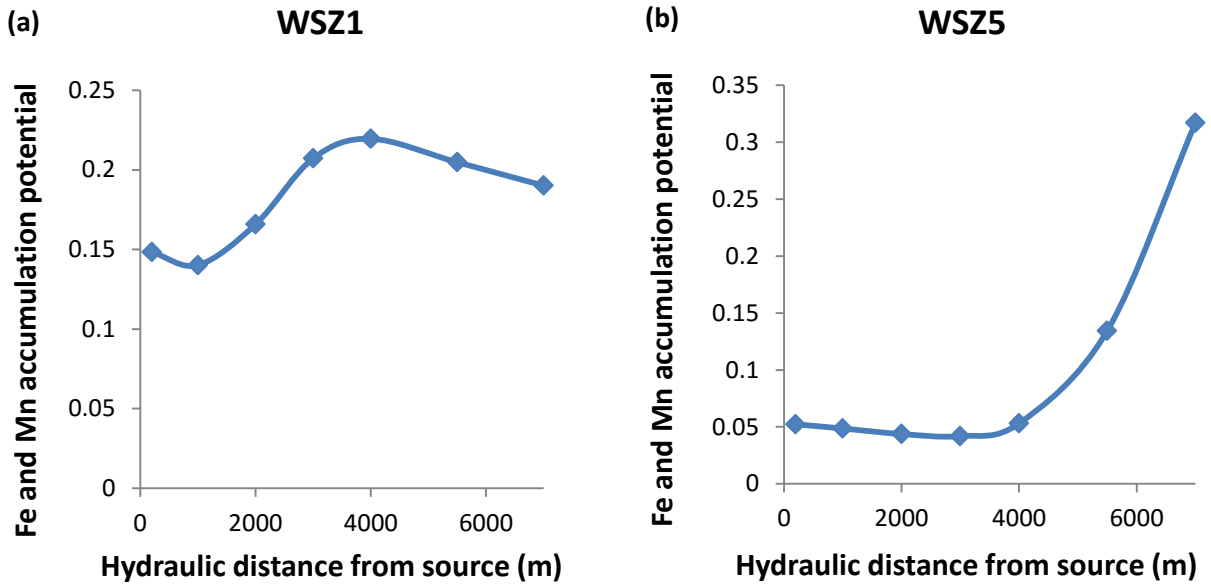


Figure 5.15 Relationship between Fe and Mn accumulation potential and hydraulic distance from source of water supply

5.5.4 Effect of some combined model variables on Fe and Mn accumulation potential

After varying the values of individual variables to determine their effects on Fe and Mn accumulation potential, various combinations of variables were grouped by properties and varied to also find their effect on it. Computations were performed using extreme values of the combined variables that are known to either increase or decrease Fe and Mn accumulation potential. Keeping all other variables at their constant default (average) values, variables that are known to influence biological oxidation were varied. Initially, Fe and Mn accumulation potential were computed with values of variables that are known to reduce biological oxidation. Fig 5.16 (a) shows the variable values known to influence biological oxidation and the predicted value (highlighted by red ovals) in the developed software. The Fe and Mn accumulation potential were then computed with values of variables that are known to increase biological oxidation (see Fig 5.16 (b)). It was observed that Fe and Mn accumulation potential increased from 0.10 (low-risk) to 0.36 (high-risk). This is an indication of how significant biological oxidation is in WSZ1.

With the exception of FCR which reduces biological oxidation as its concentration increases, the other variables; colour, P and average water age all increase with increasing biological oxidation. Researchers have identified TOC (colour) and P as significant bioavailable forms of nutrients that bacteria in WDNs need for growth and reproduction (CRCWQT, 2005). This explains why high levels of these variables increase biological oxidation. As average water age increases, residual chlorine decreases and the quality of water deteriorates, which creates a conducive environment for bacteria growth in the network. Increased average water age can give water poor taste and bad odour. Conversely, increasing FCR concentration reduces biological oxidation because it is a disinfectant which kills the oxidation bacteria.

(a) (b)

Figure 5.16 Screen shots of the developed software to show the effect of biological oxidation on Fe and Mn accumulation potential in WSZ1

Fe and Mn accumulation potential were computed with values of chemical variables that are known to reduce chemical oxidation (see Fig 5.17 (a)). The same computations were made with chemical parameters that are known to increase chemical oxidation (see Fig

5.17 (b)). The increase in predicted Fe and Mn accumulation potential value from 0.06 (low-risk) to 0.47 (high-risk) indicates the importance of chemical oxidation in WSZ5. Unlike corrosion which occurs mainly in cast iron pipes, chemical oxidation takes place in almost all types of pipes. Chemical oxidation of Fe and Mn occurs when soluble Fe^{2+} and Mn^{2+} from the source of water supply are converted to insoluble Fe^{3+} and Mn^{4+} in the presence of oxidising agents such as DO and FCR. Increase in alkalinity levels is known to generally reduce chemical oxidation, whereas increase in both FCR and hardness generally increase chemical oxidation.

(a) (b)

Figure 5.17 Screen shots of the developed software to show the effect of chemical oxidation on Fe and Mn accumulation potential in WSZ5

Corrosion, which is the most common cause of drinking water discolouration, mainly occurs in regions of WDNs with cast iron pipes. For corrosion to take place, an oxidising agents has to come in contact with the inner surface of the cast iron pipes to oxidise Fe. Generally, increase in FCR concentration and pipe age increases corrosion rates in cast

iron pipes (Knocke et al., 1990). Conversely, increase in Ca concentration reduces corrosion rates in cast iron pipes. Keeping the values of all other variables in the model constant at their default values, Fe and Mn accumulation potential were computed with chemical values that are known to reduce corrosion. The model gave a prediction of 0.08 (low-risk) (Fig 5.18 (a)). On the other hand, when it was computed with chemical parameter values that are known to increase corrosion, the model's prediction was 0.71 (high-risk) (Fig 5.18 (b)). The high predicted value shows that corrosion is very significant in the prediction of Fe and Mn accumulation potential in WSZ4.

(a) (b)

Figure 5.18 Screen shots of the developed software to show the effect of corrosion on Fe and Mn accumulation potential in WSZ4

Increase in Al, colour and hardness concentrations are known to generally increase sorption (Wang et al., 2012). When low values of these variables were used to predict Fe and Mn accumulation potential keeping all other variables at their constant default values,

the model's prediction was 0.04 (low-risk) (see Fig 5.19 (a)). However, when the values of sorption variables were increased with the other variables values still constant, the model's prediction was 0.90 (high-risk) (see Fig 5.19 (b)). The high predicted value could be due to the attachment of Fe and Mn particles on amorphous $Al(OH)_3$ and other sorption parameters. This also indicates that sorption is an important process that influences Fe and Mn accumulation in WSZ1.

(a) (b)

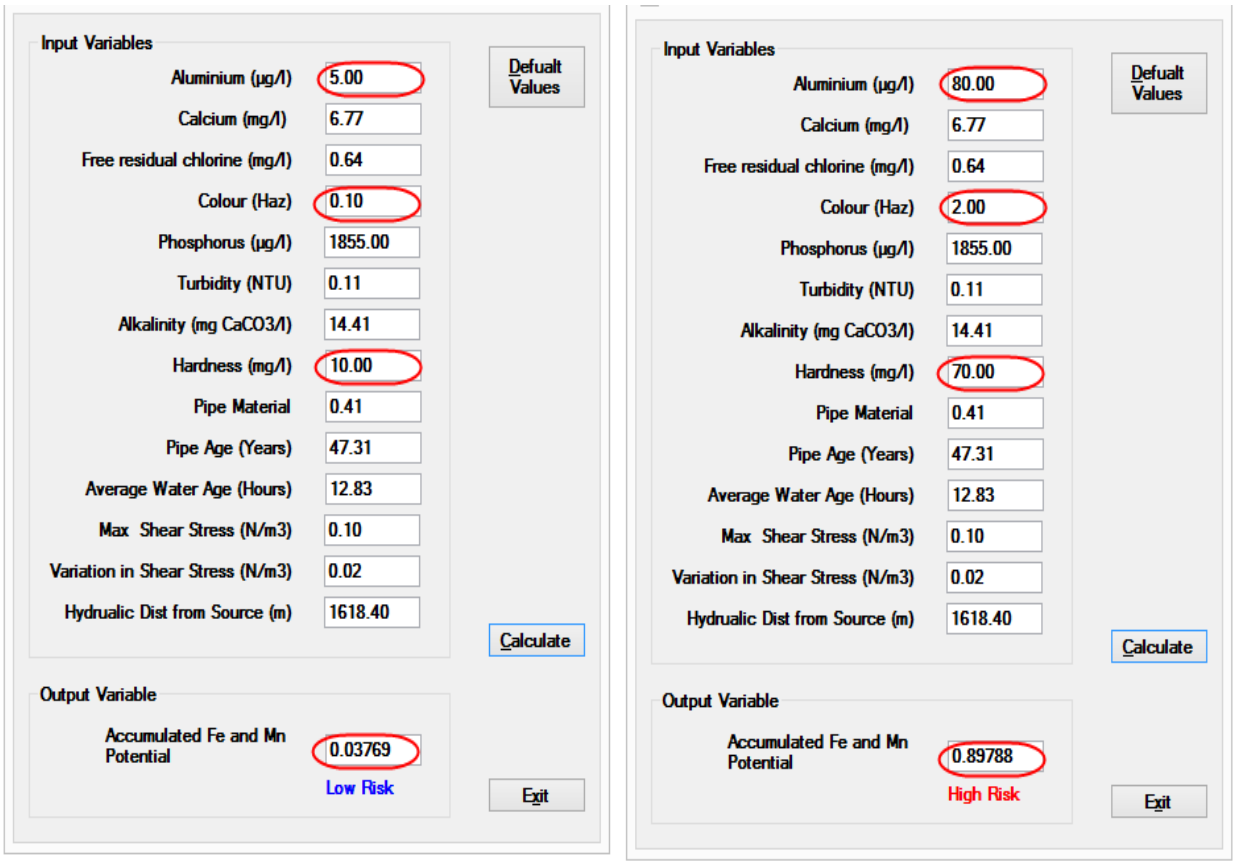


Figure 5.19 Screen shots of the developed software to show the effect of sorption on Fe and Mn accumulation potential in WSZ1

The variables variation of daily shear stress at node and maximum daily shear stress at node have similar characteristics. Both variables negatively correlate with Fe and Mn accumulation. Low values of these variables are likely to be found in regions with dead ends and redundant loops in WDNs. These regions are more susceptible to increased microbial growth and discolouration. On the other hand, high values of the variables are likely to be found on trunk mains and regions with high water demand. Running the model with high values of variation of daily shear stress and maximum daily shear stress while

keeping the other parameters at constant at their respective default values gave a predicted Fe and Mn accumulation value of 0.09 (low-risk). Whereas a predicted Fe and Mn accumulation value of 0.22 (medium-risk) was obtained when the model was run with low values of variation of daily shear stress and maximum daily shear stress (see Fig. 5.20). This observation conforms to research by Boxall et al. (2001, 2003), who suggested that discolouration materials are more likely to accumulate in networks that are more subjected to low conditioning daily shear stress than networks with high conditioning daily shear stress.

(a) (b)

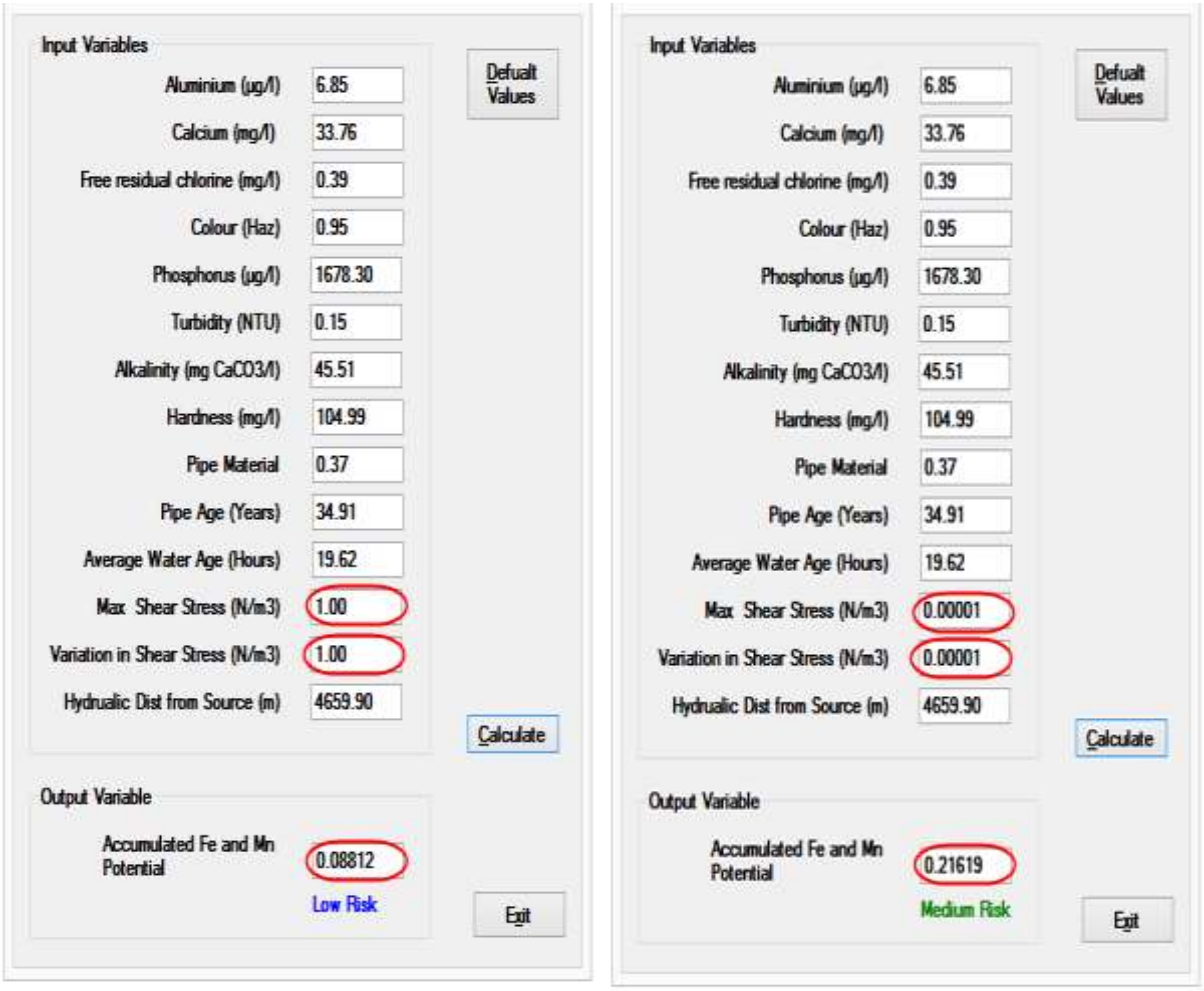


Figure 5.20 Screen shots of the developed software to show the effect of shear stress on Fe and Mn accumulation potential in WSZ2

5.6 Results and discussion of the ANN(t,ψ) models

Due to the enormous sizes of WSZs, it is almost impossible to sample every node. One of the dilemmas drinking water companies face is to estimate the concentrations of Fe and Mn at nodes that have not been sampled. If they are wrongly estimated, they may give misleading results, which will subsequently lead to making incorrect analyses and drawing wrong conclusions. As mentioned in Section 5.4, unlike the ANN(t) models, the ANN(t,ψ) models requires base data for every node in the WSZs to make predictions. The assumptions given in Section 5.4 were used to estimate the base data of yearly average water quality variables within each DMA as input variables for the ANN(t,ψ) models. However, the same pipe-related and hydraulic input variables used in the ANN(t) models were also used to develop the ANN(t,ψ) models.

Although the ANN(t) models gave slightly better predictions of Fe and Mn accumulation potential because there were no assumptions made in obtaining the base data, it could not make predictions for every node in WSZs. ANN(t) are more useful in investigating the correlation between the input and output variables. Contrary to this, ANN(t,ψ) models are able to predict Fe and Mn accumulation potential for every node as well as generate risk maps for the WSZs.

5.6.1 Performance of the ANN(t,ψ) models

The ANN(t,ψ) models developed also use CA and RMSE as performance indicators for the evaluation of the models. Six models were developed; five of the models used their respective WSZs data sets for the modelling, whereas the last model used the combined data sets from all the five WSZs for the modelling. The models were developed using linear transformed data, logarithmic transformed data, and untransformed data. The models gave poor predictions when the linear transformed data was used for the modelling (see Table 5.13). They could not predict high-risk values of Fe and Mn accumulation potential on the testing data sets very well. This could be due to distortions in the predicted values during the back-transformation.

Table 5.13 Performance of the ANN(t,ψ) models with linear transformed data

Performance indicator	WSZ1	WSZ2	WSZ3	WSZ4	WSZ5	WSZAll
Overall Training CA (%)	68.10	85.40	85.21	72.66	92.13	75.49
Overall Testing CA (%)	62.79	65.85	61.54	53.33	91.49	68.66
Training CA - low (%)	90.26	95.17	92.44	90.00	99.18	93.62
Training CA - medium (%)	31.58	61.70	50.00	39.53	23.08	28.95
Training CA - high (%)	4.76	76.47	0.00	37.21	0.00	14.29
Testing CA - low (%)	86.67	83.33	80.00	62.96	100.00	90.28
Testing CA - medium (%)	12.50	33.33	0.00	58.33	0.00	16.13
Testing CA - high (%)	0.00	50.00	0.00	0.00	0.00	11.54
Training RMSE	0.2097	0.1236	0.1105	0.1529	0.1830	0.1886
Validation RMSE	0.2126	0.1280	0.1661	0.2147	0.1874	0.2095
Testing RMSE	0.2668	0.2111	0.2802	0.2538	0.2763	0.2534
Training data points	185	180	114	205	214	901
Validation data points	47	46	28	51	53	225
Testing data points	43	41	26	45	47	201

Table 5.14 presents the performance of the ANN(t,ψ) models when untransformed data were used for the modelling. From the results obtained, it was observed that the models that used individual WSZs data for modelling outperformed the models that used the combined data sets from all the five WSZs. Unlike the ANN(t) models which gave good predictions when untransformed data was used for the modelling, it was observed that the ANN(t,ψ) models gave relatively poor predictions when untransformed data was used for the modelling. This is because the yearly averaged water quality variable values that were used to develop the ANN(t,ψ) models were highly skewed. This resulted in making the models difficult to converge at their respective global minimum and subsequently gave poor performances.

Table 5.14 Performance of the ANN(t,ψ) models with untransformed data

Performance indicator	WSZ1	WSZ2	WSZ3	WSZ4	WSZ5	WSZAll
Overall Training CA (%)	70.69%	73.01%	92.96%	67.19%	94.01%	73.45%
Overall Testing CA (%)	67.44%	65.85%	65.38%	57.78%	82.98%	65.67%
Training CA - low (%)	79.87	86.21	98.32	80.00	98.37	89.06
Training CA - medium (%)	59.65	34.04	63.64	30.23	38.46	23.78
Training CA - high (%)	33.33	70.59	100.00	53.49	55.56	38.53
Testing CA - low (%)	76.67	78.26	85.00	74.07	90.70	83.33
Testing CA - medium (%)	62.50	33.33	0.00	41.67	0.00	19.35
Testing CA - high (%)	20.00	66.67	0.00	16.67	0.00	23.08
Training RMSE	0.0307	0.0302	0.0170	0.0622	0.0600	0.1034
Validation RMSE	0.0491	0.0593	0.0201	0.0549	0.0856	0.0622
Testing RMSE	0.1413	0.1132	0.1324	0.1270	0.1476	0.1115
Training data points	185	181	114	205	214	901
Validation data points	47	45	28	51	53	225
Testing data points	43	41	26	45	47	201

Table 5.15 presents the performance of the ANN(t,ψ) models using logarithmic transformed data for the modelling. Again, it was observed that the models that used their respective individual WSZs data for the modelling gave better predictions than the model that used the combined data from the five WSZs. The combined model was able to predict only 23.08% of high-risk Fe and Mn accumulation potential. As explained in Section 5.5.2, this could be due to not having enough instances of data to represent the entire search space from the combined five WSZs or due to the formation of Fe and Mn accumulation under slightly different conditions for each WSZ. It was also observed that some of the model for the individual WSZs predicted well than others. For instance, the model for WSZ2 gave very good predictions, classifying 77.78% of the high-risk Fe and Mn accumulation potential values. This may be because the nodes where the water quality variables were sampled were uniformly distributed throughout the network. This made the yearly average concentrations of the water quality parameters at the DMA level a true representation at the node level in that DMA. On the other hand, if the nodes where the water quality variables were sampled are not well distributed at the DMA level, taking the yearly average concentrations of the water quality parameters may not be a true representation of every node in that DMA.

From Table 5.15, it was observed that the five models that used their individual WSZs logarithmic transformed data sets for the modelling had relatively low RMSE and high CA compared to the models that used untransformed and linear transformed data for modelling.

This implies that the models that used logarithmic transformed data for the modelling are more likely to predict Fe and Mn accumulation potential reasonably well on new data sets, and were therefore used in generating the risk maps in this research.

Table 5.15 Performance of the ANN(t,ψ) models with logarithmic transformed data

Performance indicator	WSZ1	WSZ2	WSZ3	WSZ4	WSZ5	WSZAll
Overall Training CA (%)	73.71	79.65	95.07	79.61	95.13	76.29
Overall Testing CA (%)	74.42	68.29	73.08	69.67	85.11	66.67
Training CA - low (%)	85.71	88.97	96.64	85.88	99.59	90.43
Training CA - medium (%)	61.40	57.45	86.36	64.19	77.46	38.82
Training CA - high (%)	19.05	70.59	100.00	70.47	75.56	26.53
Testing CA - low (%)	86.67	73.91	95.00	85.19	90.70	83.33
Testing CA - medium (%)	62.50	44.44	0.00	46.82	57.46	25.81
Testing CA - high (%)	20.00	77.78	0.00	62.29	65.57	23.08
Training RMSE	0.1792	0.1463	0.0765	0.1539	0.1660	0.2350
Validation RMSE	0.1795	0.2065	0.0870	0.1828	0.2113	0.2523
Testing RMSE	0.2502	0.2191	0.1499	0.2722	0.2805	0.2726
Training data points	185	181	114	205	213	901
Validation data points	47	45	28	51	54	225
Testing data points	43	41	26	45	47	201

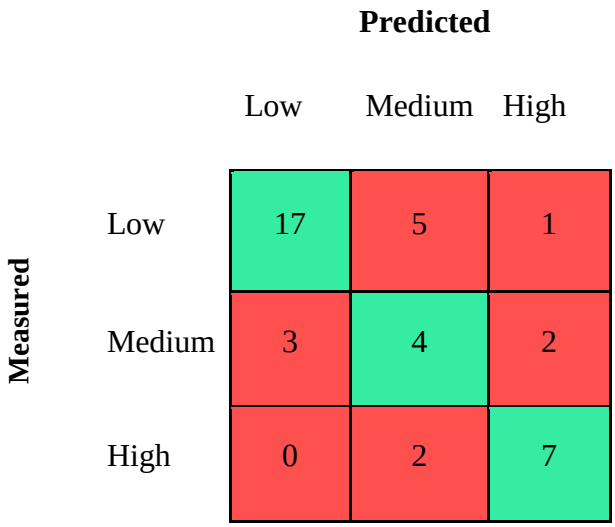


Figure 5.21 Testing data confusion matrix after predictions from the ANN(t,ψ) model for WSZ2 using logarithmic data

Figure 5.21 shows the confusion matrix of the untransformed testing data after predictions from the ANN(t) model for WSZ2. The model correctly predicted 17 out of 23 (77.78 %) high-risk values, 4 out of 9 (44.44 %) medium-risk values and 19 out of 25 (73.91 %) low-

risk values. The overall classification accuracy of 68.29% on the testing data set indicates that the ANN(t) model for WSZ2 is a good model which will make good predictions on new data sets.

It was observed from Table 5.15 that the models for WSZ1 and WSZ3 could correctly classify only 20 and 0 %, respectively of the high-risk values of Fe and Mn accumulation potential from the testing data set. This poor performance was due to the numerous sources of water supplied to these two WSZs. As a result of this, a few water quality variables values from these two WSZs had large variations and high standard deviations in each DMA. The water quality variables from the two WSZs may not have been well represented using the assumption that yearly average water quality variables at the nodes in each of the DMAs were approximately the same. In view of this, a multiple linear regression model was used to predict the measured water quality variables values at every node in each DMA for WSZ1 and WSZ3 in order to capture the variations. A multiple linear regression is a linear statistical technique that is used to establish a linear relationship between a dependent variable and several independent variables (Agha & Alnahhal, 2012). It can be mathematically expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon \tag{5.12}$$

where Y = dependent variable; X = set of independent variables;
β_n = n^{th} regression coefficient; and ε = error term.

To predict the value of the yearly average water quality variable at a given node, the independent variables used were the yearly average water quality variable of the DMA where the node is being estimated, maximum daily shear stress at the node, variation of daily shear stress at the node, hydraulic distance from source of water supply to the node, pipe age, pipe material index, and yearly average water quality variable in that DMA. For example, to estimate the yearly average turbidity level (independent variable) at a node in DMA1-12, known data of the independent variables from the DMA are used to develop a multiple linear regression model. The multiple linear regression model for estimating yearly average turbidity level at a given node can be mathematically expressed as Eqns. 5.13. An R^2 of 0.86 was obtained for this model (see Table 5.16). The sample size for the data of each DMA used for the regression model was approximately 35. Tables 5.16 and

5.17 show the coefficient of determination values after using multiple linear regression to estimate water quality data at every node in each DMA for WSZ1 and WSZ3, respectively.

Table 5.16 Coefficient of determination values after using multiple linear regression to estimate water quality data at every node in each DMA for WSZ1

DMA / Variable	DMA 1-01	DMA 1-02	DMA 1-05	DMA 1-06	DMA 1-07	DMA 1-08	DMA 1-10	DMA 1-11	DMA 1-12
Al*	0.58	0.88	0.27	0.48	0.99	0.30	0.54	0.38	0.89
Ca*	0.43	0.82	0.52	0.52	0.99	0.62	0.67	0.30	0.90
FCR	0.48	0.50	0.29	0.72	0.98	0.18	0.53	0.29	0.66
Colour	0.81	0.74	0.34	0.36	0.99	0.23	0.43	0.58	0.90
P*	0.61	0.89	0.43	0.65	0.82	0.45	0.54	0.49	0.77
Turbidity	0.34	0.66	0.38	0.54	0.69	0.42	0.30	0.73	0.88
Alkalinity	0.87	0.77	0.29	0.23	0.71	0.42	0.24	0.50	0.92
Hardness Total as CaCO3	0.42	0.84	0.50	0.53	0.98	0.64	0.64	0.30	0.86

* These are measured totals.

$$
\begin{aligned}
&Estimated\ yearly\ average\ turbidity\ at\ a\ node \\
&= -0.1022 + 0.1708(pipe\ material\ index) + 0.0035(pipe\ age) \\
&- 0.0126(average\ water\ age) + 0.3873(maximum\ daily\ shear\ stress\ at\ a\ node) \\
&- 3.7927(ariation\ of\ daily\ shear\ stress\ at\ a\ node) \\
&+ 0.0001(hydraulic\ distance\ from\ source\ of\ water\ supply) \\
&+ 1.0893(yearly\ average\ turbidity\ in\ that\ DMA) \qquad (5.13)
\end{aligned}
$$

Table 5.17 The coefficient of determination values after using multiple linear regression to estimate water quality data at every node in each DMA for WSZ3

Variable / DMA	DMA3-04	DMA3-05	DMA3-08	DMA3-09
Al*	0.95	0.36	0.59	0.34
Ca*	0.63	0.58	0.48	0.64
FCR	0.36	0.24	0.87	0.5
Colour	0.8	0.33	0.72	0.29
P*	0.6	0.26	0.43	0.53
Turbidity	0.42	0.42	0.47	0.66
Alkalinity	0.51	0.16	0.27	0.72
Hardness Total as CaCO3	0.63	0.63	0.52	0.64

* These are measured totals.

Table 5.18 presents the performance of the ANN(t,ψ) models when hydraulic, pipe-related, and estimated water quality data from the multiple linear regression model was used for the ANN modelling. It was observed that the predicted percentage of high-risk values of Fe and Mn accumulation potential from the testing data in WSZ1 improved from 20% to 66.42%. Similarly, the predicted percentage of high-risk values of Fe and Mn accumulation potential from the testing data in WSZ3 improved from 0% to 100%. These results indicate that, for WSZs with many sources of water supply, the model is able to predict high-risk values of Fe and Mn accumulation potential better when the input water quality variables for every node are estimated using multiple linear regression than assuming that yearly average water quality variable values at every node within each of the DMAs were approximately the same.

Table 5.18 The performance of the ANN(t,ψ) models using pipe-related, hydraulic and estimated water quality data from the multiple linear regression model

Performance indicator	**WSZ1**	**WSZ3**
Overall Training CA (%)	79.26	85.17
Overall Testing CA (%)	77.74	76.64
Training CA - low (%)	89.39	94.76
Training CA - medium (%)	74.73	79.91
Training CA - high (%)	69.58	75.82
Testing CA - low (%)	86.47	71.43
Testing CA - medium (%)	77.85	60.00
Testing CA - high (%)	66.42	100.00
Training RMSE	0.1591	0.0251
Validation RMSE	0.1642	0.0569
Testing RMSE	0.2349	0.1140
Training data points	185	114
Validation data points	47	28
Testing data points	43	26

5.6.2 Risk indexes for the ANN(t,ψ) models

Risk management has been successfully applied in several areas including corporate finance, project management, medicine, and engineering. Hubbard (2009) defines risk management as “the identification, assessment, and prioritisation of risks, followed by coordinated and economical application of resources to minimise, monitor, and control the probability and/or impact of unfortunate events”. In risk management, the potential of

unfortunate events likely to occur are identified and steps are taking to eliminate or reduce the impact of these likely events.

Every WDN has some level of risk of Fe and Mn failures in some regions in the network. Failure to identify or ignoring these high-risk regions will cause more Fe and Mn particles to accumulate on the pipe walls in the network, which will eventually lead to discolouration, customer complaints and Fe and Mn failures. Since the evaluation of risk is not a one-time process, but continuous, and the monitoring of water quality variables in the network is an expensive and laborious task, there is the need to devise a cost-effective method of identifying the high-risk regions.

The developed ANN(t,ψ) models can predict and classify various levels (high, low and medium) of Fe and Mn accumulation potential in WDNs. However, to effectively monitor the risk levels of each WSZ, there is a need to develop a risk index to quantify these levels. As explained in Section 5.3, the top 10% of all measured Fe and Mn accumulation potential values were classified as high-risk. If more than 10% of the predicted Fe and Mn accumulation potential by the model are high in a given WSZ, that WSZ is classified as a high-risk WSZ. If the predicted high values by the model in a WSZ are between 5 and 10% of all the model's predictions, it is classified as medium-risk WSZ. Finally, if the predicted high values by the model in a WSZ are less than 5% of all the model's predictions, it is classified as low-risk WSZ.

Table 5.19 presents the results of risk levels of five WSZs between the year 2005 and 2009 generated by the ANN(t,ψ) models. From the results, it was observed that the risk levels of some WSZs were not constant. For instance, from 2005 to 2006, the risk level of WSZ1 reduced from medium to low, and from 2008 to 2009 the risk level of WSZ4 increased from low to high. There are a number of reasons why these risk levels varied. It could be due to months of accumulation of Fe and Mn particles or network cleaning through flushing to remove accumulated sediments.

Table 5.19 Risk levels of five WSZs between 2005 and 2009 generated by the ANN(t,ψ) models at the WSZ level

WSZ \ Year	2005	2006	2007	2008	2009
WSZ1	Medium	Low	Low	Low	Low
WSZ2	High	High	High	High	High
WSZ3	Low	Medium	High	Medium	Medium
WSZ4	High	Medium	High	Low	High
WSZ5	Low	Low	Low	Low	Low

Although it is good to identify risk at the WSZ level, it is even better to identify risk at the DMA level. This is because a few high-risk DMAs often cause WSZs to be classified as high-risk. For example, WSZ2 was categorised as a high-risk WSZ in 2006 (see Table 5.19) at the WSZ level. However, the risk levels of WSZ2 in 2006 generated by the model at the DMA level categorised only 5 out of the 12 DMAs as high-risk (see Table 5.20). There were even four low-risk level DMAs found at WSZ2 in 2006, although WSZ2 was classified as high-risk at the WSZ level that year. This shows that classifying risk at the WSZ level does not always give a true picture of it. Furthermore, narrowing the risk to the DMA level makes it easier to identify and investigate the causes of the failures.

Table 5.20 Risk levels of WSZ2 in 2006 generated by the ANN(t,ψ) model at the DMA level

DMA	Percentage of high-risk nodes	Risk level
131-T2	15.38	High
128-01	0.00	Low
128-14	8.09	Medium
128-16	11.39	Medium
128-13	11.33	Medium
128-02	35.44	High
131-22	39.50	High
128-20	22.34	High
131-T5	0.00	Low
128-15	16.91	High
128-21	0.92	Low
128-T2	0.00	Low

5.6.3 Risk maps generated by the ANN(t,ψ) models

Presently, most water companies identify high discolouration risk regions in WDNs by selecting areas in the network with high Fe and Mn concentrations from their random sampling or by using customer complaints data relating to discolouration. As indicated in Chapter 1, these methods can be imprecise for two main reasons. First, with about 315,000 km of water mains in England and Wales, monitoring Fe and Mn concentrations will always be a very difficult and expensive task. This means that it may be impossible to sample every node in large WSZs. In view of this, regions which have high Fe and Mn concentrations that are not sampled will not be detected. Secondly, according to studies conducted by Ewan and Williams (1986) in the United Kingdom, approximately 30% of customers that experience discoloured water actually complain. This means that there is a high tendency that some regions in WSZs with high discolouration risk (Fe and Mn accumulation potential) can go undetected.

Although the ANN(t,ψ) models' risk indexes are able to help in identifying high-risk WSZs or DMAs, they have a limitation of not being able to determine the exact location of the high-risk nodes that contribute to making a WSZ or DMA high-risk. This is because not every node in the high-risk WSZs or DMAs contributes to making them high-risk. To overcome this limitation, risk maps were generated by the ANN(t,ψ) models to predict Fe and Mn accumulation potential for every node in a given WSZ. Narrowing the risk of Fe and Mn accumulation potential from the DMA level to the node level makes it easier to investigate the causes of high-risk Fe and Mn accumulation potential in the network.

Figure 5.22 (a) shows a risk map of predicted Fe and Mn accumulation potential for WSZ2 in 2009 generated by the model, whereas Fig. 5.22 (b) shows a risk map of the corresponding measured Fe and Mn accumulation potential. Three service reservoirs supply WSZ2 with water. Service reservoir A supplies water to the southern, north-eastern, and eastern regions of the reservoir, service reservoir B supplies water to the north-eastern and south-western regions of the reservoir, and service reservoir C supplies water to the north-western region of the reservoir. The predicted risk maps show that, generally, Fe and Mn accumulation potential increases with increasing hydraulic distance from the source of water supply.

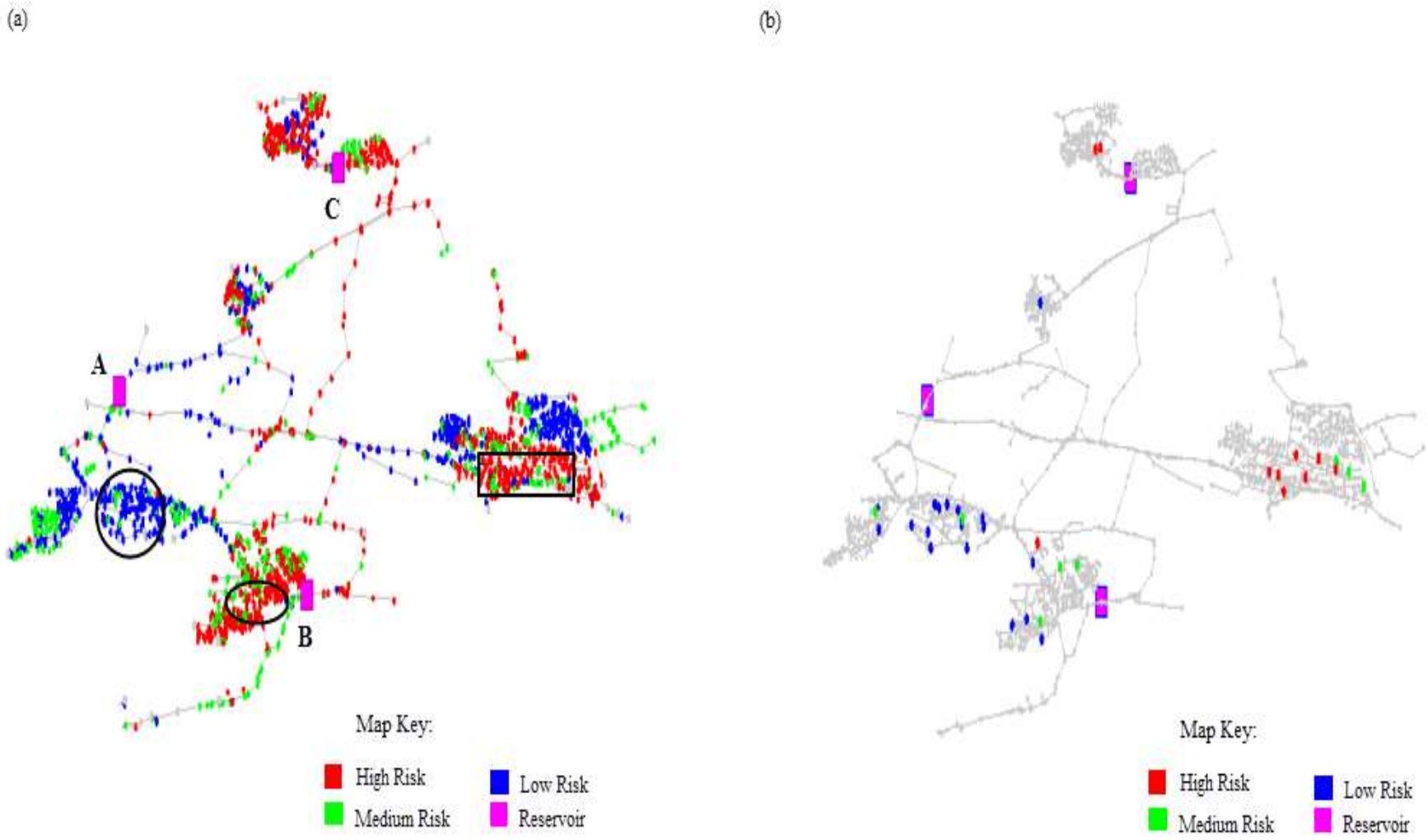


Figure 5.22 ANN(t,ψ) model risk maps showing (a) Predicted and (b) measured Fe and Mn accumulation potential at WSZ2 in 2009

From Fig. 5.22 (a), it was observed that DMA2-02 (region highlighted by a black oval) had high Fe and Mn accumulation potential values even though it is very close to service reservoir C. This is because DMA2-02 receives water from service reservoir A, which has a long hydraulic distance from this region. It was also observed that DMA2-01 (region highlighted by a black circle), which also receives water from service reservoir A, had very low Fe and Mn accumulation potential values. This is due to the short hydraulic distance from service reservoir A to DMA2-01.

In general, as hydraulic distance increases, average water age increases, chlorine dissipation increases, and microbial growth increases. Biological oxidation then becomes dominant, which subsequently leads to the biological oxidation of soluble Fe and Mn to insoluble Fe and Mn. The measured yearly average FCR concentrations at DMA2-01 and DMA2-02 for 2009 were 0.41 and 0.28 mg/L, respectively. This means that more biological oxidation occurred at DMA2-02 than DMA2-01 in 2009. The measured average water ages of 14.66 and 21.35 hours at DMA2-01 and DMA2-02, respectively observed are a further indication that the latter DMA provides a more conducive environment for microbial growth and is more prone to biological oxidation.

From Fig 5.22(a), the high predicted Fe and Mn accumulation potential values observed at DMA2-16 in 2009 (highlighted by a black rectangle) resulted from long hydraulic distance (approximately 6.5 km) from service reservoir A, low yearly average FCR (0.27 mg/L), and high yearly average water age (approximately 20 hours). Comparing the measured and predicted risk maps, it was observed that most of the regions in the network with measured high-risk of Fe and Mn accumulation potential were also predicted as high-risk regions by the model. Similarly, most of the regions with measured medium- and low-risk Fe and Mn accumulation potential were also predicted as medium- and low-risk regions by the model, respectively. These risk maps generated by the model will be able to help drinking water companies predict high-risk regions of Fe and Mn accumulation potential; including regions that have not been sampled for water quality variables.

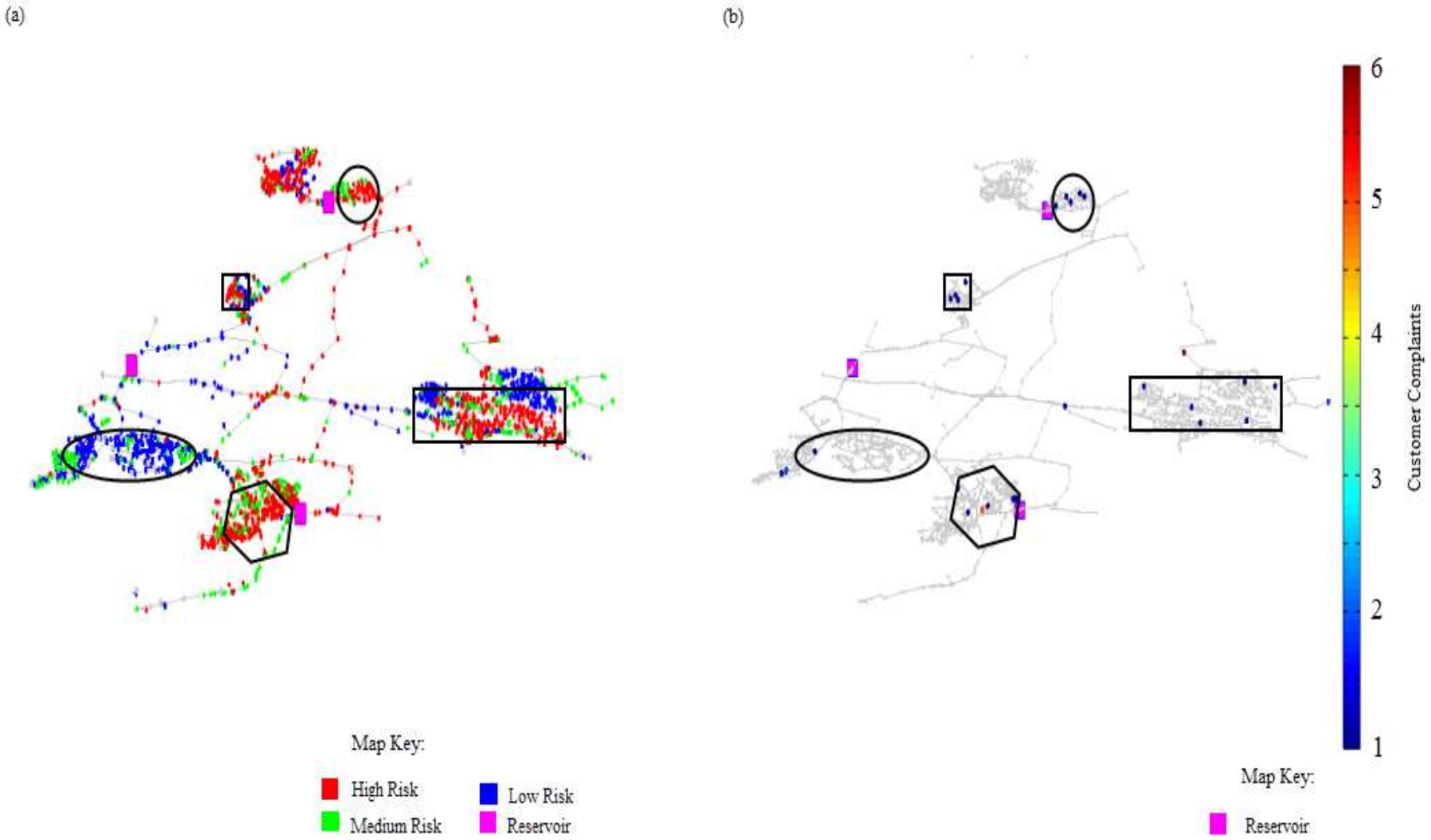


Figure 5.23 ANN(t,ψ) model risk maps showing (a) measured Fe and Mn accumulation potential and (b) customer complaints for WSZ2 in 2009

The predicted risk map for WSZ2 in 2009 was also compared with the customer complaints related to discolouration data for the same WSZ that year (Fig 5.23). It was observed that most regions in the network with high Fe and Mn accumulation potential also had high customer complaints. These results conforms to studies by Slaats (2002), who observed that the gradual accumulation or sudden increase of Fe and Mn particles in WDNs was the most common cause of water discolouration and customer complaints. This also explains why some researches have used Fe and Mn concentrations as KPIs in customer complaints studies (Bernal, Cardenoso, Babrellas, Matia, & Salvatella, 1999; Ewan & Williams, 1986; Gauthier et al., 1999).

With approximately 30% of all customer complaints related to water quality in UK being as a result of discolouration according to a study conducted by (Ewan & Williams, 1986), it is very important to identify regions in the network with potential high customer complaints. Furthermore, the introduction of the Service Incentive Mechanism (SIM) in April 2010 by the Ofwat has made it extremely important for water companies to reduce the number of customer complaints due to drinking water discolouration. The SIM rates water companies on their performance based on customer satisfaction, and either rewards or penalises them. The high-risk regions identified by the model can help to reduce customer complaints, and Fe and Mn compliant failures by periodically flushing the identified high-risk regions to clean up the accumulated Fe and Mn.

A few of the model predictions did not correlate very well with customer complaints. For instance, although the ANN(t,ψ) model for WSZ5 has an overall testing CA of 85.11% for Fe and Mn accumulation potential, it did not correlate very well with customer complaints due to discolouration in 2009. Some of the high-risk regions predicted at WSZ5 in 2009 did not correspond to its customer complaints due to discolouration that year. Instead, most of the customers complained in regions where there were pipe bursts in DMA5-06 (highlighted by a red circle), as shown in Fig. 5.24 (b) and (c). This is because the ANN(t,ψ) models have a limitation of not being able to predict high-risk levels in events such as pipe burst and opening of fire hydrants during fire extinguishing exercises. These unpredictable event variables which can also cause discolouration were not included in the model and so their effects on the risk of Fe and Mn accumulation potential or customer complaints were subsequently not captured. The model was developed to predict Fe and Mn accumulation potential, but not discolouration.

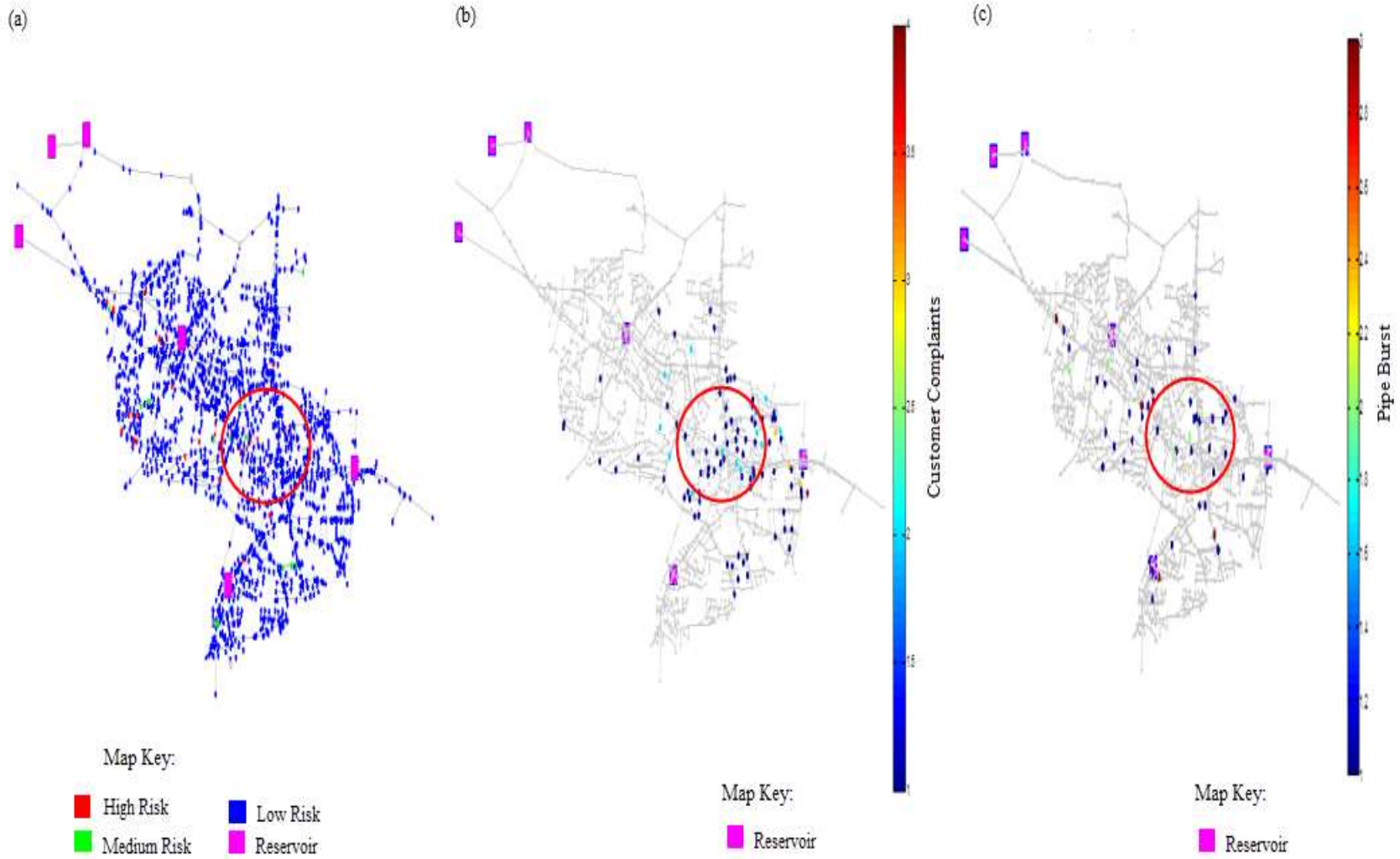


Figure 5.24 ANN(t,ψ) model risk maps showing (a) predicted Fe and Mn accumulation potential (b) customer complaints and (c) pipe burst at WSZ5 in 2009

To investigate the causes of customer complaints at DMA5-06, 2009 pipe burst data were plotted on a map of WSZ5 (see Fig 24 (c)). From the map, it was observed that there were a high number of burst in and around DMA5-06 (highlighted by a red circle). Water discolouration as a result of these pipe bursts may have contributed to the increased number of customer complaints. These complaints could also have been as a result of events such as the reinstatement of pipe mains after repairs or the opening of fire hydrants during fire extinguishing exercises. To investigate the dates on which the customers complained, monthly customer complaints data were plotted against the months of the year at DMA5-06 (see Fig 5.25). The graph showed that an unusually high number of customer complaints were recorded at DMA5-06 in July 2009. A graph of Fe and Mn concentrations against date at DMA5-06 in 2009 also showed relatively high Fe and Mn concentrations in July. This may be due to mobilised Fe and Mn particles caused by above mentioned hydraulic events in July, which were not captured by the ANN(t,ψ) model.

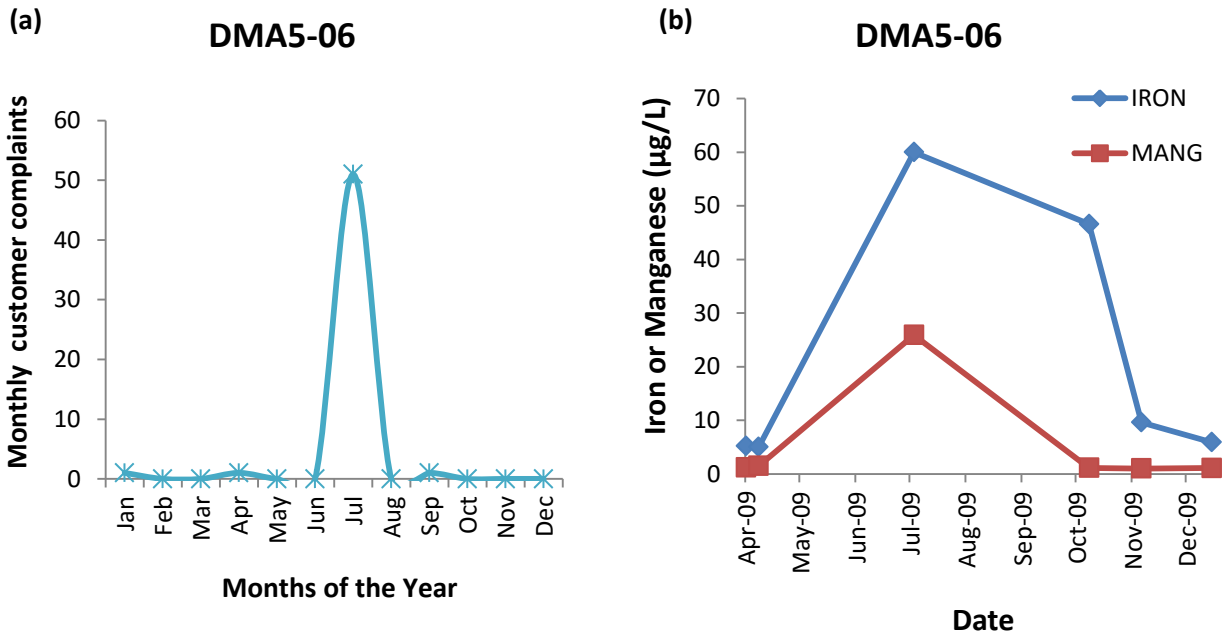


Figure 5.25 (a) Monthly variations of customer complaints and (b) variation of iron and manganese concentrations date at DMA5-06 in 2009

5.7 Summary

Two ANN models to predict Fe and Mn accumulation potential were developed in this chapter. The first model, ANN(t), was used as a sensitivity tool to select relevant input variables that influence Fe and Mn accumulation potential. It was also used to predict Fe and Mn accumulation potential. Furthermore, it could be used as an optimiser for water companies to find optimal input variable values to reduce Fe and Mn accumulation potential. From the prediction profiler graphs generated by ANN(t) model the following observations were made:

- Increased in Al concentration generally increased Fe and Mn accumulation potential.
- There was a high positive correlation between Fe and Mn accumulation potential and turbidity.
- In general, long hydraulic distance from source of water supply increased Fe and Mn accumulation potential.
- There was a negative correlation between Fe and Mn accumulation potential and Ca concentration.
- A highly positive correlation between hardness and Fe and Mn accumulation potential was observed.
- It was observed that chemical oxidation, corrosion, biological oxidation, sorption, shear stress effect and distance effect where all very important processes that influenced Fe and Mn accumulation potential.

The ANN(t,ψ) model was used to predict Fe and Mn accumulation potential for each node and to determine high-risk DMAs and WSZs. In addition it was used to generate risk maps to visually show the distribution of Fe and Mn accumulation potential in WSZs. From the risk maps generated by the model, it was observed that most of the regions in the network with high Fe and Mn accumulation potential for each of the WSZs also had high number of customer complaints due to discolouration. There were a few years the high-risk regions predicted by the model did not correlate well with customer complaints. These was because events such as pipe burst and opening of fire hydrants during flushing were not included in the models, and were therefore not captured when they caused discolouration. However, it should be noted that the ANN(t,ψ) model was developed to predict Fe and Mn

accumulation potential, which is the main cause of drinking water discolouration, but not customer complaints or discolouration in general.

The high CA and low RMSE values observed in the testing data sets show that the models are likely to predict well on new datasets. Just like all ANN models, the ANN(t) and ANN(t,ψ) models can be used to make predictions on new data sets from different WSZs. However, it should first be trained with data from the new WSZs. The developed models can be used as tools to assist in reducing discolouration and customer complaints by helping water resource engineers to identify high-risk regions, investigate the causes of high Fe and Mn accumulation potential in those regions, and if possible, find solutions to them. Although the ANN(t,ψ) was able to predict Fe and Mn accumulation reasonably well and identify high-risk zones, they could not programmatically determine the causes of the failures. It had to be manually investigated to find the reasons for the failures. The black-box nature of ANNs make it difficult to programmatically trace back the causes of the failures from the output variable to the input variables. With so many nodes in WSZs, manually investigating the causes of failures can be a laborious task. It is envisaged that a hierarchical fuzzy logic model will be able to overcome this limitation by predicting Fe and Mn accumulation potential as well as explaining the causes of failures in the network.

CHAPTER 6: Fuzzy Inference System for Predicting Accumulation Potential

6.1 Introduction

In the UK, high Fe and Mn concentrations in WDNs can lead to penalisation by the DWI and the Ofwat. To prevent this, water companies need a model that can identify both high-risk regions and the causes of failures in these regions. The causes of Fe and Mn failures are difficult to determine by mathematical formulae or traditional models. Data-driven FIS approaches are more appropriate to solve such problems because of their learning capabilities and their ability to cope well with uncertainties (Cox, 1992). Knowing the exact cause(s) of the risk will determine what appropriate measures that can be taken to reduce it. In Chapter 5, two ANN models for predicting Fe and Mn accumulation potential were developed. Although they could identify various risk levels of Fe and Mn failures in WSZs, the black-box nature of ANNs made them unable to explain the causes for these failures unless they were manually investigated. With thousands of nodes in every WSZ, manually investigating the causes of Fe and Mn failures at every node would be time-consuming and very laborious. To overcome this limitation, two hierarchical fuzzy inference systems (FISs) for predicting Fe and Mn accumulation potential were developed in this chapter. They are the hierarchical rule-based expert FIS, and the hierarchical data-driven FIS. The hierarchical rule-based expert FIS uses expert knowledge to formulate rules, whereas the hierarchical data-driven FIS uses genetic algorithm to optimise the rules and their weights.

Unlike ANN models, the intermediate nodes of FISs are white-boxed. Hence, the system can explain the causes of high Fe and Mn accumulation potential. In addition, the FISs can predict and classify various levels (high, medium, or low) of Fe and Mn accumulation potential in WDNs. They used the same relevant variables used in modelling the ANN models in Chapter 5. The relevant variables were categorised into hydraulic, chemical, and biological. These variables undergo complex processes as water travels through WDNs. Some of the processes that influence Fe and Mn accumulation include corrosion, chemical oxidation, sorption, and biological oxidation. The hierarchical FISs then capture all the processes and use them to make predictions. The remaining parts of this chapter are arranged as follows. Section 6.2 explains how the data was prepared for the FIS. This

section describes how all the data were transformed between zero and one. Sections 6.3 and 6.4 discuss the development of the hierarchical rule-based expert FIS and the hierarchical data-driven FIS, respectively. The results of the hierarchical rule-based expert FIS and hierarchical data-driven FIS are presented in Sections 6.5 and 6.6, respectively. Finally, the summary of this chapter is presented in Section 6.7.

6.2 Data preparation

The success of the FIS, just like any model, is dependent on how well the data are prepared. The same prepared five-year data set used for the ANN model was also used for the FIS. Yearly averages of water quality variables were used in the FIS. It would have been ideal to use monthly or quarterly averages as input water quality variables, since Fe and Mn accumulation potential exhibits seasonal variations. However, because some water quality variables were not sampled at sufficient frequency, the data would have had many gaps if monthly or quarterly averages were used.

The same assumption used in Chapter 5 in calculating the yearly averages of water quality variables at each node was used to prepare the data for the FIS. The 14 independent variables used for the modelling were Al, alkalinity, turbidity, hardness, calcium, FCR, colour, phosphorus, average water age, maximum daily shear stress at a node, variation of daily shear stress at a node, hydraulic distance from source of water supply to a node, pipe material index, and pipe age. Because the structure of the developed FIS is hierarchical, there were intermediate nodes to link the input variables to the output variable. The eight intermediate nodes used were chemical oxidation, corrosion, sorption, chemical effect, biological effect, shear stress effect, distance effect, and hydraulic effect. The dependent variable, measured Fe and Mn accumulation potential, was calculated using Eqn. 5.4 in Chapter 5. Data for each variable were linearly transformed between zero and one using Eqn. 5.1 in Chapter 5.

The same classification levels of Fe and Mn accumulation potential used in Chapter 5 were used in developing the FISs. Measured values of Fe and Mn accumulation potential above 90^{th} percentile were classified as high-risk, between 70^{th} and 90^{th} percentile as medium-risk and below 70^{th} percentile as low-risk.

6.3 Model development of the hierarchical rule-based expert FIS

Water companies generally set post-treatment targets of Fe and Mn to approximately 3% of their respective MCLs to reduce the concentrations in WDNs. Irrespective of how well water is treated, very low concentrations of Fe and Mn may still enter the network from water treatment plants and gradually accumulate on pipe walls. During events such as high flows, bursts, or high diurnal consumption of drinking water, these accumulated particles re-suspend and subsequently cause water discolouration. If the re-suspended particles end up in customers' taps, it prompts customers to complain. These complaints greatly undermine customers' confidence in water companies. Discolouration can also occur as a result of increased chemical oxidation, corrosion, sorption, biological oxidation, water age, and hydraulic distance from source of water supply.

A FIS is a system that makes predictions or decisions by mapping a set of given inputs to a given output using fuzzy logic. Two hierarchical FISs were developed to capture the processes that occur in WDNs. The first FIS developed, hierarchical rule-based expert FIS, uses knowledge from human experts to form rules that describe the data used for the modelling. The second FIS, the hierarchical data-driven FIS, uses genetic algorithm to generate rules for the system.

Unlike ANN models which are regarded as black-box because their internal operations are difficult to explain and rely heavily on the data that describes the input and output variables, the membership functions which make up the fuzzy sets can easily be defined. The transparency and interpretability of FISs are their main advantages over ANN models. Although fuzzy logic is a powerful tool that can be used to solve many control problems, it may not be applicable in solving some problems. Some advantages and disadvantages of fuzzy logic have been listed by Robert (1989). Cox (1992) suggested that FISs can be used to solve problems that:

(a) Have one or more continuous input variables.
(b) Cannot be solved mathematically
(c) Are difficult to solve mathematically because of computational memory.
(d) Experts can identify rules that define the behaviour of the system.

There are several types of FIS, namely the Mamdani fuzzy model, Sugeno fuzzy model, and Tsukamoto fuzzy model. However, the most commonly used FIS, Mamdani fuzzy logic approach, was used in this research because of its simplicity and effectiveness in handling linguistic variables. The following steps show how the hierarchical rule-based expert FIS was developed.

(a) Knowledge acquisition – acquiring expect knowledge for the formation of fuzzy rules.
(b) Choosing appropriate membership functions – selecting appropriate membership functions that define the points in the universe of discourse.
(c) Fuzzification – converts the input data into fuzzy representations.
(d) Fuzzy logic rules – helps in the processing of the data.
(e) Aggregation – combines outputs of each rule into a single fuzzy set.
(f) Defuzzification – converts the output data into a crisp value.
(g) Membership function tuning – manually tuning the membership functions.

6.3.1 Summary of knowledge acquired

Knowledge acquisition is a very important step in the fuzzy modelling process. It is defined as the process of gathering relevant information about a domain (De Kork, 2003). There are several ways that relevant information can be gathered. This could be deductively by human experts, inductively by learning from examples, from historical database, or by data-driven approach using modelling data (Oladipupo, Ayo, & Uwadia, 2012). Data-driven approach of acquiring knowledge can be through genetic algorithm, artificial neural network, clustering, machine language or classification, whereas expert knowledge is acquired through human experts.

An expert system is a computer program that uses knowledge from human experts to solve control problems, usually with a small number of input variables (Feigenbaum, 1982). They are normally used to solve complicated problems that cannot be solved using algorithm and therefore requires human intelligence. A fuzzy rule-based expert system is an expert system that uses human knowledge to form rules and tune membership functions to reason about data in inference mechanism (Neshat & Yaghobi, 2009). The prediction accuracy of an expert system is highly dependent on the accuracy of the knowledge-base that is used to define the rules. Therefore, the knowledge-base must be accurate to make the system predictions credible. Since the number of fuzzy logic rules increase

exponentially with input variables, it becomes difficult to manually define the rules in fuzzy rule-based expert systems with many input variables. Hence, data-driven approach of knowledge acquisition is the preferred technique for fuzzy control problems with many input variables. Unlike data-driven approach of modelling, expert systems usually use manual rule-based approach in making its computations.

Since fuzzy logic has the ability to express natural language as fuzzy logic rules, the knowledge acquired on relevant variables that influence Fe and Mn accumulation potential were translated into these rules. The following section gives a summary of the knowledge acquired to form rules for the hierarchical rule-based expert FIS.

6.3.1.1 Effect of chemical oxidation on Fe and Mn accumulation potential

Hardness

Increase in hardness increases chemical oxidation of Fe and Mn. Besides dissolved Mg and Ca that contributes to hardness, dissolved ions such as Fe, Mn, Al, and zinc also contribute to hardness (WHO, 2011b). The prediction profiler graphs in Section 5.5.3 also show a strong positive correlation between Fe and Mn accumulation potential and hardness.

Free chlorine residual

FCR has a positive correlation with chemical oxidation of Fe and Mn. Since it is an oxidising agent, it helps to chemically oxidise soluble Fe^{2+} and Mn^{2+} to insoluble Fe^{3+} and Mn^{4+}.

Alkalinity

Alkalinity has a negative correlation with chemical oxidation of Fe and Mn. This is because increase in alkalinity helps to increase the buffer capacity of drinking water, thus keeping the pH of water stable, and reducing chemical oxidation of Fe and Mn in WDNs. Also, studies by Naylor et al. (1993) showed a negative correlation between alkalinity and corrosion.

6.3.1.2 Effect of corrosion on Fe and Mn accumulation potential

Pipe Material

Pipe material was arranged from low to high values in order of corrosivity as:

Polyethylene (PE) → Polyvinyl chloride (PVC) → High Density Polyethylene (HDPE) → Asbestos Cement (AC) → Ductile Iron (DI) → Steel (ST) → Cast Iron (CI).

Each of them was given a value between zero and one termed pipe material index. Low corrosive materials were given values close to zero, whereas high corrosive materials were given values close to one.

Pipe Age

In iron pipes, increase in pipe age increases corrosion rates. This is because old cast iron pipes tend to be more corroded than newer ones.

6.3.1.3 Effect of sorption on Fe and Mn accumulation potential

Aluminium

Increase in Al concentration increases sorption. This is due to the formation of amorphous $Al(OH)_3$ with increasing Al concentration which tends to adsorb Fe and Mn particles (Wang et al., 2012). The prediction profiler graphs in section 5.5.3 also showed a strong positive correlation between Fe and Mn accumulation potential and Al concentration.

Calcium

Calcium has a positive correlation with sorption because adsorption of Mn and Fe on amorphous $Al(OH)_3$ is enhanced by high concentrations of Ca^{2+} and Mg^{2+} (Wang et al., 2012).

Colour

Colour has a positive correlation with sorption because increase in TOC, which is an indirect measure of colour, enhances the adsorption of Mn (Wang et al., 2012).

6.3.1.4 Effect of biological oxidation on Fe and Mn accumulation potential

Free chlorine residual

FCR has a negative correlation with biological oxidation of Fe and Mn. This is due to the killing or reduction of the growth of Fe- and Mn-oxidising bacteria with increasing free chlorine residual levels.

Colour

Increasing colour increases biological oxidation of Fe and Mn because increase in colour (total organic carbon) enhances biofilm formation. Carbon serves as a bioavailable form of nutrient for bacteria responsible for the formation of biofilms (van der Kooij, 2002).

Water age

Increase in water age increases biological oxidation of Fe and Mn. Stagnant water conditions promote the growth of bacteria, increase biological oxidation and result in the deterioration of water quality.

Turbidity

Turbidity has a positive correlation with biological oxidation of Fe and Mn. Increase in turbidity increases the concentration of suspended organic particles. Fe- and Mn-oxidising bacteria attach themselves to these suspended particles, causing microbial growth to increase. High turbidity levels also enhance the biological oxidation of Fe and Mn by serving as a shield to protect microorganisms from disinfection (WHO, 2011a).

Phosphorus

Increase in phosphorus increases biological oxidation because phosphorus is a bioavailable form of nutrient needed by bacteria in WDNs for growth and reproduction (CRCWQT, 2005).

6.3.1.5 Shear stress effect on Fe and Mn accumulation potential

Maximum daily shear stress at node

Increase in maximum daily shear stress decreases Fe and Mn accumulation potential because Fe and Mn precipitates are unable to accumulate on the pipe walls under high maximum daily shear stress. In general, high shear stress regions are subject to low accumulation potential, whereas low shear stress regions are subject to high accumulation potential.

Variation of daily shear stress at node

Increase in variation of daily shear stress decreases Fe and Mn accumulation potential. Nodes with low variation of daily shear stress generally have low disturbance in WDNs. Hence, Fe and Mn particles accumulate easily. On the other hand, nodes with high variation of daily shear stress generally have high disturbance in WDNs. Therefore, Fe and Mn particles in these regions are unable to accumulate on the pipe walls.

6.3.1.6 Distance effect on Fe and Mn accumulation potential

Water age

Increase in water age increases Fe and Mn accumulation potential. Generally, when water age is high, chlorine levels are low, resulting in creating a conducive environment for the formation of biofilms. Such regions are more susceptible to biological oxidation of Fe and Mn.

Hydraulic distance from source of water supply

Hydraulic distance from source of water supply has a positive correlation with Fe and Mn accumulation potential. Prediction profiler graphs from the ANN(t) models in Section

5.5.3 show that hydraulic distance from source of water supply has a strong positive correlation with Fe and Mn accumulation potential.

6.3.2 Choosing appropriate membership functions

A fuzzy membership function is a curve that defines how the points in the universe of discourse are mapped to a membership value between 0 and 1. These membership functions help in expressing fuzzy rules in linguistic form using linguistic words such as high, medium, cold, and hot. Membership functions were defined for each of the input, intermediate, and output fuzzy variables in the hierarchical FIS. Unlike classical sets, which have binary memberships and hard boundaries, fuzzy logic sets can have partial memberships. Fuzzy membership functions can take various forms or shapes. There are different types of fuzzy membership functions, including triangular, trapezoidal, Gaussian, S-function, Gbell, Pi-shaped, Dsigmoidal, and Psigmoidal. There are several methods that can be used to assign the appropriate membership function to fuzzy variables. These methods include inductive reasoning, neural networks, genetic algorithms, inference, rank ordering, and intuition (Ross, 2010). However, the precise shapes or types of membership functions that is used is not so important and have little effect on the performance of the model. It is rather the placement of the membership functions within the universe of discourse and how they overlap each other that significantly affects the performance of models (Ross, 2010).

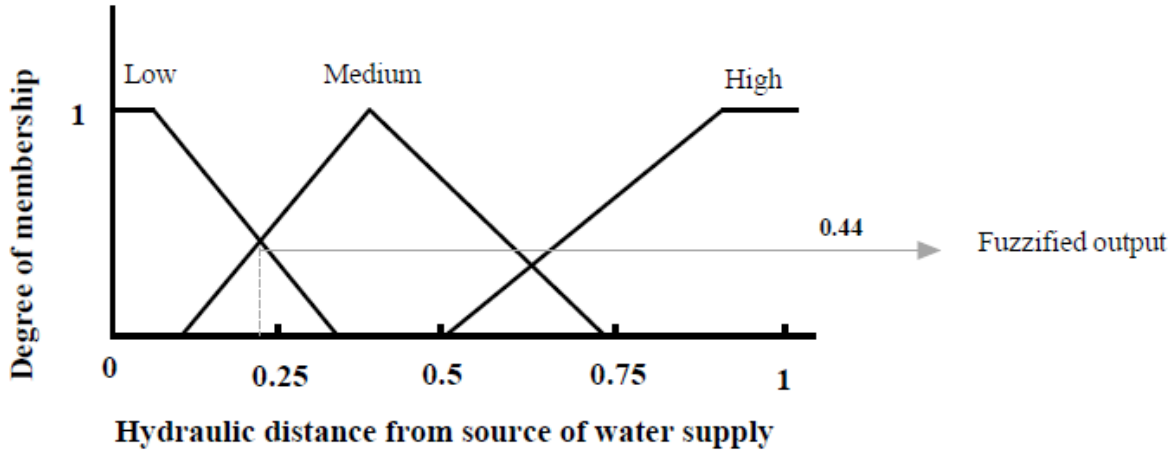


Figure 6.1 Fuzzy set for hydraulic distance from source of water supply in WSZ2 showing the membership functions

Triangular and trapezoidal membership functions is the most commonly used membership function (Nasr, Rezaei, & Dashti Barmaki, 2012). Hence, they were used to develop the FIS. The membership functions of all the fuzzy variables were partitioned into three linguistic categories, namely low, medium, and high. Trapezoidal membership functions were used for the medium category, and triangular membership functions were used for the low and high categories. The membership functions for the input variable, hydraulic distance from source of water supply in WSZ2 are presented in Fig 6.1.

6.3.3 Fuzzification

Fuzzification is a step in the FIS process where crisp values of the input variables are fuzzified. An output fuzzified value between 0 and 1 is returned irrespective of the value of the crisp input variable. Fuzzy logic sets allow partial memberships. This means that they can have intermediate values between two membership functions. For instance, in Fig 6.1, a node in WSZ2 with a hydraulic distance from source of water supply of 0.23 would have an intermediate value between 'low' and 'medium' with a fuzzified value of 0.44. In the fuzzification process, the degree to which the input values belong to each of the membership functions is determined. This concept of using linguistic terms in fuzzification is very important because it provides a way to represent real world problems which comprises of uncertainties due to imprecision or ambiguity. It also makes it possible to compute with words. More details of the fuzzification process are presented in Section 3.3.4.2.

6.3.4 Formulation of fuzzy rules from expert knowledge

The fuzzy rules were formulated from the knowledge acquired in previous chapters on variables that influence Fe and Mn accumulation potential. If the antecedent of the fuzzy rule has more than one linguistic set, the AND operator was used to combine multiple antecedents, and the Mamdani minimum implication method was used to truncate the output fuzzy sets. For details of the Mamdani minimum implication method, refer to Section 3.3.4.4. The intersection of the antecedents can be evaluated using Eqn. 6.1, whereas Fig. 6.2 illustrates the intersection of the membership functions of fuzzy sets A and B.

$$\mu_{A\cap B}(x) = \min[\mu_A(x), \mu_B(x)], \qquad x \in U \tag{6.1}$$

where μ_A = membership function that defines the fuzzy set A;
μ_B = membership function that defines the fuzzy set B; and *U* = universe of discourse.

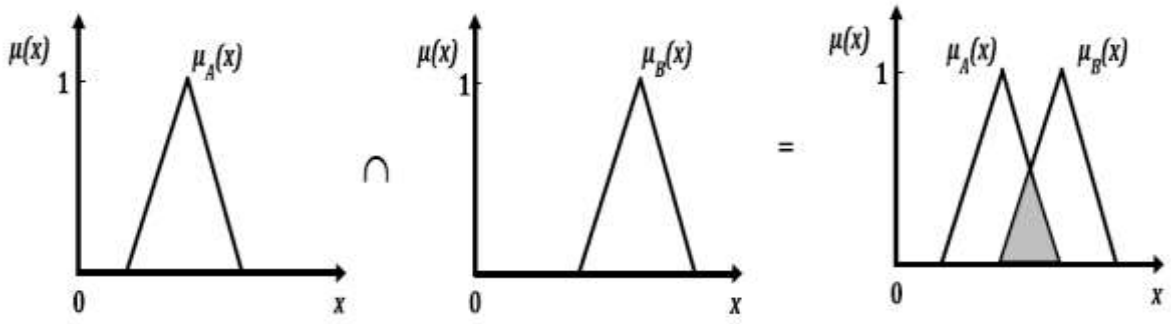


Figure 6.2 The intersection of membership functions of fuzzy sets A and B

Some of the formulated rules may have more influence on the system results than others. It is very important to ensure the system gives preference to more influential rules than the less influential rules. This was implemented by assigning weights to each of the rules. The weights are numbers between zero and one assigned to the consequent part of the rules to give them a level of importance in the FIS process. Very influential rules were assigned value one or close to one, whereas less influential rules were assigned values close to zero. From expert knowledge, it is known that corrosion is one of the most common causes of drinking water discolouration (DWI, 2007). Hence, rules associated with corrosion were assigned more weights. Similarly, it is known from expert knowledge that regions with dead-end pipes and redundant loops contribute significantly to discolouration (Boxall et al., 2001). Chlorine is also known to dissipate rapidly in these regions because of their high water age. These stagnant conditions promote the growth of bacteria, increase biological oxidation, and result in the deterioration of water quality. Therefore, rules associated with shear stress and biological oxidation were given more weights.

6.3.4.1 Reduction of the fuzzy rules

One of the dilemmas researchers face in developing FISs is the formulation of fuzzy rules with many input variables. It is well known that the number of fuzzy rules generally increases with number of input variables and/or membership functions. FIS with many input variables can reduce model performance, increase computational time, and exacerbate computational memory. It is also very difficult to manually formulate fuzzy

rules with too many input variables. A number of researchers have proposed different rule reduction techniques to mitigate this problem. Giiven and Passino (2001) proposed a fuzzy rule reduction technique in which the number of rules increases linearly with the number of input parameters. Ciliz (2005) proposed an algorithm that eliminates inconsistent and redundant fuzzy rules without affecting the performance of the FIS. In most of the reduction techniques, the performances of the FISs are reduced. These reduction techniques can be categorised into three, namely;

- the selection of most significant rules;
- the elimination of redundant rules; and
- the merging of rules with common properties (Balasubramaniam, 2006).

Since genetic algorithm was used in this research to optimise the fuzzy rules, there was the need to reduce the number of rules in order to reduce the number of subjects of the population for the genetic algorithm to find feasible solutions to the problem. The conventional method of formulating fuzzy rules also known as intersection rule configuration (IRC) increases exponentially with the number of input variables. As per this method, the consequent is obtained from the intersection of two or more antecedent part of fuzzy rules (see Eqn. 3.30). In IRC, the rules are formed for every possible combination of the membership functions. To illustrate this, suppose there is a FIS with 3 membership functions and 6 input variables, then the total number of rules formed using the IRC technique would be 3^6 (729).

To mitigate fuzzy rule explosion while maintaining a good system performance, a fuzzy rule reduction method known as the union rule configuration (URC) by (Combs & Andrews, 1998) was adopted. The URC method forms rules using only one antecedent for every consequent (see Table 6.1). This method uses simple implication to obtain the consequent part of fuzzy rules and aggregated using Eqn. 3.31. Contrary to the IRC method, the URC increases linearly with the number of input variables. For instance, the total number of rules formed for a FIS with 3 membership functions and 6 input variables using the URC method will be 3×6 (18). Comparing the illustrations from the two methods, it can be seen that the number of rules was reduced from 729 to 18. A fuzzy subsystem from the developed hierarchical FIS was used to test the accuracy of the URC. Figure 6.3 shows the fuzzy inference subsystem from the developed hierarchical FIS. The

input variables are average water age and hydraulic distance from source of water supply, whereas the output variable is distance effect.

Table 6.1 Some rules used in the hierarchical rule-based expert FIS

Rule Number	Rule
1	If pipe material index is LOW then corrosion is LOW
2	If pipe material index is MEDIUM then corrosion is MEDIUM
3	If pipe material index is HIGH then corrosion is HIGH
4	If free chlorine residual is LOW then chemical oxidation is LOW
5	If free chlorine residual is MEDIUM then chemical oxidation is MEDIUM
6	If free chlorine residual is HIGH then chemical oxidation is HIGH
7	If alkalinity is LOW then chemical oxidation is HIGH
8	If alkalinity is MEDIUM then chemical oxidation is MEDIUM
9	If alkalinity is HIGH then chemical oxidation is LOW
10	If free chlorine residual is LOW then biological oxidation is HIGH
11	If free chlorine residual is MEDIUM then biological oxidation is MEDIUM
12	If free chlorine residual is HIGH then biological oxidation is LOW
13	If colour is LOW then biological oxidation is LOW
14	If colour is MEDIUM then biological oxidation is MEDIUM
15	If colour is HIGH then biological oxidation is HIGH
16	If water age is LOW then biological oxidation is LOW
17	If water age is MEDIUM then biological oxidation is MEDIUM
18	If water age is HIGH then biological oxidation is HIGH

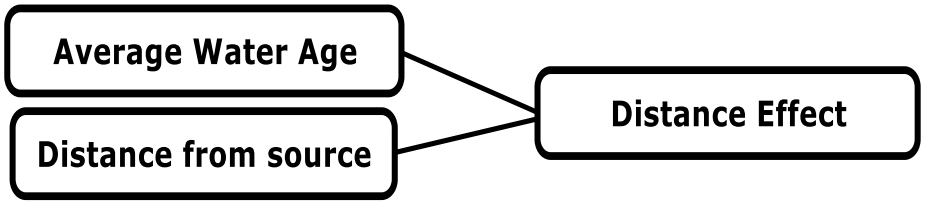
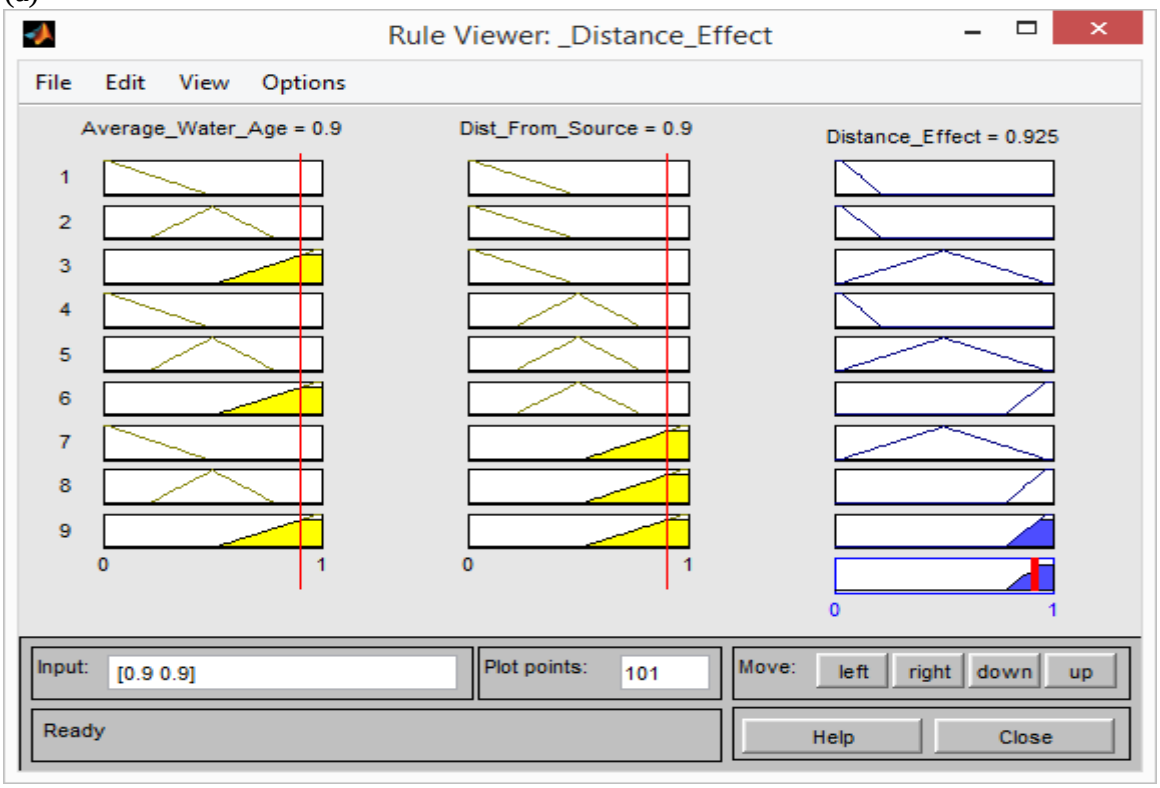


Figure 6.3 Fuzzy inference subsystem from the developed hierarchical FIS

Figure 6.4 shows screen shots of the fuzzy rule viewer and editor of the fuzzy inference subsystem when the IRC technique was used. Since there were three membership functions and two input variables, a total of nine rules were formed. Input values of 0.9 for both average water age and hydraulic distance from source of water supply gave a defuzzified crisp distance effect value of 0.925. Using the same fuzzy inference subsystem, fuzzy rules were formed using the URC method. Screen shots of the fuzzy rule viewer and editor for the fuzzy inference subsystem when the URC method was used is shown in Fig. 6.5. The total number of rules was reduced to six. Computing the distance effect value

with the same input values of 0.9 for both average water age and hydraulic distance from source of water supply gave the same result. Table 6.2 shows results from the inference subsystem when the IRC and URC methods were used with various input values. It was observed that both methods gave the same results. Since the rules reduction did not change the output results, the URC method was used to develop the hierarchical FIS in this research. A total of 78 rules were used in the FIS.

(a)



(b)

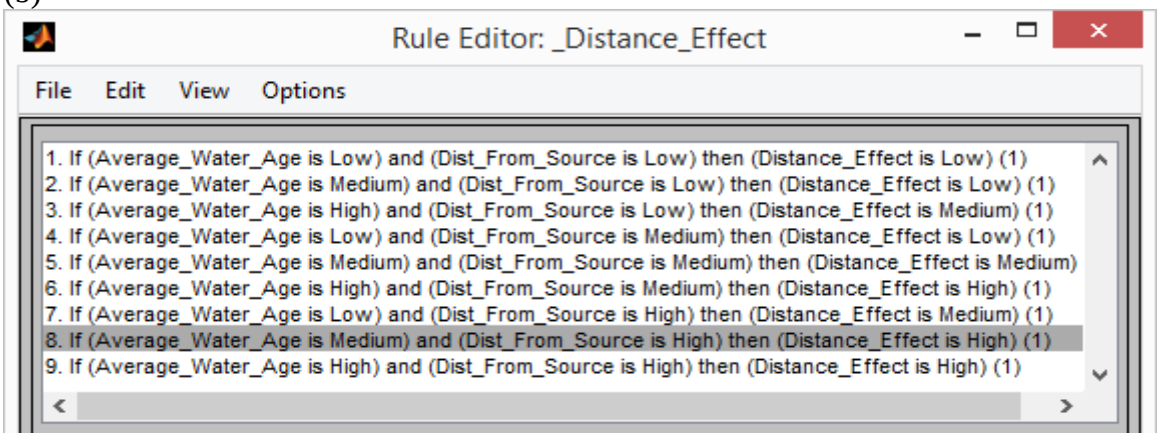
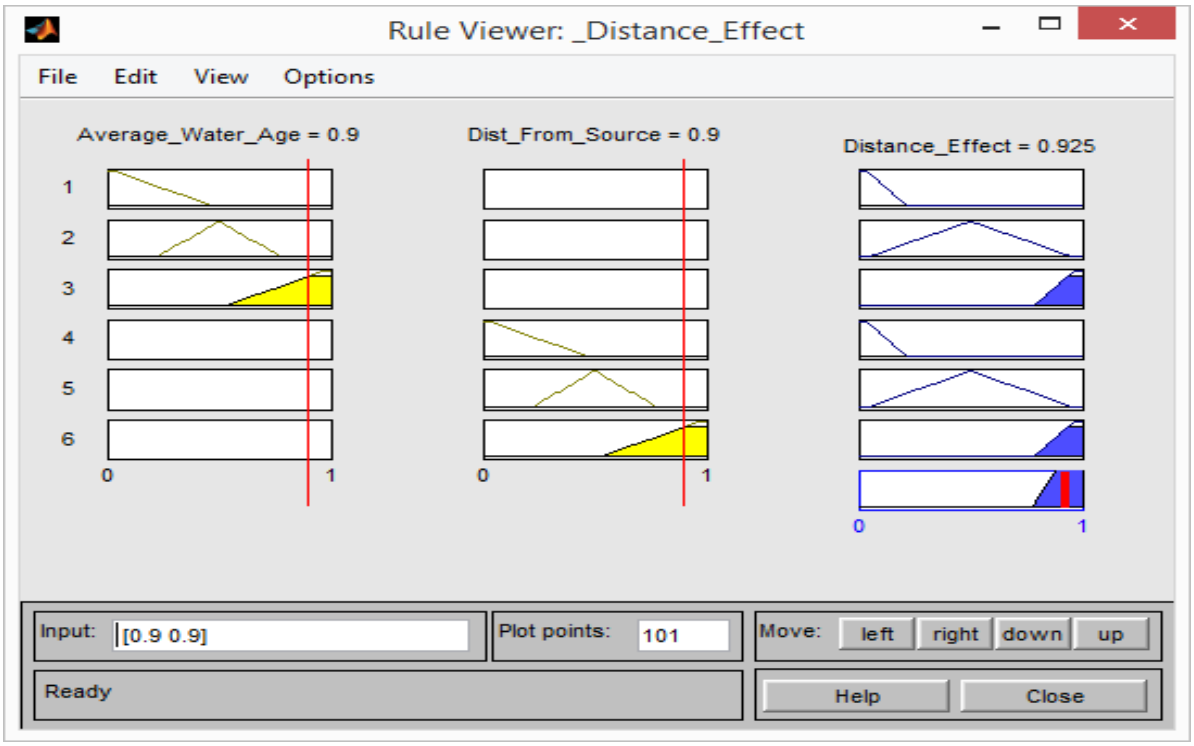


Figure 6.4 Screen shots of the (a) fuzzy rule viewer and (b) fuzzy rule editor of the fuzzy inference subsystem when the IRC technique was used

(a)



(b)

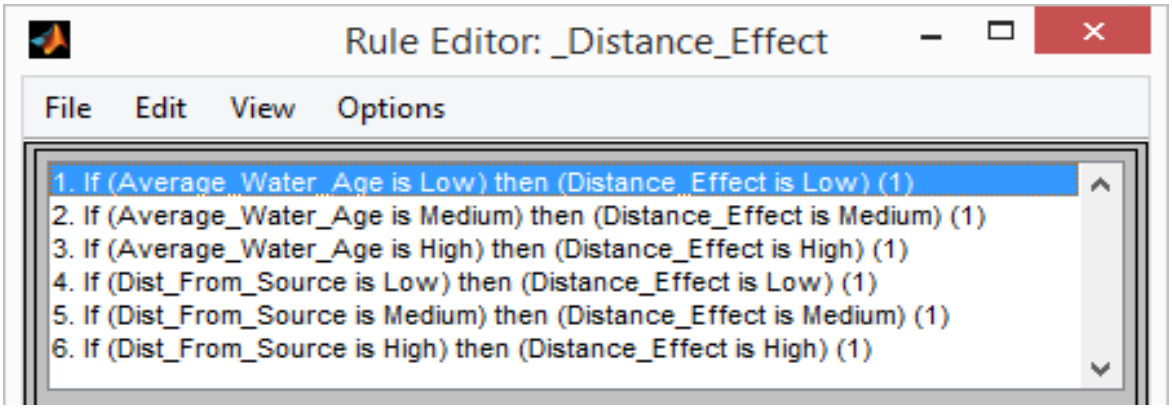


Figure 6.5 Screen shots of (a) fuzzy rule viewer and (b) fuzzy rule editor of the FIS using the URC technique

Table 6.2 Results from the inference subsystem using both IRC and URC methods

Average Water age	0.1	0.2	0.9
Hydraulic distance from source	0.1	0.8	0.9
Distance effect (using IRC method)	0.0749	0.5	0.925
Distance effect (using URC method)	0.0749	0.5	0.925

6.3.5 Rule aggregation

Since the output results of the FIS depend on all the rules, they were combined in order to make the effect of each of the rules contribute to the outcome of the predictions. The order in which the rules are executed are not important since the aggregation method is commutative. The process of combining the fuzzy output sets of each rule into a single fuzzy set is known as aggregation. The aggregated output of a fuzzy set is the union (aggregation) of two or more outputs of the rules. During the process, the truncated output functions returned by the implication process in each of the rules were aggregated using Eqn 6.2. Details on rule aggregation are presented in Section 3.3.4.5.

$$\mu_{A\cup B}(x) = \max[\mu_A(x), \mu_B(x)], \qquad x \in U \tag{6.2}$$

where

μ_A = truncated output functions returned in first rule that defines the fuzzy output set A;

μ_B = truncated output functions returned in second rule that defines the fuzzy output set B;

and U = universe of discourse; and

6.3.6 Defuzzification

The centroid method, which is also known as the centre of gravity (CoG) method, which is the most common technique of defuzzification, was used to develop the hierarchical FIS in this research because of its popularity, easiness to compute and the recommendation to use in quantitative models (Sugeno, 1985). The CoG method was also used because its deterministic response curve is smooth, continuous and gives consistent results (Pham & Castellani, 2002). Although this method gives good predictions, it has computational difficulties in processing fuzzy sets with complex membership functions (Nazz, Alam, & Biswas, 2011). The CoG method computes a crisp value representing the centre of gravity of the aggregated fuzzy outputs. It can be mathematically expressed as Eqn. 3.45 in Chapter 3. Details of the Defuzzification process are presented in Section 3.3.4.6

6.3.7 Manual tuning of membership functions

The success of a FIS also depends on how well the shapes of the membership functions are defined. The rule of thumb stipulated by Cox (1992) in defining membership functions was followed in this research. In this rule of thumb, he first suggested that the number of membership functions should be odd; between three and nine. Secondly, he proposed that

to give a smooth and stable surface to a fuzzy controller the membership functions should overlap between 10 and 50% of the space of neighbouring membership functions. Finally, he suggested that the density of the membership function should be highest around the optimal control point, and should decrease with distance from that point. These guidelines were followed in manually tuning the membership functions to improve the model performance.

6.3.8 The structure of the hierarchical rule-based expert FIS

Figure 6.6 shows the structure of the hierarchical rule-based expert FIS with 18 input nodes, 8 intermediate nodes and an output node. The input nodes represent the independent variables used in developing the model, whereas the output node represents the dependent variable Fe and Mn accumulation potential. Information flows from left to right; from the input nodes, through the intermediate nodes, to the output node. The hierarchical rule-based expert FIS is made up of 9 subsystems. Since the model is hierarchical, an output node in an outer subsystem becomes an input node in an inner subsystem. Information flows smoothly from one subsystem to another until a crisp value of the predicted Fe and Mn accumulation potential is computed at the output node.

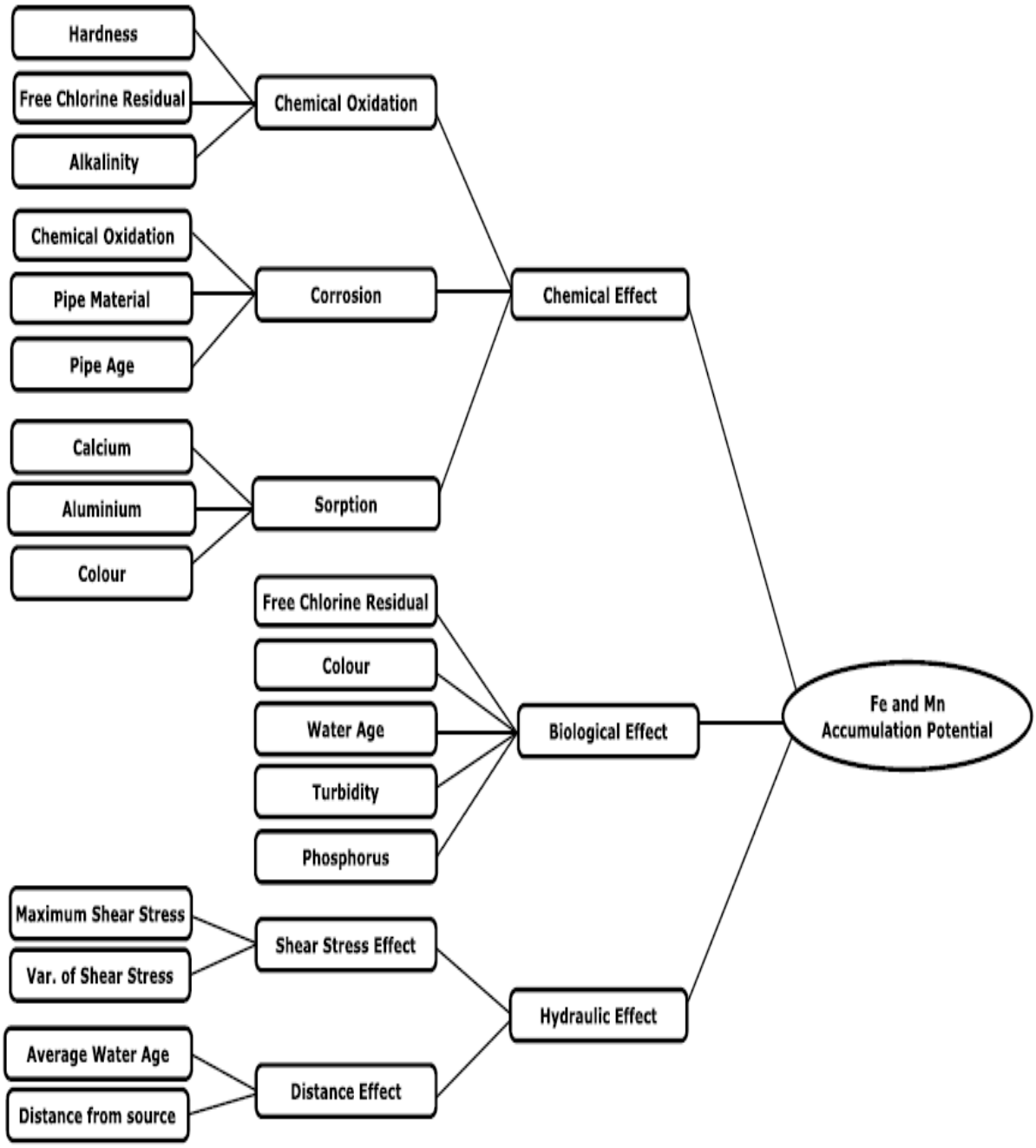


Figure 6.6 Structure of the hierarchical rule-based expert FIS

6.4 Model development of the hierarchical data-driven FIS

Although the rules of the hierarchical rule-based expert FIS can be easily formulated using expert knowledge, they may not give very accurate predictions for a number of reasons. First, the system rules may vary slightly for every WSZ. Thus, the influence of some variables that contribute to Fe and Mn accumulation may vary slightly for every WSZ. Secondly, the narrow range of the knowledge-base in most expert systems makes them give poor predictions in new situations outside this range. Finally, they are unable to learn

from data and adapt to new instances. In view of these drawbacks, a hierarchical FIS that uses data-driven approach to optimise the rules was developed. It has the same structure as the hierarchical rule-based expert FIS. However, the fuzzy rules and weights were generated automatically using a genetic algorithm. Data from five WSZs were used to optimise the rules and weights of the rules. The modelling data from each of the five WSZs were randomly divided into two sets; 80% of the data from each WSZ was used to train the FIS and the remaining 20% for testing. Using the combined measured data from all the 5 WSZs, another model was developed. Similarly, the combined data for all the five WSZs were randomly divided into training (80%) and testing data set (20%). The following sections show how the genetic algorithm was used to optimise the rules and weights of the hierarchical data-driven FIS.

6.4.1 Genetic Algorithm

Genetic algorithm is a type of evolutionary algorithm that mimics the natural selection process for a solution from a number of possible solutions. It is a search heuristic that tries to imitate the processes observed in natural evolution. Other types of evolutionary algorithm include evolution strategy (ES), evolutionary programming (EP), genetic programming (GP) and learning classifier system (LCS). Genetic algorithm is normally used in optimising solutions to problems. Optimisation is the process of searching for the best solution to a problem out of various possible solutions. Occasionally, some complex problems may have only one solution.

There are five main steps in the genetic algorithm process. These include the initialisation of the population, evaluation of the fitness function, selection of parents, cross-over, and mutation. It is an iterative process that starts with a randomly generated population. In this research, the population size was set to 78; representing the number of rules in the hierarchical data-driven FIS. The population of each iteration is known as a generation. The FIS was allowed up to 20,000 generations to optimise the rules and weights of the rules. During each iteration, the fitness of every member in the population is evaluated using Eqn. 6.3. Worst members are eliminated due to the low fitness value. As in the real-world genetic process, only the fit members are randomly selected from the current population to cross-over. The heuristic crossover method explained in Section 3.4.2.5 was used in this research (see Eqn. 3.47). The crossover ratio parameter (C_{Ratio}), which specifies how far the offspring is from the parent with a better fitness value, was set to its

default value of 0.2. Some of the offspring are randomly selected for mutation. Mutation is a very important step in the genetic algorithm process because when omitted, diversity will not be introduced into the population and would get stuck at a local minimum. During mutation, specific parts of the genetic materials of the randomly selected individual subjects are regenerated to replace lost genetic material. The Gaussian mutation method explained in Section 3.4.2.6 was used in optimising the rules and weights of the rules in this research (see Eqns. 3.48 and 3.49). The process is terminated when the satisfactory fitness level is attained, stall generation limit is exceeded, or when the maximum number of generation is reached. The stall generation limit is the stopping criterion used to stop the optimisation. If there is no improvement in the best fitness value for a number of generations, the algorithm is terminated. The stall generation limit was set to 50 in this research. Figure 6.8 shows the flow chart of the genetic algorithm that was used to optimise the rules of the hierarchical data-driven FIS.

6.4.2 Optimising the fuzzy rules

Water distribution models are more useful if they are optimised to replicate the hydraulic characteristics of real water systems. During the optimisation process, the model parameters are modified until the predicted output from the model matches the observed data. Although optimising a model helps to improve its prediction accuracy, it is plagued with a number of difficulties which includes potential uncertainties in the parameters and the need for considerable large computational resources if the input parameters are many.

Because the factors that influence Fe and Mn accumulation differ slightly in every network, using genetic algorithm to optimise the consequent part of the rules for each WSZ may help to improve the model's prediction accuracy. It is an evolutionary approach to optimisation which serves as an alternate method to the traditional methods of optimisation. This approach is data driven. Hence, it relies heavily on the modelling data to generate the optimised rules. In principle, it uses the crossover (mating) of solutions to produce new generation of solutions; which in simple terms means selecting the best solution to a problem from various possible solutions. It has the ability to solve problems that are non-parametric, multi-dimensional, and non-differentiable.

Using genetic algorithm to tune the FIS rules has several advantages. If there is enough data for the optimisation, the optimised rules generated will be a better representation of

the FIS than the rules from expert knowledge. Due to the optimisation capabilities of genetic algorithms, it is normally used to search for optimums; either global maximum or minima. Furthermore, it is able to give multiple solutions to a problem.

Initially, the rules and their corresponding weights were simultaneously optimised with a genetic algorithm. However, this gave poor model performance because the combined rules and weights increased the subjects of the population. This made it difficult for the genetic algorithm to find feasible solutions to the problem. In view of this, the optimisation was performed in two phases. A genetic algorithm was used to optimise the consequent part of the rules for the fuzzy model developed. Afterwards, another genetic algorithm was used to optimise the weights of the fuzzy rules. The source code of the genetic algorithm for optimising the rules is presented in Appendix J, whereas the flow chat for the genetic algorithm for optimising the rules is presented in Fig. 6.7. The MSE was used as the objective function (see Eqn. 6.3). The smaller the MSE value, the better the predictive power of the model. The optimisation graph for WSZ2 after 20,000 generations is presented in Fig. 6.8. Similar optimisation graphs for the remaining WSZs are presented in Appendix S. It was observed that if the fitness function reduced from 0.0795 to 0.0301. This indicates that the genetic algorithm was able to improve the performance of the model through optimisation of the rules. Results from the remaining WSZs showing how fitness function values were reduced by the genetic algorithm for optimising the rules are presented in Table 6.4.

$$Objective\ function = MSE = \frac{1}{sp}\sum_{i=1}^{sp}(Y_i - X_i)^2 \qquad (6.3)$$

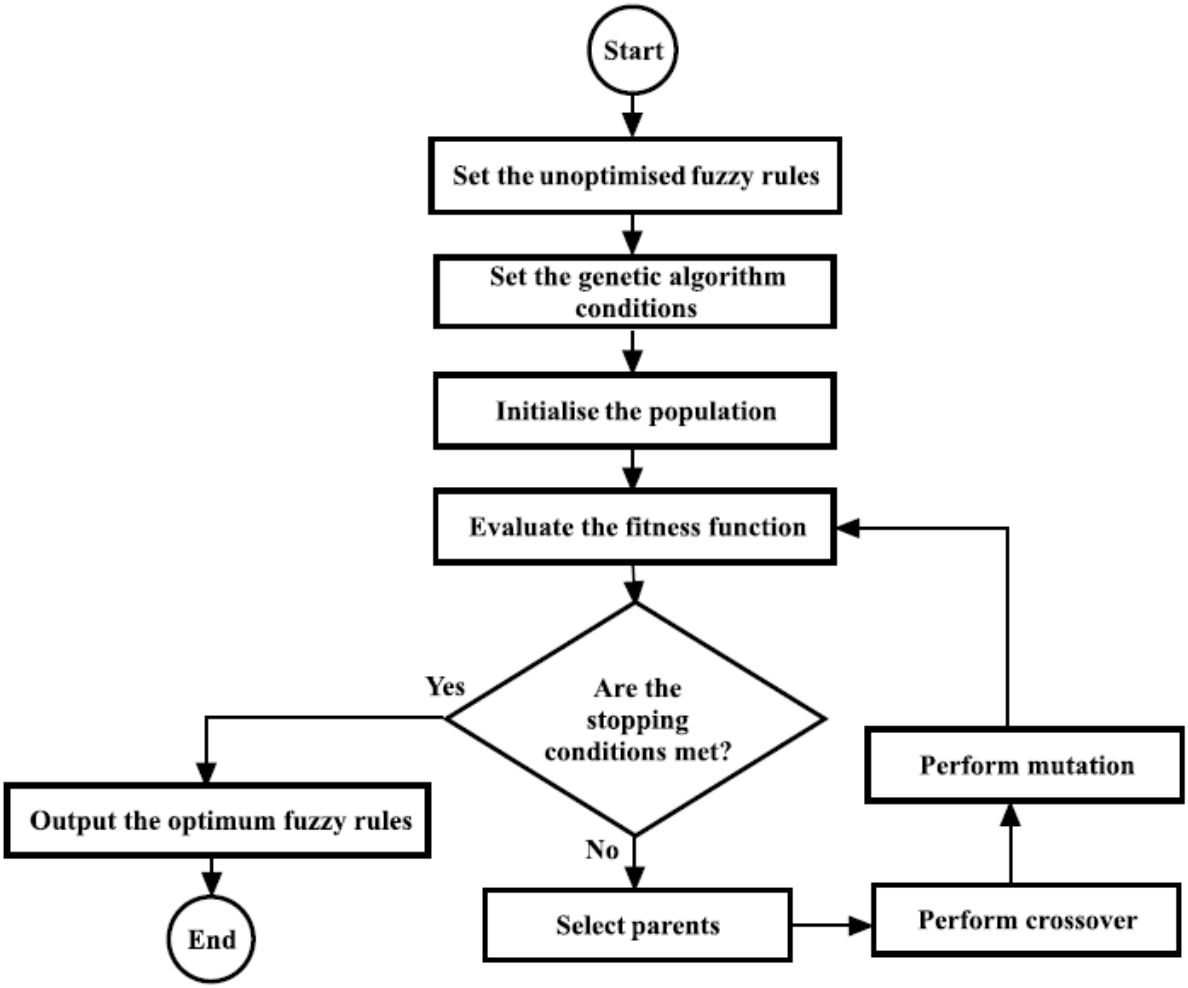


Figure 6.7 Flow chat for the genetic algorithm for optimising the rules

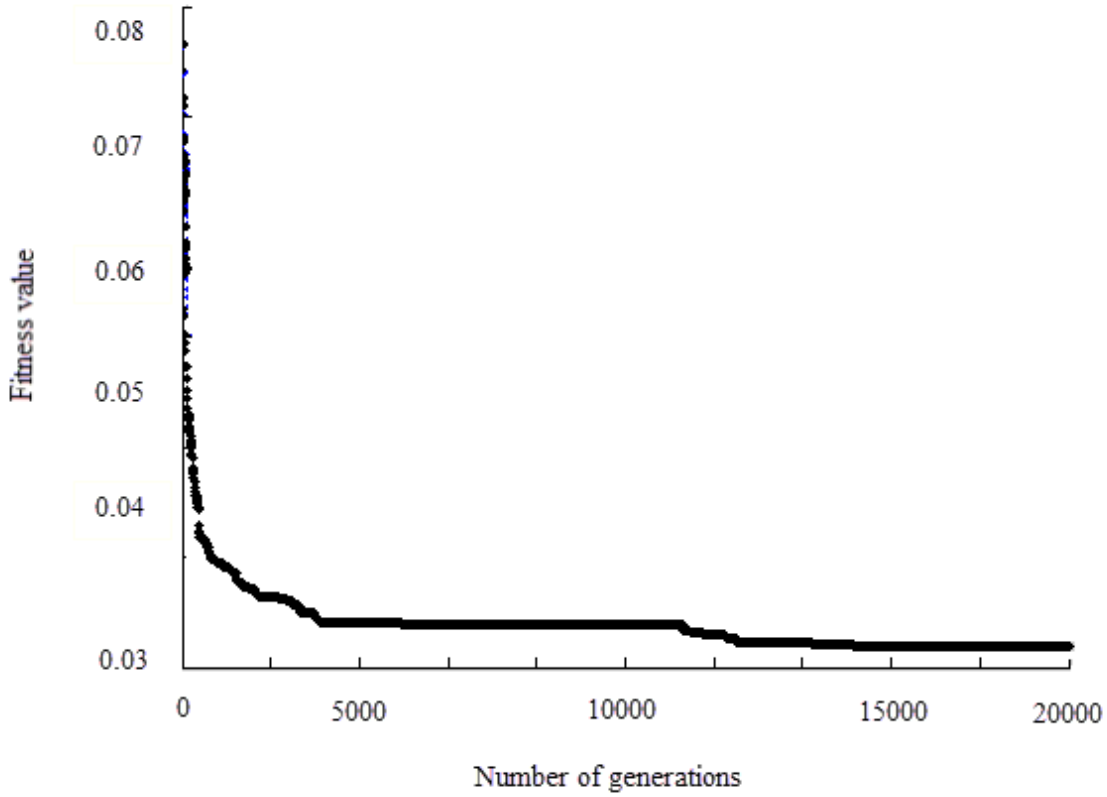


Figure 6.8 Fitness function graph for WSZ2 during rule optimisation

6.4.3 Optimising the fuzzy weights

The FIS was updated with the optimised rules and were initially given equal weights for all the rules. Another genetic algorithm was used to optimise the weights of the rules. The genetic algorithm parameters and stopping conditions were defined. Again, MSE is the objective function that was used to evaluate the fitness of individual subjects after every generation. The algorithm was run for 20,000 generations with the aim of achieving a better fitness through the generations. The optimisation graph for WSZ3 after 20,000 generations is presented in Fig. 6.9. It was observed that the fitness function reduced from 0.0532 to 0.0382. Results from the remaining WSZs showing how fitness function values were reduced by the genetic algorithm for optimising the weights of the rules are presented in Table 6.4. The algorithm is terminated when the stopping conditions are met. The flow chat of the genetic algorithm for optimising the weights of the rules is presented in Fig. 6.10.

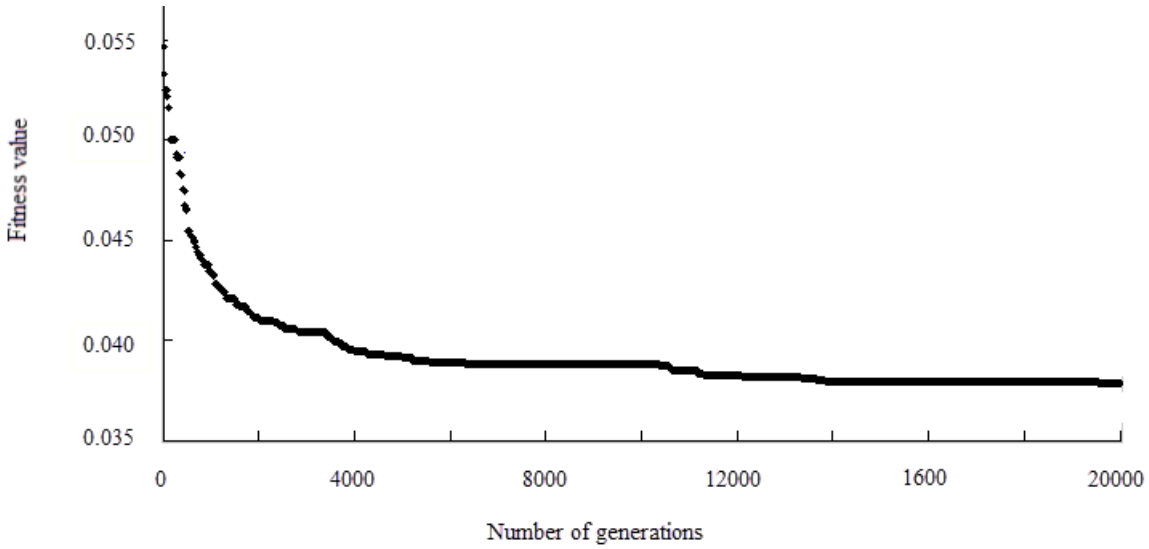


Figure 6.9 Fitness function graph for WSZ3 during weight optimisation

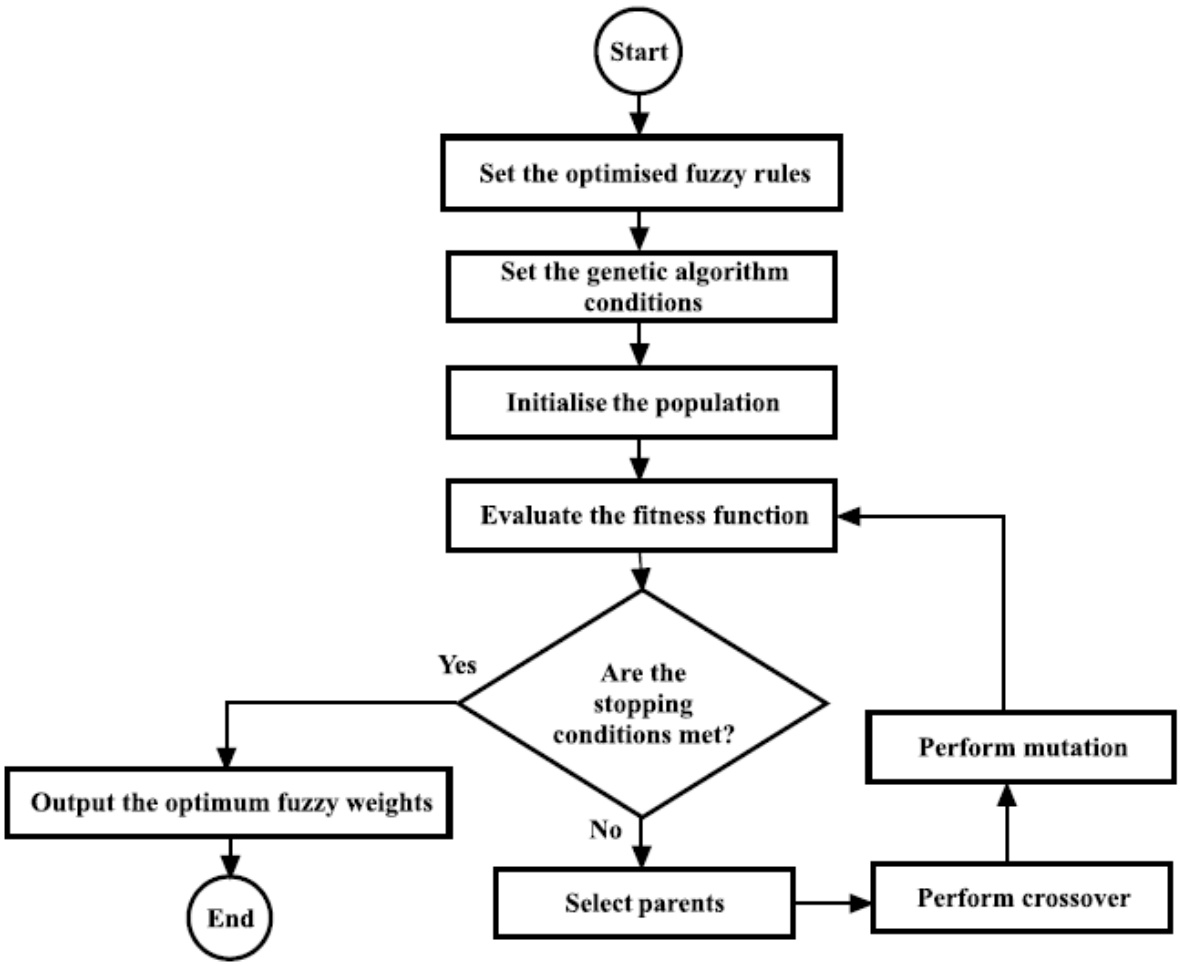


Figure 6.10 Flow chat for the genetic algorithm for optimising the weights

6.5 Results and discussion of the hierarchical rule-based expert FIS

The ANN(t,ψ) model developed in Chapter 5 was able to predict Fe and Mn accumulation potential at every node in a given WSZ. However, the black-box nature of ANNs made it difficult to access and evaluates the effect of each hidden node on Fe and Mn accumulation potential. Therefore, the causes of Fe and Mn failures were manually investigated. Due to the enormous sizes of WSZs, manually investigating the Fe and Mn failures became time-consuming and very laborious. FISs, which are regarded as white-box in nature because their intermediate nodes can be accessed and their effect on the output variables can be evaluated, was developed to overcome this limitation. The hierarchical rule-based expert FIS developed was able to indicate which intermediate nodes and input variables cause high-risk of Fe and Mn accumulation potential without manual investigation.

6.5.1 Performance of the hierarchical rule-based expert FIS

Six hierarchical rule-based expert FISs were developed; five of them used their respective WSZs data sets for the modelling, whereas the sixth FIS used the combined data sets of all the five WSZs for the modelling. CA and MSE were used as performance indicators for the evaluation of the models. Models with high CA have better predictions than those with low CA. On the other hand, models with smaller MSE have better prediction accuracy than those with larger MSE. To evaluate the performance of the FISs for each range of classified risk, the CA of each class correctly predicted were computed as an additional evaluation. This evaluation was done to ensure the FISs developed were not exhibiting accuracy paradox; which are models with lower level of accuracy that appear to have better predictive powers than models with higher accuracy (Valverde-Albacete, 2014). Table 6.3 shows the performance of the six hierarchical rule-based expert FISs. It was observed that all the six FISs had low CA. They were also unable to predict the high-risk values of Fe and Mn accumulation potential very well. The high-risk values of Fe and Mn accumulation potential correctly predicted ranged from 0–16.28%. They also had relatively high MSE.

Table 6.3 Performance of the six hierarchical rule-based expert FISs

Performance indicator	WSZ1	WSZ2	WSZ3	WSZ4	WSZ5	WSZAll
Overall CA (%)	51.64	54.68	58.33	48.84	55.10	56.07
CA - low (%)	66.49	75.15	67.15	57.65	62.13	71.74
CA - medium (%)	23.44	21.43	23.08	43.55	13.33	16.99
CA - high (%)	15.38	16.67	0.00	16.28	0.00	11.11
Mean square error	0.0782	0.0836	0.0586	0.0724	0.0804	0.1585
Sample size	275	267	168	301	314	1327

Figure 6.11 shows the confusion matrix generated by the hierarchical rule-based expert FIS for WSZ2. The left diagonal cells of the confusion matrix (highlighted in green) are the correctly predicted values from the FIS. It correctly predicted only 7 out of 42 (16.67 %) high-risk values, 12 out of 56 (21.43 %) medium-risk values and 31 out of 169 (75.21 %) low-risk values. The FIS for WSZ2 has a poor prediction power because it has an overall classification accuracy of 54.68% and was able to predict only 16.67% of the high classified values of Fe and Mn accumulation potential.

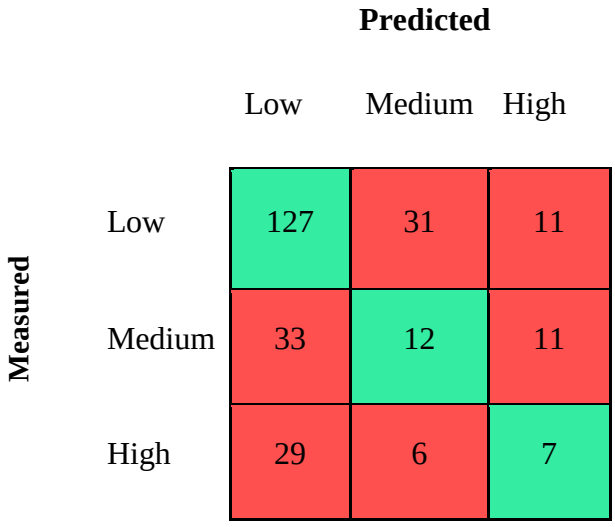


Figure 6.11 Confusion matrix generated by the hierarchical rule-based expert FIS for WSZ2

There are three main reasons why the models may have given poor performances. First, as indicated in Section 6.4, Fe and Mn accumulation are formed under slightly different conditions for each WSZ. Therefore, the use of the same expert system rules for each of

the WSZs was not an accurate representation of the FISs. Secondly, some of the weights assigned to the rules to handle more or less influential variables that influence Fe and Mn accumulation potential may not have been correct. This is because it is very difficult to find which variable will be highly influential just by inspection of the data. Moreover, the highly influential variables vary slightly for each WSZ making it difficult to assign the correct weights for the rules. Finally, since the rules formed did not learn from the data, they are unable to adapt to new instances. The first 15 rules and their corresponding weights from the hierarchical rule-based expert FIS for WSZ2 are presented in Table 6.4. All the 78 rules with their corresponding weights for WSZ2 and that for the remaining WSZs are presented in Appendix S.

Table 6.4 The first 15 rules and their corresponding weights from the hierarchical rule-based expert FIS for WSZ2

Rule Number	Rules from expert knowledge	Weights from expert knowledge
1	If Hardness is LOW then Chemical oxidation is LOW	0.9
2	If Hardness is MEDIUM then Chemical oxidation is MEDIUM	0.9
3	If Hardness is HIGH then Chemical oxidation is HIGH	0.9
4	If FCR is LOW then Chemical oxidation is LOW	0.9
5	If FCR is MEDIUM then Chemical oxidation is MEDIUM	0.9
6	If FCR is HIGH then Chemical oxidation is HIGH	0.9
7	If Alkalinity is LOW then Chemical oxidation is HIGH	0.9
8	If Alkalinity is MEDIUM then Chemical oxidation is MEDIUM	0.9
9	If Alkalinity is HIGH then Chemical oxidation is LOW	0.9
10	If Chemical oxidation is LOW then Corrosion is LOW	1.0
11	If Chemical oxidation is MEDIUM then Corrosion is MEDIUM	1.0
12	If Chemical oxidation is HIGH then Corrosion is HIGH	1.0
13	If Pipe material index is LOW then Corrosion is LOW	1.0
14	If Pipe material index is MEDIUM then Corrosion is MEDIUM	1.0
15	If Pipe material index is HIGH then Corrosion is HIGH	1.0

6.6 Results and discussion of the hierarchical data-driven FIS

A major problem in the development of FISs is how to correctly define the fuzzy rules and assign appropriate weights to them in order to make good predictions. As can be seen from

the hierarchical rule-based expert FIS results in Section 6.5.1, when solving complex problems with many input variables, generalising the rules for every WSZ resulted in poor predictions. It is known that using expert knowledge to form fuzzy rules with many input variables can be very difficult and complicated (Babusk, 1998). Hence, the data-driven approach which automates the generation of the fuzzy rules is often preferred. The following sections show the performance of the hierarchical data-driven FIS developed which used genetic algorithm to optimise the fuzzy rules and weights.

6.6.1 Performance of the hierarchical data-driven FIS

Again, six models were developed this time using hierarchical data-driven FIS. Five of the FISs used their respective WSZs data sets for the modelling, whereas the sixth used the combined data from all the WSZs. The performance indicators CA and MSE were used in evaluating the models. Table 6.5 shows the performance of the six hierarchical data-driven FISs. The MSE of the FISs were recorded during the first generation of the rules optimisation and after the rule optimisation was completed. It was observed that for all the six FISs, the optimisation was able to successfully improve the performance. Likewise, it was observed that the performance of all the FISs further improved after the weights assigned to the fuzzy rules were also optimised. For instance the MSE of the FIS for WSZ2 reduced from 0.0795 to 0.0263 after the rules and weights optimisation. This is an indication of a good model performance. Results of the remaining WSZs showing how fitness function values were reduced by the genetic algorithm are presented in Table 6.5.

From Table 6.5, it was observed that the overall CA on the testing data set gave better predictions than the hierarchical rule-based expert FIS shown in Table 6.3. This may be due to the slight variation of the factors that influence Fe and Mn accumulation potential for every WSZ; which makes generalisation of the rules not a good representation of the hierarchical rule-based expert FIS. It was also observed that the FIS for the combined data set gave relatively poor prediction. As explained in Chapter 5, this could be due to not having enough instances of data to represent the entire search space for the combined five water supply zones. It could also be due to the fact that Fe and Mn accumulation are formed under slightly different conditions for each WSZ. Therefore, combining the data sets resulted in having too many sources of water supply which confused the training process and subsequently gave relatively poor predictions. With the exception of the FISs

for WSZ1 and WSZ3 which gave poor predictions of high-risk Fe and Mn accumulation potential, the FISs for the three remaining WSZs gave relatively good predictions. The FISs for WSZ1 and WSZ3 could correctly classify only 40 and 33.33% respectively of the high-risk values of Fe and Mn accumulation potential on the testing data set. This was due to too many sources of water supplied to these two WSZs as explained in Section 5.6.1. The solution to improving these poor predictions is presented in Section 6.6.2.

Table 6.5 Performance of the six hierarchical data-driven FISs

Performance indicator	WSZ1	WSZ2	WSZ3	WSZ4	WSZ5	WSZAll
MSE after first generation	0.0832	0.0795	0.0780	0.0813	0.0766	0.0784
MSE after rules optimisation	0.0624	0.0301	0.0532	0.0589	0.0639	0.0542
MSE after weights optimisation	0.0494	0.0263	0.0382	0.0439	0.4907	0.0475
Overall training CA (%)	64.22	69.91	66.20	66.41	76.40	68.21
Overall testing CA (%)	65.12	68.29	61.54	60.78	68.09	62.19
Testing CA - low (%)	73.33	83.33	65.00	76.92	75.00	77.78
Testing CA - medium (%)	50.00	22.22	66.67	36.67	55.00	15.63
Testing CA - high (%)	40.00	75.00	33.33	58.00	61.00	32.00
Training sample size	232	226	142	256	267	940
Testing sample size	43	41	26	45	47	201

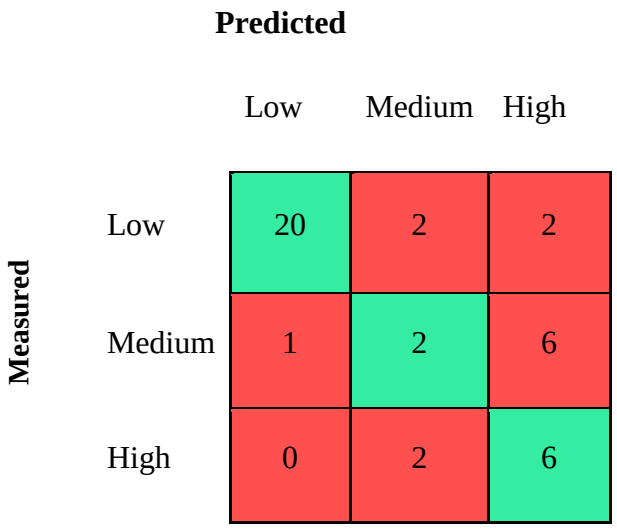


Figure 6.12 Testing data confusion matrix after predictions from the hierarchical data-driven FIS for WSZ2

Figure 6.12 shows the testing data confusion matrix after predictions from the hierarchical data-driven FIS for WSZ2. The model was able to correctly predict 83.33%, 22.22% and

75% of its low-, medium- and high-risk values, respectively. The overall classification accuracy on the testing data set of 68.29% indicates that the hierarchical data-driven FIS for WSZ2 is a good model which will make good predictions on new data sets.

The hierarchical data-driven FIS gave better results than the hierarchical rule-based expert FIS because the genetic algorithm was able to optimise the rules and weights of the rules. The first 15 rules and their corresponding weights from the hierarchical data-driven FIS for WSZ2 are presented in Table 6.6. All the 78 rules with their corresponding weights for WSZ2 and that for the remaining WSZs are presented in Appendix S.

Table 6.6 The first 15 rules and their corresponding weights from the hierarchical data-driven FIS for WSZ2

Rule Number	Rules after optimisation	Weights after optimisation
1	If Hardness is LOW then Chemical oxidation is MEDIUM	0.4107
2	If Hardness is MEDIUM then Chemical oxidation is LOW	0.3711
3	If Hardness is HIGH then Chemical oxidation is LOW	0.7952
4	If FCR is LOW then Chemical oxidation is HIGH	0.5760
5	If FCR is MEDIUM then Chemical oxidation is MEDIUM	0.5613
6	If FCR is HIGH then Chemical oxidation is MEDIUM	0.3699
7	If Alkalinity is LOW then Chemical oxidation is HIGH	0.5656
8	If Alkalinity is MEDIUM then Chemical oxidation is HIGH	0.4491
9	If Alkalinity is HIGH then Chemical oxidation is MEDIUM	0.3667
10	If Chemical oxidation is LOW then Corrosion is MEDIUM	0.4668
11	If Chemical oxidation is MEDIUM then Corrosion is MEDIUM	0.2811
12	If Chemical oxidation is HIGH then Corrosion is LOW	0.7013
13	If Pipe material index is LOW then Corrosion is MEDIUM	0.5214
14	If Pipe material index is MEDIUM then Corrosion is LOW	0.2452
15	If Pipe material index is HIGH then Corrosion is MEDIUM	0.4890

6.6.2 Improving the performance of the hierarchical data-driven FIS

As mentioned in Section 5.4, In order to predict Fe and Mn accumulation potential for every node, there is the need to have a base data set which consists of measured data for all the nodes. However, due to the enormous sizes of WSZs, it was impossible to have measured data of water quality variables for every node. It was observed that majority of the water quality variable values had small standard deviations. It was therefore assumed that at any given time, concentrations of chemical variables and variables that influence

biological processes in a given DMA were approximately the same. Yearly average water quality variables value at all the nodes in a given DMA were subsequently assumed to be approximately the same. However, there were a few water quality variables in WSZ1 and WSZ3 that had high standard deviations because there were too many sources of water supply to these WSZs. This means that the water quality variables from these two WSZs may not have been well represented using this assumption. This resulted in poor performance from the FIS for WSZ1 and WSZ3.

Table 6.7 Performance of the hierarchical data-driven FIS using water quality variables estimates from the multiple linear regression models

Performance indicator	**WSZ1**	**WSZ3**
MSE before rules optimisation	0.0826	0.0766
MSE after rules optimisation	0.0639	0.0498
MSE after weights optimisation	0.0436	0.0364
Overall training CA (%)	70.95	65.49
Overall testing CA (%)	69.77	65.38
Testing CA - low (%)	75.00	69.42
Testing CA - medium (%)	62.50	66.67
Testing CA - high (%)	63.25	59.33
Training sample size	232	142
Testing sample size	43	26

In view of the poor performance, the results obtained from the multiple linear regression model developed in Section 5.6.1 to predict the measured water quality variables values of every node in each DMA for WSZ1 and WSZ2 was used to develop the FISs in order to capture the variations. The measured water quality variables predicted by the multiple linear regression model, which were used as new input water quality variables for the FISs, were transformed between zero and one using Eqn. 5.1 in Chapter 5. Table 6.7 shows the performance of the hierarchical data-driven FISs when the input water quality variables estimates from the multiple linear regression model was used in the modelling. It was observed that the predicted percentage of high classified values of Fe and Mn accumulation potential from the testing data in WSZ1 improved from 40% to 63.25%. The MSE, after the rules and weights optimisation, improved from 0.0494 to 0.0436. Also, the overall CA on the testing data set improved from 65.12 to 69.77. Similarly, the predicted percentage of high classified values of Fe and Mn accumulation potential from the testing data in WSZ3 improved from 33.33% to 59.33%. Both the MSE and CA on the testing

data set in WSZ3 also improved. This indicates that the models are able to give better predictions in WSZs with too many sources when the input water quality variables for every node are estimated using multiple linear regression than assuming that yearly average water quality variable values at every node within each of the DMAs were approximately the same.

6.6.3 Risk index of the hierarchical data-driven FIS

Drinking water companies have a duty to routinely sample a number of water quality variables which includes Fe and Mn. The tests are mainly done at the treatment plants, service reservoirs, and customer taps. The results of these tests are electronically transferred to DWI monthly. Drinking water companies are required to send annual monitoring programme to DWI. They are also required to send data of customer complaints due to drinking water discolouration to the Ofwat. There are appropriate sanctions in place by these water regulatory authorities to penalise drinking water companies if they fail to comply with regulations. Presently, most water companies identify high discolouration risk regions in water distribution networks (WDNs) by either selecting areas in the network with high Fe and Mn concentrations from their routine sampling or using customer complaints data due to discolouration. However, as indicated in Section 5.6.3, these risk assessment methods are imprecise because only selected few nodes are sampled and not all customers that experience water discolouration complain. Hence, there is a high likelihood that certain regions in WSZs with high customer complaints or Fe and Mn concentrations can go undetected.

To overcome the above-mentioned limitations, a risk index that uses the predicted Fe and Mn accumulation potential by the FIS at every node was developed to quantify the various levels of risk. The risk levels defined in Chapter 5 were used to develop the FISs. If more than 10% of all the predicted Fe and Mn accumulation potential by the model are high in a given WSZ, that WSZ is classified as a high-risk WSZ. If the predicted high values by the model in a WSZ are between 5 and 10% of all the model's predictions, it was classified as medium-risk WSZ. WSZs with less than 5% of all the model's predictions which are high were classified as low-risk WSZ. The risk levels of the five WSZs between the year 2005 and 2009 generated by the hierarchical data-driven FIS are presented in Table 6.8. It was observed that there were variations in risk levels for each of the WSZs. As indicated in Section 5.6.2, these variations could be as a result of months of accumulation of Fe and

Mn particles on the pipe walls of WDNs or network cleaning through flushing, to remove accumulated sediments.

Table 6.8 Risk levels of the WSZs between the year 2005 and 2009 generated by the hierarchical data-driven FIS

WSZ\Year	2005	2006	2007	2008	2009
WSZ1	Medium	Medium	Medium	Low	Medium
WSZ2	High	Medium	High	High	High
WSZ3	Medium	Medium	High	Low	Medium
WSZ4	High	Medium	High	High	High
WSZ5	Low	Low	Low	Medium	High

Table 6.9 Customer complaints levels of the five WSZs from 2005 to 2009

WSZ\Year	2005	2006	2007	2008	2009
WSZ1	Medium	Medium	Medium	Low	Medium
WSZ2	High	High	Medium	Medium	High
WSZ3	Low	Medium	Low	Low	Medium
WSZ4	Medium	High	High	High	High
WSZ5	Medium	Low	Low	Low	High

A cumulative frequency curve for all customer complaints per 1000 properties from all the WSZs was plotted. The 90^{th} and 70^{th} percentile of the customer complaints per 1000 properties corresponded to 2.5 and 1.2, respectively. Using these percentiles, WSZs with customer complaints per 1000 properties greater than 2.5 were classified as high, between 2.5 and 1.2 as medium, and below 1.2 as low. Table 6.7 shows the customer complaints levels of the five WSZs from 2005 to 2009. Comparing Table 6.8 with Table 6.9, it was observed that most WSZs with high customer complaints also had high-risk levels of Fe and Mn concentrations potential predicted by the FIS. However, it was observed that there were a few number of years the high-risk levels predicted by the FIS did not match high customer complaints in some WSZs. There are a number of reasons for this disparity. First, it should be noted that the aim of this research is not to predict drinking water discolouration, but to predict Fe and Mn accumulation potential. In view of this, only variables that influence Fe and Mn accumulation were included in the FIS. Although increased Fe and Mn concentrations (accumulation) are the main causes of drinking water discolouration, there are other factors that also cause water to discolour. Hydraulic events such as opening of fire hydrants during flushing operations or fire extinguishing exercises, and increase in flow due to pipe burst can all cause drinking water discolouration and

prompt customers to complain. Secondly, as indicated in Chapter 2, approximately 30% of customers that experienced discoloured water in the United Kingdom actually complain (Ewan & Williams, 1986). This explains why customer complaints can sometimes be ineffective in identifying high-risk regions in WSZs.

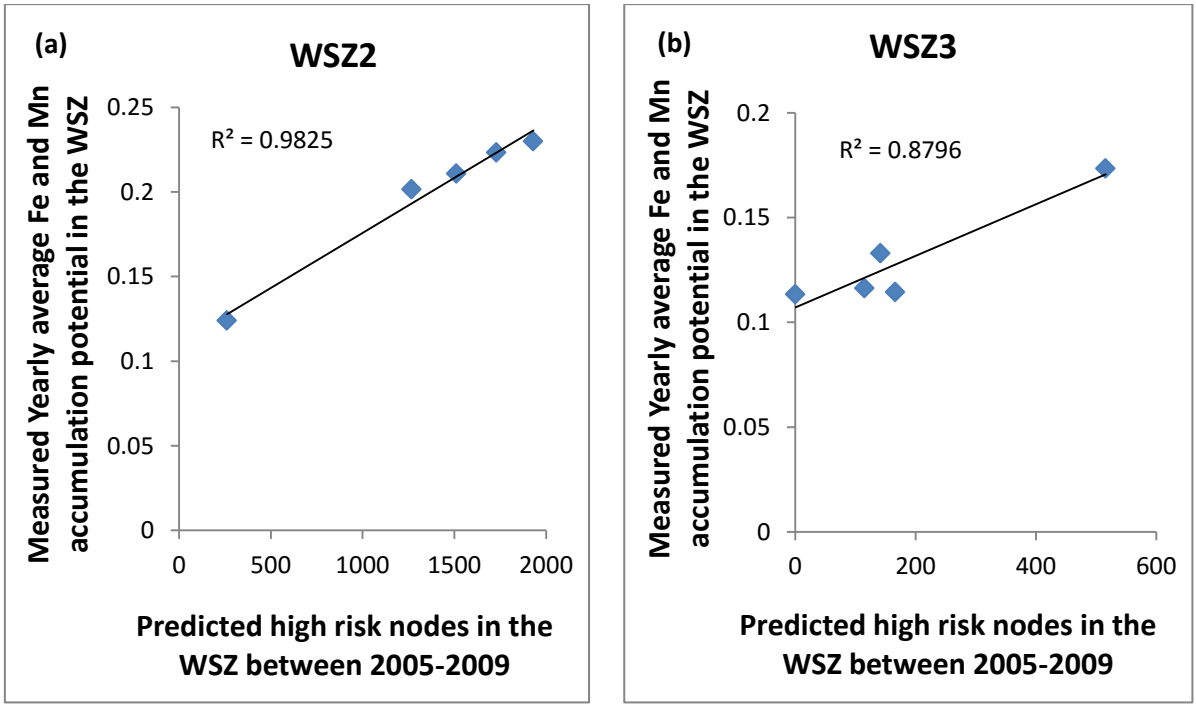


Figure 6.13 Correlation between measured yearly average Fe and Mn accumulation potential and predicted high-risk nodes from 2005-2009

To further investigate the performance of the hierarchical data-driven FISs, Graphs of measured yearly average Fe and Mn accumulation potential in each WSZ were plotted against the corresponding number of high-risk nodes predicted by the FISs between 2005 and 2009. Graphs of measured yearly average Fe and Mn accumulation potential plotted against number of high-risk nodes predicted by the FISs for WSZ2 and WSZ3 had an R^2 of 0.98 and 0.88 respectively (see Fig 6.13). Similar graphs from the remaining WSZs are presented in Appendix R. The strong positive correlations observed is an indication that the hierarchical data-driven FISs are predicting well.

6.6.4 Risk maps generated by hierarchical data-driven FIS

Due to the non-uniform distribution of risk levels of Fe and Mn accumulation potential in WSZs, narrowing the risk index from WSZ level to node level makes it easier to identify high-risk regions and investigate the causes of the failures. The ability of the hierarchical

data-driven FISs developed to generate risk maps which visually show all the various levels of Fe and Mn accumulation potential made this possible. In addition, it has the ability to backtrack from the output node, through the intermediate nodes, to the input nodes and automatically indicates which intermediate nodes and input variables cause high-risk of Fe and Mn accumulation potential.

Figure 6.14(a) shows a risk map of predicted Fe and Mn accumulation potential for WSZ2 in 2005 generated by the hierarchical data-driven FIS. While Fig. 6.14(b) shows a risk map of its corresponding measured Fe and Mn accumulation potential. Three service reservoirs (labelled A, B and C) supply WSZ2 with water. Comparing the measured and predicted risk maps, it was observed that most of the regions in the network with measured high-risk of Fe and Mn accumulation potential were also predicted as high-risk regions by the model. Similarly, most of the regions in the network with measured medium-risk of Fe and Mn accumulation potential were also predicted as medium-risk regions by the model. It was observed that DMA2-02 (highlighted by a black circle) which receives water from service reservoir A had low Fe and Mn accumulation potential. This was mainly due to biological oxidation of Fe and Mn. In 2005, the biological oxidation of Fe and Mn in that region was low as a result of very low phosphorus concentrations. As indicated in Section 6.3.1.4, phosphorus is a bioavailable form of nutrients that bacteria in WDNs need for growth and reproduction (CRCWQT, 2005). Therefore, low concentrations of it will help to reduce the growth or kill the bacteria responsible for oxidising Fe and Mn.

DMA2-16 (highlighted by a black rectangle) receives water from service reservoir A (see Fig. 6.14(a)). There were a number of factors that contributed to the high Fe and Mn accumulation potential at this DMA. Tracing from the output node to the intermediate nodes revealed that high values of the intermediate nodes; biological oxidation and hydraulic effect, were the main causes of the high-risk experienced in this region. Further backtracking from the hydraulic effect intermediate node to the input nodes showed that high values of the variables hydraulic distance from service reservoir A and average water age contributed to the high Fe and Mn accumulation potential observed at DMA2-16. As explained in Section 6.3.1, increase levels of these two variables increases Fe and Mn accumulation potential. Also, high levels of phosphorus, turbidity, and average water age all contributed to increased Fe and Mn accumulation potential at DMA2-16 as explained in section 6.3.1.

Service reservoir A in Fig. 6.14(a) supplies water to DMA2-13 (highlighted by a black square). It was observed that nodes in this DMA had very low Fe and Mn accumulation potential. The low-risk levels experienced in this region can be attributed to low intermediate values of hydraulic effect, biological oxidation, chemical oxidation, and corrosion. Tracing the intermediate nodes back to their input nodes, it was observed that the low values of hydraulic effect were as a result of low average water age and short hydraulic distance from service reservoir A. In general, it is known that low water age and short hydraulic distance from source of water supply reduces Fe and Mn accumulation potential. The short retention time of water under these conditions prevents disinfectants such as chlorine from dissipating, which helps to prevent microbial growth and eventually leads to the reduction of biological oxidation of Fe and Mn. Very low levels of colour were observed at DMA2-13 in 2005. It was the lowest level recorded from 2005 to 2009 for the entire WSZ2. Since carbon is the main bioavailable form of nutrients for the bacteria responsible for oxidising Fe and Mn, low concentrations of it reduces microbial growth and subsequently reduces biological oxidation of Fe and Mn (CRCWQT, 2005). The low corrosion and chemical oxidation levels observed were as result of high alkalinity concentrations in this region. As explained in Section 6.3.1, research has shown that there is a negative correlation between alkalinity and corrosion (Naylor et al., 1993).

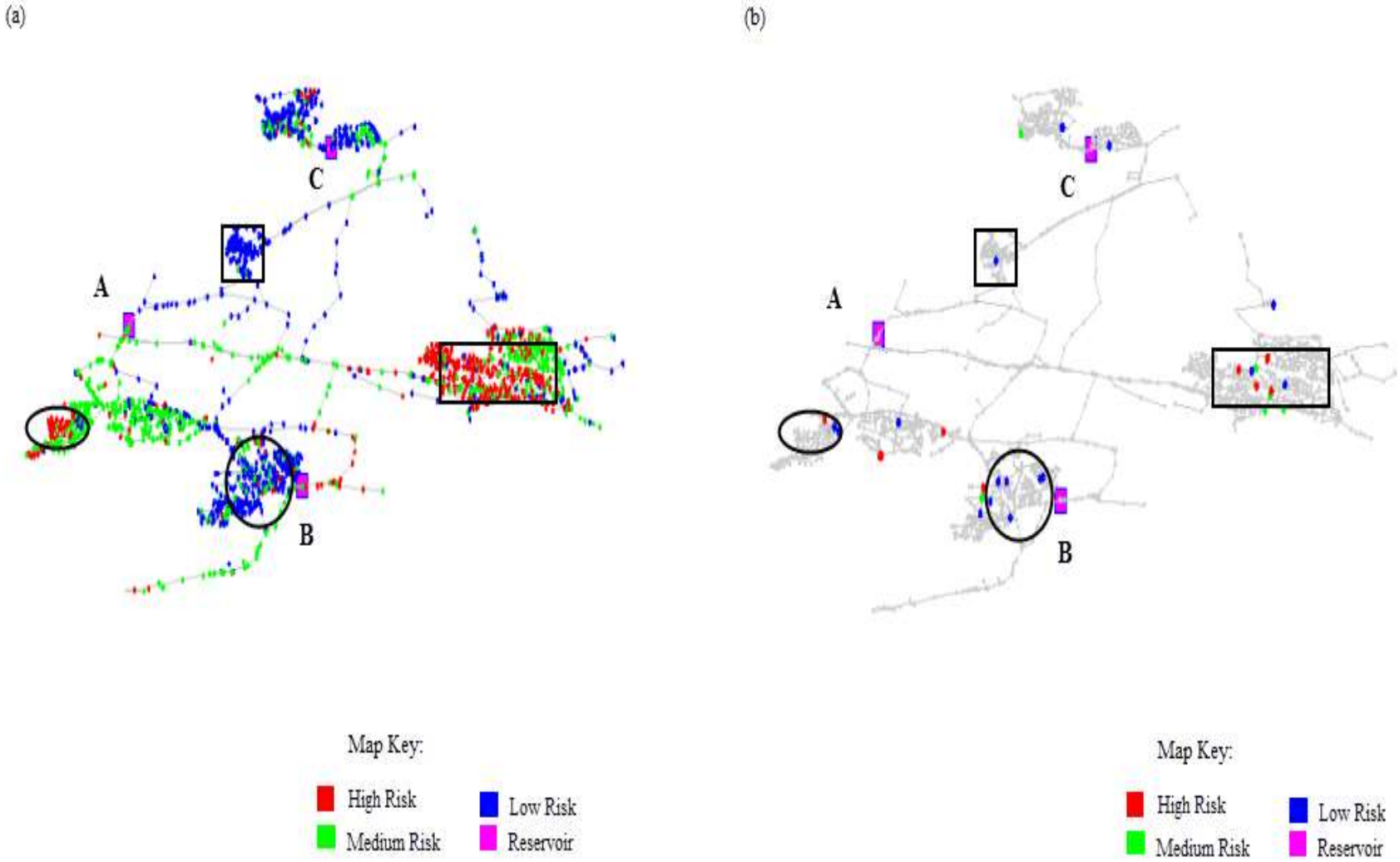


Figure 6.14 Hierarchical data-driven FIS risk maps showing (a) predicted and (b) measured Fe and Mn accumulation potential at WSZ2 in 2005

From Fig. 6.14 (a), the high predicted Fe and Mn accumulation potential values observed at DMA2-01 in 2005 (highlighted by a black oval) was as a result of high values of the intermediate nodes; biological oxidation and hydraulic effect. The high values of biological oxidation observed were as a result of high levels of average water age, colour, and turbidity. As already explained in Section 6.3.1, increase in average water age and colour increases biological oxidation. It is also known that increased levels of suspended organic and inorganic particles increase turbidity levels. This enhances biological oxidation by allowing Fe- and Mn-oxidising bacteria to attach themselves to these suspended particles. Furthermore, high levels of turbidity aid the biological oxidation process by serving as a shield to inhibit microorganisms from disinfection (WHO, 2011a). The increased hydraulic effect was as a result of low maximum daily shear stress and high average water age. Since DMA2-01 is at the periphery of WSZ2, it has many dead ends. These regions are more susceptible to accumulation of Fe and Mn particles on the pipe walls. This observation conforms to research by (Boxall et al., 2001) which suggested that discolouration materials are more likely to accumulate in networks that are less subjected to low conditioning daily shear stress than networks with high conditioning daily shear stress.

Comparing the predicted risk map with customer complaints risk map for WSZ2 in 2005, it was observed that most of the regions in the network with high predicted Fe and Mn accumulation potential also had high customer complaints (see Fig 6.15 (a) and (b)). A significant number of high-risk nodes predicted by the FIS in the region highlighted by a black rectangle also had high number of customer complaints. It was observed that a few number of high-risk nodes were predicted by the FIS in the regions highlighted with black circle and hexagon. Similarly, there were few number of customer complaints observed in the same regions. There were no high-risk nodes predicted by the FIS in the region highlighted with black square in 2005. Likewise, there were no customer complaints in the same region that year. The region highlighted with black oval had high customer complaints in 2005. However, the FIS predicted many medium-risk nodes and a few high-risk nodes in the region that year. The customers may have complained as a result of water discolouration from hydraulic events such as pipe burst or opening of fire hydrants during flushing exercises, which were not included in the FIS.

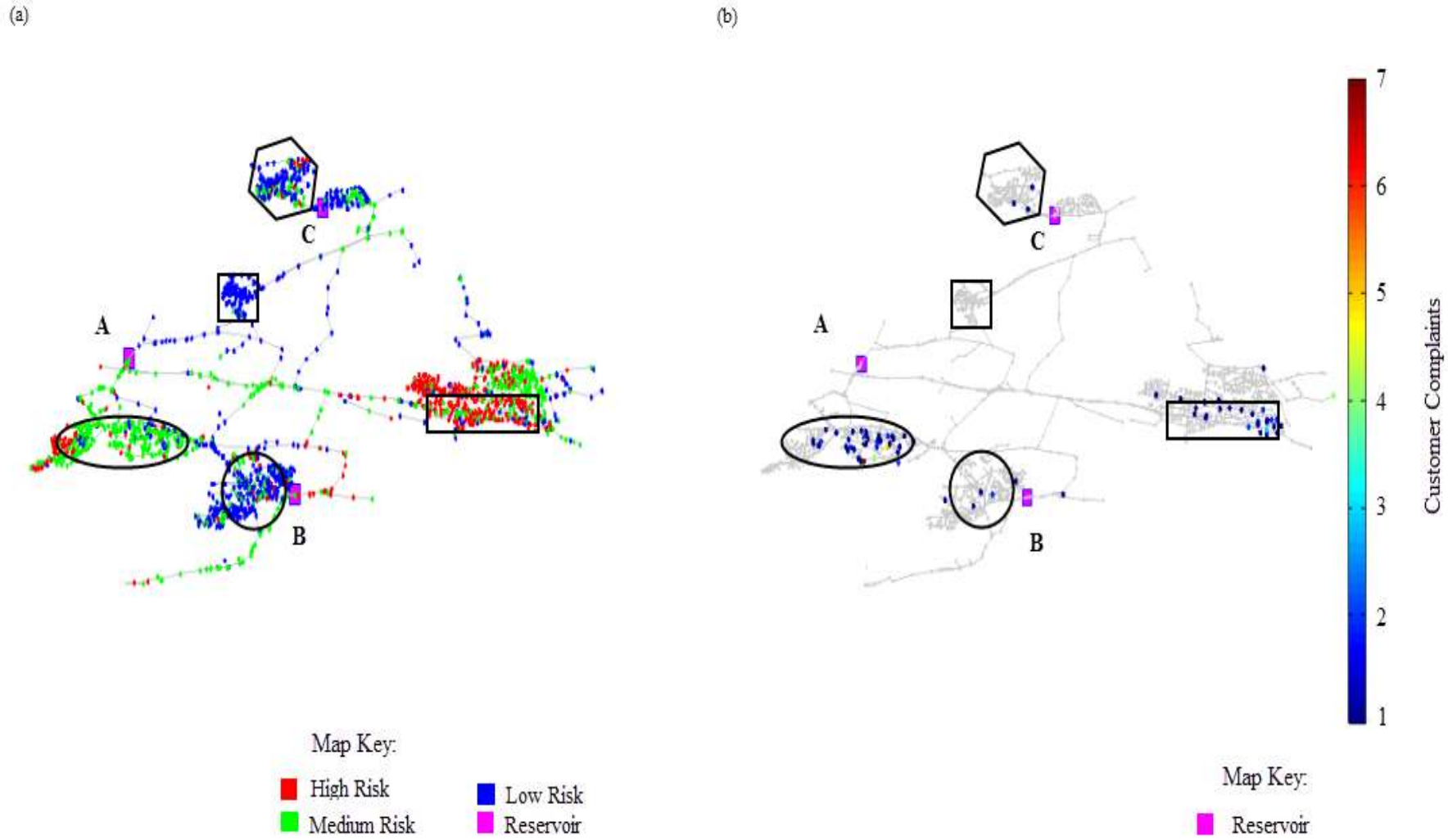


Figure 6.15 Hierarchical data-driven FIS risk maps showing (a) predicted Fe and Mn accumulation potential and (b) customer complaints for WSZ2 in 2005

6.7 Summary

In this chapter, two FISs were developed to predict Fe and Mn accumulation potential by using relevant chemical, biological, and physical/hydraulic variables. The first FIS developed, the hierarchical rule-based expert FIS, used expert knowledge to formulate rules and assigned weights to them in making its predictions. The hierarchical rule-based expert FIS gave relatively poor results because the same rules were used in the prediction for each of the WSZ. Since Fe and Mn accumulation are formed under slightly different conditions for each WSZ, using the same expert system rules and weights for all the WSZs may not be an accurate representation of the FISs.

The second FIS, the hierarchical data-driven FIS, was developed to overcome this limitations of the hierarchical rule-based expert FIS. It uses genetic algorithm to optimise its rules and weights. It was observed that the hierarchical data-driven FIS gave better predictions than the hierarchical rule-based expert FIS. A risk index that uses the predicted Fe and Mn accumulation potential by the FIS was developed to rank the risk levels of the WSZs. The developed hierarchical data-driven FIS was also able to determine the location of the high-risk regions by generating risk maps that predict Fe and Mn accumulation potential for every node in the WSZs. Comparing the predicted risk maps generated to the measured risk maps, it was observed that most regions with high predicted Fe and Mn accumulation potential also had high measured Fe and Mn accumulation potential. Similarly, most of the regions predicted by the FIS as having medium- and low-risk Fe and Mn accumulation potential also had measured medium- and low-risk Fe and Mn accumulation potential, respectively.

Comparing the predicted risk map with customer complaints risk map, it was observed that most of the regions in the network with high predicted Fe and Mn accumulation potential also had high customer complaints. However, a few regions did not follow this pattern because not all customers that experience water discolouration complained. This could also be due to hydraulic events such as opening of fire hydrants and pipe burst that can cause drinking water discolouration and prompt customers to complain, which were not used as variables in the FISs. Unlike the ANN models developed in Chapter 5, the white-box nature of FISs makes their intermediate nodes accessible for the evaluation. Hence, they are able to automatically indicate which intermediate nodes and input variables are

causes of high-risk of Fe and Mn accumulation potential. The developed hierarchical data-driven FIS could be of great benefit to water resource engineers and drinking water supply companies by using it as an important tool to identify high-risk regions and also explain the causes of the risk.

CHAPTER 7: Conclusions and recommendations

7.1 Conclusions

Although only small concentrations of Fe and Mn enter WDNs after water has been treated, years of accumulation of Fe and Mn particles and other adsorbed compounds associated with the accumulation process can cause drinking water discolouration. This may prompt customers to complain and lead to penalisation by Ofwat. A comprehensive literature review on drinking water discolouration models and the factors that influence the formation of drinking water discolouration showed that:

- Researchers have only studied each of the factors that influence Fe and Mn accumulation either partially or separately, but not in combination.
- The physical, chemical and/or biological processes that lead to formation of Fe and Mn accumulation/drinking water discolouration are very complex and interrelated. Hence, it is very difficult to use mathematical formulae or traditional models to solve such problems. AI-based methods of modelling such as ANNs and FISs are more appropriate to solve these complex problems because of their learning capabilities and ability to cope well with uncertainties in data.
- Most of the reviewed models mainly used physical/hydraulic variables in predicting drinking water discolouration. Hence, they could not capture all the factors that influence the formation of discoloured water in WDNs and therefore may not properly explain the processes and mechanisms that lead to Fe and Mn accumulation.
- The current practices by drinking water companies to identify regions with high-risk of discolouration or Fe and Mn failures includes (a) identifying regions in the network with high Fe and Mn concentrations and (b) identifying regions in the network with high number of customer complaints due to discolouration. However, these methods are ineffective because:
    - Not all customers who experience discolouration complain. In the UK, studies have shown that approximately 30% of customers that experience discoloured water event actually complain, whereas in Australia studies have shown that only 15% of customers who experience water discolouration complain. This means that high discolouration risk regions where customers do not complain may not be detected.

- The large sizes of WSZs make it impossible to sample all regions in the network. Hence, regions which have high Fe and Mn concentrations that are not sampled may not be detected.

The main observations and conclusions drawn from the analysis of the customer complaints, water quality and hydraulic/physical data are as follows:

- There were high number of customer complaints during the second and third quarters of the year. These spikes in complaints observed could be attributed to high water consumption during this period. Excess demand for water during this period increases flow velocity and shear stress, which causes accumulated Fe and Mn particles to dislodge from the pipe walls, and subsequently leads to water discolouration. The seasonal variations observed could also be due to high temperatures during this period. High temperatures promote bacterial growth, which cause biological oxidation of soluble Fe and Mn to their precipitate/insoluble form in WDNs. High temperatures are also known to expedite the chemical oxidation of Fe and Mn.
- Fe and Mn concentrations plotted against FCR concentrations for all 176 DMAs showed that when FCR concentrations were greater than 0.8 mg/L, neither Fe nor Mn exceeded their respective MCLs. This indicates that most of the oxidation that occurred within the distribution system may be microbial-induced, and that FCR concentrations above 0.8 mg/l were able to kill or reduce the growth of Fe- and Mn-oxidising bacteria. An optimum level of FCR is needed in the water distribution system to prevent the growth of microorganisms and preserve water quality.
- Fe and Mn concentrations plotted against maximum shear stress at nodes showed that areas with high maximum daily shear stress had low Fe and Mn concentrations. This is because Fe and Mn precipitates are unable to accumulate on pipe walls under high shear stress conditions. On the other hand, it was observed that regions with low daily maximum shear stress had high concentrations of Fe and Mn. Generally, low shear stress regions are subjected to high Fe and Mn accumulation because the shear stress exerted on the pipe walls in these regions are not high enough to dislodge any deposits of Fe and Mn particles. Low shear stress

also increases water age, reduces FCR, increases microbial growth, and subsequently leads to the deterioration of water quality.

- Fe and Mn concentrations were found to gradually increase with hydraulic distance from source of water supply. This is because as water travels through WDNs, water age generally increases and chlorine levels decrease. Since chlorine is a disinfectant, it supresses the growth of or kill Fe- and Mn-oxidising bacteria in regions with high hydraulic distance from source of water supply, preventing them to biologically oxidise soluble Fe and Mn to insoluble Fe and Mn. Conversely, regions with long hydraulic distance from source have low concentrations of FCR. Hence, such regions promote microbial growth, increase biological oxidation of Fe and Mn and subsequently increase Fe and Mn concentrations.

The main aim of this research was to use AI-based models to predict Fe and Mn accumulation potential with relevant biological, chemical and hydraulic/physical variables. Two ANN models were developed to overcome the limitations of the models and current methods used by drinking water companies to identify regions with high-risk of discolouration or Fe and Mn compliance failures. The first ANN model developed, ANN(t), was used as a sensitivity tool to select relevant input variables that influenced Fe and Mn accumulation potential, and also as a tool to investigate the relationship between the input variables and the predicted Fe and Mn accumulation potential. The following observations were made from this model:

- Increased concentrations of Al generally increased Fe and Mn accumulation potential. This is due to the formation of amorphous $Al(OH)_3$ with increasing Al concentration which tends to adsorb Fe and Mn particles.
- Increased turbidity levels generally increased Fe and Mn accumulation potential because increase in turbidity increases suspended organic particles. Fe- and Mn-oxidising bacteria attach themselves to these suspended particles, causing microbial growth to increase. High levels of turbidity also enhance the biological oxidation of Fe and Mn by serving as a shield to inhibit Fe- and Mn-oxidising bacteria from disinfection.
- In general, as hydraulic distance from source of water supply increases, Fe and Mn accumulation potential also increases. Generally, increase in hydraulic distance from source of water supply increases water age and reduces FCR concentration.

Since chlorine is a disinfectant which suppresses the growth of Fe- and Mn-oxidising bacteria, increase in hydraulic distance from source of water supply reduces chlorine levels, which subsequently leads to increase in biological oxidation.

- There was a negative correlation between Fe and Mn accumulation potential and Ca concentration. Increase in Ca leads to the formation of calcium carbonate ($CaCO_3$) in the presence of DO in WDNs. $CaCO_3$ serves as a corrosion inhibitor by forming protective scales on the inner walls of ferrous pipes which prevents drinking water from coming into direct contact with these pipes in WDNs; thereby reducing Fe failures.
- There was a high positive correlation between hardness and Fe and Mn accumulation potential because increase in hardness increases chemical oxidation of Fe and Mn.
- Increased alkalinity levels reduced Fe and Mn accumulation potential. Increase in alkalinity helps to increase the buffer capacity of drinking water by keeping the pH of water stable, thereby reducing chemical oxidation of Fe and Mn in WDNs.
- Increased colour levels increased Fe and Mn accumulation potential. This is because increase in TOC, which is an indirect measure of colour, enhances adsorption of Mn. It also promotes microbial growth and increases biological oxidation of Fe and Mn because carbon is a bioavailable form of nutrients for Fe- and Mn-oxidising bacteria.

The second ANN model, ANN(t,ψ), uses biological, chemical, and hydraulic/physical variables to predict Fe and Mn accumulation potential for every node in a given WSZ. It can also be used to generate risk maps to visually see the distribution of the predicted Fe and Mn accumulation potential in WSZs in order to determine the high Fe and Mn accumulation potential risk regions. From the risk maps generated by the model, it was observed that:

- Most of the regions in the network with high Fe and Mn accumulation potential also had high customer complaints due to discolouration.
- There were a few years the high-risk regions predicted by the model did not correlate well with customer complaints. This was because events such as pipe bursts and the opening of fire hydrants during flushing, which can also cause water

discolouration and prompt customers to complain, were not included in the model. The aim of this research was not to predict water discolouration, but to predict Fe and Mn accumulation potential.

- Although ANN(t,ψ) model was able to predict Fe and Mn accumulation reasonably well and identify high-risk regions, the causes of failures had to be manually investigated. With so many nodes in WSZs, manually investigating the causes of failures can be a laborious task.

The hierarchical rule-based expert FIS and the hierarchical data-driven FIS were developed to overcome the limitations of the ANN(t,ψ) model. Unlike the developed ANN models, the FISs were able to automatically indicate the causes of high-risk of Fe and Mn accumulation potential. The hierarchical rule-based expert FIS used expert knowledge to formulate its rules and assign weights to the rules, whereas the hierarchical data-driven FIS used genetic algorithm to optimise the rules and weights of the rules. Results from both FISs showed that:

- The hierarchical data-driven FIS performed better than the hierarchical rule-based expert FIS. The relatively poor results observed in the hierarchical rule-based expert FIS were because the same rules formed from expert knowledge were used to model each of the WSZs. These rules did not accurately represent this FIS because Fe and Mn accumulation are formed under slightly different conditions in every WSZ.
- The hierarchical data-driven FIS gave good predictions because the rules and weights of the rules were optimised with a genetic algorithm for each WSZ. From the generated risk maps, it was observed that most regions with high customer complaints also had high Fe and Mn accumulation potential.

The developed ANN models and FISs can be used as tools to assist drinking water companies and water resource engineers in reducing discolouration and customer complaints by identifying high-risk Fe and Mn accumulation potential regions and explaining the causes of the risks. Since the models are able to predict Fe and Mn accumulation potential at every node, they can be used to identify high-risk regions including regions where water quality variables have not been sampled. In addition, the developed models can help in the development of cleaning protocols, maintenance of

water mains, and development of operational and management strategies for water distribution at the national and international levels.

7.2 Limitations of the developed models

The models developed in this research have a few limitations, just like every model. The following sections list some of these limitations.

7.2.1 Limitations of the ANN(t) model

- The ANN(t) model is only able to predict Fe and Mn accumulation potential at nodes where past sampling data exist. Thus, it is unable to make predictions for every node in a WSZ.
- ANNs require large data sets that are sampled adequately from the entire search space in order to have enough instances to make good predictions.

7.2.2 Limitations of the ANN(t,ψ) model

- Although ANN(t,ψ) model is able to predict Fe and Mn accumulation potential for every node in a WSZ, including regions where no past sampling data exist, it is unable to predict high-risk levels caused by hydraulic events such as pipe bursts and the opening of fire hydrants during fire extinguishing exercises because such hydraulic events were not included in the model.
- It requires large data sets from all regions in the WSZ to improve its prediction capabilities.
- Although the ANN(t,ψ) model gave better predictions than the FIS, its black-box nature make it unable to explain the causes of high-risk of Fe and Mn accumulation potential unless it is investigated manually.

7.2.3 Limitation of the hierarchical rule-based expert FIS

- Since Fe and Mn accumulation are formed under slightly different conditions for each WSZ, using the same rules formulated from expert knowledge to model every WSZ resulted in relatively poor model performance.

7.2.4 Limitation of the hierarchical data-driven FIS

- The ANN(t,ψ) model performed better than the hierarchical data-driven FIS because, generally, FISs are not used in solving problems that require high level precision solutions.

7.3 Recommendations and future work

It would have been ideal to use monthly or quarterly averages as input water quality variables because Fe and Mn accumulation potential exhibits seasonal variations. However, because some water quality variables were not sampled frequently, the data would have had many gaps if monthly or quarterly averages were used. It is therefore recommended that the water quality variables be sampled frequently in order to make monthly or weekly predictions possible.

DO is a very important variable that chemically oxidises Fe and Mn. However, it was not included in the models because it was not sampled. Nevertheless, it can be assumed that dissolved oxygen will be available in abundance in drinking water systems. Temperature is another important variable which aids the formation of biofilms and expedites the chemical oxidation of Fe and Mn. However, it was also not included because there was not enough data of it. TOC is another variable that influences Fe and Mn accumulation that was not included in the models because it was not sampled. Instead, colour, which is an indirect measure of TOC, was used as a variable in the model. Flushing frequency is a very important variable because flushing reduces or cleans years of accumulation of Fe and Mn particles in WDNs. However, no data for this variable were available. These variables would have greatly improved the models' predictions. It is therefore recommended that they are added to future models.

To make the developed models more useful to drinking water companies, a user friendly interface (software) for the two models need to be developed. This would enable engineers with little or no knowledge in ANN or fuzzy logic to use the models effectively.

REFERENCES

Aafjes, J., Verberne, A. J., Hendrix, L. J., & Vingerhoeds, R. (1997). Voorspellen van het drinkwaterverbruik met een neuraal netwerk. *H20, 30*(15), 484-487.

Ackers, J., Brandt, M., & Powell, J. (2001). *Hydraulic characterisation of deposits and review of sediment modelling. Drinking water quality and health - distribution systems.* London, UK: UK Water Industry Research.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9*(1), 147-169.

Agha, S. R., & Alnahhal, M. J. (2012). Neural network and multiple linear regression to predict school children dimensions for ergonomic school furniture design. *Applied Ergonomics, 43*(6), 979-984.

American Water Works Association. (1999). *Water quality and treatment: a handbook of community water supplies* (Fifth Edition ed.). (R. D. Letterman, Ed.) New York: McGraw-Hill.

Anderson, D., & McNeill, G. (1992). *Artificial neural networks technology.* Kaman Sciences Corporation, Utica, New York.

Augasta, M. G., & Kathirvalavakumar, T. (2012). Reverse engineering the neural networks for rule extraction in classification problems. *Neural Processing Letters, 35*(2), 131-150.

Australian National Health and Medical Research Council & Australian National Resource Management Ministerial Council [ANHMRC & ANRMMC]. (2004). Australian Drinking Water Guidelines 6, National Health and Medical Research Council and the National Resource Management Ministerial Council. Retrieved May 2, 2012, from http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/eh52_aust_drinking_water_guidelines_update_120710_0.pdf

Babusk, R. (1998). *Fuzzy modeling for control.* Boston: Kluwer Academic Publishers.

Balasubramaniam, J. (2006). Conditions for inference invariant rule reduction in FRBS by combining rules with identical consequents. *Acta Polytechnica Hungarica, 3*(4).

Bean, E. L. (1974). Potable water quality goals. *Journal of the American Water Works Association, 66*(4), 221-230.

Becker, K. (1998). Detachment studies on microfouling in natural biofilms on substrata with different surface tensions. *International Biodeterioration and Biodegradation, 41*, 93-100.

Benjamin, M. M., Sontheimer, H., & Leroy, P. (1996). Corrosion of iron and steel. In *Internal Corrosion of Water Distribution Systems.* Denver, Colorado: American Water Works Association Research Foundation and Deutsche Vereinigung des Gas- und Wasserfaches – Technologiezentrum Wasser.

Bernal, A., Cardenoso, R., Fabrellas, C., Matia, L., & Salvatella, N. (1999). An aesthetic quality index for Barcelona's water supply. *Water Science and Technology, 40*(6), 23-29.

Bhagwan, J. (2009). *Compendium of best practices in water infrastructure asset management.*

Bhalla, D., Bansal, R. K., & Gupta, H. (2012). Function analysis based rule extraction from artificial neural networks for transformer incipient fault diagnosis. *International Journal of Electrical Power & Energy Systems, 43*(1), 1196-1203.

Bhave, P. R. (1991). *Analysis of flow in water distribution systems.* Lancaster, PA, USA: Technomic Publishing.

Black, M. (1937). Vagueness: An exercise in logical analysis. *Philosophy of Science, 4*(4), 427-455.

Blokker, E., & Vreeburg, J. (2005). Monte Carlo simulation of residential water demand: a stochastic end-use model. *7th Annual Symposium on Water Distribution Systems Analysis* (pp. 15-19). Alaska, USA: ASCE.

Bologna, G. (2001). A study on rule extraction from several combined neural networks. *International Journal of Neural Systems, 11*(3), 247-255.

Bouslama, F., & Ichikawa, A. (1992). Fuzzy control rules and their natural control laws. *Fuzzy Sets and Systems, 48*(1), 65-86.

Bowden, G. J., Dandy, G. C., & Maier, H. R. (2003). Data transformation for neural network models in water resources applications. *Journal of Hydroinformatics, 5*(4), 245-258.

Bowden, G., Dandy, G., & Maier, H. (2005). Forecasting cyanobacteria (blue-green algae) using artificial neural networks. In S. Lingireddy, & G. M. Brion (Eds.), *Artificial neural networks in water supply engineering* (pp. 71-96). Reston, Virginia: American Society of Civil Engineers Press.

Boxall, J., & Husband, S. (2005). *An introduction and guide to the PODDS model.* University of Sheffield, Sheffield.

Boxall, J., & Prince, R. (2006). Modelling discolouration in a Melbourne (Australia) potable water distribution system. *Journal of Water Supply Research, 55*, 207-219.

Boxall, J., & Saul, A. (2005). Modeling discoloration in potable water distribution systems. *Journal of Environmental Engineering, ASCE, 131*(5), 716-725.

Boxall, J., Skipworth, P., & Saul, A. (2001). A novel approach to modelling sediment movement in distribution mains based on particle characteristics. *Computer and Control in Water Industry Conference.* De Montfort University, UK.

Boxall, J., Skipworth, P., & Saul, A. (2003). Aggressive flushing for discolouration event mitigation in water distribution networks. *Water Science and Technology: Water Supply, 3*(1), 179-186.

Broomhead, D. S., & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems, 2*, 321-355.

Bryson, A. E., & Ho, Y. C. (1969). *Applied optimal control: optimization, estimation and control.* Waltham, Massachusetts: Blaisdell Publishing Company.

Cerrato, J. M., Falkinham, J. O., Dietrich, A. M., Knocke, W. R., McKinney, C. W., & Pruden, A. (2010). Manganese-oxidizing and -reducing microorganisms isolated from biofilms in chlorinated drinking water systems. *Water Research, 44*, 3935-3945.

Cerrato, J. M., Reyes, L. P., Alvarado, C., & Dietrich, A. (2006). Effect of PVC and iron materials on Mn(II) deposition in drinking water distribution systems. *Water Research, 40*, 2720-2726.

Chadderton, R. A., Christensen, G. L., & Henry-Unrath, P. (1992). *Implementation and optimization of distribution flushing programs.* Project Number 90600. AWWA Research Foundation and American Water Works Association, Denver, USA.

Chang, F.-J., Chung, C.-H., Chen, P.-A., Liu, C.-W., Coynel, A., & Vachaud, G. (2014). Assessment of arsenic concentration in stream water using neuro fuzzy networks with factor analysis. *Science of the Total Environment, 494*, 202-210.

Christensen, R. T. (2009). *Age effects on iron-based pipes in water distribution systems.* PhD. thesis, Civil and Environmental Engineering, Utah State University, Logan, Utah.

Christensen, V., Rasmussen, P., & Ziegler, A. (2002). Real-time water-quality monitoring and regression analysis to estimate nutrient and bacteria concentrations in Kansas streams. *Water Science Technology, 45*(9), 205-211.

Ciliz, M. K. (2005). Rule base reduction for knowledge-based fuzzy controllers with application to a vacuum cleaner. *Expert Systems with Applications, 28*(1), 175-184.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

Combs, W. E., & Andrews, J. E. (1998). Combinatorial rule explosion eliminated by a fuzzy rule configuration. *IEEE Transactions on Fuzzy Systems, 6*(1), 1-11.

Cook, D. (2007). Field investigation of discolouration material accumulation rates in live drinking Water distribution systems. PhD. thesis, University of Sheffield.

Cook, D. M., Boxall, J. B., Hall, S. J., & Styan, E. (2005). Structural integrity and water quality in water distribution networks. *Proceedings of the 8th International Conference on Computing and Control for the Water Industry*, (pp. 205- 210).

Cooperative Research Centre for Water Quality and Treatment [CRCWQT]. (2005). *Biofilms: Understanding the impact on water quality and water treatment processes: management implications from the research programs of the Cooperative Research Centre for Water Quality and Treatment.* Australia.

Cox, E. (1992). Fuzzy fundamentals. *IEEE Spectrum, 29*(10), 58-61.

Cruse, H. (1971). Dissolved-copper effect on iron pipe. *Journal of the American Water Works Association, 79*(3), 79-81.

Dayton, E. A., & Basta, N. T. (2005). A method for determining the phosphorus sorption capacity and amorphous aluminum of aluminum-based drinking water treatment residuals. *Journal of Environmental Quality, 34*, 1112-1118.

De Kork, E. (2003). *Decentralising the codification of rules in a decision support expert knowledge base.* (MSc thesis), University of Pretoria etd, Pretoria.

Deborde, M., & Von Gunten, U. (2008). Reactions of chlorine with inorganic and organic compounds during water treatment – kinetics and mechanisms: a critical review. *Water Research, 42*, 13-52.

Decho, A. (2000). Microbial biofilms in intertidal systems. *Continental Shelf Research, 20*, 1257-1273.

Degnan, M. (1994). Recent work in Aristotle's logic. *Philosophical Books, 35*(2), 81-89.

Deines, P., Sekar, R., Jensen, H., Tait, S., Boxall, J., Osborn, A., & Biggs, C. (2010). MUWS (Microbiology in Urban Water Systems) – An interdisciplinary approach

to study microbial communities in urban water systems. *Drinking Water Engineering and Science*, *3*(2), 91-99.

Department of Human Services and Department of Natural Resources and Environment. (2000). *A new regulatory framework for drinking water quality in Victoria - consultation paper.* East Melbourne, Australia: Water Sector Services.

DeSilets, L., Golden, B., Wang, Q., & Kumar, R. (1992). Predicting salinity in the Chesapeake Bay using backpropagation. *Computers and Operations Research, 19*(3-4), 277-285.

Devi, R., Rani, B. S., & Prakash, V. (2012). Role of hidden neurons in an elman recurrent neural network in classification of cavitation signals. *International Journal of Computer Applications, 37*(7), 9-13.

Dewis, N., & Randall-Smith, M. (2005). Discolouration risk modelling. *Proceedings of the 8th International Conference on Computing and Control for the Water Industry.* Exeter, UK.

Deyo, R. A. (2010). Manipulation of knowledge. In E. De Corte, & J. E. Fenstad (Eds.), *Information to knowledge* (pp. 105-114). London: Portland Press.

Di Cristo, C., Esposito, G., & Leopardi, A. (2013). Modelling trihalomethanes formation in water supply systems. *Environmental Technology, 34*(1), 61-70.

Donlan, R., & Pipes, W. (1986). Pipewall biofilm in drinking water mains. Portland, Oregon: Water Quality and Technology Conference .

DWI. (2007). *Information leaflet discoloured water*. Retrieved March 10, 2012, from Drinking Water Inspectorate: http://www.dwi.gov.uk/consumer/faq/discolour.htm

Effler, S. W., Schafran, G. C., & Driscoll, C. T. (1985). Partitioning light attenuation in an acidic lake. *Canadian Journal of Fisheries and Aquatic Sciences, 42*, 1707-1711.

Ellison, D. (2003). *Investigation of pipe cleaning methods.* Denver, USA: American Water Works Association Research Foundation.

Emde, K. M., Smith, D. W., & Facey, R. (1992). Initial investigation of microbially influenced corrosion (mic) in a low temperature water distribution system. *Water Research, 26*(2), 169-175.

Evans, H. E. (1988). The binding of three PCB congeners to dissolved organic carbon in freshwaters. *Chemosphere, 17*(12), 2235-2238.

Evins, C., Liebeschuetz, J., & Williams, S. M. (1990). *Aesthetic water quality problems in distribution systems.* Water Research Centre. Buckinghamshire, UK.

Ewan, V. J., & Williams, S. (1986). *DoE Contract E15 interim report, monitoring water quality deterioration in distribution systems, final report.* Water Research Centre, Engineering. Wiltshire, UK.

Feigenbaum, E. A. (1982). *Knowledge engineering in the 1980s.* Stanford, CA: Deptartment of Computer Science, Stanford University.

Flemming, H.-C. (1998). Biofilme in Trinkwassersystemen, Teil. I. Übersicht. *Wasser Abwasser, 139*, 95-119.

Gautam, D. (1999). *Classification of wind events Using Kohonen networks.* Joint workshop on neural networks in civil engineering, Delft.

Gauthier, V., Barbeau, B., Milette, R., Block, J., & Prevost, M. (2001). Suspended particles in the drinking water of two distribution systems. *Water Science and Technology: Water Supply, 1*(4), 237-245.

Gauthier, V., Gérard, B., Portal, J.-M., Block, J.-C., & Gatel, D. (1999). Organic matter as loose deposits in a drinking water distribution system. *Water Research, 33*(4), 1014-1026.

Gauthier, V., Rosin, C., Mathieu, L., Portal, J., Block, J., Chaix, P., & Gatel, D. (1996). Characterization of the deposits in drinking water distribution systems. *Proceedings of Water Quality Technology Conference of the American.* USA.

Gedge, G. (1992). Corrosion of cast iron in potable water service. *Corrosion and related aspects of materials for potable water supplied.* London, UK: Proceedings of the Institute of Materials Conference.

Geldreich, E. (1996). *Microbial quality of water supply in distribution systems.* Boca Raton (FL): CRC Lewis Publishers.

Gerke, T. L., Maynard, J. B., Schock, M. R., & Lytle, D. L. (2008). Physiochemical characterization of five iron tubercles from a single drinking water distribution system: possible new insights on their formation and growth. *Corrosion Science, 50*(7), 2030-2039.

Gharibi, H., Sowlat, M. H., Mahvi, A. H., Mahmoudzadeh, H., Arabalibeik, H., Keshavarz, M., . . . Hassani, G. (2012). Development of a dairy cattle drinking water quality index (DCWQI) based on fuzzy inference systems. *Asian Biomedicine, 20*, 228-237.

Giiven, M. K., & Passino, K. M. (2001). Avoiding exponential parameter growth in fuzzy systems. *IEEE Transactions on Fuzzy Systems, 9*(1), 194 - 199.

Ginart, A., Sanchez, G., Links, I., & Back, G. (2002). Fast defuzzification method based on centroid estimation. *Applied Modelling and Simulation*.

Ginige, M., Wylie, J., & Plumb, J. (2011). Influence of biofilms on iron and manganese deposition in drinking water distribution systems. *Biofouling, 27*, 151-163.

Gorham, E., Underwood, J. K., Martin, F. B., & Ogden, G. J. (1986). Natural and anthropogenic causes of lake acidification in Nova Scotia. *Nature, 324*(4), 451-453.

Grainger, C., Wu, J., Nguyen, B., Ryan, G., Jayaratne, A., & Mathes, P. (2002). *Particles in water distribution system, 4th Progress Report; Part 1: Settling, resuspension and transport.* CRC of Water Quality and Treatment, Project Number 4.3.6., Melbourne, Australia.

Gray, N. (1994). *Drinking water quality - problems and solutions.* Chichester: John Wiley and Sons.

Grissom, R. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68*(1), 155-165.

Gschneidner, K. A. (1996). *CRC handbook of chemistry and physics* (77 ed.). Boca Raton, Florida: CRC Press.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. (1998). *Multivariate data analysis.* New Jersey: Prentice Hall.

Hanrahan, G. (2011). *Artificial neural networks in biological and environmental analysis.* Boca Raton, Florida: CRC Press.

Haudidier, K., Paquin, J. L., Francais, T., Hartemann, P., Grapin, G., Colin, F., . . . Miazga, J. (1988). Biofilm growth in drinking water network: a preliminary industrial pilot plant experiment. *Water Sci. Technol, 20*, 109.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory.* New York: John Wiley And Sons, Inc.

Herculano-Houzel, S. (2009). The human brain in numbers: a linearly scaled up primate brain. *Frontiers in Human Neuroscience, 3*(31).

Herman, G. M. (1996). *Iron and Manganese in Household Water.* North Carolina Cooperative Extension Service, North Carolina.

Hidmi, L., Gladwell, D., & Edwards, M. (1994). *Water quality and lead, copper, and iron corrosion in Boulder water.* University of Colorado, Report to the City of Boulder, Colorado.

Hinton, G. E. (1992). How neural networks learn from experience. *Scientific American, 267*, 145-151.

Holland, J. H. (1975). *An introductory analysis with applications to biology, control, and artificial intelligence.* Michigan: The University of Michigan Press.

Homoncik, S., MacDonald, A., Heal, K. V., Ó Dochartaigh, B., & Ngwenya, B. (2010). Manganese concentrations in Scottish groundwater. *Science of the Total Environment, 408*, 2467-2473.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA*, (pp. 2554-2558). 79 (9).

Howell, D. C. (2007). *Statistical methods for psychology* (6 ed.). Belmont, California: Thomson Wadsworth.

Hsieh, W. W. (2008). *Machine learning methods in environmental sciences: neural networks and kernels.* New York: Cambridge University Press.

Hubbard, D. W. (2009). *The failure of risk management: Why it's broken and how to fix it.* Hoboken, New Jersey: John Wiley & Sons.

Huck, P., & Gagnon, G. (2004). Understanding the distribution system as abioreactor: a framework for managing heterorophic plate count levels. *International Journal of Food Microbiology, 92*, 347-353.

Hunter, D., Yu, H., Pukish III, M. S., Kolbusz, J., & Wilamowski, B. M. (2012). Selection of proper neural network sizes and architectures: a comparative study. *IEEE Transactions on Industrial Informatics, 8*(2), 228-240.

Husband, P., & Boxall, J. (2011). Asset deterioration and discolouration in water distribution systems. *Water Research, 45*(1), 113-124.

Husband, S., Boxall, J., & Saul, A. (2008). Laboratory studies investigating the processes leading to discolouration in water distribution networks. *Water Research, 42*(16), 4309-4318.

Hwang, Y., & Bang, S. (1997). An efficient method to construct a radial basis function neural network classifier. *Neural Networks, 10*(8), 1495-1503.

Ibrahim, A. M. (2004). *Fuzzy logic for embedded systems applications.* Burlington, USA: Elsevier Science.

Islam, N., Sadiq, R., Rodriguez, M., & Francisque, A. (2013). Evaluation of source water protection strategies: A fuzzy-based model. *Journal of Environmental Management, 121*, 191-201.

Jinchuan, K., & Xinzhe, L. (2008). Empirical analysis of optimal hidden neurons in neural network modeling for stock prediction. *Proceedings of the Pacific-Asia workshop on computational intelligence and industrial application* (pp. 828-832). Wuhan: Institute of Electrical and Electronics Engineers.

Joarder, M. A., Raihan, F., Alam, J. B., & Hasanuzzaman, S. (2008). Regression analysis of ground water quality data of Sunamganj District, Bangladesh. *International Journal of Environmental Research, 2*(3), 291-296.

Kashinkunti, R., Metz, D., DeMarco, J., & Hartman, D. (1999). How to reduce lead corrosion without increasing iron release in the distribution system. *Proceedings of the 1999 American Water Works Association Water Quality Technology Conference.* Tampa, Fla: American Water Works Association.

Kirmeyer, G., Friedman, M., Martel, K., & Howie, D. (2001). *Pathogen intrusion into distribution system.* Denver, CO, USA.: American Water Works Association Research Foundation.

Knocke, W., Van Benschoten, J., Kearney, M., & Soborski, A. (1990). *Alternative oxidants for the removal of soluble iron and manganese.* American Water Works Association Research Foundation, Denver, 132.

Kohl, P. M., Medlar, S. J., Foundation., A. R., & Agency., U. S. (2006). *Occurrence of manganese in drinking water and manganese control.* Denver, CO: Awwa Research Foundation/American Water Works Association/IWA Pub.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78*(9), 1464-1480.

Kord, M., & Asghari Moghaddam, A. (2014). Spatial analysis of Ardabil plain aquifer potable groundwater using fuzzy logic. *Journal of King Saud University - Science, 26*(2), 129-140.

LeChevallier, M. (1990). Coliform regrowth in drinking water: a review. *Journal of the American Water Works Association, 182*, 74-86.

LeChevallier, M., Babcock, T., & Lee, R. (1987). Examination and characterization of distribution system biofilms. *Applied Environmental Microbiology, 53*(12), 2714-2724.

Li, D., Chen, L., & Lin, Y. (2003). Using functional virtual population as assistance to learn scheduling knowledge in dynamic manufacturing environments. *International Journal of Production Research, 41*, 4011-4024.

Lin, J., & Coller, B. (1997). Aluminium in a water suppy, Part 3: Domestic tap waters. *Journal of the Australian Water Association, 24*(1), 11-13.

Lingireddy, S., & Brion, G. M. (2005). *Artificial neural networks in water supply engineering.* Reston, VA: American Society of Civil Engineers.

Lint, J. W., & Vonk, Z. C. (1999). Neurale netwerken voorspellen waterstanden, Simposium: Neurale netwerken in waterbeheer. Technical University Delft, Delft.

Lippmann, R. P. (1987). An introduction to computing with neural nets. *IEEE Acoustical Speech and Signal Processing Magazine, 4*, 4-22.

Lobbrecht, A., Dibike, Y., & Solomatine, D. (2002). *Applications of neural networks and fuzzy logic to integrated water management. Project report.* IHE-Delft, Nijmegen, Netherlands.

Lukasiewicz, J. (1963). *Elements of mathematical logic.* New York: Macmillan.

Maier, H. R., & Dandy, G. (1996). The use of artificial neural networks for the prediction of water quality parameters. *Water Resources Research, 33*(4), 1013–1022.

Mamdani, E. H. (1974). Application of fuzzy algorithms for the control of a simple dynamic plant. *Proceedings of the IEEE, 121*(12), 121-159.

Mao, R., Zhu, H., Zhang, L., & Chen, A. (2006). A new method to assist small data set neural network learning. *Intelligent systems design and applications,* (pp. 17-22). Washington, DC, USA.

McClymont, K., Keedwell, E. C., Savic, D., & Randall-Smith, M. (2010). Mitigating discolouration risk with optimised network design. *Proceedings of the Ninth International Conference on Hydroinformatics.* Tianjin, China.

McClymont, K., Walker, D., Keedwell, K., Everson, R., J., F., Savic, D., & Randall-Smith, M. (2011). Novel methods for ranking district metered areas for water distribution network maintenance scheduling. *Proceedings of the Eleventh International Conference on Computing and Control for the Water Industry.* Exeter.

McNeill, F., & Thro, E. (1994). *Fuzzy Logic: A Practical Approach.* Boston, MA: Academic Press.

McNeill, L. S. (2000). *Water quality factors influencing iron and lead corrosion in drinking water.* PhD. thesis, PhD. thesis, Virginia Polytechnic Institute and State University, Virginia.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas imminent in nervous activity. *Bulletin of Math. Biophys, 5*, 115-133.

Mendes, R., Cortez, P., Rocha, M., & Neves, J. (2002). Particle swarms for feedforward neural network training. *Proceedings of the International Joint Conference on*

Neural Networks. 2, pp. 1895-1899. Honolulu, USA: Institute of Electrical and Electronics Engineers.

Minsky, M. L. (1954). *Neural Nets and the Brain Model Problem.* Ph.D. dissertation in Mathematics, Princeton University.

Mirsepassi, A., Cathers, B., & Dharmappa, H. (1995). Application of artificial neural networks to the real time operation of water treatment plants. *IEEE International Conference on Neural Networks: Proceedings, Institute of Electrical and Electronics Engineers*, (pp. 516-521). Perth, Australia.

Moller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks, 6*(4), 523-533.

Molot, L., & Dillon, P. (2003). Variation in iron, aluminum and dissolved organic carbon mass transfer coefficients in lakes. *Water Research, 37*(8), 1759-1768.

Murdoch, P., & Shanley, J. (2006). Detection of water quality trends at high, median, and low flow in a Catskill mountain stream, New York, through a new statistical method. *Water Resources Research, 42*(8), 8407–8417.

Nakhaei, F., & Irannajad, M. (2013). Comparison between neural networks and multiple regression methods in metallurgical performance modeling of flotation column. *Physicochemical Problems of Mineral Processing, 49*(1), 255-266.

Nasr, A. S., Rezaei, M., & Dashti Barmaki, M. (2012). Analysis of groundwater quality using Mamdani fuzzy inference system (MFIS) in Yazd province, Iran. *International Journal of Computer Applications, 59*(7), 45-53.

National Research Council. (2005). *Public water distribution systems: assessing and reducing risks.* Water Science and Technology Board, National Research Council of the National Academies, The National Academies Press,, Washington, DC.

Naylor, R., Nicholas, D., Murry, B., & Roddy, S. (1993). Optimisation of calcium bicarbonate buffering for corrosion control in potable water. Gold Coast, Queensland, Australia: Australian Water and Wastewater Association.

Nazz, S., Alam, A., & Biswas, R. (2011). Effect of different defuzzification methods in a fuzzy based load balancing application. *International Journal of Computer Science, 8*(5), 261-267.

Neshat, M., & Yaghobi, M. (2009). Designing a fuzzy expert system of diagnosing the hepatitis B intensity rate and comparing it with adaptive neural network fuzzy system. *Proceedings of the World Congress on Engineering and Computer Science 2009 Vol II WCECS 2009*, (pp. 797-802). San Francisco, USA.

Odell, L., Cyr, R., & Prather, S. (1998). Rethinking iron and manganese removal. *Water quality technology conference* . Denever, CO: American Water Works Association - OSHA.

Oladipupo, O. O., Ayo, C. K., & Uwadia, C. O. (2012). A fuzzy association rule mining expert-driven (FARME-D) approach to knowledge acquisition. *African Journal of Computing & ICT, 5*(5), 43-60.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research & Evaluation, 9* (6), 1-12.

Ozkan, C., & Erbek, F. S. (2003). The comparison of activation functions for multispectral landsat TM image classification. *Photogrammetric Engineering and Remote Sensing, 62*, 491-499.

Peng, C., Korshin, G., Valentine, R., Hill, A., Friedman, M. J., & Reiber, S. H. (2010). Characterization of elemental and structural composition of corrosion scales and deposits formed in drinking water distribution systems. *Water Research, 44*(15), 4570-4580.

Patterson, D. W. (1996). *Artificial neural networks: theory and applications.* Singapore: Prentice Hall.

Pham, D. T., & Castellani, M. (2002). Action aggregation and defuzzification in Mamdani-type fuzzy systems. *Journal of Mechanical Engineering Science, 216*(7), 747-759.

Pitts, W., & McCulloch, W. S. (1947). How we know universals the perception of auditory and visual forms. *9*(3), 127-147.

Poggio, T., & Girosi, F. (1990). Networks for approximation and learning. *The Proceedings of the IEEE, 78*(9), 1481-1497.

Priddy, K. L., & Paul, E. K. (2005). *Artificial neural networks: an introduction.* Bellingham, USA: SPIE Press.

Prince, R. (2008). *Formation of discoloured water and turbidity in an unfiltered water distribution system.* PhD Thesis, Swinburne University of Technology, Sydney, Australia.

Prince, R., Goulter, I., & Ryan, G. (2001). Relationship betweeen velocity profiles and turbidity problems in distribution systems. *World Water and Environmental Resources Congress.* Orlando, Florida.

Prince, R., Ryan, G., & Goulter, I. (2003). Role of operational changes in forming discoloured water in water distribution system. *Computer and Control in Water Industry*, (pp. 451-457). London, UK.

Puth, M.-.. T., Neuhäuser, M., & Ruxton, G. D. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour, 102*, 77-84.

Rahnamayan, S., Tzihoosh, H. R., & Salama, M. M. (2007). A novel population initialization method for accelerating evolutionary algorithms. *Computers & Mathematics with Applications, 53*(10), 1605–1614.

Rajendra Prasad, D. S., Sadashivaiah, C., & Ranganna, G. (2011). A comparative study of techniques for prediction of water quality parameters. *International Journal of Earth Sciences and Engineering, 4*, 423-428.

Raman, H., & Sunilkumar, N. (1995). Multivariate modelling of water resources time series using artificial neural networks. *Journal of Hydrological Sciences, 40*(2), 145-163.

Robert, N. L. (1989). Applications of fuzzy sets to rule-based expert system development. *Telematics Inform, 6*(3-4), 403–406.

Rodgers, J., & Nicewander, W. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician,, 42*(1), 59-66.

Rodriguez, M., West, J., Powell, J., & Serodes, J. (1997). Application of two approaches to model chlorine residuals in Severn Trent Water Ltd. (STW) distribution systems. *Water Science and Technology, 35*(5), 317-324.

Rogers, S. K., & Kabrisky, M. (1991). *An introduction to biological and artificial neural networks for pattern recognition.* Bellingham, Washington, USA: SPIE Optical Engineering Press.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review, 65*, 386-408.

Roseth, N. (2002). Community views on drinking water quality. *Water Services Association of Australia News Letter*(23), 3.

Roseth, N., & Rock, K. (2003). *Community views on drinking water quality. A survey of people living in Australia's capital cities. Results for Melbourne - South East Water.* Project # 1301, Cooperative Research Centre for Water, Melbourne, Australia.

Ross, T. J. (2010). *Fuzzy logic with engineering applications.* (3, Ed.) London: Wiley-Blackwell.

Rothery, P. (1988). A cautionary note on data transformation: bias in back-transformed means. *Bird Study, 35*(3), 219-221.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533-536.

Sadiq, R., Kleiner, Y., & Rajani, B. (2004). Aggregative risk analysis for water quality failure in distribution networks. *Journal of Water Supply Research and Technology, 53*(4), 241-261.

Sadiq, R., Kleiner, Y., & Rajani, B. (2007). Water quality failures in distribution networks-risk analysis using fuzzy logic and evidential reasoning. *Risk Analysis, 27*(5), 1381-94.

Saletic, D. Z., Velasevic, D., & Mastorakis, N. (2002). Analysis of basic defuzzification technique. *Proceedings of the 6th WSES International Multi Conference on Circuits, Systems, Communications and Computers*, (pp. 7-14). Rethymnon, Greece.

Sarin, P., Snoeyink, V. L., Bebee, J., Jim, K. K., Beckett, M. A., Kriven, W. M., & Clement, J. A. (2004). Iron release from corroded iron pipes in drinking water distribution systems: effect of dissolved oxygen. *Water Research, 38*(5), 1259-1269.

Schintu, M., Meloni, P., & Contu, A. (2000). Aluminium fractions in drinking water from reservoirs. *Ecotoxicology and Environmental Safety, 46*, 29-33.

Seo, G., Jung, H.-R., Lee, H.-D., S., C. W., & Gee, C. S. (1998). Characteristics of water quality parameters on enhancing and inhibiting corrosion in water distribution system (Korean) . *Journal of the Korean Society of Environmental Engineering, 20*(8), 1151-1160.

Servais, P., Laurent, P., & Randon, G. (1995). Comparison of the bacterial dynamics in various french distribution systems. *Journal of Water Supply Research and Technology—AQUA, 44*(1), 10–17.

Setiono, R., Baesens, B., & Mues, C. (2008). Recursive neural network rule extraction for data with mixed attributes. *IEEE Transactions on Neural Networks, 19*(2), 299-307.

Seyoum, A. G., & Tanyimboh, T. T. (2014). Pressure dependent network water quality modelling. *Proceedings of the Institution of Civil Engineers - Water Management, 167*(6), 342-345.

Shah, H., Ghazali, R., Nawi, N., & Deris, M. (2012). Global hybrid ant bee colony algorithm for training artificial neural networks. *J. LNCS, 7333*(1), 87-100.

Shahriari, S., & Shahriari, S. (2014). Predicting ionic liquid based aqueous biphasic systems with artificial neural networks. *Journal of Molecular Liquids, 197*, 65-67.

Shang, F., Uber, J., & Polycarpou, M. (2002). Particle backtracking algorithm for water distribution system analysis. *Journal of Environmental Engineering, 128*(5), 441-450.

Sheela, K. G., & Deepa, S. N. (2013). *Review on methods to fix number of hidden neurons in neural networks.* Mathematical Problems in Engineering, Article ID: 425740.

Shi, J. J. (2000). Reducing prediction error by transforming input data for neural networks. *Journal of Computing in Civil Engineering, 14*(2), 109-116.

Shibata, K., & Ikeda, Y. (2009). Effect of number of hidden neurons on learning in large-scale layered neural networks. *Proceedings of the ICROS-SICE International Joint Conference 2009 (ICCAS-SICE '09),* (pp. 5008–5013). Fukuoka, Japan.

Shin, Y. C., & Xu, C. (2009). *Intelligent systems; modeling, optimization, and control.* New York: CRC Press.

Singh, K. P., & Gupta, S. (2012). Artificial intelligence based modeling for predicting the disinfection by-products in water. *Chemometrics and Intelligent Laboratory Systems, 114*, 122-131.

Slaats, N. (2002). *Processes involved in generation of discoloured water.* Kiwa. The Netherlands: American Water Works Association Research Fundation.

Sly, L., Hodgkindon, M., & Arunpairojana, V. (1988). Effect of water velocity on the early development of manganese-depositing biofilm in a drinking-water distribution system FEMS. *FEMS Microbiology Ecology, 53*(3-4), 175-186.

Sly, L., Hodgkinson, M., & Arunpairojana, V. (1990). Deposition of manganese in drinking water distribution system. *Applied and Evironmental Microbiology, 56*(3), 628-639.

Smith, L. C., & Subandoro, A. (2007). *Measuring food security using household expenditure surveys.* Washington D.C.: International Food Policy Research Institute.

Smith, M. (1993). *Neural networks for statistical modeling.* (V. N. Reinhold, Ed.) New York.

Smith, S. E., Bisset, A., Colbourne, J. S., Hold, D. M., & Lloyd, B. J. (1997). The occurrence and significance of particles and deposits in a drinking water distribution system. *Journal of New England Water Works Association, 111*(2), 135-150.

Sonali, B. W. (2014). Analytical study of neural network techniques: SOM, MLP and classifier-a survey. *IOSR Journal of Computer Engineering, 16*(3), 86-92.

Stanley, S., Baxter, C., Zhang, Q., & Shariff, R. (2000). Process Modelling and control of enhanced coagulation. *AWWA Research Foundation and American Water Works Association*.

Stumm, W. (1960). Investigation of the corrosive behavior of waters. *Journal of the ASCE Sanitary Engineering Division, 86*(6), 27-46.

Sugeno, M. (1985). An introductory survey of fuzzy control. *Information Sciences, 36*, 59-83.

Sugeno, M. (1985). *Industrial applications of fuzzy control.* New York: Elsevier scientific publishing company.

Swistock, B. R., Sharpe, W. E., & Robillard, P. D. (2001). *Iron and Manganese in Private Water Systems.* Penn State College of Agricultural Sciences and Cooperative Extension, F138, Pennsylvania, USA.

Teasdale, P., O'Halloran, K., Doolan, C., & Hamilton, L. (2007). *Literature review on discoloured water formation and desktop study of industry practices.* The Cooperative Research Centre for Water Quality and Treatment, Salisbury.

The PDP Research Group. (1986). *Parallel distributed processing: explorations in the microstructures of cognition, vol. 1: Foundations* (Vol. 1). (D. Rumelhart, & J. McClelland, Eds.) Cambridge: MIT Press.

USEPA. (1994). *Drinking water criteria document for manganese.* United States Environmental Protection Agency, Washington, DC.

USEPA. (2002). *Health risks from microbial growth and biofilms in the distribution system.* United States Environmental Protection Agency, Office of Groundwater and Drinking Water:, Washington, D.C. Retrieved from United States Environmental Protection Agency.

USEPA. (2009). Basic information about chloramines. Retrieved from http://www.epa.gov/ogwdw/disinfection/chloramine/pdfs/all29_q.pdf

Valverde-Albacete, F. P.-M. (2014). 100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox. *PloS One, 9*(1).

Van Benschoten, J., Lin, W., & Knocke, W. (1992). Kinetic modeling of manganese(II) oxidation by chlorine dioxide and potassium permanganate. *Environmental Science & Technology, 26*(7), 1327-1333.

van Boomen, M., & Vreeburg, J. (1999). *Nieuwe ontwerprichtlijnen voor distributienetten (New design rules for distribution networks).* Kiwa report SWE99.011, Netherlands.

van der Kooij, D. (2002). Assimilable organic carbon (AOC) in treated water: determination and significance. In G. Bitton (Ed.), *Encyclopedia of Environmental Microbiology.* Hoboken, NJ, USA: John Wiley & Sons.

van der Wende, E., Characklis, W. G., & Smith, D. B. (1989). Biofilms and bacterial drinking water quality. *Water Research, 23*, 1313.

van Leekwijck, W., & Kerre, E. E. (1999). Defuzzification: criteria and classification. *Fuzzy Sets and Systems, 108*(2), 159-178.

Vigliotta, G., Nutricati, E., Carata, E., Tredici, S., De Stefano, M., Pontieri, P., Alifano, P. (2007). Clonothrix fusca Roze 1896, a filamentous, sheathed, methanotrophic γ-proteobacterium. *Applied and Environmental Microbiology, 73*, 3556-3565.

Vinay, C., Vinay, A., & Ravindra, N. (2014). Modeling slump of ready mix concrete using genetically evolved artificial neural networks. *Advances in Artificial Neural Systems, 2014*, 1-9.

Vreeburg, J. (1996). *Brown water, cause and consequence. Efficiency of cleaning wih flushing, water/air scouring and pigging.* Kiwa report SWE 96.008, Nieuwegein, The Nettherlands.

Vreeburg, J. (2007). Discolouration in Drinking Water Systems: A Particular Approach. PhD. thesis, Department of Civil Engineering and Geosciences, Delft University of Technology,Delft, The Netherlands.

Vreeburg, J., & Boxall, J. (2007). Discolouration in potable water distribution systems: a review. *Water Research, 41*(3), 519-529.

Vreeburg, J., Schaap, P., & van Dijk, J. (2004a). Measuring resuspension risk: resuspension potential method. *IWA Edge Conference.* Prague.

Vreeburg, J., Schaap, P., & van Dijk, J. (2004b). Particles in the drinking water system: from source to discolouration. *Water Science and Technology, 4*(5), 431-438.

Wallace, L., & Campbell, H. (1991). *Iron and manganese treatment small water systems.* American Water Works Association, Denver, 1-325.

Walski, T. (1991). Understanding solids transport in water distribution systems. *Water Quality Modelling in Dstribution Systems* (pp. 305-309). Cincinnati, Ohio: American Water Works Association Research Foundation.

Walski, T., & Draus, S. (1996). Predicting water quality changes during flushing. *American Water Works Association Annual Conference*, (pp. 121-129). Toronto, Ontario.

Wang, W., Zhang, X., Wang, H., Wang, X., Zhou, L., Liu, R., & Liang, Y. (2012). Laboratory study on the adsorption of Mn2+ on suspended and deposited amorphous Al(OH)3 in drinking water distribution systems. *Water Research, 46*(13), 4063-4070.

Wasserman, G. A., Liu, X., Parvez, F., Ahsan, H., Levy, D., Factor-Litvak, P., . . . Graziano, J. H. (2006). Water manganese exposure and children's intellectual function in Araihazar, Bangladesh. *Environmental Health Perspectives, 114*(1), 124-129.

Werbos, P. J. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences.* PhD thesis, Harvard University.

WHO. (2006). *Guidelines for drinking-water quality* (4th ed.). Geneva: World Health Organization.

WHO. (2011a). *Guidelines for drinking-water quality.* World Health Organisation, Geneva.

WHO. (2011b). *Hardness in drinking-water: background document for development of WHO guidelines for drinking-water quality.* WHO/HSE/WSH/10.01/10/Rev/1. World Health Organization, Geneva.

WHO, & UNICEF. (2006). *Meeting the MDG drinking water and sanitation targets: the urban and rural challenge of the decade.* Geneva.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *IRE WESCON Convention Record, 4*, 96-104.

Widrow, B., Rumelhard, D. E., & Lehr, M. A. (1994). Neural networks: applications in industry, business and science. *Communications of the ACM, 37*, 93-105.

Williams, R. W., & Herrup, K. (1988). The control of neuron number. *Annual Review of Neuroscience, 11*, 423-53.

Wricke, B., Henning, L., Korth, A., Vreeburg, J., Schaap, P., Osterhus, S., . . . Coelho, S. (2007). *Particles in relation to water quality deterioration and problems in the network – state-of-the-art review.* Techneau report, deliverable D 5.5.1 and D 5.5.2, Dresden, Germany.

Wu, J., Noui-Mehidi, N., Grainger, G., Nguyen, B., Ryan, G., Jayaratne, A., & Mathes, P. (2003). *Particles in water distribution systems - 6th progress report. Particle sediment modelling: PSM software.* CMIT-2003-234. Cooperative research Centre for Water Quality and Treatment, Melbourne, Australia.

Xiong, Y., & Liu, Y. (2010). Biological control of microbial attachment: a promising alternative for mitigating membrane biofouling. *Applied Microbiology and Biotechnology, 86*(3), 825-837.

Yarra Valley Water. (1998). *Water Quality Report 97/98.* Melbourne, Australia: Yarra Valley Water.

Zadeh, L. (1965). Fuzzy sets. *Information and Control, 8*(3), 338–353.

Zeng, L. (1999). Prediction and classification with neural network. *Sociological Methods and Research, 27*(4), 499-524.

Zhang, J., Lok, T., & Lyu, M. (2007). A hybrid particle swarm optimization back propagation algorithm for feed forward neural network training. *J. Applied Mathematics and Computation, 185*, 1026-1037.

Zhang, Q., & Stanley., S. J. (1997). Forecasting raw-water quality parameters for the North Saskatchewan River by neural network modeling. *Water Research, 31*(9), 2340-2350.

Zhang, W., & DiGiano, F. A. (2001). Comparison of bacterial regrowth in distribution systems using free chlorine and chloramine: a statistical study of causative factors. *Water Research, 36*, 1469–1482.

Zhang, Y., Wang, S., Ji, G., & Phillips, P. (2014). Fruit classification using computer vision and feedforward neural network. *Journal of Food Engineering, 143*, 167-177.

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. *Journal of General Psychology, 121*(4), 391-401.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67*(1), 55-68.

APPENDICES

Appendix A: Source code for calculating shear stress at node

A.1 Source code for calculating minimum daily shear stress

```
function [Vmin, Tmin] = LinkCalcsMin(InputFile)

Vmin = []; Tmin=[]; Fmin = []; Hlmin = [];

%Load EPANET DLL
%loadlibrary('epanet2', 'epanet2.h');
if ~libisloaded('epanet2'), loadlibrary('epanet2', 'epanet2.h'); end
%Edited


%Open EPANET toolkit
Err = calllib('epanet2', 'ENopen', InputFile, 'report.rpt',' ');

%Supress the writing of all error messages to be written to report.rpt.
Err = calllib('epanet2', 'ENsetreport', 'MESSAGES NO');

%Make Calculations
%Calculate the number of links
Nlinks = 0;
[Err, Nlinks] = calllib('epanet2', 'ENgetcount', 2, Nlinks);

%Calculate minimum velocity, shear stress, flow and headloss
[minVel, minShear, minFlw, minHl] = HydraulicCalculations(Nlinks);
%Changed from[a,b]
Vmin = minVel;
Tmin = minShear;
Fmin = minFlw;
Hlmin = minHl;

%Read  Link Ids from EPANET
id = ' ';
LinkIds = { };
for i = 1:Nlinks
    [Err, id] = calllib('epanet2', 'ENgetlinkid', i, id);
    LinkIds{i} = id;
end

%Read Start and End node of link
S_Node_ = 0; E_Node_ = 0;
S_Node_Id_ = 'abcdefghijklmnopqrstuvwxyzaaaaa'; E_Node_Id_ =
'abcdefghijklmnopqrstuvwxyzaaaaa';

S_Node_Id = { };
E_Node_Id = { };

%S_Node_Id = zeros(1,Nlinks);
%E_Node_Id = zeros(1,Nlinks);
for i = 1:Nlinks
    [Err, S_Node_, E_Node_]  = calllib('epanet2', 'ENgetlinknodes', i,
S_Node_, E_Node_);
```

```
    [Err, S_Node_Id_] = calllib('epanet2', 'ENgetnodeid', S_Node_,
S_Node_Id_);
    [Err, E_Node_Id_] = calllib('epanet2', 'ENgetnodeid', E_Node_,
E_Node_Id_);

    S_Node_Id{i} = S_Node_Id_;
    E_Node_Id{i} = E_Node_Id_;
end


%Read Link length from EPANET
len_ = 0.0;
len = zeros(1,Nlinks);
for i = 1:Nlinks
    [Err, len_] = calllib('epanet2', 'ENgetlinkvalue', i, 1, len_);
    len(i) = len_;
end


%Read Link diameters from EPANET
dia_ = 0.0;
dia = zeros(1,Nlinks);
for i = 1:Nlinks
    [Err, dia_] = calllib('epanet2', 'ENgetlinkvalue', i, 0, dia_);
    dia(i) = dia_;
end



%%Export the values
M = [LinkIds', S_Node_Id', E_Node_Id', num2cell(len'), num2cell(dia'),
num2cell(Fmin'), num2cell(Hlmin'), num2cell(Vmin'), num2cell(Tmin')];
dlmcell('LinkParameters.csv',M, 'delimiter', ',');

%Close EPANET toolkit
Err = calllib('epanet2', 'ENclose');

%Unload EPANET DLL.
unloadlibrary('epanet2');
return


function [minVel, minShear, minFlw, minHl] = HydraulicCalculations(Nlinks)
%Solve Hydraulics and calculate parameters for current loading


minVel = ones(1,Nlinks)*200; minShear = ones(1,Nlinks)*200; minFlw =
ones(1,Nlinks)*200; minHl = ones(1,Nlinks)*200;
t = 0; tstep=1;
Err = calllib('epanet2', 'ENopenH');
Err = calllib('epanet2', 'ENinitH', 0);
while(tstep > 0)
   [Err, t] = calllib('epanet2', 'ENrunH', t);

   %Calculate various parameters at current timestep
   [minVel, minShear, minFlw, minHl] = calcparams(minVel, minShear,
minFlw, minHl, Nlinks);
```

```
   [Err, tstep] = calllib('epanet2', 'ENnextH', tstep);
end
Err = calllib('epanet2', 'ENcloseH');
return

function [v, t, f, hL] = calcparams(mV, mT, mfL, mhL, Nl)
Rho = 1000.0; g = 9.80665;
dia = 0.0; len = 0.0; vel = 100.0; Hl = 100.0;
S_Node = 0; E_Node = 0; flw = 100.0;
S_Node_Id = 'abcdefghijklmnopqrstuvwxyzaaaaa'; E_Node_Id =
'abcdefghijklmnopqrstuvwxyzaaaaa';

v = zeros(1,Nl); t = zeros(1,Nl); f = zeros(1,Nl); hL = zeros(1,Nl);
for i = 1:Nl
    [Err, dia] = calllib('epanet2', 'ENgetlinkvalue', i, 0, dia);
    [Err, len] = calllib('epanet2', 'ENgetlinkvalue', i, 1, len);
    [Err, vel] = calllib('epanet2', 'ENgetlinkvalue', i, 9, vel);
    [Err, Hl]  = calllib('epanet2', 'ENgetlinkvalue', i, 10, Hl);
    [Err, S_Node, E_Node]  = calllib('epanet2', 'ENgetlinknodes', i,
S_Node, E_Node);
    [Err, S_Node_Id] = calllib('epanet2', 'ENgetnodeid', S_Node,
S_Node_Id);
    [Err, E_Node_Id] = calllib('epanet2', 'ENgetnodeid', E_Node,
E_Node_Id);
    [Err, flw]  = calllib('epanet2', 'ENgetlinkvalue', i, 8, flw);

    v(i) = min(vel, mV(i));
    shear = Rho*g*(dia/4000)*(Hl/Len);
    t(i) = min(shear, mT(i));
    f(i) = min(flw, mfL(i));
    hL(i) = min(Hl, mhL(i));
end
return
```

A.2 Source code for calculating maximum daily shear stress

```
function [Vmax, Tmax] = LinkCalcsMax(InputFile)

Vmax = []; Tmax=[]; Fmax = []; Hlmax = [];

%Load EPANET DLL
%loadlibrary('epanet2', 'epanet2.h');
if ~libisloaded('epanet2'), loadlibrary('epanet2', 'epanet2.h'); end
%Edited


%Open EPANET toolkit
Err = calllib('epanet2', 'ENopen', InputFile, 'report.rpt',' ');

%Supress the writing of all error messages to be written to report.rpt.
Err = calllib('epanet2', 'ENsetreport', 'MESSAGES NO');

%Make Calculations
%Calculate the number of links
Nlinks = 0;
[Err, Nlinks] = calllib('epanet2', 'ENgetcount', 2, Nlinks);

%Calculate maximum velocity and shear stress
```

```
[maxVel, maxShear, maxFlw, maxHl] = HydraulicCalculations(Nlinks);
%Changed from[a,b]
Vmax = maxVel;
Tmax = maxShear;
Fmax = maxFlw;
Hlmax = maxHl;

%Read  Link Ids from EPANET
id = ' ';
LinkIds = { };
for i = 1:Nlinks
    [Err, id] = calllib('epanet2', 'ENgetlinkid', i, id);
    LinkIds{i} = id;
end

%Read Start and End node of link
S_Node_ = 0; E_Node_ = 0;
S_Node_Id_ = 'abcdefghijklmnopqrstuvwxyzaaaaa'; E_Node_Id_ =
'abcdefghijklmnopqrstuvwxyzaaaaa';

S_Node_Id = { };
E_Node_Id = { };

%S_Node_Id = zeros(1,Nlinks);
%E_Node_Id = zeros(1,Nlinks);
for i = 1:Nlinks
    [Err, S_Node_, E_Node_]  = calllib('epanet2', 'ENgetlinknodes', i,
S_Node_, E_Node_);
    [Err, S_Node_Id_] = calllib('epanet2', 'ENgetnodeid', S_Node_,
S_Node_Id_);
    [Err, E_Node_Id_] = calllib('epanet2', 'ENgetnodeid', E_Node_,
E_Node_Id_);

    S_Node_Id{i} = S_Node_Id_;
    E_Node_Id{i} = E_Node_Id_;
end


%Read Link length from EPANET
len_ = 0.0;
len = zeros(1,Nlinks);
for i = 1:Nlinks
    [Err, len_] = calllib('epanet2', 'ENgetlinkvalue', i, 1, len_);
    len(i) = len_;
end


%Read Link diameters from EPANET
dia_ = 0.0;
dia = zeros(1,Nlinks);
for i = 1:Nlinks
    [Err, dia_] = calllib('epanet2', 'ENgetlinkvalue', i, 0, dia_);
    dia(i) = dia_;
end


%%Export the values
M = [LinkIds', S_Node_Id', E_Node_Id', num2cell(len'), num2cell(dia'),
num2cell(Fmax'), num2cell(Hlmax'), num2cell(Vmax'), num2cell(Tmax')];
```

```
dlmcell('LinkParameters.csv',M, 'delimiter', ', ');

%Close EPANET toolkit
Err = calllib('epanet2', 'ENclose');

%Unload EPANET DLL.
unloadlibrary('epanet2');
return


function [maxVel, maxShear, maxFlw, maxHl] = HydraulicCalculations(Nlinks)
%Solve Hydraulics and calculate parameters for current loading

maxVel = zeros(1,Nlinks); maxShear = zeros(1,Nlinks); maxFlw =
zeros(1,Nlinks); maxHl = zeros(1,Nlinks);
t = 0; tstep=1;
Err = calllib('epanet2', 'ENopenH');
Err = calllib('epanet2', 'ENinitH', 0);
while(tstep > 0)
   [Err, t] = calllib('epanet2', 'ENrunH', t);

   %Calculate various parameters at current timestep
   [maxVel, maxShear, maxFlw, maxHl] = calcparams(maxVel, maxShear,
maxFlw, maxHl, Nlinks);

   [Err, tstep] = calllib('epanet2', 'ENnextH', tstep);
end
Err = calllib('epanet2', 'ENcloseH');
return

function [v, t, f, hL] = calcparams(mV, mT, mfL, mhL, Nl)
Rho = 1000.0; g = 9.80665;
dia = 0.0; len = 0.0; vel = 0.0; Hl = 0.0;
S_Node = 0; E_Node = 0; flw = 0.0;
S_Node_Id = 'abcdefghijklmnopqrstuvwxyzaaaaa'; E_Node_Id =
'abcdefghijklmnopqrstuvwxyzaaaaa';

v = zeros(1,Nl); t = zeros(1,Nl); f = zeros(1,Nl); hL = zeros(1,Nl);
for i = 1:Nl
    [Err, dia] = calllib('epanet2', 'ENgetlinkvalue', i, 0, dia);
    [Err, len] = calllib('epanet2', 'ENgetlinkvalue', i, 1, len);
    [Err, vel] = calllib('epanet2', 'ENgetlinkvalue', i, 9, vel);
    [Err, Hl]  = calllib('epanet2', 'ENgetlinkvalue', i, 10, Hl);
    [Err, S_Node, E_Node]  = calllib('epanet2', 'ENgetlinknodes', i,
S_Node, E_Node);
    [Err, S_Node_Id] = calllib('epanet2', 'ENgetnodeid', S_Node,
S_Node_Id);
    [Err, E_Node_Id] = calllib('epanet2', 'ENgetnodeid', E_Node,
E_Node_Id);
    [Err, flw]  = calllib('epanet2', 'ENgetlinkvalue', i, 8, flw);

    v(i) = max(vel, mV(i));
    shear = Rho*g*(dia/4000)*(Hl/Len);
    t(i) = max(shear, mT(i));
    f(i) = max(flw, mfL(i));
    hL(i) = max(Hl, mhL(i));
end
return
```

A.3 Source code for calculating variation of daily shear stress

```
function [Vavg, Tavg, Favg, Hlavg, varShearOutPut] = CalcsVar(InputFile)

Vavg = []; Tavg=[]; Favg = []; Hlavg = [];

%Load EPANET DLL
%loadlibrary('epanet2', 'epanet2.h');
if ~libisloaded('epanet2'), loadlibrary('epanet2', 'epanet2.h'); end
%Edited

%Open EPANET toolkit
Err = calllib('epanet2', 'ENopen', InputFile, 'report.rpt',' ');

%Supress the writing of all error messages to be written to report.rpt.
Err = calllib('epanet2', 'ENsetreport', 'MESSAGES NO');

%Make Calculations
%Calculate the number of links
Nlinks = 0;
[Err, Nlinks] = calllib('epanet2', 'ENgetcount', 2, Nlinks);

%Calculate average velocity and shear stress
[avgVel, avgShear, avgFlw, avgHl,varShear] =
HydraulicCalculations(Nlinks);   %Changed from[a,b]
Vavg = avgVel;
Tavg = avgShear;
Favg = avgFlw;
Hlavg = avgHl;
varShearOutPut = varShear';

%Read  Link Ids from EPANET
id = ' ';
LinkIds = { };
for i = 1:Nlinks
    [Err, id] = calllib('epanet2', 'ENgetlinkid', i, id);
    LinkIds{i} = id;
end

%Read Start and End node of link
S_Node_ = 0; E_Node_ = 0;
S_Node_Id_ = 'abcdefghijklmnopqrstuvwxyzaaaaa'; E_Node_Id_ =
'abcdefghijklmnopqrstuvwxyzaaaaa';

S_Node_Id = { };
E_Node_Id = { };

for i = 1:Nlinks
    [Err, S_Node_, E_Node_]  = calllib('epanet2', 'ENgetlinknodes', i,
S_Node_, E_Node_);
    [Err, S_Node_Id_] = calllib('epanet2', 'ENgetnodeid', S_Node_,
S_Node_Id_);
    [Err, E_Node_Id_] = calllib('epanet2', 'ENgetnodeid', E_Node_,
E_Node_Id_);

    S_Node_Id{i} = S_Node_Id_;
    E_Node_Id{i} = E_Node_Id_;
end
```

```
%Read Link length from EPANET
len_ = 0.0;
len = zeros(1,Nlinks);
for i = 1:Nlinks
    [Err, len_] = calllib('epanet2', 'ENgetlinkvalue', i, 1, len_);
    len(i) = len_;
end

%Read Link diameters from EPANET
dia_ = 0.0;
dia = zeros(1,Nlinks);
for i = 1:Nlinks
    [Err, dia_] = calllib('epanet2', 'ENgetlinkvalue', i, 0, dia_);
    dia(i) = dia_;
end

%%Export the values
M = [LinkIds', S_Node_Id', E_Node_Id', num2cell(len'), num2cell(dia'),
num2cell(Favg'), num2cell(Hlavg'), num2cell(Vavg'), num2cell(Tavg')];
dlmcell('LinkParameters.csv',M, 'delimiter', ',');

%Close EPANET toolkit
Err = calllib('epanet2', 'ENclose');

%Unload EPANET DLL.
unloadlibrary('epanet2');
return

function [avgVel, avgShear, avgFlw, avgHl,varShear] =
HydraulicCalculations(Nlinks)
%Solve Hydraulics and calculate parameters for current loading

sumVel = zeros(1,Nlinks); sumShear = zeros(1,Nlinks); sumFlw =
zeros(1,Nlinks); sumHl = zeros(1,Nlinks);
%avgVel = zeros(1,Nlinks); avgShear = zeros(1,Nlinks); avgFlw =
zeros(1,Nlinks); avgHl = zeros(1,Nlinks);
t = 0; tstep=1;
Err = calllib('epanet2', 'ENopenH');
Err = calllib('epanet2', 'ENinitH', 0);
tCount = 0;
%sumShear = zeros(1,5000);
while(tstep > 0)
   [Err, t] = calllib('epanet2', 'ENrunH', t);
   sumShear = zeros(1,Nlinks);
   %Calculate various parameters at current timestep
   [sumVel, sumShear, sumFlw, sumHl] = calcparams(sumVel, sumShear,
sumFlw, sumHl, Nlinks);

   [Err, tstep] = calllib('epanet2', 'ENnextH', tstep);

   if tCount ==0
       sumShear_ = sumShear;
   else
   sumShear_ =[sumShear_;sumShear];
   end
   tCount = tCount+1;
end

varShear = nanstd(sumShear_,0,1);
```

```
%save varShear.csv;
dbltCount= double(tCount);
avgVel = sumVel/dbltCount;
avgShear = sumShear/dbltCount;
avgFlw = sumFlw/dbltCount;
avgHl = sumHl/dbltCount;

Err = calllib('epanet2', 'ENcloseH');
return

function [v, t, f, hL] = calcparams(mV, mT, mfL, mhL, Nl)
Rho = 1000.0; g = 9.80665;
dia = 0.0; len = 0.0; vel = 0.0; Hl = 0.0;
S_Node = 0; E_Node = 0; flw = 0.0;
S_Node_Id = 'abcdefghijklmnopqrstuvwxyzaaaaa'; E_Node_Id =
'abcdefghijklmnopqrstuvwxyzaaaaa';

v = zeros(1,Nl); t = zeros(1,Nl); f = zeros(1,Nl); hL = zeros(1,Nl);
for i = 1:Nl
    [Err, dia] = calllib('epanet2', 'ENgetlinkvalue', i, 0, dia);
    [Err, len] = calllib('epanet2', 'ENgetlinkvalue', i, 1, len);
    [Err, vel] = calllib('epanet2', 'ENgetlinkvalue', i, 9, vel);
    [Err, Hl]  = calllib('epanet2', 'ENgetlinkvalue', i, 10, Hl);
    [Err, S_Node, E_Node]  = calllib('epanet2', 'ENgetlinknodes', i,
S_Node, E_Node);
    [Err, S_Node_Id] = calllib('epanet2', 'ENgetnodeid', S_Node,
S_Node_Id);
    [Err, E_Node_Id] = calllib('epanet2', 'ENgetnodeid', E_Node,
E_Node_Id);
    [Err, flw]  = calllib('epanet2', 'ENgetlinkvalue', i, 8, flw);

    mV(i)= mV(i)+ vel;
    v(i) = mV(i);
    shear = Rho*g*(dia/4000)*(Hl/Len);
    mT(i) = mT(i)+ shear;
    t(i) = mT(i);
    mfL(i) = mfL(i)+ flw;
    f(i) = mfL(i);
    mhL(i) = mhL(i)+ Hl;
    hL(i)= mhL(i);
end
return
```

Appendix B: Microsoft visual basic source code for the ANN(t) model

```
Public Class Form1
    Dim H1_1 As Double
    Dim H1_2 As Double
    Dim H1_3 As Double
    Dim H1_4 As Double
    Dim H1_5 As Double
    Dim H1_6 As Double



'Calculate accumulation potential
Private Sub cmdCalculate_Click(ByVal sender As System.Object, ByVal e As
       System.EventArgs) Handles cmdCalculate.Click

       'Hidden node1
       H1_1 = Math.Tanh(0.5 * (0.463284479120175 + 0.00949247676167365 *
       Me.txtALUM.Text + 0.0126597034177741 * Me.txtCALC.Text + -0.241611053047977 *
       Me.txtFCR.Text + 0.0868278598646471 * Me.txtCOLO.Text + -0.0565284114139191 *
       Me.txtpHEstimate.Text + 0.0255275956774647 * Me.txtMAGN.Text +
       0.0000766274482116695 * Me.txtPHUS.Text + -0.617791898552024 *
       Me.txtTURB.Text))

       'Hidden node2
       H1_2 = Math.Tanh(0.5 * (0.524089025180102 + -0.0093927263963424 *
       Me.txtALUM.Text + -0.00429357092183455 * Me.txtCALC.Text + 0.184590911540848
       * Me.txtFCR.Text + -0.0975214033083568 * Me.txtCOLO.Text + -
       0.0932180605196759 * Me.txtpHEstimate.Text + -0.0679959768791871 *
       Me.txtMAGN.Text + 0.0000531507655027744 * Me.txtPHUS.Text + -
       0.591656495168582 * Me.txtTURB.Text))

       'Hidden node3
       H1_3 = Math.Tanh(0.5 * (0.592972546773058 + 0.00910191563652503 *
       Me.txtALUM.Text + -0.0125474935526051 * Me.txtCALC.Text + -0.104847701371256
       * Me.txtFCR.Text + 0.0837164595497211 * Me.txtCOLO.Text + -
       0.00928227514352046 * Me.txtpHEstimate.Text + -0.116632230957134 *
       Me.txtMAGN.Text + -0.0000363256013144583 * Me.txtPHUS.Text +
       0.534025280758758 * Me.txtTURB.Text))

       'Hidden node4
       H1_4 = Math.Tanh(0.5 * ((-1.24855417584072) + -0.0100469592521452 *
       Me.txtALUM.Text + 0.00226351139847208 * Me.txtCALC.Text + 0.160204621314887 *
       Me.txtFCR.Text + -0.0388309125044971 * Me.txtCOLO.Text + 0.14773421980523 *
       Me.txtpHEstimate.Text + 0.12642136437944 * Me.txtMAGN.Text + -
       0.0000601494808437273 * Me.txtPHUS.Text + 0.595142615935719 *
       Me.txtTURB.Text))

       'Hidden node5
       H1_5 = Math.Tanh(0.5 * ((-0.388120283539696) + -0.0101866006155669 *
       Me.txtALUM.Text + -0.00331745165021977 * Me.txtCALC.Text + 0.108475546713748
       * Me.txtFCR.Text + -0.00751084934619563 * Me.txtCOLO.Text +
       0.0618187020514616 * Me.txtpHEstimate.Text + 0.0177091783385372 *
       Me.txtMAGN.Text + 0.0000155636972814439 * Me.txtPHUS.Text + 0.180649323547993
       * Me.txtTURB.Text))

       'Hidden node6
       H1_6 = Math.Tanh(0.5 * ((-2.51392454307646) + -0.0176116384731722 *
       Me.txtALUM.Text + 0.0223087999712135 * Me.txtCALC.Text + 1.28423784732323 *
       Me.txtFCR.Text + 0.0634063145648184 * Me.txtCOLO.Text + 0.292916543065033 *
```

```
        Me.txtpHEstimate.Text + 0.0179912457938581 * Me.txtMAGN.Text + -
        0.000210485948113387 * Me.txtPHUS.Text + 0.514295980691018 * Me.txtTURB.Text))

        'Fe and Mn acummulated
        Me.txtFeandMnAccum.Text = (-0.102988239949188) + 1.65672756816735 * H1_1 +
        1.94026905238202 * H1_2 + 1.69574405883207 * H1_3 + 2.47113093924084 * H1_4 +
        -1.15005573105549 * H1_5 + -0.0517311032644249 * H1_6
End Sub



'Exit
Private Sub cmdExit_Click(ByVal sender As System.Object, ByVal e As
        System.EventArgs) Handles cmdExit.Click
          End
        End Sub
End Class
```

Appendix C: seasonal variations of customer complaints

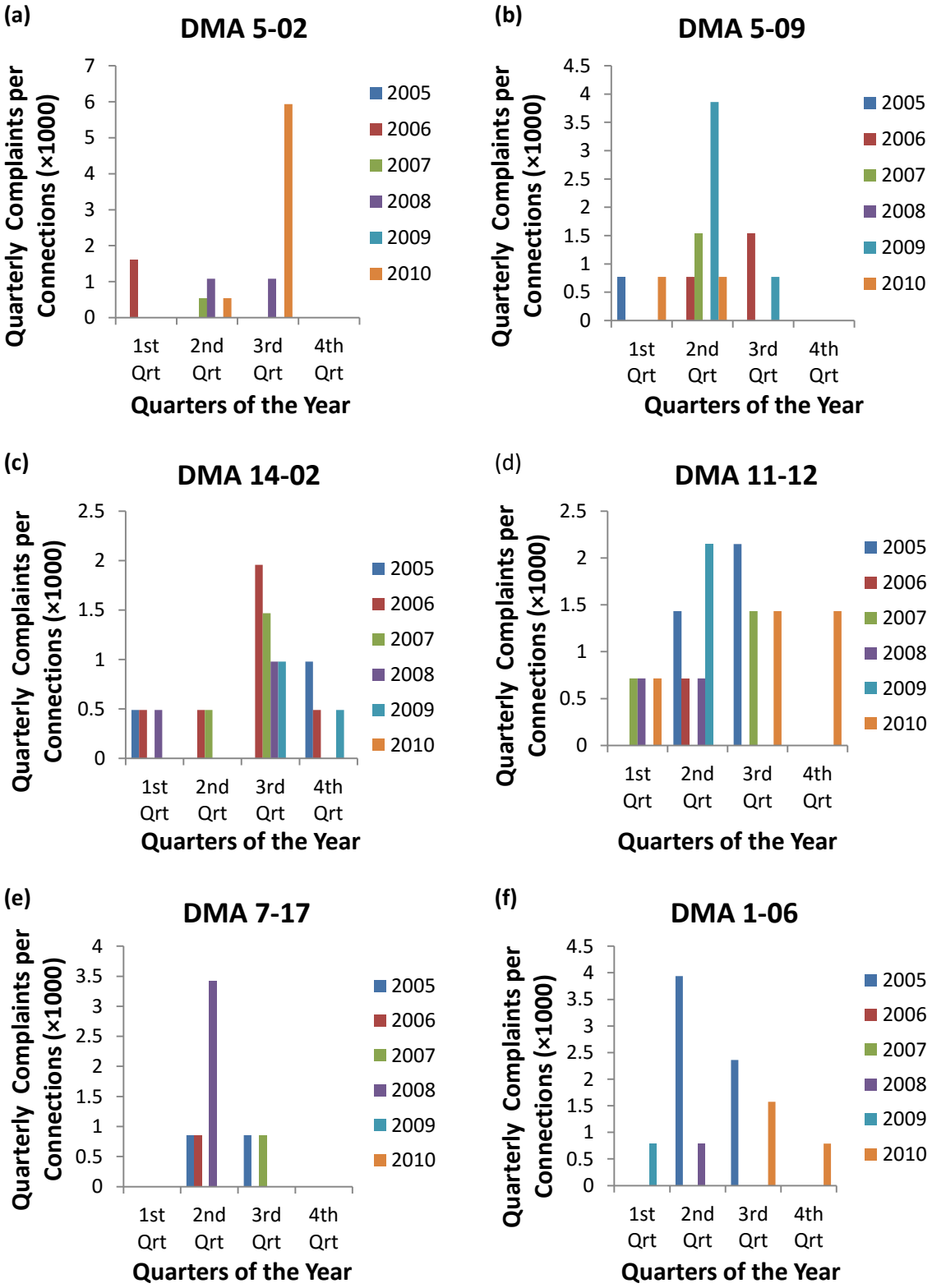


Figure B.1 Seasonal variations of customer complaints from some DMAs

Appendix D: Source code for calculating the shortest distance from reservoir to node

```
function [] = CalcDistFromReservoirToNode(InputFile)

%Load EPANET DLL
if ~libisloaded('epanet2'), loadlibrary('epanet2', 'epanet2.h'); end

%Get the excel tab
ExcelTab = InputFile(1:6);

%Open EPANET toolkit
Err = calllib('epanet2', 'ENopen', InputFile, 'report.rpt',' ');

%Supress the writing of all error messages to be written to report.rpt.
Err = calllib('epanet2', 'ENsetreport', 'MESSAGES NO');

%Calculate the number of links
Nlinks = 0;
[Err, Nlinks] = calllib('epanet2', 'ENgetcount', 2, Nlinks);

%Calculate the number of nodes
Nnodes = 0;
[Err, Nnodes] = calllib('epanet2', 'ENgetcount', 0, Nnodes);

%Get  Node Ids from EPANET
id = 'abcdefghijklmnopqrstuvwxyzaaaaa';
NodeIds = { };
for i = 1:Nnodes
    [Err, id] = calllib('epanet2', 'ENgetnodeid', i, id);
    NodeIds{i} = id;
end

len_ = 0.0; len = zeros(1,Nlinks);
S_Node_ = 0; E_Node_ = 0;
S_Node_Id_ = 'abcdefghijklmnopqrstuvwxyzaaaaa'; E_Node_Id_ =
'abcdefghijklmnopqrstuvwxyzaaaaa';
S_Node_Id = { }; E_Node_Id = { }; id = ' '; LinkIds = { };
for i = 1:Nlinks
    [Err, len_] = calllib('epanet2', 'ENgetlinkvalue', i, int32(1), len_);
    [Err, S_Node_, E_Node_]  = calllib('epanet2', 'ENgetlinknodes', i,
S_Node_, E_Node_);
    [Err, S_Node_Id_] = calllib('epanet2', 'ENgetnodeid', S_Node_,
S_Node_Id_);
    [Err, E_Node_Id_] = calllib('epanet2', 'ENgetnodeid', E_Node_,
E_Node_Id_);
    [Err, id] = calllib('epanet2', 'ENgetlinkid', i, id);

    LinkIds{i} = id;
    if len_ == 0
        len_ = 0.3;
    end

    len(i) = len_;
    S_Node_Id{i} = S_Node_Id_;
    E_Node_Id{i} = E_Node_Id_;
end
```

```
%Give nodes integer names
NodeNum_ = 0;
for i = 1:Nnodes
    NodeNum_ = NodeNum_ +1;
    NodeNum(i) = NodeNum_;
end

%Givelink integer names
LinkNum_ = 0;
for i = 1:Nlinks
    LinkNum_ = LinkNum_ +1;
    LinkNum(i) = LinkNum_;
end

%Put node properties in a Matrix
NodeProperties = [NodeIds', num2cell(NodeNum')];

%Get the integer values for StartNodes
SNodeNum = zeros(1,Nlinks);
for i=1:Nlinks
    [SNodeNum_] = vlookup(NodeProperties, cell2mat(S_Node_Id(i)), 2, 1);
    SNodeNum(i) = cell2mat(SNodeNum_);
end

%Get the numeric values for EndNodes
ENodeNum = zeros(1,Nlinks);
for i=1:Nlinks
    [ENodeNum_] = vlookup(NodeProperties, cell2mat(E_Node_Id(i)), 2, 1);
    ENodeNum(i) = cell2mat(ENodeNum_);
end

%Convert the number of links and nodes from integer to double
dblNnodes = double(Nnodes);
dblNlinks = double(Nlinks);
%Create a sparse matrix and force it into a square matrix
Msparse = sparse(SNodeNum', ENodeNum', len',  dblNnodes, dblNnodes,
dblNlinks);

%Create a bidirectional link between the nodes
M = Msparse + Msparse';
 %Calculate maximum velocity and shear stress
[maxVel, maxShear, maxFlw, maxHl] = Hydraul
%Create an adjancy matrix
MSparseAdj = (M>0);
Madj = full(MSparseAdj);

%Find the dead ends nodes (in integers)
DeadEndsAllInt=leaf_nodes(Madj);

%calculate the number of dead ends
NDeadEndsAll = numel(DeadEndsAllInt);

%List the the dead end node names (original ie. in characters) which
comprises of nodes tanks
%and reservoirs
for i=1:NDeadEndsAll
    [DeadEndNodesAll_] = vlookup(NodeProperties, DeadEndsAllInt(i), 1, 2);
    DeadEndNodesAllChar(i) = DeadEndNodesAll_;
end
```

```
%Find all reservoirs and tanks
ReservoirIds_= 'abcdefghijklmnopqrstuvwxyzaaaaa'; ReservoirIds = { };
TankIds_= 'abcdefghijklmnopqrstuvwxyzaaaaa'; TankIds = { };
NodeType_ = int32(0);
for i = 1:Nnodes
    %type=int32(0);
    [Err, NodeType_] = calllib('epanet2', 'ENgetnodetype', i, NodeType_);
    NodeType(i) = NodeType_;
    if NodeType_ == 1
        [Err, ReservoirIds_] = calllib('epanet2', 'ENgetnodeid', i,
ReservoirIds_);
        ReservoirIds{i} = ReservoirIds_;
    elseif NodeType_ == 2
        [Err, TankIds_] = calllib('epanet2', 'ENgetnodeid', i, TankIds_);
        TankIds{i} = TankIds_;
    end
end

%Remove empty cells and add the ReservoirIds to the TankIds
ReservoirIds = ReservoirIds(~cellfun('isempty',ReservoirIds));
TankIds = TankIds(~cellfun('isempty',TankIds));
ReservoirsAndTanksIDs = {ReservoirIds{:}, TankIds{:}};

%Remove Reservoirs and Tanks from dead ends
DeadEndNodes = setdiff(DeadEndNodesAllChar,ReservoirsAndTanksIDs);

%Find the corresponding numeric values for the dead end
NDeadEndNodes = numel(DeadEndNodes);
for i=1:NDeadEndNodes
    [DeadEndNodesNum_] = vlookup(NodeProperties,
cell2mat(DeadEndNodes(i)), 2, 1);
    DeadEndNodesNum(i) = DeadEndNodesNum_;
end

%Find the numeric values for the reservoirs and tanks
NReservoirsAndTanks = numel(ReservoirsAndTanksIDs);
for i=1:NReservoirsAndTanks
    [ReservoirsAndTanksNum_] = vlookup(NodeProperties,
cell2mat(ReservoirsAndTanksIDs(i)), 2, 1);
    ReservoirsAndTanksNum(i) = ReservoirsAndTanksNum_;
end

%Calculate the shortest distance between reservoirs and nodes
[ShortDistReservoirToNode, CorrShortestNodeNum, CorrReservoirsAndTanksNum]
= GetDistSampledToDeadEnd(M,ReservoirsAndTanksNum,NodeNum);

%Find the corresponding character values for the nodes with shortest
distance from reservoir
disp('Finding the corresponding character values for the nodes with
shortest distance from reservoir');
NCorrShortestNodeNum = numel(CorrShortestNodeNum);
for i=1:NCorrShortestNodeNum
%for i=1:Nnodes
    CorrShortestNodeNum = CorrShortestNodeNum';
    [CorrShortestNodeChar_] = vlookup(NodeProperties,
CorrShortestNodeNum(i), 1, 2);
    CorrShortestNodeChar(i) = CorrShortestNodeChar_;
    disp(['Calculating ', num2str(i), ' Out of ',
num2str(NCorrShortestNodeNum)]);
```

```
end
disp('Finished Finding the corresponding character values for the nodes
with shortest distance from reservoir');

%Find the corresponding character values for reservoirs and tanks
disp('Finding the corresponding character values for reservoirs and
tanks');
NCorrReservoirsAndTanksNum = numel(CorrReservoirsAndTanksNum);
for i=1:NCorrReservoirsAndTanksNum
%for i=1:NReservoirsAndTanks
    CorrReservoirsAndTanksNum = CorrReservoirsAndTanksNum';
    [CorrReservoirsAndTanksChar_] = vlookup(NodeProperties,
CorrReservoirsAndTanksNum(i), 1, 2);
    CorrReservoirsAndTanksChar(i) = CorrReservoirsAndTanksChar_;
    disp(['Calculating ', num2str(i), ' Out of ',
num2str(NCorrReservoirsAndTanksNum)]);
end
disp('Finished Finding the corresponding character values for reservoirs
and tanks');

%Write the values into excel sheet
rangeStr = sprintf('A2:A%d', length(CorrReservoirsAndTanksNum)+1);
xlswrite('ReservoirNodesCalculations.xlsx',CorrReservoirsAndTanksChar',Ex
celTab,rangeStr);

rangeStr = sprintf('B2:B%d', length(CorrShortestNodeNum)+1);
xlswrite('ReservoirNodesCalculations.xlsx',CorrShortestNodeChar',ExcelTab
,rangeStr);

rangeStr = sprintf('C2:C%d', length(ShortDistReservoirToNode)+1);
xlswrite('ReservoirNodesCalculations.xlsx',ShortDistReservoirToNode',Exce
lTab,rangeStr);

%Close EPANET toolkit
Err = calllib('epanet2', 'ENclose');

%Unload EPANET DLL.
unloadlibrary('epanet2');

return



function [ShortDistReservoirToNode, CorrShortestNodeNum,
CorrReservoirsAndTanksNum] =
GetDistSampledToDeadEnd(M,ReservoirsAndTanksNum,NodeNum)
   %%%%%%-----  This function calculates the shortest distance from the
sampled nodes to dead ends -----%%%%
    NReservoirsAndTanksNum = numel(ReservoirsAndTanksNum); NNodeNum =
numel(NodeNum);
    CorrShortestNodeNum = zeros(1,NNodeNum); CorrShortestNodeNum_ =
zeros(1,NNodeNum);
    CorrReservoirsAndTanksNum = zeros(1,NNodeNum);
CorrReservoirsAndTanksNum_ = zeros(1,NNodeNum);
    ShortDistReservoirToNode = zeros(1,NNodeNum);

    for i = 1:NReservoirsAndTanksNum
          disp(['Calculating ReservoirsAndTanks: ', num2str(i), ' Out of
', num2str(NReservoirsAndTanksNum)]);
```

```
            ShortDistReservoirToNode__ = zeros(1,NNodeNum);
            for j = 1:NNodeNum
                [ShortDistReservoirToNode_, path_, pred_] =
graphshortestpath(M,cell2mat(ReservoirsAndTanksNum(i)),NodeNum(j));
                ShortDistReservoirToNode__(j) =
ShortDistReservoirToNode__(j) + ShortDistReservoirToNode_;
                CorrReservoirsAndTanksNum_(j) =
cell2mat(ReservoirsAndTanksNum(i));
                CorrShortestNodeNum_(j)= NodeNum(j);
            end
            if i==1
                CorrReservoirsAndTanksNum = CorrReservoirsAndTanksNum_;
                CorrShortestNodeNum = CorrShortestNodeNum_;
                ShortDistReservoirToNode = ShortDistReservoirToNode__;
            else
            CorrReservoirsAndTanksNum =
[CorrReservoirsAndTanksNum,CorrReservoirsAndTanksNum_];
            CorrShortestNodeNum =
[CorrShortestNodeNum,CorrShortestNodeNum_];
            ShortDistReservoirToNode =
[ShortDistReservoirToNode,ShortDistReservoirToNode__];
            end
    end
 return
```

Appendix E: Source code to determine which of the reservoirs / tanks supply the nodes with water

```
function SupplySources = CalcPathDelaysResToNode1(epainpfile, outputfile)

%Open epanet library
if ~libisloaded('epanet2'), loadlibrary('epanet2', 'epanet2.h'); end

%Open EPANET toolkit
Err = calllib('epanet2', 'ENopen', epainpfile, 'report.rpt',' ');
%***Added****

%Supress the writing of all error messages to be written to report.rpt.
Err = calllib('epanet2', 'ENsetreport', 'MESSAGES NO'); %***Added****

%Retrieve the number of nodes in the network (junctions+reservoirs+tanks)
Nnodes = 0; Err =0;
[Err, Nnodes] = calllib('epanet2', 'ENgetcount', 0, Nnodes);

%Retrieve the number of tanks (reservoirs+tanks) in the network
Ntanks = 0; Err =0;
[Err, Ntanks] = calllib('epanet2', 'ENgetcount', 1, Ntanks);


%NOTE: epanet numbers juctions from 1 to Njuncs
%and (reservoirs + tanks) from (Njuncs+1) to Nnodes
%Calculate No. of Junctions in the network
Njuncs = Nnodes - Ntanks;

%use a large sample time to achieve steady state
sampletime = 241;

%Complete hydraulic analysis and save hydraulics data
hyddata = 'hydraulicsfile';
Err = calllib('epanet2', 'ENsettimeparam', 0, (sampletime+40)*3600); %set
duration EN_DURATION
Err = calllib('epanet2', 'ENsolveH');
Err = calllib('epanet2', 'ENsavehydfile', hyddata);

disp('Hydraulic analysis is Completed.....');
disp('  ');

%Call the water quality function for each junction (output node)
SupplySources = zeros(Njuncs, Ntanks);

for i = 1:Njuncs
    %Retrieve id of junction i(output node)
    outputid = 'abcdefghijklmnopqrstuvwxyzaaaaa';
    [Err, outputid] = calllib('epanet2', 'ENgetnodeid', i, outputid);
    disp(['Calculating Delays for junction: ', num2str(i), ' Out of ',
num2str(Njuncs)]);

    %find which source nodes (reservoirs and tanks) supplying water to
output node i (junction)
    SourcesSupplyingOrNot = zeros(1,Ntanks);  %initially assume no source
is supplying (0: false)
```

```
    for j = 1:Ntanks
        %retrie the id of source node j
        sourceid = 'abcdefghijklmnopqrstuvwxyzaaaaa';
        [Err, sourceid] = calllib('epanet2', 'ENgetnodeid', j+Njuncs,
sourceid);

        %Check if the source node j is supplying water to output node i
        IsSupplying = CalcPathDelaysResToNode2(hyddata, sourceid,
outputid, sampletime);

        %if source node j is supply node to junction i, store that
information
        SourcesSupplyingOrNot(j) = IsSupplying; % 1 means true (supplying)
and 0 means false
    end

    %Store whether each source node is suplying or not to junction i
    SupplySources(i,:) = SourcesSupplyingOrNot;
end

%Print output to a file
fid = fopen(outputfile, 'w');
fprintf(fid, '%s', 'JuncId');
for j = Njuncs+1:Nnodes
    sourceid = 'abcdefghijklmnopqrstuvwxyzaaaaa';
    [Err, sourceid] = calllib('epanet2', 'ENgetnodeid', j, sourceid);
    fprintf(fid, '\t%s', sourceid);
end
for i = 1:Njuncs %***Changed****
    juncid = 'abcdefghijklmnopqrstuvwxyzaaaaa';
    [Err, juncid] = calllib('epanet2', 'ENgetnodeid', i, juncid);
    fprintf(fid, '\n%s', juncid);
    fprintf(fid, '\t%d', SupplySources(i,:));
end
fclose(fid);

disp('  ');
disp('Completed calculating Delays....');

%Close & unload epanet library
Err = calllib('epanet2', 'ENclose');
unloadlibrary('epanet2');
return
```

Appendix F: Source code for plotting the ANN(t,ψ) risk maps

```
function [] = RiskMaps_Hierarchical_MShear_VarShear_Wage_DFS(WSZ)
ExcelTab = WSZ;
%%%%%   MaxShear and VarShear input Parameters   %%%%%%
FeAndMnAcummPFuzzy_MaxShear_VarShear_Input=readfis('FeandMnAcummPotential
_Shear_VarShear.fis');
%Read the input variables from excel sheet
InputVariables_MaxShear_VarShear = xlsread('TableForRiskMaps.xlsx',
ExcelTab, 'E:F');
%compute the Accumulation potential
FeAndMnAcummPInterm_MaxShear_VarShear=evalfis(InputVariables_MaxShear_Var
Shear,FeAndMnAcummPFuzzy_MaxShear_VarShear_Input);
%Write the values of the Accumulation potential into excel sheet
rangeStr = sprintf('BN2:BN%d',
length(InputVariables_MaxShear_VarShear)+1);
xlswrite('TableForRiskMaps.xlsx',FeAndMnAcummPInterm_MaxShear_VarShear,Ex
celTab,rangeStr);

%%%%%   WaterAge and Distance froemSource input Parameters   %%%%%%
FeAndMnAcummPFuzzy_WAge_DFS_Input=readfis('FeandMnAcummPotential_WaterAge
_DistFrmSource.fis');
%Read the input variables from excel sheet
InputVariables_WAge_DFS = xlsread('TableForRiskMaps.xlsx', ExcelTab,
'G:H');
%compute the Accumulation potential
FeAndMnAcummPInterm_WAge_DFS=evalfis(InputVariables_WAge_DFS,FeAndMnAcumm
PFuzzy_WAge_DFS_Input);
%Write the values of the Accumulation potential into excel sheet
rangeStr = sprintf('BO2:BO%d', length(InputVariables_WAge_DFS)+1);
xlswrite('TableForRiskMaps.xlsx',FeAndMnAcummPInterm_WAge_DFS,ExcelTab,ra
ngeStr);

ExcelTab = WSZ;
%%%%%  Intermediate Parameters and Output  %%%%%%
FeAndMnAcummPFuzzy_MaxShearVarShear_WAgeDFS_Interm=readfis('FeandMnAcummP
otential_Hierarchical_Shear_VarShear_WaterAge_DistFrmSource.fis');
%Read the input variables from excel sheet
IntermVariables_MaxShearVarShear_WAgeDFS =
xlsread('TableForRiskMaps.xlsx', ExcelTab, 'BN:BO');
%compute the Accumulation potential
FeAndMnAcummPOutPut_MaxShearVarShear_WAgeDFS=evalfis(IntermVariables_MaxS
hearVarShear_WAgeDFS,FeAndMnAcummPFuzzy_MaxShearVarShear_WAgeDFS_Interm);
%Write the values of the Accumulation potential into excel sheet
rangeStr = sprintf('BP2:BP%d',
length(IntermVariables_MaxShearVarShear_WAgeDFS)+1);
xlswrite('TableForRiskMaps.xlsx',FeAndMnAcummPOutPut_MaxShearVarShear_WAg
eDFS,ExcelTab,rangeStr);

%Read the x and y-coordinates (all nodes) from excel sheet
x_ = xlsread('TableForRiskMaps.xlsx', ExcelTab, 'B:B');
y_ = xlsread('TableForRiskMaps.xlsx', ExcelTab, 'C:C');
z_ = xlsread('TableForRiskMaps.xlsx', ExcelTab, 'BP:BP'); %Risk parameter

%Read the x and y-coordinates for reservoirs and tanks
x_rt = xlsread('TableForRiskMaps.xlsx', ExcelTab, 'BC:BC');
y_rt = xlsread('TableForRiskMaps.xlsx', ExcelTab, 'BD:BD');
```

```
%Read the x and y-coordinates and Customer complaints nodes from excel
sheet
xCC = xlsread('TableForRiskMaps.xlsx', ExcelTab, 'AH:AH');
yCC = xlsread('TableForRiskMaps.xlsx', ExcelTab, 'AI:AI');
zCC = xlsread('TableForRiskMaps.xlsx', ExcelTab, 'AG:AG'); %Customer
complaints

disp('Finished Reading');

%Plot the network Nodes
Fig.;
%Plot the reservoirs
scatter(x_rt,y_rt,250,'s', 'markerfacecolor',[1 0 1]);
%scatter(x_rt,y_rt,1000,'.');
hold on;
scatter(x_, y_, 181, z_, '.');
colorbar;
hold on;
%Plot the network diagram for risk map
PlotNetworkDiagram(ExcelTab);
title(['Predicted Fe and Mn Accumulation Potential Risk Map using Input
Parameters Maximum Shear Stress, Variation in Shear Stress, Water Age and
Distance from Source  for ',WSZ]);
%Remove x and y ticks
set(gca,'xtick',[]);
set(gca,'ytick',[]);
hold off;

%Plot the contour map - Customer complaints
Fig.;
%Plot the reservoirs
scatter(x_rt,y_rt,250,'s', 'markerfacecolor',[1 0 1]);
%scatter(x_rt,y_rt,1000,'.');
hold on;
PlotContourMap(xCC,yCC,zCC)
colorbar;
hold on;
%Plot the network Nodes for contour map - Customer complaints
plot(x_, y_,'k.');
hold on;
%Plot the network diagram for contour map - Customer complaints
PlotNetworkDiagram(ExcelTab);
title(['Customer Complaints Contour Map for ',WSZ]);
%Remove x and y ticks
set(gca,'xtick',[]);
set(gca,'ytick',[]);
hold off;

%Plot the network Nodes (All Nodes) for risk map - Customer complaints
Fig.;
%Plot the reservoirs
scatter(x_rt,y_rt,250,'s', 'markerfacecolor',[1 0 1]);
%scatter(x_rt,y_rt,1000,'.');
hold on;
%Plot the network Nodes for contour map - Customer complaints
plot(x_, y_,'k.');
hold on;
scatter(xCC, yCC, 250, zCC, '.');
colorbar;
hold on;
```

```
%Plot the network diagram for risk map
PlotNetworkDiagram(ExcelTab);
title(['Customer Complaints Risk Map for ',WSZ]);
%Remove x and y ticks
set(gca,'xtick',[]);
set(gca,'ytick',[]);
hold off;

%Plot the 3D risk map - Customer complaints
Fig.;
%Plot the reservoirs
scatter(x_rt,y_rt,250,'s', 'markerfacecolor',[1 0 1]);
%scatter(x_rt,y_rt,1000,'.');
hold on;
%Plot the network Nodes for contour map - Customer complaints
plot(x_, y_,'k.');
hold on;
%Plot the network diagram for risk map
PlotNetworkDiagram(ExcelTab);
stem3 (xCC, yCC, zCC, 'marker', 'none', 'linewidth',3)
hold on;
title(['Customer Complaints 3D Risk Map for ',WSZ]);
%Remove x and y ticks
set(gca,'xtick',[]);
set(gca,'ytick',[]);
hold off;
return

function PlotContourMap(x,y,z)
a = size([x,y,z]);
b=a(:,1);
xlin=linspace(min(x),max(x),b);    %xlin=linspace(min(x_),max(x_),b);
ylin=linspace(min(y),max(y),b);    %ylin=linspace(min(y_),max(y_),b);
[X,Y]=meshgrid(xlin,ylin);
uint8(x); uint8(y); uint8(z); uint8(X); uint8(Y);
Z=griddata(x,y,z,X,Y);
%mesh(X,Y,Z);
contourf(X,Y,Z);
%colorbar;
return




function PlotNetworkDiagram(ExcelTab)

%Read the node properties (x and y-coordinates) from excel sheet
XYCoor= xlsread('TableForRiskMaps.xlsx', ExcelTab, 'B:C');
%Give nodes integer names
NodeNum_ = 0; Nnodes = length(XYCoor);
for i = 1:Nnodes
    NodeNum_ = NodeNum_ +1;
    NodeNum(i) = NodeNum_;
end
Node_XY_MatrixNum = [NodeNum' XYCoor];

%Read Node Ids
[~,NodeIds,~] = xlsread('TableForRiskMaps.xlsx', ExcelTab, 'A:A');
```

```
NodeIds(1) = [];

%Put node properties in a Matrix
NodeMatrix = [NodeIds, num2cell(NodeNum')];

%Read the link properties from excel sheet
[~,LinkProperties,~] = xlsread('TableForRiskMaps.xlsx', ExcelTab,
'AA:AC');
LinkProperties(1,:)=[];
S_Node_Id = LinkProperties(:,2);
%S_Node_Id = S_Node_Id';
E_Node_Id = LinkProperties(:,3);
%E_Node_Id = E_Node_Id';
%Givelink integer names
LinkNum_ = 0; Nlinks = int32(length(LinkProperties));
for i = 1:Nlinks
    LinkNum_ = LinkNum_ +1;
    LinkNum(i) = LinkNum_;
end

%Get the corresponding integer values for StartNodes
SNodeNum = zeros(1,Nlinks);
for i=1:Nlinks
    [SNodeNum_] = vlookup(NodeMatrix, cell2mat(S_Node_Id(i)), 2, 1);
    SNodeNum(i) = cell2mat(SNodeNum_);
end

%Get the corresponding numeric values for EndNodes
ENodeNum = zeros(1,Nlinks);
for i=1:Nlinks
    [ENodeNum_] = vlookup(NodeMatrix, cell2mat(E_Node_Id(i)), 2, 1);
    ENodeNum(i) = cell2mat(ENodeNum_);
end

LinkMatrixNum = [LinkNum' SNodeNum' ENodeNum'];
%Plot network diagram
for i = 1:Nlinks
%    if (i <= Nnodes)
%        text(Node_XY_MatrixNum(i,2),Node_XY_MatrixNum(i,3),[' '
num2str(i)]);
%    end

plot(Node_XY_MatrixNum(LinkMatrixNum(i,2:3)',2),Node_XY_MatrixNum(LinkMat
rixNum(i,2:3)',3),'k');
end
```

Appendix G: Source code to read data for fuzzy model

```
function [inputData, FeAndMnMeasured, fSys, ruleData_ante, ruleData_cons,
ruleData_wt, ruleData_conn, outputCol] = ReadData()
%%%%%% NOTE!!!!
%Run this function 1st before you run the Genetic Algorithm function
%(GAfcn)to write the values consequents into excel sheet

 ExcelTab = 'WSZ2_YrAvg';

%Read the column to determine which rules need to be optimised
outputCol = xlsread('_OptimisedRules.xlsx', 'Rules4', 'C:C');

%Read actual Fe and Mn accumulation potential
FeAndMnMeasured = xlsread('_TableForRiskMaps.xlsx', 'WSZ2_YrAvgOptAll',
'AC:AC');

%FeAndMnMeasured = xlsread('_TableForRiskMaps.xlsx', ExcelTab, 'AS:AS');

 disp('Half way through');

inputData =  xlsread('_TableForRiskMaps.xlsx', 'WSZ2_YrAvgOptAll',
'F2:Y219');

 %Put the fuzzy files into a cell
fSys_ =
{'_Chemical_Oxidation.fis','_Corrosion_With_Pipe_Age.fis','_Sorption.fis'
,'_Shear_Effect.fis','_Distance_Effect.fis','_Chemical_Effect.fis', ...

'_Biological_Effect_With_WaterAge4.fis','_Hydraulic_Effect.fis','_Fe&Mn_A
ccum_Potential.fis'};

%Create an empty cell to store the fuzzy system.
fSys = cell(9,1);
for i=1:9
    fSys{i}=readfis(fSys_{i});
end

%Create an empty cell to store rules for the fuzzy system
ruleData_cons = cell(9,1);
ruleData_ante = cell(9,1);
ruleData_conn = cell(9,1);
ruleData_wt = cell(9,1);

%%%%%%%%% Read the 1st fuzzy sub-system (Chemical Oxidation
Parameters)%%%%%%%%%%%
Chemical_Oxidation_Fuzzy=readfis(fSys_{1});
%determine the number of rules
n = getfis(Chemical_Oxidation_Fuzzy,'numRules');
rules1_cons = zeros(n,1);
rules1_wt = zeros(n,1);
rules1_conn = zeros(n,1);
%determine the number of antecedents
m = length(Chemical_Oxidation_Fuzzy.rule(1).antecedent);
rules1_ante = zeros(n,m);
for j=1:n
    rules1_cons(j) = Chemical_Oxidation_Fuzzy.rule(j).consequent;
    rules1_ante(j,:)= Chemical_Oxidation_Fuzzy.rule(j).antecedent;
```

```
    rules1_wt(j) = Chemical_Oxidation_Fuzzy.rule(j).weight;
    rules1_conn(j) = Chemical_Oxidation_Fuzzy.rule(j).connection;
end
ruleData_cons{1} = rules1_cons;
ruleData_ante{1} = rules1_ante;
ruleData_wt{1} = rules1_wt;
ruleData_conn{1} = rules1_conn;

%%%%%%%%%% Read the 2nd fuzzy sub-system (Corrosion Parameters)
%%%%%%%%%%%%%%%
Corrosion_Fuzzy=readfis(fSys_{2});
%determine the number of rules
n = getfis(Corrosion_Fuzzy,'numRules');
rules2_cons = zeros(n,1);
rules2_wt = zeros(n,1);
rules2_conn = zeros(n,1);
%determine the number of antecedents
m = length(Corrosion_Fuzzy.rule(1).antecedent);
rules2_ante = zeros(n,m);
for j=1:n
    rules2_cons(j) = Corrosion_Fuzzy.rule(j).consequent;
    rules2_ante(j,:)= Corrosion_Fuzzy.rule(j).antecedent;
    rules2_wt(j) = Corrosion_Fuzzy.rule(j).weight;
    rules2_conn(j) = Corrosion_Fuzzy.rule(j).connection;
end
ruleData_cons{2} = rules2_cons;
ruleData_ante{2} = rules2_ante;
ruleData_wt{2} = rules2_wt;
ruleData_conn{2} = rules2_conn;

%%%%%%%%%% Read the 3rd fuzzy sub-system (Sorption Parameters)
%%%%%%%%%%%%%%%%
Sorption_Fuzzy=readfis(fSys_{3});
%determine the number of rules
n = getfis(Sorption_Fuzzy,'numRules');
rules3_cons = zeros(n,1);
rules3_wt = zeros(n,1);
rules3_conn = zeros(n,1);
%determine the number of antecedents
m = length(Sorption_Fuzzy.rule(1).antecedent);
rules3_ante = zeros(n,m);
for j=1:n
    rules3_cons(j) = Sorption_Fuzzy.rule(j).consequent;
    rules3_ante(j,:)= Sorption_Fuzzy.rule(j).antecedent;
    rules3_wt(j) = Sorption_Fuzzy.rule(j).weight;
    rules3_conn(j) = Sorption_Fuzzy.rule(j).connection;
end
ruleData_cons{3} = rules3_cons;
ruleData_ante{3} = rules3_ante;
ruleData_wt{3} = rules3_wt;
ruleData_conn{3} = rules3_conn;

%%%%%%%%%% Read the 4th fuzzy sub-system (Shear Stress Effect Parameters)
%%%%%%%%%%%%%%%%%%%%%%
Shear_Stress_Effect_Fuzzy=readfis(fSys_{4});
%determine the number of rules
n = getfis(Shear_Stress_Effect_Fuzzy,'numRules');
rules4_cons = zeros(n,1);
rules4_wt = zeros(n,1);
rules4_conn = zeros(n,1);
%determine the number of antecedents
```

```
m = length(Shear_Stress_Effect_Fuzzy.rule(1).antecedent);
rules4_ante = zeros(n,m);
for j=1:n
    rules4_cons(j) = Shear_Stress_Effect_Fuzzy.rule(j).consequent;
    rules4_ante(j,:)= Shear_Stress_Effect_Fuzzy.rule(j).antecedent;
    rules4_wt(j) = Shear_Stress_Effect_Fuzzy.rule(j).weight;
    rules4_conn(j) = Shear_Stress_Effect_Fuzzy.rule(j).connection;
end
ruleData_cons{4} = rules4_cons;
ruleData_ante{4} = rules4_ante;
ruleData_wt{4} = rules4_wt;
ruleData_conn{4} = rules4_conn;

%%%%%%%%%% Read the 5th fuzzy sub-system (Distance Effect Parameters)
%%%%%%%%%%%%%%%%
Distance_Stress_Effect_Fuzzy=readfis(fSys_{5});
%determine the number of rules
n = getfis(Distance_Stress_Effect_Fuzzy,'numRules');
rules5_cons = zeros(n,1);
rules5_wt = zeros(n,1);
rules5_conn = zeros(n,1);
%determine the number of antecedents
m = length(Distance_Stress_Effect_Fuzzy.rule(1).antecedent);
rules5_ante = zeros(n,m);
for j=1:n
    rules5_cons(j) = Distance_Stress_Effect_Fuzzy.rule(j).consequent;
    rules5_ante(j,:)= Distance_Stress_Effect_Fuzzy.rule(j).antecedent;
    rules5_wt(j) = Distance_Stress_Effect_Fuzzy.rule(j).weight;
    rules5_conn(j) = Distance_Stress_Effect_Fuzzy.rule(j).connection;
end
ruleData_cons{5} = rules5_cons;
ruleData_ante{5} = rules5_ante;
ruleData_wt{5} = rules5_wt;
ruleData_conn{5} = rules5_conn;

%%%%%%%%%% Read the 6th fuzzy sub-system (Chemical Effect Intermidiate
Parameters) %%%%%%%%%%%%%%%%
Chemical_Effect_Fuzzy=readfis(fSys_{6});
%determine the number of rules
n = getfis(Chemical_Effect_Fuzzy,'numRules');
rules6_cons = zeros(n,1);
rules6_wt = zeros(n,1);
rules6_conn = zeros(n,1);
%determine the number of antecedents
m = length(Chemical_Effect_Fuzzy.rule(1).antecedent);
rules6_ante = zeros(n,m);
for j=1:n
    rules6_cons(j) = Chemical_Effect_Fuzzy.rule(j).consequent;
    rules6_ante(j,:)= Chemical_Effect_Fuzzy.rule(j).antecedent;
    rules6_wt(j) =  Chemical_Effect_Fuzzy.rule(j).weight;
    rules6_conn(j) =  Chemical_Effect_Fuzzy.rule(j).connection;
end
ruleData_cons{6} = rules6_cons;
ruleData_ante{6} = rules6_ante;
ruleData_wt{6} = rules6_wt;
ruleData_conn{6} = rules6_conn;

%%%%%%%%%% Read the 7th fuzzy sub-system (Biological Effect Intermidiate
Parameters) %%%%%%%%%%%%%%%%
Biological_Effect_Fuzzy=readfis(fSys_{7});
%determine the number of rules
```

```
n = getfis(Biological_Effect_Fuzzy,'numRules');
rules7_cons = zeros(n,1);
rules7_wt = zeros(n,1);
rules7_conn = zeros(n,1);
%determine the number of antecedents
m = length(Biological_Effect_Fuzzy.rule(1).antecedent);
rules7_ante = zeros(n,m);
for j=1:n
    rules7_cons(j) = Biological_Effect_Fuzzy.rule(j).consequent;
    rules7_ante(j,:)= Biological_Effect_Fuzzy.rule(j).antecedent;
    rules7_wt(j) = Chemical_Oxidation_Fuzzy.rule(j).weight;
    rules7_conn(j) = Chemical_Oxidation_Fuzzy.rule(j).connection;
end
ruleData_cons{7} = rules7_cons;
ruleData_ante{7} = rules7_ante;
ruleData_wt{7} = rules7_wt;
ruleData_conn{7} = rules7_conn;

%%%%%%%%%%%%%%% Read the 8th fuzzy sub-system (Hydraulic Effect
Intermidiate Parameters) %%%%%%%%%%%%%%%%%
Hydraulic_Effect_Fuzzy=readfis(fSys_{8});
%determine the number of rules
n = getfis(Hydraulic_Effect_Fuzzy,'numRules');
rules8_cons = zeros(n,1);
rules8_wt = zeros(n,1);
rules8_conn = zeros(n,1);
%determine the number of antecedents
m = length(Hydraulic_Effect_Fuzzy.rule(1).antecedent);
rules8_ante = zeros(n,m);
for j=1:n
    rules8_cons(j) = Hydraulic_Effect_Fuzzy.rule(j).consequent;
    rules8_ante(j,:)= Hydraulic_Effect_Fuzzy.rule(j).antecedent;
    rules8_wt(j) = Hydraulic_Effect_Fuzzy.rule(j).weight;
    rules8_conn(j) = Hydraulic_Effect_Fuzzy.rule(j).connection;
end
ruleData_cons{8} = rules8_cons;
ruleData_ante{8} = rules8_ante;
ruleData_wt{8} = rules8_wt;
ruleData_conn{8} = rules8_conn;

%%%%%%%%%%%% Read the 9th fuzzy sub-system (Accumulation Potential
Parameters) %%%%%%%%%%%%%%
Accumulation_Potential_Fuzzy=readfis(fSys_{9});
%determine the number of rules
n = getfis(Accumulation_Potential_Fuzzy,'numRules');
rules9_cons = zeros(n,1);
rules9_wt = zeros(n,1);
rules9_conn = zeros(n,1);
%determine the number of antecedents
m = length(Accumulation_Potential_Fuzzy.rule(1).antecedent);
rules9_ante = zeros(n,m);
for j=1:n
    rules9_cons(j) = Accumulation_Potential_Fuzzy.rule(j).consequent;
    rules9_ante(j,:)= Accumulation_Potential_Fuzzy.rule(j).antecedent;
    rules9_wt(j) = Accumulation_Potential_Fuzzy.rule(j).weight;
    rules9_conn(j) = Accumulation_Potential_Fuzzy.rule(j).connection;
end
ruleData_cons{9} = rules9_cons;
ruleData_ante{9} = rules9_ante;
ruleData_wt{9} = rules9_wt;
ruleData_conn{9} = rules9_conn;
```

```
%Put the fuzzy system into a cell
 fSys = {Chemical_Oxidation_Fuzzy, Corrosion_Fuzzy, Sorption_Fuzzy,
Shear_Stress_Effect_Fuzzy, Distance_Stress_Effect_Fuzzy, ...
     Chemical_Effect_Fuzzy, Biological_Effect_Fuzzy,
Hydraulic_Effect_Fuzzy, Accumulation_Potential_Fuzzy};

%Write the values consequents into excel sheet
myCons = cell2mat(ruleData_cons);
rangeStr = sprintf('D2:D%d', length(myCons)+1);
xlswrite('_OptimisedRules.xlsx',myCons,'Rules4',rangeStr);

return
```

Appendix H: Source code to evaluate the fuzzy system

```
function [FeAndMnAccumP_Predicted] = EvalFuzzySytem(fSys,inputData)

%%%%%%%%% Evaluate the 1st fuzzy sub-system (Chemical Oxidation
Parameters) %%%%%%%%%%%
%Compute the Chemical Oxidation
inputData(:,5)=evalfis(inputData(:,[1 2 3 4]),fSys{1});

%%%%%%%%%% Evaluate the 2nd fuzzy sub-system (Corrosion Parameters)
%%%%%%%%%%%%%%%
%Compute the Corrosion
inputData(:,8)=evalfis(inputData(:,[5 6 7]),fSys{2});

%%%%%%%%%%%%%%%%%% Evaluate the 3rd fuzzy sub-system (Sorption Parameters)
%%%%%%%%%%%%%%%
%Compute the Sorption
inputData(:,11)=evalfis(inputData(:,[1 9 10]),fSys{3});

%%%%%%%%%% Evaluate the 4th fuzzy sub-system (Shear Stress Effect
Parameters) %%%%%%%%%%%%%%%%%%%%%%%%
inputData(:,17)=evalfis(inputData(:,[15 16]),fSys{4});

%%%%%%%%%% Evaluate the 5th fuzzy sub-system (Distance Effect Parameters)
%%%%%%%%%%%%%%%
%Compute the Distance Effect
inputData(:,19)=evalfis(inputData(:,[13 18]),fSys{5});

%%%%%%%%%% Evaluate the 6th fuzzy sub-system (Chemical Effect
Intermidiate Parameters) %%%%%%%%%%%%%%%
%Compute the Chemical Effect
inputData(:,12)=evalfis(inputData(:,[5 8 11]),fSys{6});

%%%%%%%%%% Evaluate the 7th fuzzy sub-system (Biological Effect
Intermidiate Parameters) %%%%%%%%%%%%%%%
%Compute the Biological Effect
inputData(:,14)=evalfis(inputData(:,[2 10 13 3]),fSys{7});

%%%%%%%%%%%%%%% Evaluate the 8th fuzzy sub-system (Hydraulic Effect
Intermidiate Parameters) %%%%%%%%%%%%%%%%%
%Compute the Hydraulic Effect
inputData(:,20)=evalfis(inputData(:,[17 19]),fSys{8});

%%%%%%%%%%% Evaluate the 9th fuzzy sub-system (Accumulation Potential
Parameters) %%%%%%%%%%%%%%
%Compute the Accumulation Potential
FeAndMnAccumP_Predicted=evalfis(inputData(:,[12 14 20]),fSys{9});

myData = inputData(:,:);
return
```

Appendix I: Source code to assign rule to the fuzzy system

```
function [fSys] = AssignRules(fSys, optData, ruleData_ante, ruleData_cons,
ruleData_wt, ruleData_conn, outputCol)

%Assign rules to 1st fuzzy sub-system
rules_cons1 = ModifyRules(ruleData_cons{1}, optData(1:9), outputCol(1:9));
[fSys{1}] = ReAssignRules(fSys{1}, ruleData_ante{1}, rules_cons1,
ruleData_wt{1}, ruleData_conn{1});

%Assign rules to 2nd fuzzy sub-system
rules_cons2 = ModifyRules(ruleData_cons{2}, optData(10:18),
outputCol(10:18));
[fSys{2}] = ReAssignRules(fSys{2}, ruleData_ante{2}, rules_cons2,
ruleData_wt{2}, ruleData_conn{2});

%Assign rules to 3rd fuzzy sub-system
rules_cons3 = ModifyRules(ruleData_cons{3}, optData(19:27),
outputCol(19:27));
[fSys{3}] = ReAssignRules(fSys{3}, ruleData_ante{3}, rules_cons3,
ruleData_wt{3}, ruleData_conn{3});

%Assign rules to 4th fuzzy sub-system
rules_cons4 = ModifyRules(ruleData_cons{4}, optData(28:33),
outputCol(28:33));
[fSys{4}] = ReAssignRules(fSys{4}, ruleData_ante{4}, rules_cons4,
ruleData_wt{4}, ruleData_conn{4});

%Assign rules to 5th fuzzy sub-system
rules_cons5 = ModifyRules(ruleData_cons{5}, optData(34:39),
outputCol(34:39));
[fSys{5}] = ReAssignRules(fSys{5}, ruleData_ante{5}, rules_cons5,
ruleData_wt{5}, ruleData_conn{5});

%Assign rules to 6th fuzzy sub-system
rules_cons6 = ModifyRules(ruleData_cons{6}, optData(40:48),
outputCol(40:48));
[fSys{6}] = ReAssignRules(fSys{6}, ruleData_ante{6}, rules_cons6,
ruleData_wt{6}, ruleData_conn{6});

%Assign rules to 7th fuzzy sub-system
rules_cons7 = ModifyRules(ruleData_cons{7}, optData(49:63),
outputCol(49:63));
[fSys{7}] = ReAssignRules(fSys{7}, ruleData_ante{7}, rules_cons7,
ruleData_wt{7}, ruleData_conn{7});

%Assign rules to 8th fuzzy sub-system
rules_cons8 = ModifyRules(ruleData_cons{8}, optData(64:69),
outputCol(64:69));
[fSys{8}] = ReAssignRules(fSys{8}, ruleData_ante{8}, rules_cons8,
ruleData_wt{8}, ruleData_conn{8});

%Assign rules to 9th fuzzy sub-system
rules_cons9 = ModifyRules(ruleData_cons{9}, optData(70:78),
outputCol(70:78));
[fSys{9}] = ReAssignRules(fSys{9}, ruleData_ante{9}, rules_cons9,
ruleData_wt{9}, ruleData_conn{9});
return
```

```
 function [rules_cons] = ModifyRules(rules_cons, optData, outputCol)
n=length(rules_cons);
j = 1;
for i = 1:n
    if(outputCol(i) > 0)
        x = 0;
    else
        rules_cons(i) = optData(j);
        j = j + 1;
    end
end

return

function [fSystems] = ReAssignRules(fSystems, rules_ante, rules_cons,
rules_wt, rules_conn)
   fSystems.rule=[];
   %merge the antecedent, consequent, weight  and connective
   rules_All = [rules_ante,rules_cons,rules_wt,rules_conn];
   fSystems = addrule(fSystems,rules_All);
return
```

Appendix J: Source code for the genetic algorithm

```
function [optData, fval, exitflag, output, finalpop, finalscore] = GAfnc()
%Read the data to be optimised
optData = xlsread('_OptimisedRules.xlsx', 'Rules4', 'E:E');
%delete all zeroes
optData = optData(optData~=0);
optData = optData';

%Read the data from excel sheet and fuzy system
[inputData, FeAndMnMeasured, fSys, ruleData_ante, ruleData_cons,
ruleData_wt, ruleData_conn, outputCol] = ReadData();

%Define number of variables
Nvars = length(optData);

%Define bounds
LB = ones(1,Nvars);
UB = 5*ones(1,Nvars);

%define anonymous objective function and number of variables
objfun = @(optData)EvalRows(optData, inputData, FeAndMnMeasured, fSys,
ruleData_cons, ruleData_ante, ruleData_wt, ruleData_conn, outputCol);

%************************************************************************
**
%*************Enter Algorithm Options
Here*********************************
%************************************************************************
**
%define Genetic Algorithm options
gaoptions = gaoptimset(@ga);
gaoptions.PlotFcns = @gaplotbestf;
gaoptions.PopulationType = 'doubleVector';
gaoptions.PopulationSize = [100];
gaoptions.PopInitRange = [LB; UB];
gaoptions.InitialPopulation = [];
gaoptions.EliteCount = 1;
gaoptions.CreationFcn = @int_pop;
gaoptions.MutationFcn = @int_mutation;
%gaoptions.MutationFcn = {@mutationgaussian, 0.2, 0.8};
gaoptions.CrossoverFcn = @crossoverscattered;
%gaoptions.CrossoverFcn = {@crossoverheuristic, 1.2};
gaoptions.CrossoverFraction = 0.8 + 0.2*rand;
gaoptions.MigrationDirection = 'both';
gaoptions.MigrationInterval = 20;
gaoptions.MigrationFraction = 0.03;
gaoptions.Generations = 4000;
gaoptions.StallGenLimit = gaoptions.Generations;
gaoptions.TolFun = 1.0e-100;
gaoptions.Display = 'iter';
gaoptions.Vectorized = 'off';
%************************************************************************
**
%*************Enf of Algorithm
Options**************************************
%************************************************************************
**
```

```
% Run the Genetic Algorithm
[optData, fval, exitflag, output, finalpop, finalscore] =
ga(objfun,Nvars,[],[],[],[],LB,UB,[],gaoptions);

%************************************************************************
**
%*******************************Post
Processing************************
%*************************End of Post
Processing************************
%************************************************************************
**
return;

%---------------------------------------------------
% Mutation function to generate childrens satisfying the range and
integer
% constraints on decision variables.
function mutationChildren = int_mutation(parents,options,GenomeLength, ...
    FitnessFcn,state,thisScore,thisPopulation)
shrink = .01;
scale = 1;
scale = scale - shrink * scale * state.Generation/options.Generations;
range = options.PopInitRange;
lower = range(1,:);
upper = range(2,:);
scale = scale * (upper - lower);
mutationPop =  length(parents);
% The use of ROUND function will make sure that childrens are integers.
mutationChildren =  repmat(lower,mutationPop,1) +  ...
    round(repmat(scale,mutationPop,1) .* rand(mutationPop,GenomeLength));
return;
% End of mutation function
%---------------------------------------------------
function Population = int_pop(GenomeLength,FitnessFcn,options)

totalpopulation = sum(options.PopulationSize);
range = options.PopInitRange;
lower= range(1,:);
span = range(2,:) - lower;
% The use of ROUND function will make sure that individuals are integers.
Population = repmat(lower,totalpopulation,1) +  ...
    round(repmat(span,totalpopulation,1) .*
rand(totalpopulation,GenomeLength));
return;
% End of creation function
```

Appendix K: SQL code to retrieve customer complaints data in WSZ1

```
SELECT
    tblCCData.[Model Node],
    Year([Date]) AS [Year],
    Sum(tblCCData.NumberOfCC) AS NumberOfCC, tblCCData.DMA
FROM tblWSZ1_Hyd INNER JOIN tblCCData ON tblWSZ1_Hyd.Node =
tblCCData.[Model Node]
WHERE (((tblCCData.WSZ)="WSZ1") AND
    ((tblCCData.Contact_reason)="Discoloured Water" OR
    (tblCCData.Contact_reason)="Slime"))
GROUP BY tblCCData.[Model Node], Year([Date]), tblCCData.DMA
ORDER BY Year([Date]), tblCCData.[Model Node];
```

Appendix L: SQL code to retrieve hydraulic, Fe and Mn data in WSZ2

```
SELECT
    qryWSZ2_YearlyAveragesWQ_AtNodes.WSZ, tblWSZ2_Hyd.Node,
    qryWSZ2_YearlyAveragesWQ_AtNodes.Year,
    qryWSZ2_YearlyAveragesWQ_AtNodes.AvgHARD_Node, tblWSZ2_Hyd.[Pipe
    Material], tblWSZ2_Hyd.[Pipe Age], tblWSZ2_Hyd.AvgWaterAge,
    tblWSZ2_Hyd.MaxShearStressAtNdode, tblWSZ2_Hyd.VarShearStressAtNode,
    tblWSZ2_Hyd.[Hydraulic Dist From Source],
    qryWSZ2_YearlyAveragesWQ_AtNodes.DMA,
    qryWSZ2_YearlyAveragesWQ_AtNodes.AvgIRON_Node,
    qryWSZ2_YearlyAveragesWQ_AtNodes.AvgMANG_Node
FROM tblWSZ2_Hyd INNER JOIN qryWSZ2_YearlyAveragesWQ_AtNodes
ON tblWSZ2_Hyd.Node = qryWSZ2_YearlyAveragesWQ_AtNodes.Model_node
ORDER BY qryWSZ2_YearlyAveragesWQ_AtNodes.Year;
```

Appendix M: Algorithm for estimating missing pipe age data

Open the file with the pipe IDs, pipe roughness and missing pipe age data.

Open the report file.

Read pipe IDs and pipe roughness from the data file.

for *pipe = 1 to number of records in the file*

if *pipe =Polyethylene*

Use the linear regression equation for Polyethylene to compute pipe age.

else if *pipe = Polyvinyl Chloride*

Use the linear regression equation for Polyvinyl Chloride to compute pipe age.

else if *pipe = High Density Polyethylene*

Use the linear regression equation for High Performance Polyethylene to compute pipe age.

else if *pipe = Asbestos Cement*

Use the linear regression equation for Asbestos Cement to compute pipe age.

else if *pipe = Ductile Iron*

Use the linear regression equation for Ductile Iron to compute pipe age.

else if *pipe = Steel*

Use the linear regression equation for Steel to compute pipe age.

else

Use the linear regression equation for Cast Iron to compute pipe age.

end if

Print the computed pipe age in the report file.

end for

Close the report file.

Close the data file.

Appendix N: Algorithm for choosing appropriate number of hidden nodes and layers

Initialise random number generator

Load the data

Open the report file

for *hidden nodes in layer1 = 1 to 15*

    for *hidden nodes in layer2 = 3 to 8*

        for *iteration = 1 to 30*

            Create a network

            Set the default network parameters

            Initialise and train network

            Save the network and training data

            Compute for the performance of the model

            Print the performance of the model into the report file

        end

    end

end

Close the report file.

Appendix O: Algorithm for tuning the network parameters

Initialise random number generator

Load the data

Open the report file

for *network parameter = Xi to Xn*

    for *iteration = 1 to 30*

        Create a network

        Set the net n^{th} *value of the network parameter*

        Initialise and train network

        Save the network and training data

        Compute for the performance of the model

        Print the performance of the model into the report file

    end

end

Close the report file.

Appendix P: Results of tuned parameters

Table P.1 Average performance of the ANN(t) model on the test data set for WSZ1

Average RMSE on testing data set	Average CA on testing data set (%)	Hidden nodes in 1st layer	Hidden nodes in 2nd layer
0.1551	50.91	1	3
0.1745	48.97	2	3
0.1747	52.69	3	3
0.1734	48.82	4	3
0.1648	48.34	5	3
0.1740	49.94	6	3
0.1778	50.60	7	3
0.1620	49.89	8	3
0.1702	48.46	9	3
0.1569	49.49	10	3
0.1759	51.09	11	3
0.1550	**52.84**	**12**	**3**
0.1755	49.72	13	3
0.1607	51.57	14	3
0.1735	49.05	15	3

Table P.2 Average performance of the ANN(t) model using three different activation functions for WSZ1

	Sigmoid activation function	Linear activation function	Hyperbolic activation function
Average RMSE on testing data set	**0.1526**	0.1942	0.1800
Average CA on testing data set (%)	**53.05**	42.57	46.43

Table P.3 Average performance of the ANN(t) model on the testing data set using different minimum gradient values for WSZ1

Average RMSE on testing data set	Average CA on testing data set (%)	Minimum gradient magnitude
0.1615	51.52	0.01
0.1532	52.55	0.001
0.1524	51.42	0.0001
0.1466	53.41	1.00E-05
0.1455	**55.07**	**1.00E-06**
0.1618	53.09	1.00E-07
0.1534	52.18	1.00E-08
0.1617	54.92	1.00E-09
0.1583	54.05	1.00E-10
0.1495	54.50	1.00E-11

Table P.4 Average performance of the ANN(t) model on the testing data set using different learning rate values for WSZ1

Average RMSE on testing data set	Average CA on testing data set (%)	Learning rate
0.1563	55.22	0.001
0.1537	55.74	0.008
0.1489	55.53	0.005
0.1550	55.15	0.01
0.1413	55.87	0.08
0.1407	**56.46**	**0.05**
0.1462	55.06	0.1
0.1509	55.98	0.15
0.1446	54.16	0.2
0.1563	55.62	0.3

Table P.5 Average performance of the ANN(t) model on the testing data set using different initial Mu values for WSZ1

Average RMSE on testing data set	Average CA on testing data set (%)	Initial Mu
0.1408	56.38	1.00E-05
0.1464	56.15	5.00E-05
0.1464	56.15	0.0001
0.1529	54.21	0.001
0.1505	55.39	0.01
0.1408	56.36	0.04
0.1387	**57.76**	**0.08**
0.1459	54.71	0.1
0.1432	56.40	0.2
0.1437	55.04	0.5

Table P.6 Average performance of the ANN(t) model on the testing data set using different Mu increase factor values for WSZ1

Average RMSE on testing data set	Average CA on testing data set (%)	Mu increase factor
0.1450	56.43	0.01
0.1424	57.91	0.1
0.1309	**60.92**	**1**
0.1446	60.18	3
0.1496	59.54	7
0.1339	59.55	10
0.1341	59.87	15
0.1481	60.58	20
0.1380	57.72	30
0.1444	57.67	50

Table P.7 Average performance of the ANN(t) model on the testing data set using different Mu decrease factor values for WSZ1

Average RMSE on testing data set	Average CA on testing data set (%)	Mu decrease factor
0.1307	61.56	0.001
0.1344	62.83	0.01
0.1248	**63.26**	**0.05**
0.1283	62.75	0.08
0.1384	60.30	0.1
0.1294	60.25	0.12
0.1330	59.24	0.15
0.1344	62.63	0.2
0.1355	59.41	0.5
0.1343	60.72	1

Table P.8 Average performance of the ANN(t) model on the testing data set using the scaled conjugate gradient backpropagation algorithm for WSZ1

Performance indicator	Scaled conjugate gradient backpropagation
Average RMSE on testing data set	0.1539
Average CA on testing data set (%)	58.92

Table P.9 The tuned ANN(t) model parameter values for WSZ1

Name	Tuned value	Description of parameter
Show	5	The display of epochs within display
Epochs	1000	The maximum number of iteration
Goal	0	Performance goal
Min_grad	1.00E-06	Minimum gradient magnitude
Mu	0.08	Initial Mu
Mu_inc	1	Mu increase factor
Mu_dec	0.05	Mu decrease factor
η	0.05	Learning rate
1^{st} layer nodes	12	The number of nodes in 1^{st} layer
2^{nd} layer nodes	3	The number of nodes in 2^{nd} layer
Sigmoid activation function		The activation functions used in the model
Levenberg–Marquardt algorithm		Optimisation algorithm used in the model

Table P.10 Average performance of the ANN(t) model on the test data set for WSZ3

Average RMSE on testing data set	Average CA on testing data set (%)	Hidden nodes in 1st layer	Hidden nodes in 2nd layer
0.1719	51.70	1	5
0.1760	48.33	2	5
0.1670	50.64	3	5
0.1782	50.66	4	5
0.1551	48.66	5	5
0.1511	**53.48**	**6**	**5**
0.1536	51.73	7	5
0.1728	48.70	8	5
0.1782	51.82	9	5
0.1674	50.02	10	5
0.1799	49.57	11	5
0.1764	49.88	12	5
0.1796	52.14	13	5
0.1595	51.31	14	5
0.1762	53.00	15	5

Table P.11 Average performance of the ANN(t) model using three different activation functions for WSZ3

	Sigmoid activation function	Linear activation function	Hyperbolic activation function
Average RMSE on testing data set	**0.1502**	0.1802	0.1795
Average CA on testing data set (%)	**53.89**	42.10	47.27

Table P.12 Average performance of the ANN(t) model on the testing data set using different minimum gradient values for WSZ3

Average RMSE on testing data set	Average CA on testing data set (%)	Minimum gradient magnitude
0.1466	**55.01**	**0.01**
0.1544	53.25	0.001
0.1631	52.42	0.0001
0.1526	50.00	1.00E-05
0.1548	52.51	1.00E-06
0.1601	52.82	1.00E-07
0.1540	54.18	1.00E-08
0.1551	52.64	1.00E-09
0.1508	52.11	1.00E-10
0.1634	52.96	1.00E-11

Table P.13 Average performance of the ANN(t) model on the testing data set using different learning rate values for WSZ3

Average RMSE on testing data set	Average CA on testing data set (%)	Learning rate
0.1550	54.84	0.001
0.1515	54.52	0.008
0.1419	**56.22**	**0.005**
0.1487	55.60	0.01
0.1558	54.74	0.08
0.1596	54.12	0.05
0.1513	54.15	0.1
0.1461	55.72	0.15
0.1556	55.78	0.2
0.1439	55.10	0.3

Table P.14 Average performance of the ANN(t) model on the testing data set using different initial Mu values for WSZ3

Average RMSE on testing data set	Average CA on testing data set (%)	Initial Mu
0.1435	56.24	1.00E-05
0.1376	54.63	5.00E-05
0.1549	57.78	0.0001
0.1451	57.67	0.001
0.1355	**58.19**	**0.01**
0.1465	54.15	0.04
0.1477	55.93	0.08
0.1418	54.76	0.1
0.1549	54.75	0.2
0.1438	55.59	0.5

Table P.15 Average performance of the ANN(t) model on the testing data set using different Mu increase factor values for WSZ3

Average RMSE on testing data set	Average CA on testing data set (%)	Mu increase factor
0.1355	59.23	0.01
0.1302	**60.99**	**0.1**
0.1468	59.48	1
0.1378	57.82	3
0.1419	56.82	7
0.1413	57.95	10
0.1361	59.88	15
0.1366	57.44	20
0.1493	58.06	30
0.1331	57.60	50

Table P.16 Average performance of the ANN(t) model on the testing data set using different Mu decrease factor values for WSZ3

Average RMSE on testing data set	Average CA on testing data set (%)	Mu decrease factor
0.1280	58.76	0.001
0.1267	61.93	0.01
0.1214	**62.55**	**0.05**
0.1315	60.51	0.08
0.1341	58.67	0.1
0.1222	61.09	0.12
0.1314	60.85	0.15
0.1335	59.44	0.2
0.1256	60.89	0.5
0.1301	60.62	1

Table P.17 Average performance of the ANN(t) model on the testing data set using the scaled conjugate gradient backpropagation algorithm for WSZ3

Performance indicator	Scaled conjugate gradient backpropagation
Average RMSE on testing data set	0.1414
Average CA on testing data set (%)	55.09

Table P.18 The tuned ANN(t) model parameter values for WSZ3

Name	Tuned value	Description of parameter
Show	5	The display of epochs within display
Epochs	1000	The maximum number of iteration
Goal	0	Performance goal
Min_grad	0.01	Minimum gradient magnitude
Mu	0.01	Initial Mu
Mu_inc	0.1	Mu increase factor
Mu_dec	0.05	Mu decrease factor
η	0.005	Learning rate
1^{st} layer nodes	6	The number of nodes in 1^{st} layer
2^{nd} layer nodes	5	The number of nodes in 2^{nd} layer
Sigmoid activation function		The activation functions used in the model
Levenberg–Marquardt algorithm		Optimisation algorithm used in the model

Table P.19 Average performance of the ANN(t) model on the test data set for WSZ4

Average RMSE on testing data set	Average CA on testing data set (%)	Hidden nodes in 1^{st} layer	Hidden nodes in 2^{nd} layer
0.1919	48.60	1	4
0.1991	45.59	2	4
0.1791	46.38	3	4
0.1839	49.45	4	4
0.1703	48.66	5	4
0.1943	48.91	6	4
0.1778	46.30	7	4
0.1906	45.52	8	4
0.1849	47.27	9	4
0.1912	48.94	10	4
0.1897	47.46	11	4
0.1937	48.66	12	4
0.1965	48.82	13	4
0.1807	49.90	14	4
0.1702	**49.98**	**15**	**4**

Table P.20 Average performance of the ANN(t) model using three different activation functions for WSZ4

	Sigmoid activation function	Linear activation function	Hyperbolic activation function
Average RMSE on testing data set	0.1691	0.2095	0.1711
Average CA on testing data set (%)	50.32	41.42	47.06

Table P.21 Average performance of the ANN(t) model on the testing data set using different minimum gradient values for WSZ4

Average RMSE on testing data set	Average CA on testing data set (%)	Minimum gradient magnitude
0.1742	50.96	0.01
0.1666	50.45	0.001
0.1775	51.27	0.0001
0.1714	49.78	1.00E-05
0.1697	51.12	1.00E-06
0.1847	51.18	1.00E-07
0.1798	49.03	1.00E-08
0.1688	52.18	1.00E-09
0.1667	48.67	1.00E-10
0.1650	**52.86**	**1.00E-11**

Table P.22 Average performance of the ANN(t) model on the testing data set using different learning rate values for WSZ4

Average RMSE on testing data set	Average CA on testing data set (%)	Learning rate
0.1767	55.00	0.001
0.1705	53.24	0.008
0.1721	52.00	0.005
0.1767	51.85	0.01
0.1665	50.85	0.08
0.1731	50.41	0.05
0.1616	**54.29**	**0.1**
0.1702	51.57	0.15
0.1748	51.63	0.2
0.1673	53.07	0.3

Table P.23 Average performance of the ANN(t) model on the testing data set using different initial Mu values for WSZ4

Average RMSE on testing data set	Average CA on testing data set (%)	Initial Mu
0.1718	52.65	1.00E-05
0.1741	54.12	5.00E-05
0.1652	56.13	0.0001
0.1739	52.66	0.001
0.1632	52.52	0.01
0.1568	55.50	0.04
0.1726	56.50	0.08
0.1662	56.31	0.1
0.1613	55.08	0.2
0.1567	**56.65**	**0.5**

Table P.24 Average performance of the ANN(t) model on the testing data set using different Mu increase factor values for WSZ4

Average RMSE on testing data set	Average CA on testing data set (%)	Mu increase factor
0.1619	54.50	0.01
0.1566	55.46	0.1
0.1684	56.86	1
0.1553	56.36	3
0.1566	58.55	7
0.1505	**59.00**	**10**
0.1538	57.83	15
0.1630	58.10	20
0.1601	55.14	30
0.1569	57.60	50

Table P.25 Average performance of the ANN(t) model on the testing data set using different Mu decrease factor values for WSZ4

Average RMSE on testing data set	Average CA on testing data set (%)	Mu decrease factor
0.1498	57.69	0.001
0.1584	57.27	0.01
0.1571	59.99	0.05
0.1534	57.36	0.08
0.1489	59.99	0.1
0.1552	59.49	0.12
0.1506	56.22	0.15
0.1554	59.54	0.2
0.1473	**60.59**	**0.5**
0.1595	58.05	1

Table P.26 Average performance of the ANN(t) model on the testing data set using the scaled conjugate gradient backpropagation algorithm for WSZ4

Performance indicator	Scaled conjugate gradient backpropagation
Average RMSE on testing data set	0.1715
Average CA on testing data set (%)	55.44

Table P.27 The tuned ANN(t) model parameter values for WSZ4

Name	Tuned value	Description of parameter
Show	5	The display of epochs within display
Epochs	1000	The maximum number of iteration
Goal	0	Performance goal
Min_grad	1.00E-11	Minimum gradient magnitude
Mu	0.5	Initial Mu
Mu_inc	10	Mu increase factor
Mu_dec	0.5	Mu decrease factor
η	0.1	Learning rate
1^{st} layer nodes	15	The number of nodes in 1^{st} layer
2^{nd} layer nodes	4	The number of nodes in 2^{nd} layer
Sigmoid activation function		The activation functions used in the model
Levenberg–Marquardt algorithm		Optimisation algorithm used in the model

Table P.28 Average performance of the ANN(t) model on the test data set for WSZ5

Average RMSE on testing data set	Average CA on testing data set (%)	Hidden nodes in 1st layer	Hidden nodes in 2nd layer
0.1742	50.40	1	5
0.1561	50.78	2	5
0.1678	50.76	3	5
0.1523	50.25	4	5
0.1568	48.64	5	5
0.1506	**53.51**	**6**	**5**
0.1542	52.32	7	5
0.1710	48.68	8	5
0.1721	48.38	9	5
0.1789	51.51	10	5
0.1719	48.46	11	5
0.1704	49.58	12	5
0.1531	48.44	13	5
0.1516	50.36	14	5
0.1609	49.41	15	5

Table P.29 Average performance of the ANN(t) model using three different activation functions for WSZ5

	Sigmoid activation function	Linear activation function	Hyperbolic activation function
Average RMSE on testing data set	0.1469	0.1910	0.1571
Average CA on testing data set (%)	55.98	42.04	46.30

Table P.30 Average performance of the ANN(t) model on the testing data set using different minimum gradient values for WSZ5

Average RMSE on testing data set	Average CA on testing data set (%)	Minimum gradient magnitude
0.1307	56.10	0.01
0.1259	55.36	0.001
0.1332	56.09	0.0001
0.1359	54.74	1.00E-05
0.1286	53.04	1.00E-06
0.1257	**57.70**	**1.00E-07**
0.1381	53.05	1.00E-08
0.1366	53.75	1.00E-09
0.1306	54.07	1.00E-10
0.1329	55.75	1.00E-11

Table P.31 Average performance of the ANN(t) model on the testing data set using different learning rate values for WSZ5

Average RMSE on testing data set	Average CA on testing data set (%)	Learning rate
0.1353	56.90	0.001
0.1270	59.35	0.008
0.1363	57.24	0.005
0.1297	56.17	0.01
0.1335	57.11	0.08
0.1314	58.53	0.05
0.1317	58.02	0.1
0.1369	58.66	0.15
0.1248	**60.42**	**0.2**
0.1394	56.11	0.3

Table P.32 Average performance of the ANN(t) model on the testing data set using different initial Mu values for WSZ5

Average RMSE on testing data set	Average CA on testing data set (%)	Initial Mu
0.1149	61.36	1.00E-05
0.1149	59.00	5.00E-05
0.1095	58.68	0.0001
0.1067	57.26	0.001
0.1084	60.77	0.01
0.1033	57.78	0.04
0.1005	**62.92**	**0.08**
0.1154	57.80	0.1
0.1042	57.70	0.2
0.1009	59.44	0.5

Table P.33 Average performance of the ANN(t) model on the testing data set using different Mu increase factor values for WSZ5

Average RMSE on testing data set	Average CA on testing data set (%)	Mu increase factor
0.0854	64.07	0.01
0.0811	**65.24**	**0.1**
0.0864	64.72	1
0.0820	61.48	3
0.0878	62.18	7
0.0911	63.28	10
0.0956	62.19	15
0.0891	62.20	20
0.0974	61.03	30
0.0827	60.77	50

Table P.34 Average performance of the ANN(t) model on the testing data set using different Mu decrease factor values for WSZ5

Average RMSE on testing data set	Average CA on testing data set (%)	Mu decrease factor
0.0634	69.46	0.001
0.0627	70.39	0.01
0.0590	72.67	0.05
0.0608	69.38	0.08
0.0554	69.89	0.1
0.0610	73.52	0.12
0.0571	72.50	0.15
0.0578	73.15	0.2
0.0509	**74.07**	**0.5**
0.0637	71.04	1

Table P.35 Average performance of the ANN(t) model on the testing data set using the scaled conjugate gradient backpropagation algorithm for WSZ5

Performance indicator	Scaled conjugate gradient backpropagation
Average RMSE on testing data set	0.1194
Average CA on testing data set (%)	63.45

Table P.36 The tuned ANN(t) model parameter values for WSZ5

Name	Tuned value	Description of parameter
Show	5	The display of epochs within display
Epochs	1000	The maximum number of iteration
Goal	0	Performance goal
Min_grad	1.00E-07	Minimum gradient magnitude
Mu	0.08	Initial Mu
Mu_inc	0.1	Mu increase factor
Mu_dec	0.5	Mu decrease factor
η	0.2	Learning rate
1^{st} layer nodes	6	The number of nodes in 1^{st} layer
2^{nd} layer nodes	5	The number of nodes in 2^{nd} layer
Sigmoid activation function		The activation functions used in the model
Levenberg–Marquardt algorithm		Optimisation algorithm used in the model

Appendix Q: Prediction profiler graphs from the ANN(t) model

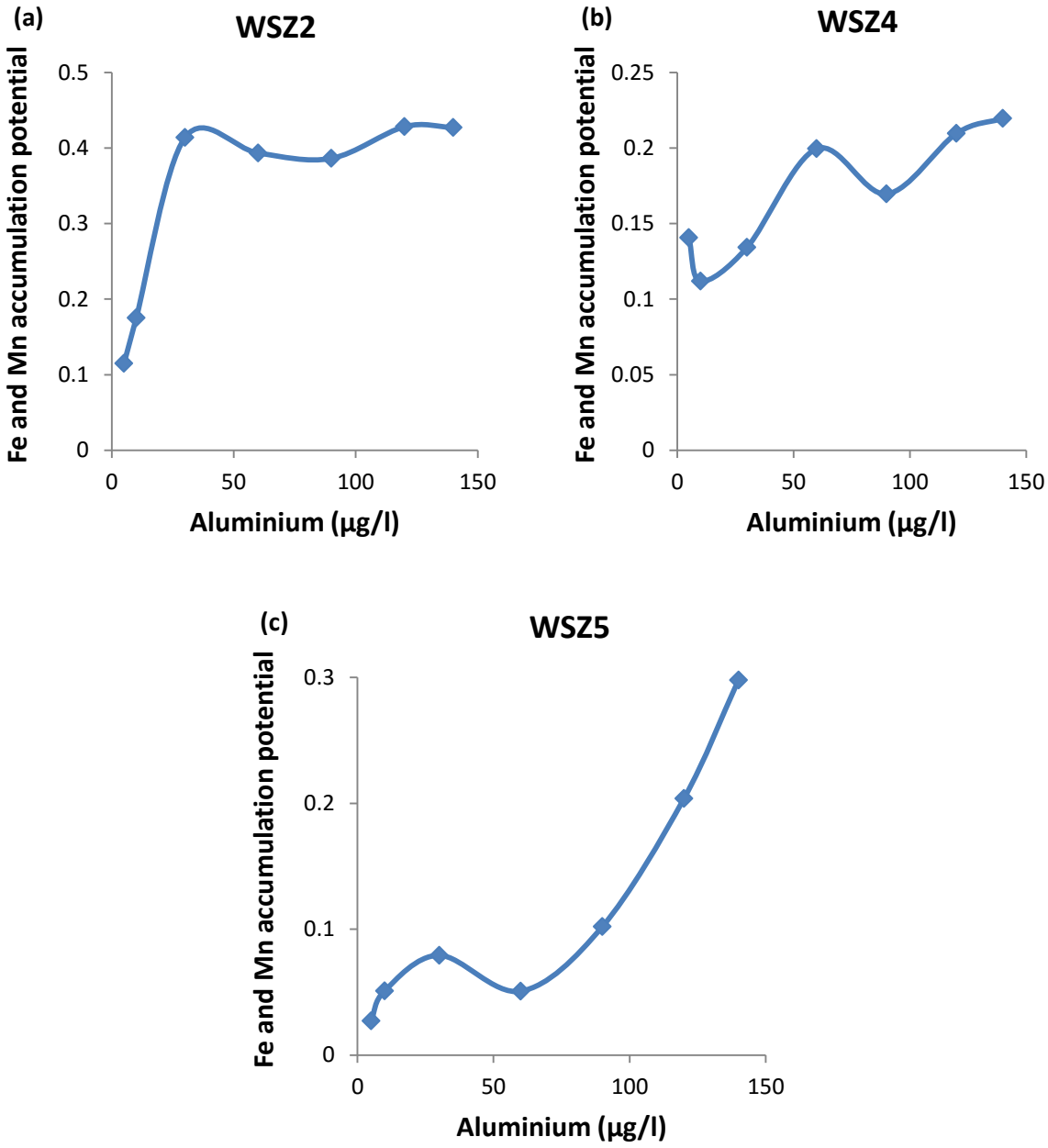


Figure Q.1 Relationship between Fe and Mn accumulation potential and aluminium

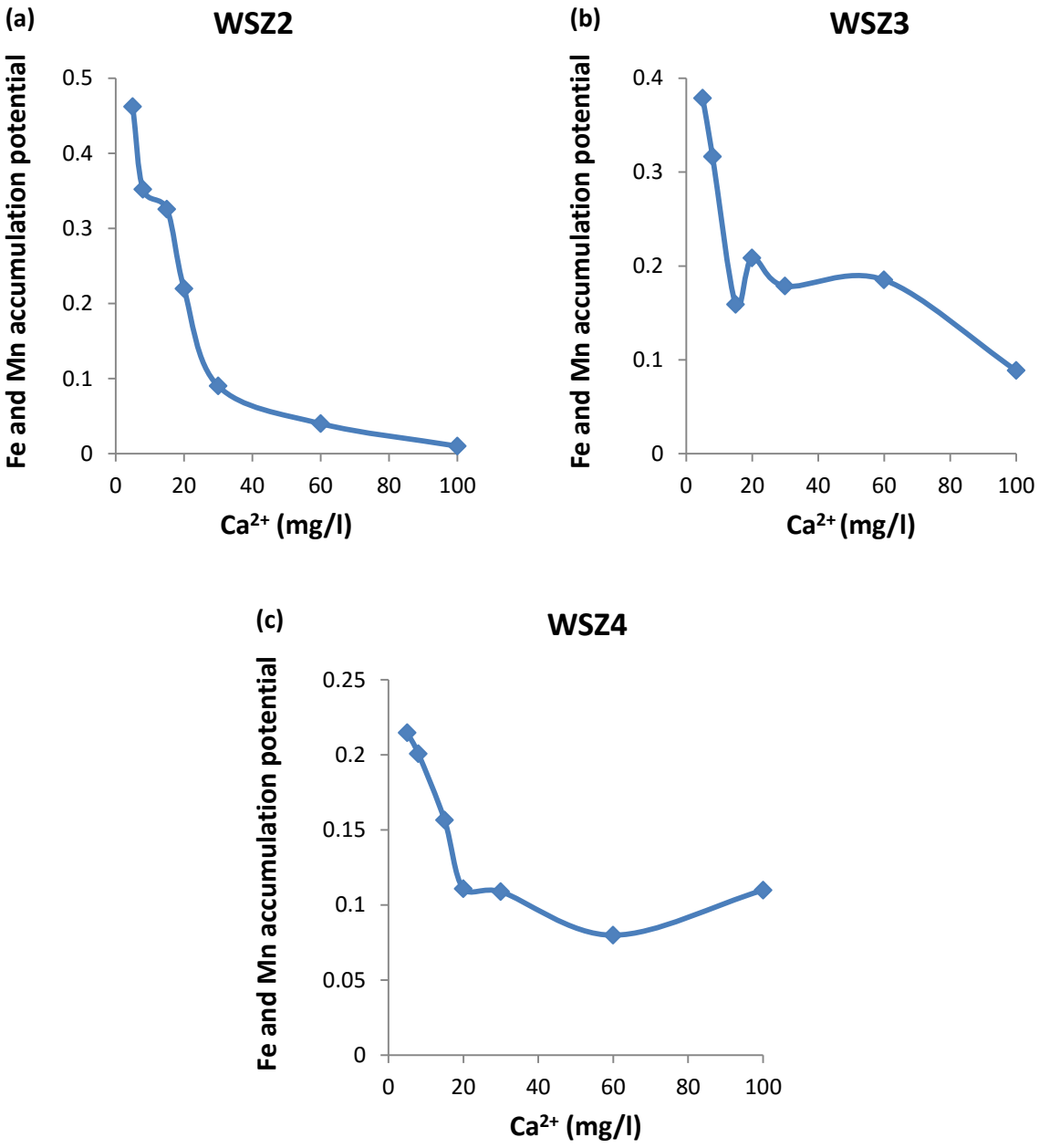


Figure Q.2 Relationship between Fe and Mn accumulation potential and Ca^{2+}

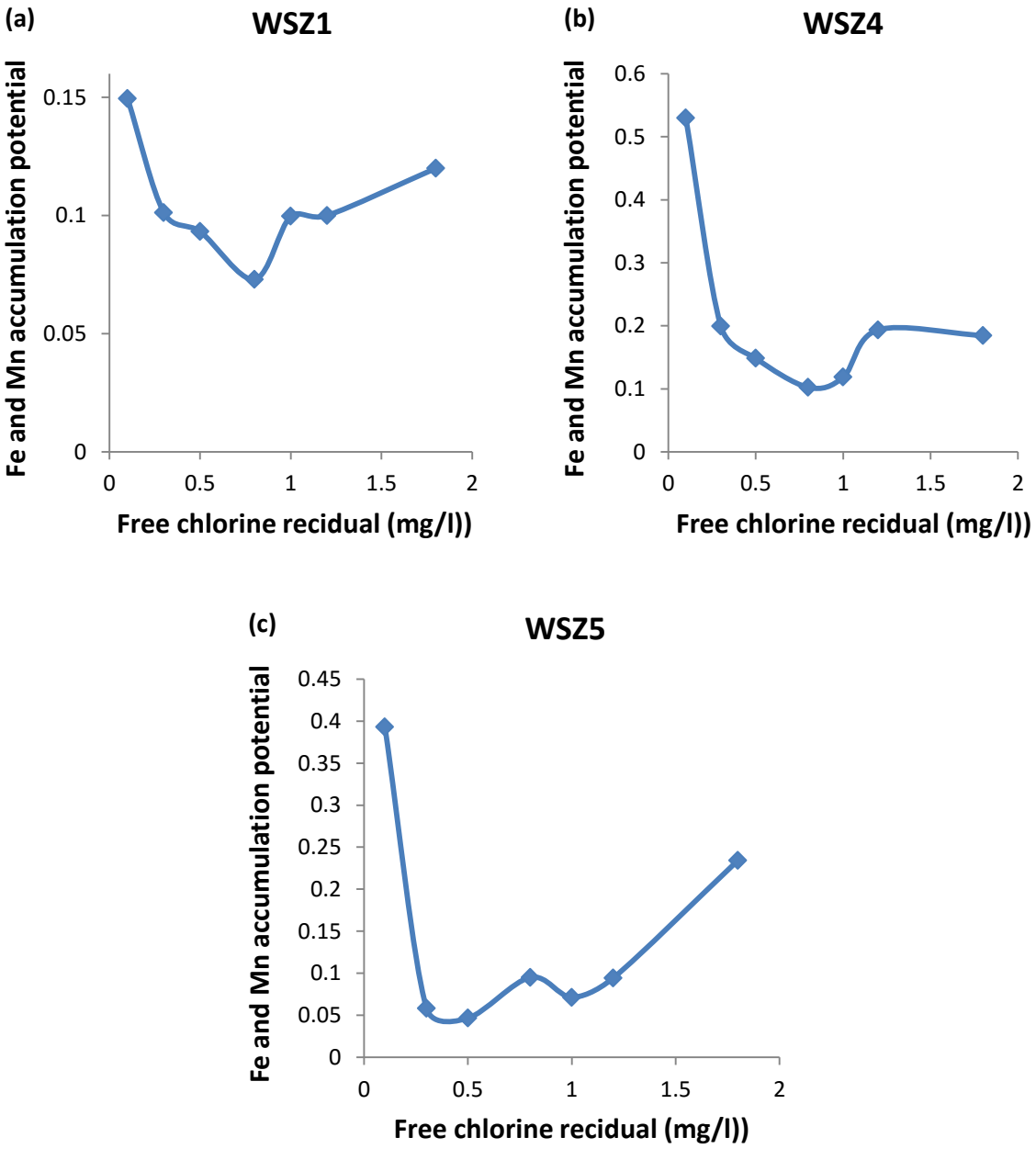


Figure Q.3 Relationship between Fe and Mn accumulation potential and free chlorine recidual

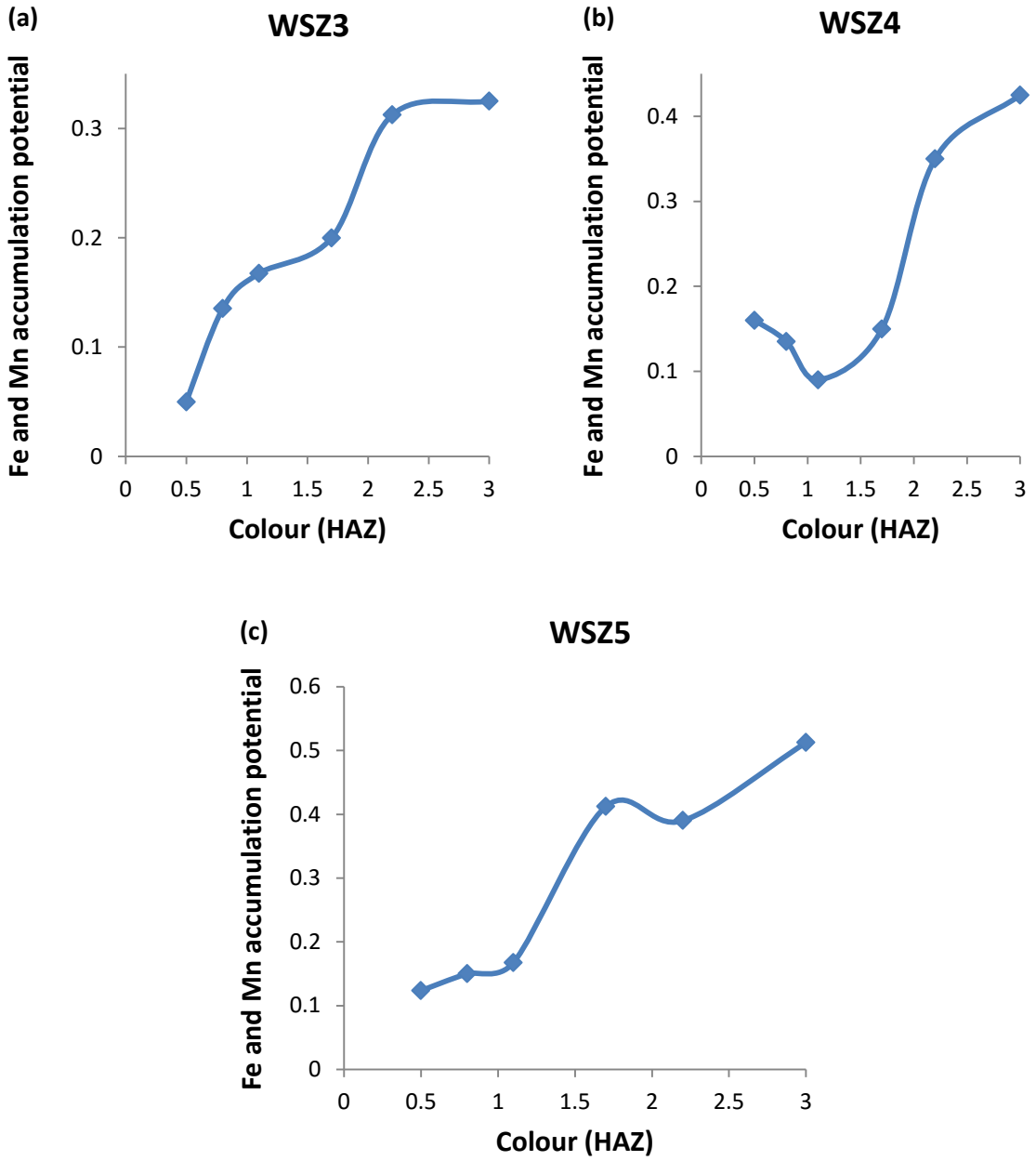


Figure Q.4 Relationship between Fe and Mn accumulation potential and colour

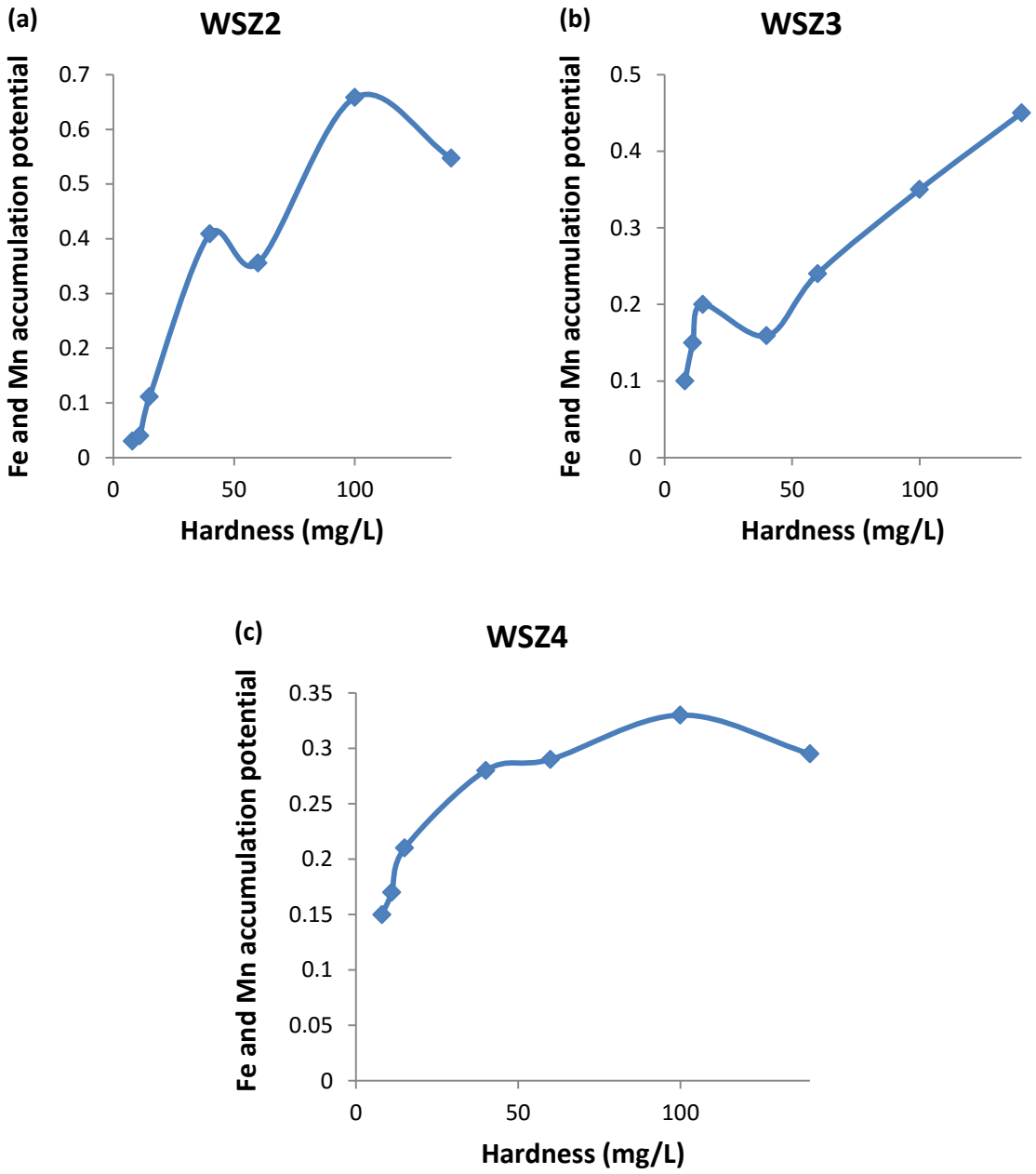


Figure Q.5 Relationship between Fe and Mn accumulation potential and hardness

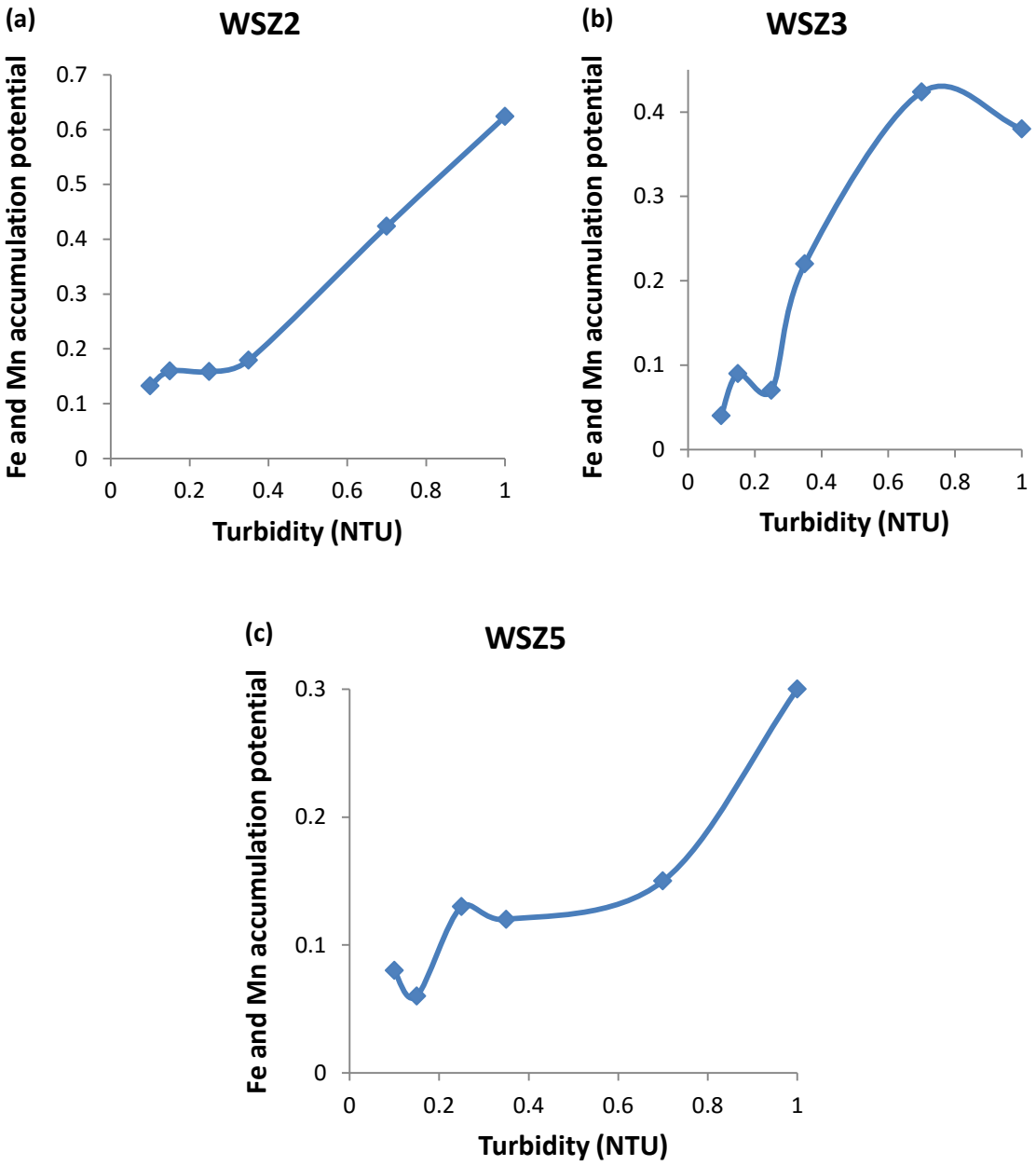


Figure Q.6 Relationship between Fe and Mn accumulation potential and tubidity

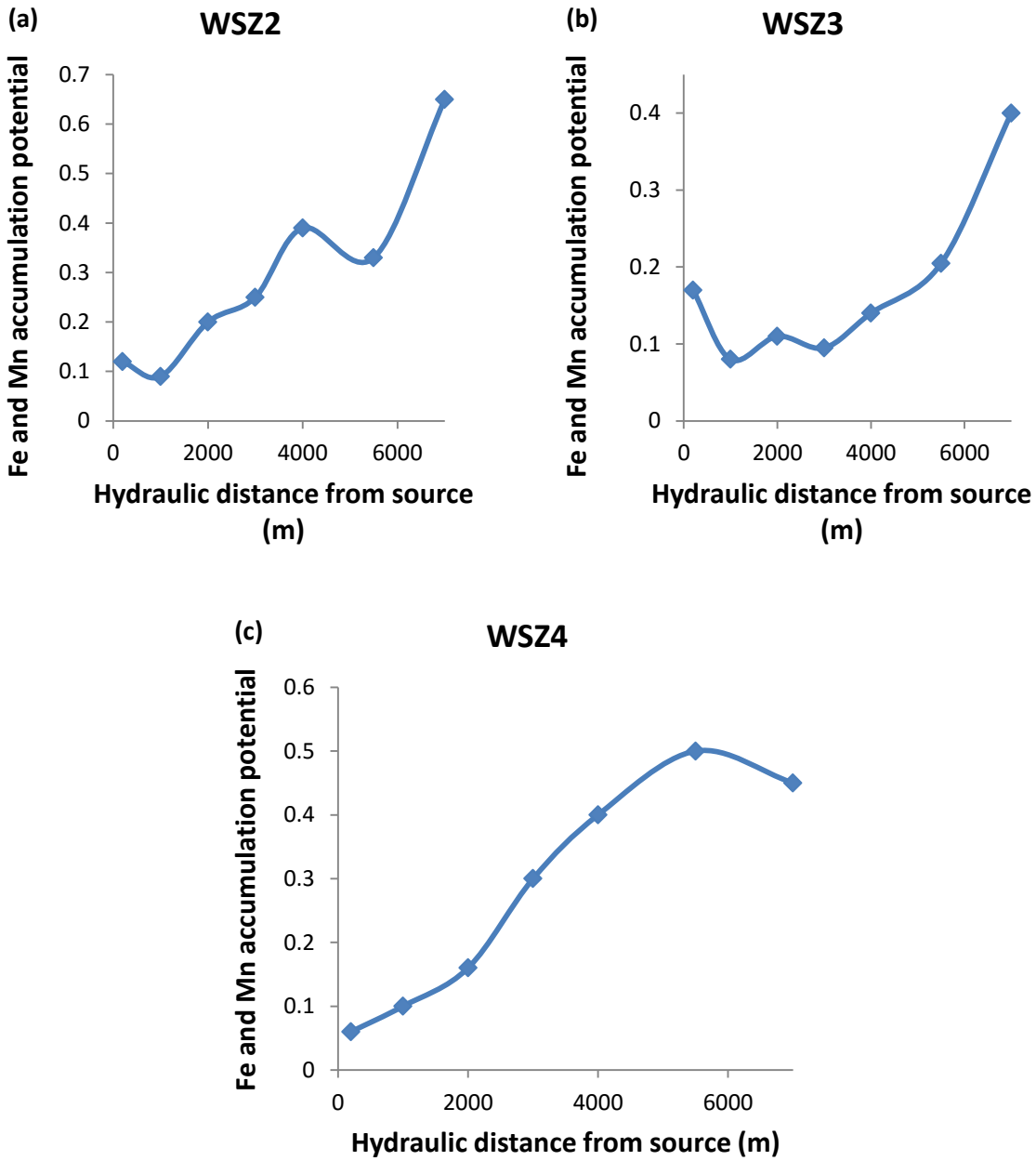


Figure Q.7 Relationship between Fe and Mn accumulation potential and hydraulic distance from source of water supply

Appendix R: Graphs of measured yearly average Fe and Mn accumulation potential and predicted high-risk nodes

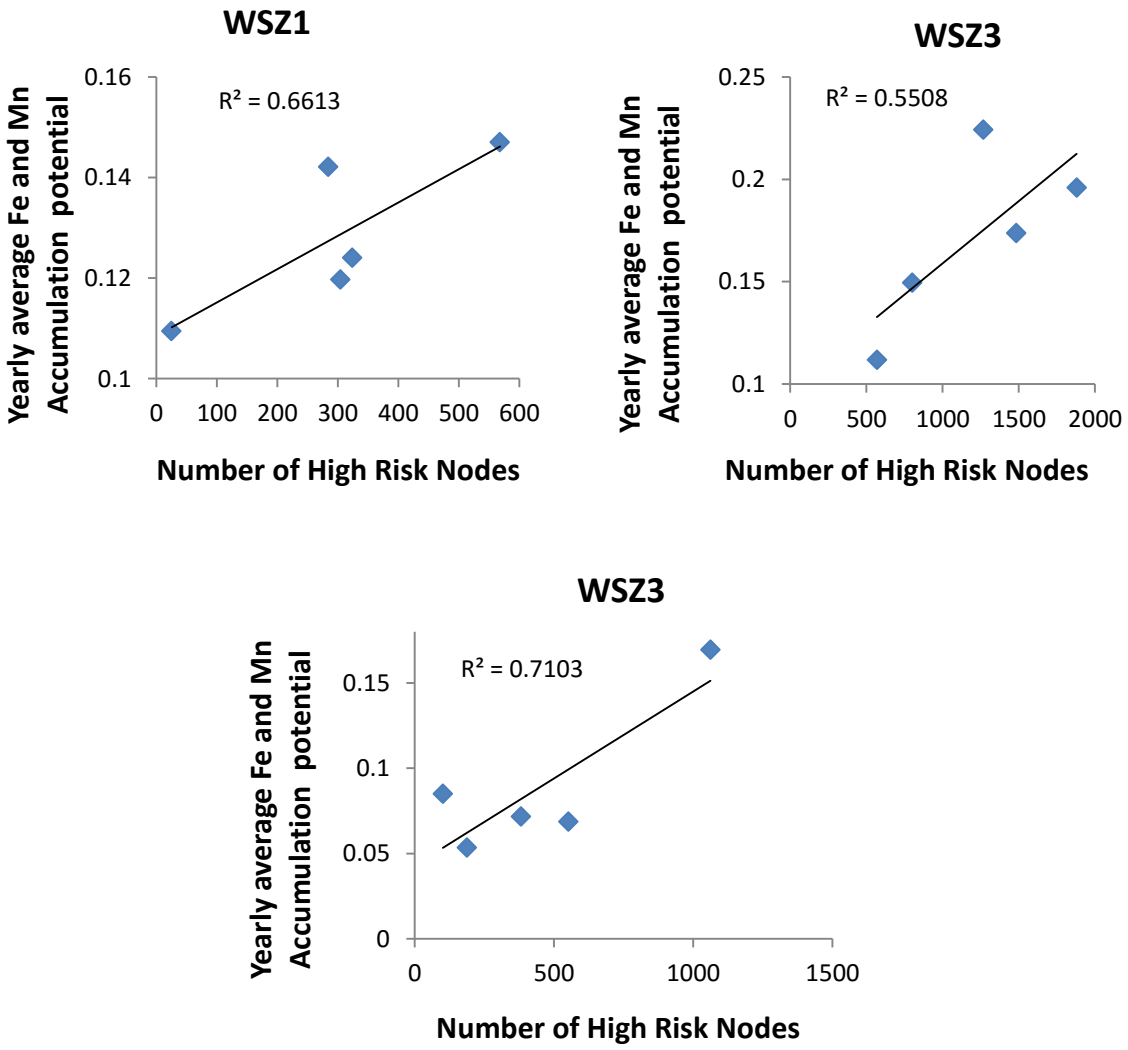


Figure R.1 Correlation between measured yearly average Fe and Mn accumulation potential and predicted high-risk nodes from 2005-2009

Appendix S: Results from the FIS

Table S.1 Rules and their corresponding weights from the hierarchical rule-based expert FIS for WSZ2

Rule Number	Rules from expert knowledge	Weights from expert knowledge
1	If Hardness is LOW then Chemical oxidation is LOW	0.9
2	If Hardness is MEDIUM then Chemical oxidation is MEDIUM	0.9
3	If Hardness is HIGH then Chemical oxidation is HIGH	0.9
4	If FCR is LOW then Chemical oxidation is LOW	0.9
5	If FCR is MEDIUM then Chemical oxidation is MEDIUM	0.9
6	If FCR is HIGH then Chemical oxidation is HIGH	0.9
7	If Alkalinity is LOW then Chemical oxidation is HIGH	0.9
8	If Alkalinity is MEDIUM then Chemical oxidation is MEDIUM	0.9
9	If Alkalinity is HIGH then Chemical oxidation is LOW	0.9
10	If Chemical oxidation is LOW then Corrosion is LOW	1.0
11	If Chemical oxidation is MEDIUM then Corrosion is MEDIUM	1.0
12	If Chemical oxidation is HIGH then Corrosion is HIGH	1.0
13	If Pipe material index is LOW then Corrosion is LOW	1.0
14	If Pipe material index is MEDIUM then Corrosion is MEDIUM	1.0
15	If Pipe material index is HIGH then Corrosion is HIGH	1.0
16	If Pipe age is LOW then Corrosion is LOW	1.0
17	If Pipe age is MEDIUM then Corrosion is MEDIUM	1.0
18	If Pipe age is HIGH then Corrosion is HIGH	1.0
19	If Calcium is LOW then Sorption is LOW	0.6
20	If Calcium is MEDIUM then Sorption is MEDIUM	0.6
21	If Calcium is HIGH then Sorption is HIGH	0.6
22	If Aluminium is LOW then Sorption is LOW	0.6
23	If Aluminium is MEDIUM then Sorption is MEDIUM	0.6
24	If Aluminium is HIGH then Sorption is HIGH	0.6
25	If Colour is LOW then Sorption is LOW	0.6
26	If Colour is MEDIUM then Sorption is MEDIUM	0.6
27	If Colour is HIGH then Sorption is HIGH	0.75
28	If Maximum shear stress is LOW then Shear stress effect is HIGH	0.75
29	If Maximum shear stress is MEDIUM then Shear stress effect is MEDIUM	0.75
30	If Maximum shear stress is HIGH then Shear stress effect is LOW	0.75
31	If Variation of shear stress is LOW then Shear stress effect is HIGH	0.75
32	If Variation of shear stress is MEDIUM then Shear stress effect is MEDIUM	0.75
33	If Variation of shear stress is HIGH then Shear stress effect is LOW	0.75

Table S.1 Rules and their corresponding weights from the hierarchical rule-based expert FIS for WSZ2 continued

Rule Number	Rules from expert knowledge	Weights from expert knowledge
34	If Average water age is LOW then Distance effect is LOW	0.9
35	If Average water age is MEDIUM then Distance effect is MEDIUM	0.9
36	If Average water age is HIGH then Distance effect is HIGH	0.9
37	If Distance from source is LOW then Distance effect is LOW	0.9
38	If Distance from source is MEDIUM then Distance effect is MEDIUM	0.9
39	If Distance from source is HIGH then Distance effect is HIGH	0.9
40	If Chemical oxidation is LOW then Chemical effect is LOW	0.7
41	If Chemical oxidation is MEDIUM then Chemical effect is MEDIUM	0.7
42	If Chemical oxidation is HIGH then Chemical effect is HIGH	0.7
43	If Corrosion is LOW then Chemical effect is LOW	1.0
44	If Corrosion is MEDIUM then Chemical effect is MEDIUM	1.0
45	If Corrosion is HIGH then Chemical effect is HIGH	1.0
46	If Sorption is LOW then Chemical effect is HIGH	0.7
47	If Sorption is MEDIUM then Chemical effect is MEDIUM	0.7
48	If Sorption is HIGH then Chemical effect is LOW	0.7
49	If FCR is LOW then Biological effect is HIGH	0.9
50	If FCR is MEDIUM then Biological effect is MEDIUM	0.9
51	If FCR is HIGH then Biological effect is LOW	0.9
52	If Colour is LOW then Biological effect is LOW	0.9
53	If Colour is MEDIUM then Biological effect is MEDIUM	0.9
54	If Colour is HIGH then Biological effect is HIGH	0.9
55	If Average water age is LOW then Biological effect is LOW	0.9
56	If Average water age is MEDIUM then Biological effect is MEDIUM	0.9
57	If Average water age is HIGH then Biological effect is HIGH	0.9
58	If Turbidity is LOW then Biological effect is LOW	0.9
59	If Turbidity is MEDIUM then Biological effect is MEDIUM	0.9
60	If Turbidity is HIGH then Biological effect is HIGH	0.9
61	If Phosphorus is LOW then Biological effect is LOW	0.9
62	If Phosphorus is MEDIUM then Biological effect is MEDIUM	0.9
63	If Phosphorus is HIGH then Biological effect is HIGH	0.9
64	If Shear stress effect is LOW then Hydraulic effect is LOW	0.6
65	If Shear stress effect is MEDIUM then Hydraulic effect is MEDIUM	0.6
66	If Shear stress effect is HIGH then Hydraulic effect is HIGH	0.6
67	If Distance effect is LOW then Hydraulic effect is LOW	0.6

Table S.1 Rules and their corresponding weights from the hierarchical rule-based expert FIS for WSZ2 continued

Rule Number	Rules from expert knowledge	Weights from expert knowledge
68	If Distance effect is MEDIUM then Hydraulic effect is MEDIUM	0.6
69	If Distance effect is HIGH then Hydraulic effect is HIGH	0.6
70	If Chemical effect is LOW then Fe and Mn Accumulation Potential is LOW	0.8
71	If Chemical effect is MEDIUM then Fe and Mn Accumulation Potential is MEDIUM	0.8
72	If Chemical effect is HIGH then Fe and Mn Accumulation Potential is HIGH	0.8
73	If Biological effect is LOW then Fe and Mn Accumulation Potential is LOW	0.9
74	If Biological effect is MEDIUM then Fe and Mn Accumulation Potential is MEDIUM	0.9
75	If Biological effect is HIGH then Fe and Mn Accumulation Potential is HIGH	0.9
76	If Hydraulic effect is LOW then Fe and Mn Accumulation Potential is LOW	0.8
77	If Hydraulic effect is MEDIUM then Fe and Mn Accumulation Potential is MEDIUM	0.8
78	If Hydraulic effect is HIGH then Fe and Mn Accumulation Potential is HIGH	0.8

Table S.2 Rules and their corresponding weights from the hierarchical data-driven FIS for WSZ2

Rule Number	Rules after optimisation	Weights after optimisation
1	If Hardness is LOW then Chemical oxidation is MEDIUM	0.4107
2	If Hardness is MEDIUM then Chemical oxidation is LOW	0.3711
3	If Hardness is HIGH then Chemical oxidation is LOW	0.7952
4	If FCR is LOW then Chemical oxidation is HIGH	0.5760
5	If FCR is MEDIUM then Chemical oxidation is MEDIUM	0.5613
6	If FCR is HIGH then Chemical oxidation is MEDIUM	0.3699
7	If Alkalinity is LOW then Chemical oxidation is HIGH	0.5656
8	If Alkalinity is MEDIUM then Chemical oxidation is HIGH	0.4491
9	If Alkalinity is HIGH then Chemical oxidation is MEDIUM	0.3667
10	If Chemical oxidation is LOW then Corrosion is MEDIUM	0.4668
11	If Chemical oxidation is MEDIUM then Corrosion is MEDIUM	0.2811
12	If Chemical oxidation is HIGH then Corrosion is LOW	0.7013
13	If Pipe material index is LOW then Corrosion is MEDIUM	0.5214
14	If Pipe material index is MEDIUM then Corrosion is LOW	0.2452
15	If Pipe material index is HIGH then Corrosion is MEDIUM	0.4890
16	If Pipe age is LOW then Corrosion is LOW	0.5701
17	If Pipe age is MEDIUM then Corrosion is LOW	0.4383
18	If Pipe age is HIGH then Corrosion is MEDIUM	0.2985
19	If Calcium is LOW then Sorption is HIGH	0.6080
20	If Calcium is MEDIUM then Sorption is MEDIUM	0.7014
21	If Calcium is HIGH then Sorption is MEDIUM	0.4569
22	If Aluminium is LOW then Sorption is HIGH	0.6724
23	If Aluminium is MEDIUM then Sorption is HIGH	0.6577
24	If Aluminium is HIGH then Sorption is LOW	0.4827
25	If Colour is LOW then Sorption is MEDIUM	0.6550
26	If Colour is MEDIUM then Sorption is HIGH	0.5176
27	If Colour is HIGH then Sorption is HIGH	0.2943
28	If Maximum shear stress is LOW then Shear stress effect is LOW	0.3714
29	If Maximum shear stress is MEDIUM then Shear stress effect is HIGH	0.5035
30	If Maximum shear stress is HIGH then Shear stress effect is HIGH	0.4864
31	If Variation of shear stress is LOW then Shear stress effect is LOW	0.4596
32	If Variation of shear stress is MEDIUM then Shear stress effect is HIGH	0.6906
33	If Variation of shear stress is HIGH then Shear stress effect is HIGH	0.4778
34	If Average water age is LOW then Distance effect is MEDIUM	0.5690

Table S.2 Rule2 and their corresponding weights from the hierarchical data-driven FIS for WSZ2 continued

Rule Number	Rules after optimisation	Weights after optimisation
35	If Average water age is MEDIUM then Distance effect is LOW	0.4344
36	If Average water age is HIGH then Distance effect is LOW	0.6127
37	If Distance from source is LOW then Distance effect is LOW	0.4960
38	If Distance from source is MEDIUM then Distance effect is MEDIUM	0.7344
39	If Distance from source is HIGH then Distance effect is LOW	0.3634
40	If Chemical oxidation is LOW then Chemical effect is HIGH	0.5635
41	If Chemical oxidation is MEDIUM then Chemical effect is HIGH	0.2577
42	If Chemical oxidation is HIGH then Chemical effect is MEDIUM	0.2917
43	If Corrosion is LOW then Chemical effect is MEDIUM	0.6755
44	If Corrosion is MEDIUM then Chemical effect is LOW	0.2321
45	If Corrosion is HIGH then Chemical effect is MEDIUM	0.5092
46	If Sorption is LOW then Chemical effect is MEDIUM	0.5641
47	If Sorption is MEDIUM then Chemical effect is HIGH	0.3497
48	If Sorption is HIGH then Chemical effect is HIGH	0.5422
49	If FCR is LOW then Biological effect is HIGH	0.3163
50	If FCR is MEDIUM then Biological effect is HIGH	0.6469
51	If FCR is HIGH then Biological effect is HIGH	0.6256
52	If Colour is LOW then Biological effect is MEDIUM	0.4055
53	If Colour is MEDIUM then Biological effect is HIGH	0.6338
54	If Colour is HIGH then Biological effect is HIGH	0.5545
55	If Average water age is LOW then Biological effect is HIGH	0.7453
56	If Average water age is MEDIUM then Biological effect is MEDIUM	0.4455
57	If Average water age is HIGH then Biological effect is HIGH	0.6680
58	If Turbidity is LOW then Biological effect is MEDIUM	0.6479
59	If Turbidity is MEDIUM then Biological effect is HIGH	0.3436
60	If Turbidity is HIGH then Biological effect is HIGH	0.7268
61	If Phosphorus is LOW then Biological effect is LOW	0.3801
62	If Phosphorus is MEDIUM then Biological effect is HIGH	0.6458
63	If Phosphorus is HIGH then Biological effect is MEDIUM	0.2964
64	If Shear stress effect is LOW then Hydraulic effect is LOW	0.6986
65	If Shear stress effect is MEDIUM then Hydraulic effect is LOW	0.5170
66	If Shear stress effect is HIGH then Hydraulic effect is LOW	0.3959
67	If Distance effect is LOW then Hydraulic effect is MEDIUM	0.4881
68	If Distance effect is MEDIUM then Hydraulic effect is MEDIUM	0.4031
69	If Distance effect is HIGH then Hydraulic effect is MEDIUM	0.4769

Table S.2 Rules and their corresponding weights from the hierarchical data-driven FIS for WSZ2 continued

Rule Number	Rules after optimisation	Weights after optimisation
70	If Chemical effect is LOW then Fe and Mn Accumulation Potential is MEDIUM	0.6556
71	If Chemical effect is MEDIUM then Fe and Mn Accumulation Potential is HIGH	0.5015
72	If Chemical effect is HIGH then Fe and Mn Accumulation Potential is MEDIUM	0.5852
73	If Biological effect is LOW then Fe and Mn Accumulation Potential is MEDIUM	0.2359
74	If Biological effect is MEDIUM then Fe and Mn Accumulation Potential is LOW	0.6829
75	If Biological effect is HIGH then Fe and Mn Accumulation Potential is LOW	0.4641
76	If Hydraulic effect is LOW then Fe and Mn Accumulation Potential is HIGH	0.6766
77	If Hydraulic effect is MEDIUM then Fe and Mn Accumulation Potential is HIGH	0.4635
78	If Hydraulic effect is HIGH then Fe and Mn Accumulation Potential is LOW	0.5634

As observed in Tables S.1 and S.2, the antecedent parts of the rules do not change in both the hierarchical rule-based expert FIS and hierarchical data-driven FIS. Also, the same weights based on expert knowledge were used for all WSZs in the hierarchical rule-based expert FIS. Hence, they will be omitted from subsequent results. The chromosomes in the genetic algorithm of the hierarchical data-driven FIS consist of the numbers 1, 2, and 3 representing LOW, MEDIUM, and HIGH. Due to limited space, the results for the remaining WSZs will be reported with these numbers.

Table S.3 Rules and their corresponding weights from the hierarchical data-driven FIS for WSZ1, WSZ3, WSZ4, and WSZ5

Rule Number	Rule consequents after optimisation				Weights after optimisation			
	WSZ1	WSZ3	WSZ4	WSZ5	WSZ1	WSZ3	WSZ4	WSZ5
1	1	3	1	2	0.7289	0.9510	0.5122	0.2118
2	3	1	1	1	0.4254	0.4709	0.6109	0.4682
3	2	2	3	3	0.5231	0.1983	0.7537	0.4509
4	1	2	3	3	0.7340	0.0181	0.6177	0.3947
5	3	3	3	2	0.3267	0.3870	0.4352	0.4028
6	1	2	1	1	0.4317	0.7710	0.4920	0.2471
7	2	3	3	3	0.4756	0.6247	0.5829	0.3474
8	3	3	1	1	0.4981	0.0873	0.2237	0.1268
9	3	2	1	3	0.4983	0.1342	0.4678	0.6359
10	2	3	1	2	0.8036	0.1770	0.4468	0.3355
11	1	1	1	1	0.2587	0.4785	0.3839	0.4514
12	1	3	1	2	0.6652	0.2563	0.4944	0.1941
13	3	1	3	3	0.8447	0.5324	0.6292	0.2357
14	2	3	3	2	0.6981	0.3859	0.4932	0.0100
15	1	3	2	2	0.4000	0.3765	0.2731	0.7438
16	2	3	3	3	0.4651	0.2862	0.5036	0.6295
17	1	3	3	3	0.3896	0.0779	0.5969	0.1698
18	3	3	2	3	0.1241	0.6980	0.3048	0.0660
19	1	2	2	3	0.3485	0.6448	0.3839	0.3784
20	3	2	2	1	0.4450	0.4231	0.3693	0.5624
21	2	3	1	3	0.6740	0.6200	0.6081	0.3312
22	3	3	1	2	0.5935	0.1302	0.4536	0.6525
23	2	3	3	1	0.5382	0.6104	0.6524	0.4171
24	1	3	2	2	0.6200	0.5838	0.6470	0.4536
25	1	1	1	2	0.2600	0.5763	0.7375	0.7070
26	3	2	3	1	0.3954	0.6104	0.4580	0.7900
27	3	2	2	3	0.7142	0.1476	0.5760	0.5129
28	3	2	2	3	0.3226	0.5208	0.3511	0.7615
29	3	1	2	1	0.5130	0.7225	0.3619	0.9287
30	1	1	1	2	0.5745	0.5726	0.4556	0.7100
31	2	3	3	2	0.7625	0.1294	0.5583	0.7158
32	3	2	1	3	0.2952	0.5971	0.5728	0.5104
33	2	1	1	2	0.2466	0.8732	0.7520	0.0259
34	2	1	3	1	0.3193	0.8417	0.4356	0.5766
35	3	2	3	2	0.4409	0.2704	0.2989	0.6259
36	3	3	1	2	0.5147	0.2908	0.6425	0.8606
37	1	3	1	1	0.3015	0.4056	0.5822	0.3344
38	3	3	3	2	0.5123	0.6277	0.8013	0.1631
39	3	2	2	1	0.2644	0.1142	0.6555	0.5810

Table S.3 Rules and their corresponding weights from the hierarchical data-driven FIS for WSZ1, WSZ3, WSZ4, and WSZ5 continued

Rule Number	Rule consequents after optimisation				Weights after optimisation			
	WSZ1	WSZ3	WSZ4	WSZ5	WSZ1	WSZ3	WSZ4	WSZ5
40	3	1	3	1	0.9541	0.4028	0.7595	0.4194
41	3	3	1	3	0.6431	0.7589	0.4538	0.7093
42	3	3	2	1	0.1760	0.6774	0.3432	0.7464
43	1	3	2	3	0.4619	0.3546	0.5698	0.9224
44	1	3	3	3	0.3735	0.0780	0.5304	0.3642
45	1	1	2	1	0.7647	0.6069	0.4723	0.0100
46	3	3	1	3	0.4449	0.3307	0.4539	0.5973
47	3	3	1	2	0.1247	0.6971	0.5995	0.0001
48	2	2	1	3	0.4393	0.3715	0.4548	0.4991
49	1	1	3	3	0.7035	0.2781	0.5448	0.7844
50	3	3	2	1	0.4101	0.4817	0.1032	0.1588
51	3	1	3	1	0.6666	0.3031	0.5716	0.4799
52	1	3	1	1	0.7279	0.5665	0.5424	0.4874
53	1	3	2	1	0.4430	0.0428	0.5198	0.5313
54	1	3	3	3	0.1145	0.5847	0.6609	0.8102
55	3	1	3	1	0.4666	0.8799	0.2596	0.6595
56	1	1	3	2	0.4090	0.4032	0.3552	0.5715
57	2	1	2	3	0.4748	0.4055	0.3418	0.2711
58	3	1	1	1	0.4836	0.4092	0.5796	0.7178
59	2	3	3	2	0.4077	0.4157	0.5077	0.8541
60	1	3	3	1	0.7159	0.6356	0.6415	0.2200
61	3	3	1	2	0.5755	0.6125	0.5230	0.8865
62	1	3	2	1	0.3151	0.8438	0.6647	0.5987
63	1	3	3	1	0.8668	0.0005	0.7091	0.4773
64	1	1	3	3	0.5476	0.1432	0.3609	0.6436
65	3	2	3	1	0.3639	1.0000	0.9980	0.8827
66	2	2	1	2	0.4109	0.7251	0.5225	0.4012
67	1	1	3	3	0.4196	0.1684	0.8146	0.4714
68	2	3	1	3	0.5876	0.9223	0.5206	0.2910
69	1	3	3	3	0.5575	0.9503	0.8327	0.3564
70	1	2	1	1	0.3827	0.2563	0.8031	0.4736
71	3	3	3	1	0.3386	0.4262	0.2604	0.4833
72	3	1	3	3	0.3661	0.6626	0.5101	0.6920
73	3	1	2	2	0.2384	0.2050	0.6044	0.4187
74	1	1	1	1	0.5687	0.0482	0.4311	0.4067
75	2	3	1	1	0.5584	0.4187	0.5018	0.7447
76	1	2	1	1	0.2612	0.7867	0.1941	0.6721
77	1	1	1	3	0.0122	0.0244	0.5840	0.6800
78	3	3	2	2	0.5879	0.5315	0.7107	0.3815

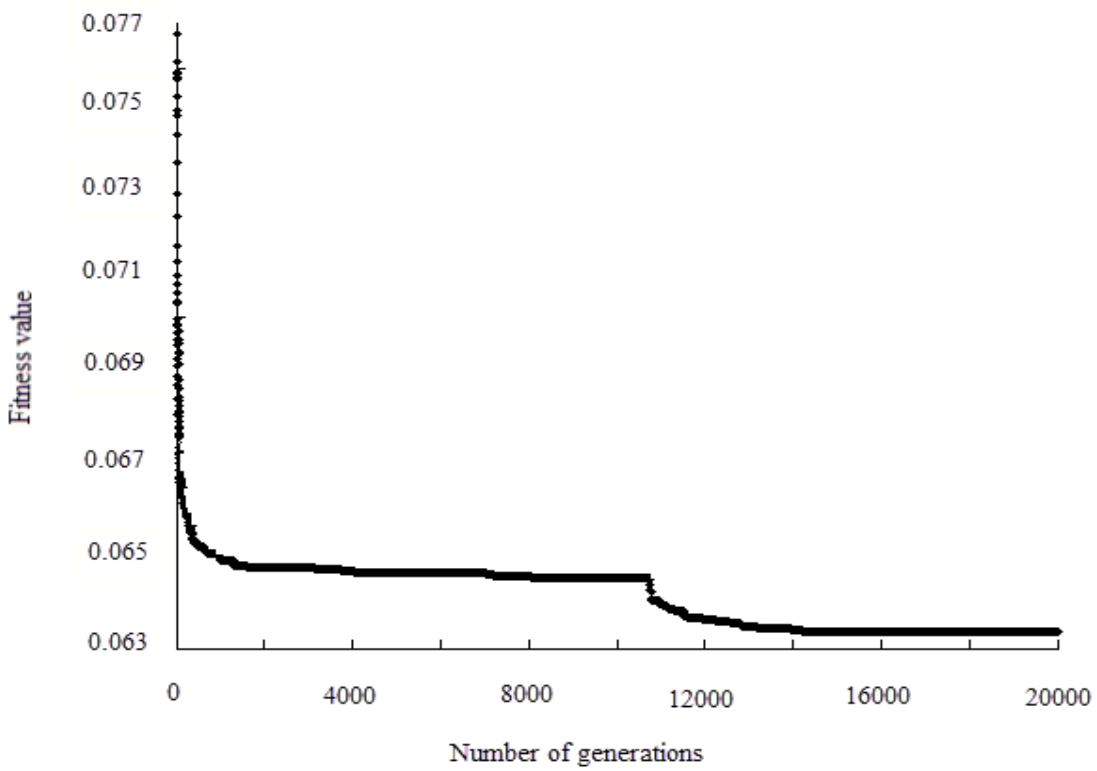


Figure S.1 Fitness function graph for WSZ5 during rule optimisation

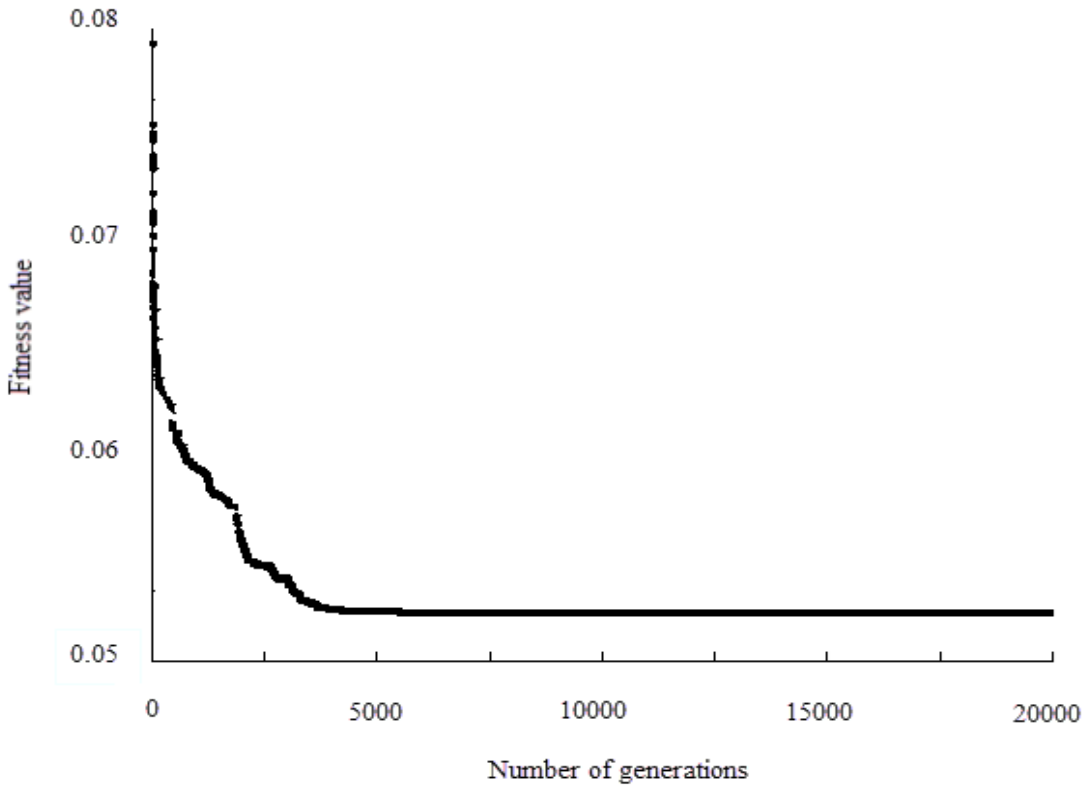


Figure S.2 Fitness function graph for WSZ3 during rule optimisation

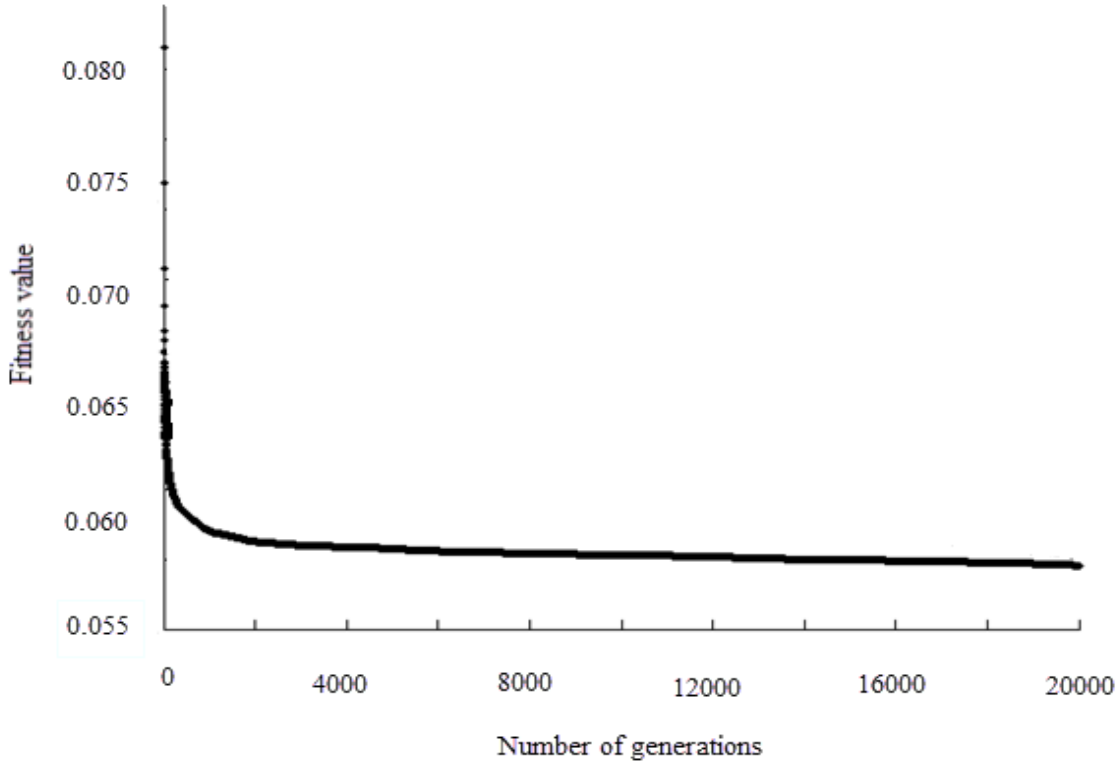


Figure S.3 Fitness function graph for WSZ4 during rule optimisation

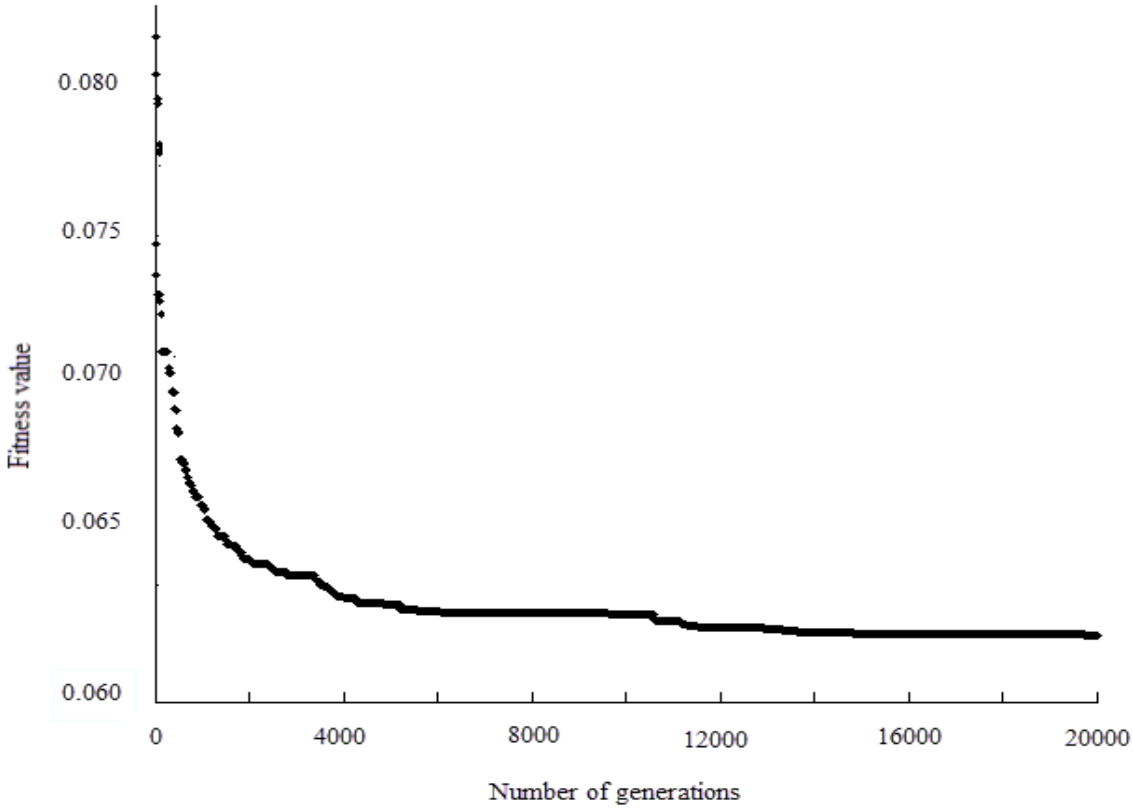


Figure S.4 Fitness function graph for WSZ1 during rule optimisation

Appendix T: Risk maps generated by the ANN(t,ψ) model

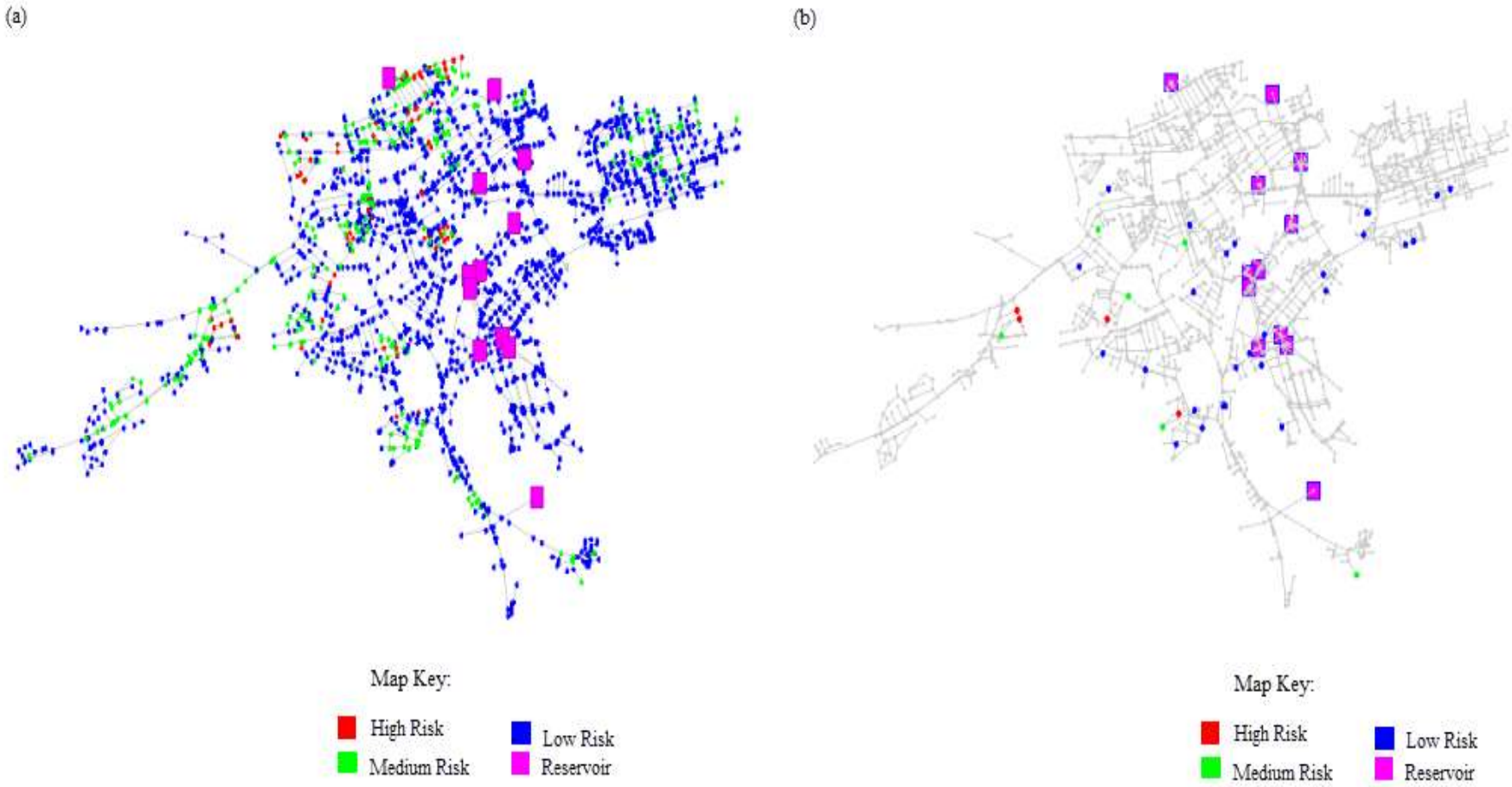


Figure T.1 ANN(t,ψ) model risk maps showing (a) Predicted and (b) measured Fe and Mn accumulation potential at WSZ1 in 2009

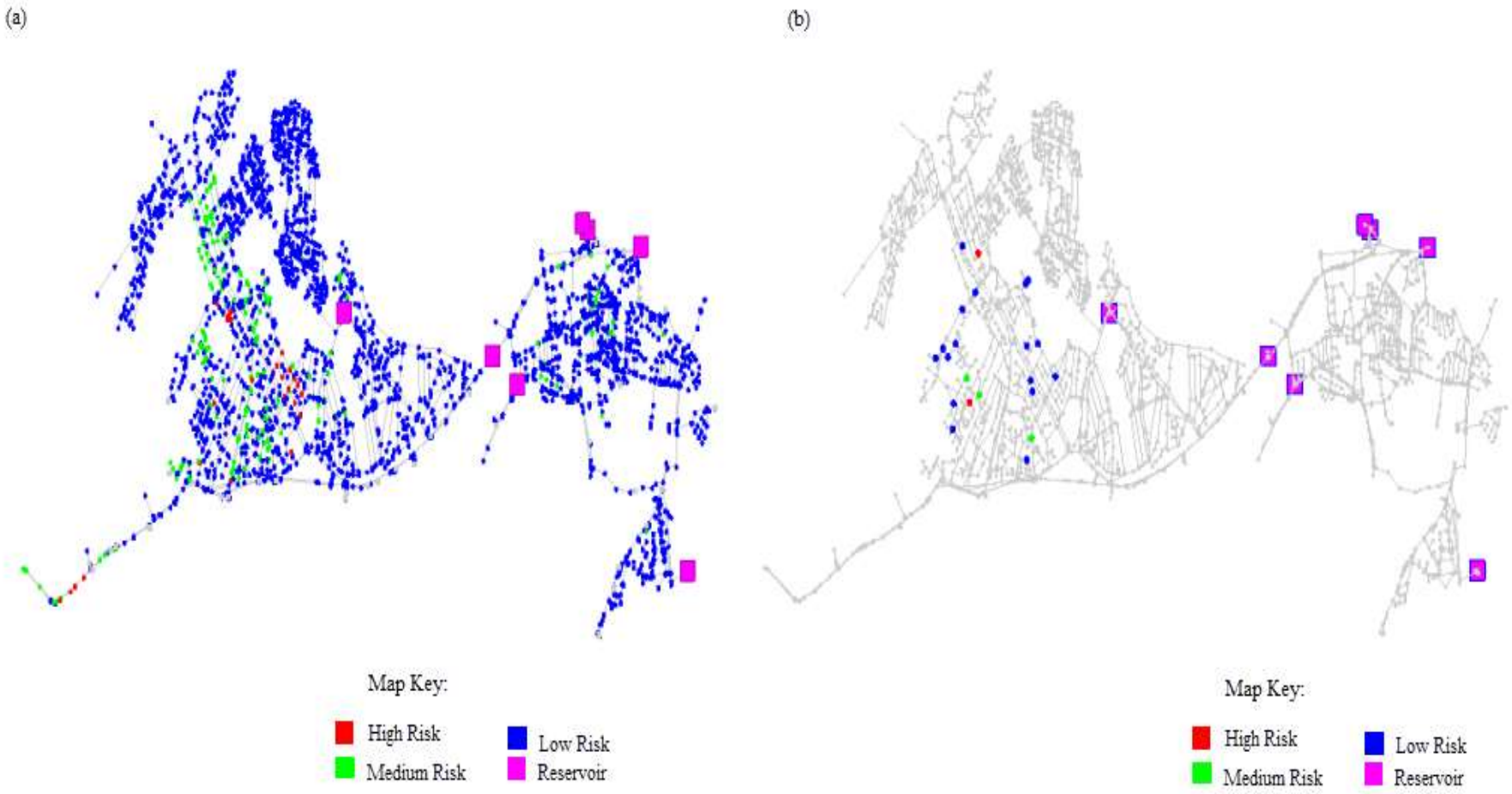


Figure T.2 ANN(t,ψ) model risk maps showing (a) Predicted and (b) measured Fe and Mn accumulation potential at WSZ3 in 2006

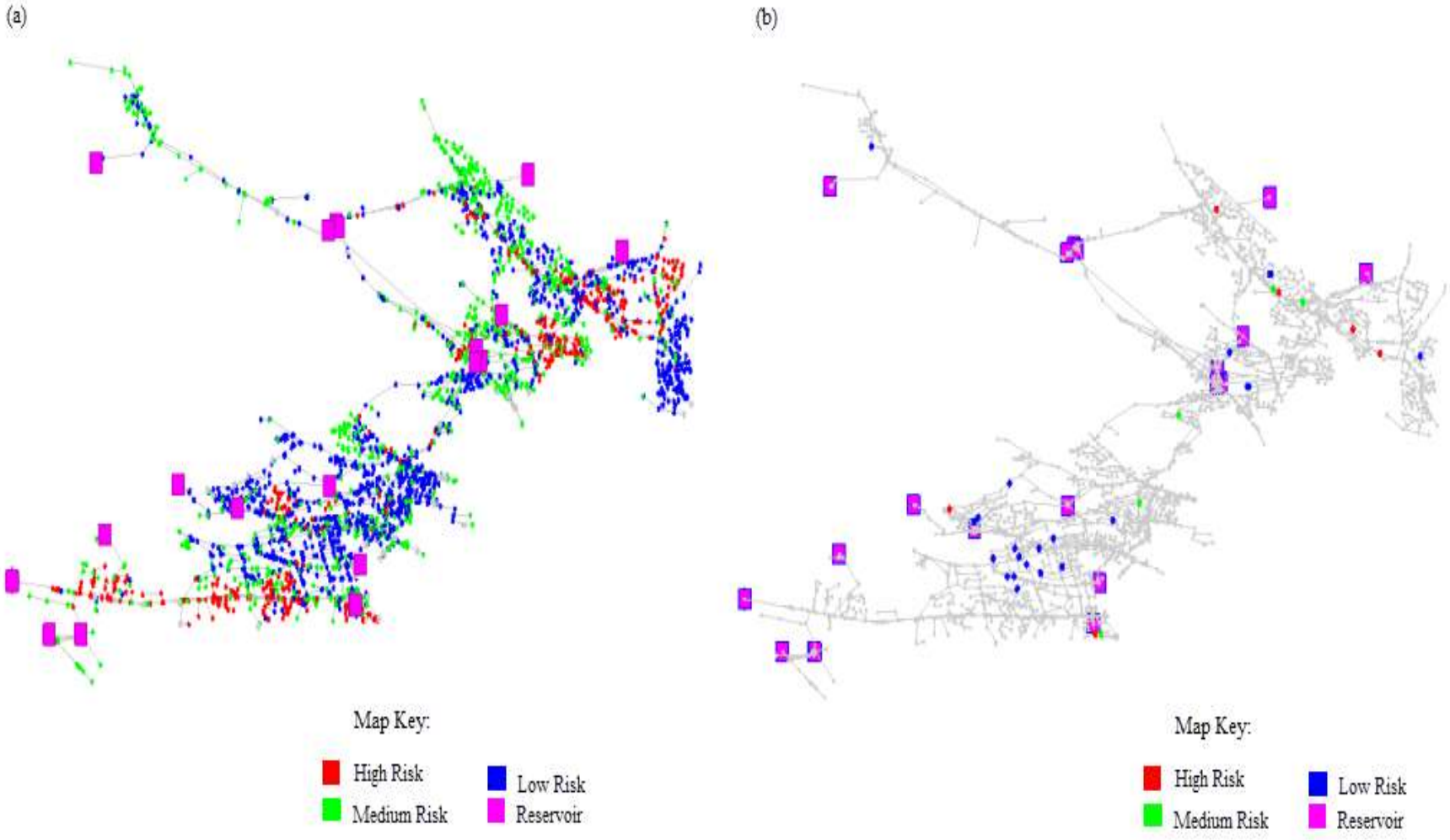


Figure T.3 ANN(t,ψ) model risk maps showing (a) Predicted and (b) measured Fe and Mn accumulation potential at WSZ4 in 2005

Appendix U: Risk maps generated by the Hierarchical data-driven FIS

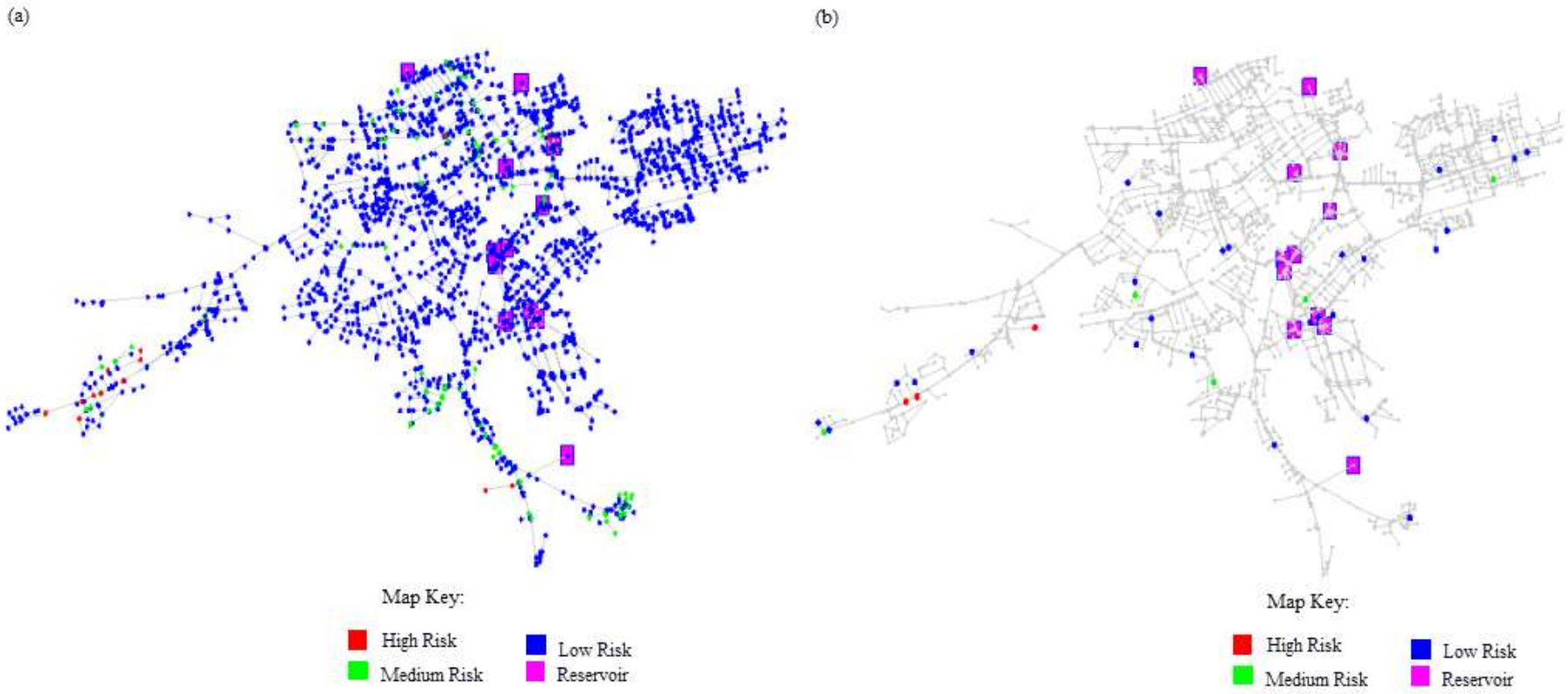


Figure U.1 Hierarchical data-driven FIS risk maps showing (a) predicted and (b) measured Fe and Mn accumulation potential at WSZ1 in 2008

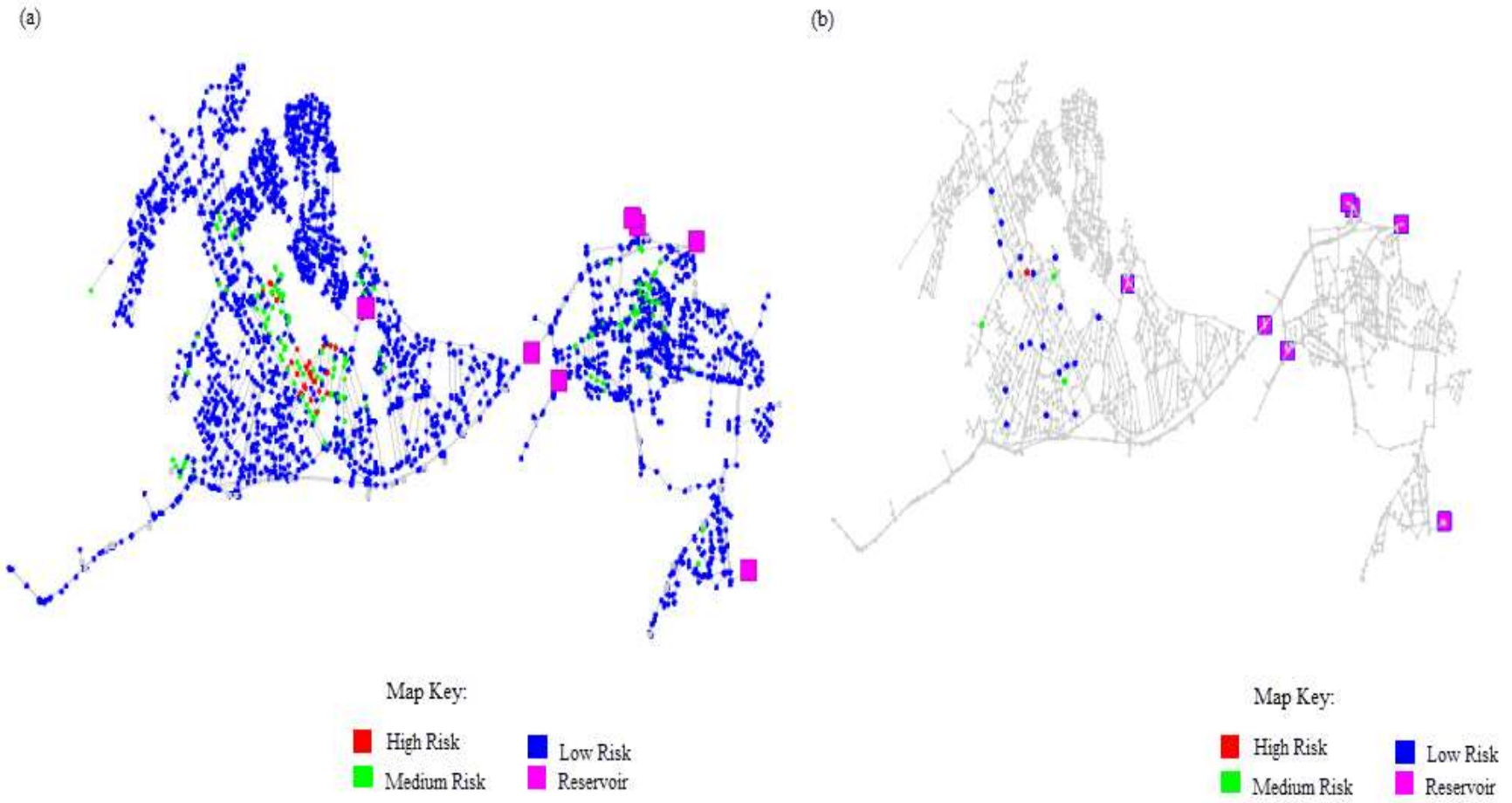


Figure U.2 Hierarchical data-driven FIS risk maps showing (a) predicted and (b) measured Fe and Mn accumulation potential at WSZ3 in 2008

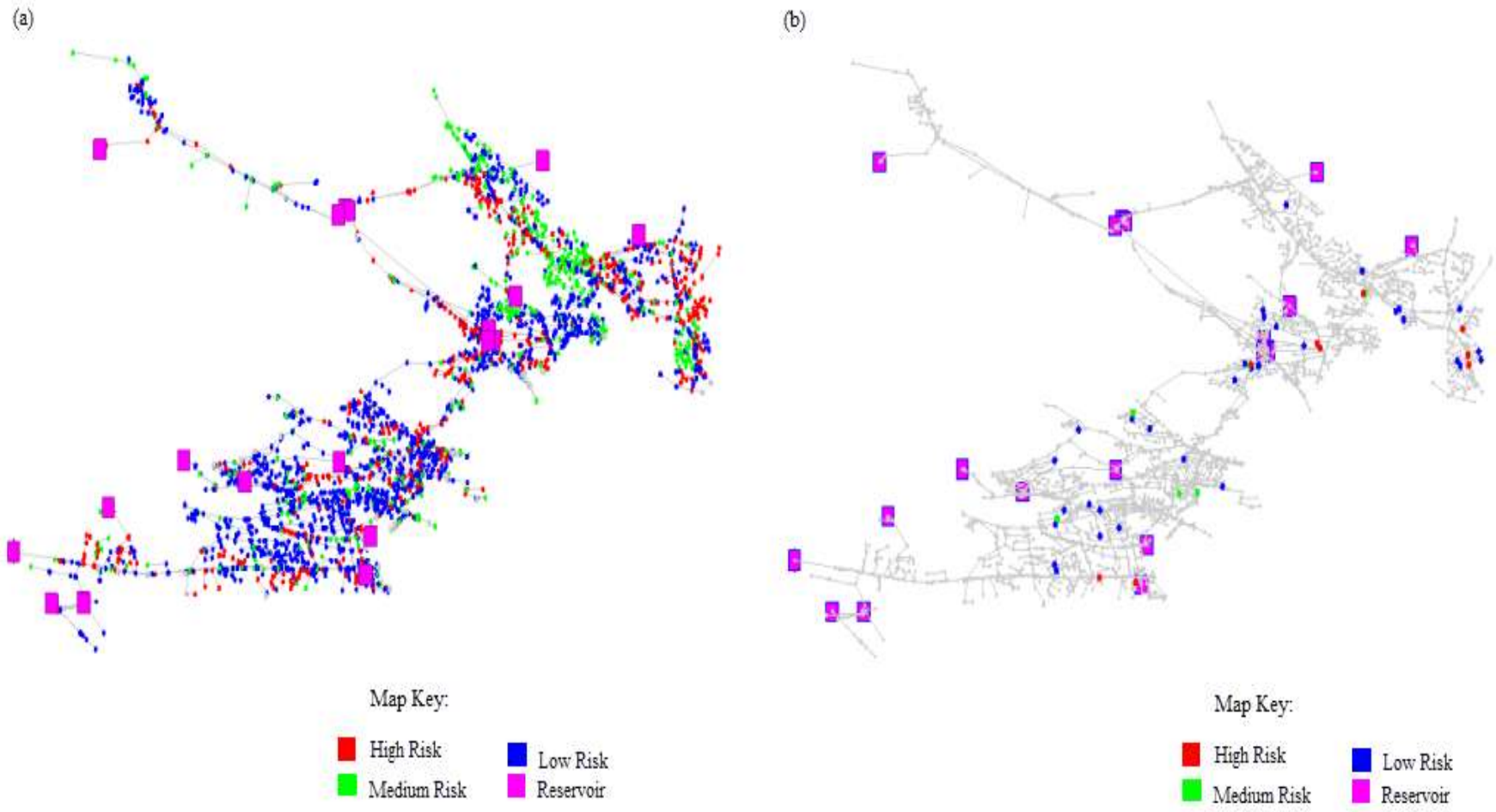


Figure U.3 Hierarchical data-driven FIS risk maps showing (a) predicted and (b) measured Fe and Mn accumulation potential at WSZ4 in 2009

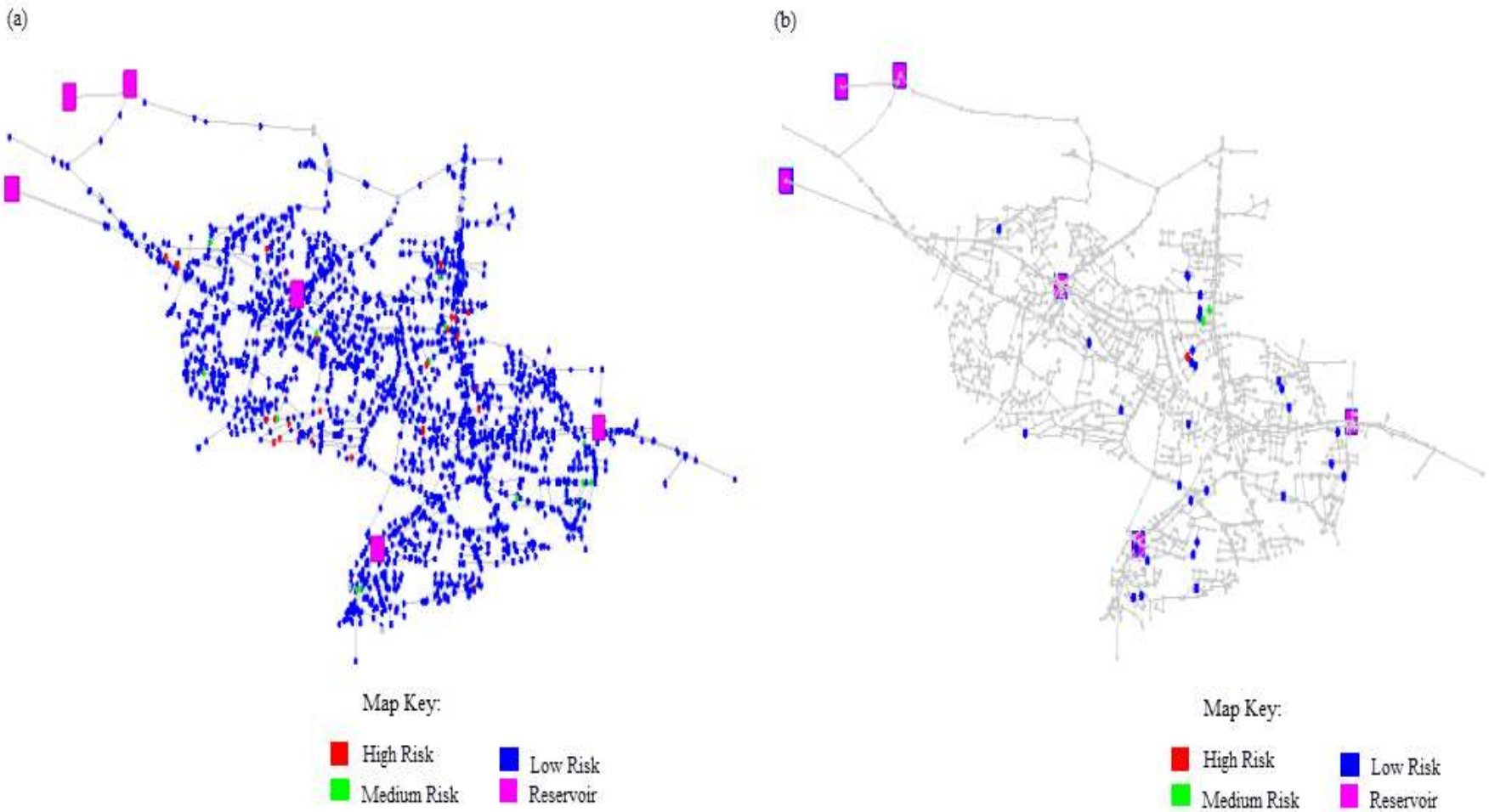


Figure U.4 Hierarchical data-driven FIS risk maps showing (a) predicted and (b) measured Fe and Mn accumulation potential at WSZ5 in 2006