CRANFIELD UNIVERSITY

Matthew Robert North

The Impact of Soils, Weather and Trees on Water Infrastructure Failure

School of Water, Energy and Environment

PhD
Academic Year: 2018 - 2019

Supervisor: Dr Timothy Farewell
Associate Supervisor: Dr Stephen Hallett
September 2018

CRANFIELD UNIVERSITY

School of Water, Energy and Environment

PhD

Academic Year 2018 - 2019

Matthew Robert North

The Impact of Soils, Weather and Trees on Water Infrastructure Failure

Supervisor: Dr Timothy Farewell
Associate Supervisor: Dr Stephen Hallett
September 2018

This thesis is submitted in partial fulfilment of the requirements for the degree of PhD

ABSTRACT

The uninterrupted supply and reliable distribution of drinking water is fundamental in a modern society; however, water pipelines are subject to a range of operational and environmental factors which can lead to asset failure. For the privatised water-sector in the UK, utility companies are moving towards the deployment of statistical models for proactive asset management. For some companies, statistical models have facilitated the migration away from static annual burst targets, to targets which are dynamic and adjusted to observed environmental conditions. There is an increasing need for the development of accurate pipeline failure prediction models to support asset management and regulatory reporting. This thesis evaluates several quantitative measures to improve current methods of pipeline failure prediction. The aim of this thesis is to establish the impact of soils, weather and trees on water infrastructure failure and to develop a series of material-specific drinking water pipeline failure models for an entire distribution network.

A quantitative assessment investigating the impact of data cleaning on the attained model error of a series of previously developed models was conducted. Material-specific variable selection and step-wise modelling approaches was used to construct a series of improved statistical models, which have an increased representation of the environmental factors leading to pipeline failure. A detailed national tree inventory was investigated for its use in enhancing pipeline failure predictions and for calculating failure rates of different pipe materials under varying soil shrink swell and tree density conditions. The value in creating separate winter and summer pipeline failure models was also evaluated, to increase representation of the highly seasonal nature of pipeline failure. Finally, a satellite approach was used to generate soil-related land surface deformation measurements across a regional area was investigated. The result is a series of enhanced statistical models for the prediction of water pipeline failure and a greater understanding into the environmental drivers leading to asset failure.

Keywords: Statistical modelling; pipeline failure; prediction; environmental risk; water utilities

ACKNOWLEDGEMENTS

I would like to thank the Natural Environment Research Council (NERC) for providing funding for this project and also to the Data, Risks and Environmental Analytical Methods (DREAM) centre for doctoral training. I would also like to thank Anglian Water plc., Bluesky ltd., and the Met Office for the provision of data used throughout the research project.

To my supervisors, Tim and Steve, who have guided me and given me the opportunity to undertake this PhD research, thank you. Your patience and kindness have made the process highly rewarding and enjoyable. And to my other colleagues at Cranfield, particularly to Angela and Juan who have consistently supported me.

Above anyone else, I would like to thank my mother for being my inspiration to work hard and achieve my goals. Also, to my amazing brother and gran. I would also like to personally thank Michelle and Maurice for making me feel incredibly welcomed and like family during my time living with them in Milton Keynes.

I would also like to acknowledge all of my friends, who have given me an immeasurable amount of support throughout my studies and have motivated me to work hard. Robbie, Phil, Dave, Luke, Julie, PJ, James H., Matt, Georgia, James R., Henry, Ben, Al, Clay, Tom, Griff, Joe, George, Hannah, and many more – thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AC	Asbestos Cement
AIC	Akaike's Information Criterion
BMA	Bayesian Model Averaging
DEM	Digital Elevation Model
D-InSAR	Differential Synthetic Aperture Radar Interferometry
ENVISAT	Environmental Satellite
ERS	European Remote-sensing Satellite
GCP	Ground Control Point
GIS	Geographical Information System
GLM	Generalised Linear Model
HOST	Hydrology of Soils Type
HYDROCK	Hydrological Rock Type
I	Iron
ISBAS	Intermittent Small Baseline Subset
LASSO	Least Absolute Shrinkage and Subset Operator
LOS	Line Of Sight
MAE	Mean Absolute Error
MORECS	Met Office Rainfall and Evapotranspiration Calculation System
NJUG	National Joint Utilities Group
NSRI	National Soils Resources Institute
O	Other
Ofwat	Office of Water Services
OS	Ordinance Survey
PE	Polyethylene
PS	Persistent Scatterer
PSI	Persistent Scatterers Interferometry
PVC	Polyvinylchloride
R^2	Correlation Coefficient of Determination
RMSE	Root Mean Squared Error
SAR	Synthetic Aperture Radar
SBAS	Small Baseline Subset
SDI	Steel and Ductile Iron
SLC	Single Look Complex
SMD	Soil Moisture Deficit
SRTM	Shuttle Research Topography Mission
UK	United Kingdom
VV	Vertical Transmit Vertical Receive
VWC	Volumetric Water Content
WISPA	Water Infrastructure Serviceability Performance Assessment

1 Introduction

1.1 Overview

This thesis investigates the complex interactions between soil, weather and trees and their impact on the failure rates of common pipe materials across the geographically largest water distribution network in the United Kingdom (UK). The evaluation of new datasets and data processing techniques has resulted in an improved understanding of the operational and environmental conditions leading to infrastructure failure. This has led to the development of a series of statistical models for the prediction of pipeline bursts, which builds on and improves a series of existing models. Opportunities and recommendations for future research are highlighted throughout, drawing upon the lessons learnt from the present research.

This chapter describes the context of the study, underlining the importance of the topic and the potential wider impacts in a global setting. An overview of previous efforts and techniques used to model the interactions between soil, weather and trees is detailed in Section 1.3. Section 1.4 discusses the motivation of the research and helps position the thesis in the context of the current knowledge gap. Section 1.5 details the overall research aim, discussing five objectives developed to satisfy the research aim. The individual approaches to the overall research methodology and the format of the thesis are provided in Section 1.6. Section 1.7 provides a list of publications and details the dissemination of the thesis to date, along with a proposed publication strategy.

1.2 Research Context

There is a critical reliance across public, commercial and industry sectors on the consistent and uninterrupted provision of water. Water supply is essential for human health, commercial activity and socio-economic development. In many locations throughout the world, drinking water distribution networks are dependent on an increasingly ageing and deteriorating network of buried infrastructure, which is impacted on by numerous operational and environmental pressures (Pelletier *et al.*, 2003). Currently, in the UK alone, estimates suggest

that approximately 3 billion litres of potable and treated water is lost from leaking pipes each day, which is approximately 18% of all water distributed (Ofwat, 2018). This is despite investments of over £110 billion to date (Institution of Civil Engineers, 2014). It is also noted, that by 2050 water demand in the UK could increase up to 35% per household (Vale and Poole, 2011). With the current status of pipeline leakage, plus a potential increase in water demand, effective asset management is vital to ensure that the demand of water is met for future generations.

Globally, water pipeline failure has the potential to cause disruption to society in numerous ways. A direct impact of a burst water pipe can include water discoloration, increased turbidity and the reduction of pressure in the distribution network (Cook *et al.*, 2015). For the privatised water sector in the UK, this can lead to large fines set by the regulatory body, Ofwat. Secondary impacts of pipeline failure are numerous, and can include damage to proximal infrastructure assets (such as gas, sewer, telecommunication and transport infrastructure), the creation of sand wash-out and sinkhole events leading to cascading failures (leading to building damage, loss of electricity and gas, vehicle damage), and also the incurred costs and disruption of water pipeline repair (Farewell *et al.*, 2017; Wols and van Thienen, 2014; Yamijala *et al.*, 2009).

There is a desire in the water industry to become proactive in the way they manage their assets. In order to achieve this, an improvement in the ability to predict pipeline failure is key for this transition, along with an improved understanding into the conditions which lead to asset deterioration (Farmani *et al.*, 2017; Rostum, 2000). However, identifying the conditions which lead to infrastructure failure is inherently difficult as the failure of water pipes is controlled by a large number of complex factors with different spatio-temporal variabilities (Boxall *et al.*, 2007).

Several approaches have been developed to understand the current risks posed upon water distribution networks, the most common form of which is

mathematical based models (Francis et al., 2014; Kleiner and Rajani, 2001; Park, 2004; Rajani and Kleiner, 2001; Xu et al., 2011; Yamijala et al., 2009). Mathematical models, which have been applied to the analysis of water pipeline failure, can be broadly classified into physical models and statistical models (Wilson *et al.*, 2017). Physical models simulate the behaviour of pipe materials by analysing the mechanical performance of the pipeline itself, the internal loads due to operational practices, and pipe material deterioration due to external factors (Rajani and Kleiner, 2001). Statistical models can be used with a variety of input data and have the ability to summarise the impacts of a wide range of external variables leading to the failure of potable water pipes. For this reason, statistical models are the most widely used method for the prediction of water pipe failure globally, with many studies applying such models in a range different geographical locations (Berardi *et al.*, 2008; Boxall *et al.*, 2007; Folkman, 2018; Kabir *et al.*, 2016; Martínez-Espiñeira *et al.*, 2017; Wang *et al.*, 2013; Yamijala *et al.*, 2009).

Globally, each water distribution network has its own unique set of prevailing risks. Therefore, there is no statistical model or approach which serves as a global solution for the prediction of water pipeline failure. Typically models will include operational variables such as age, material and diameter, as these factors are highly predictive for asset failure (Kimutai *et al.*, 2015). With the development of accurate environmental datasets, such as detailed soils and weather data, more studies are now investigating the impact of environmental conditions preceding incidences of failure. This increased representation of the environment has led to the development of more complex models, which not only evaluate the operational conditions and management of the pipes, but also provide an insight into the environmental drivers of pipeline failure (Kleiner and Rajani, 2001; Makar and Kleiner, 2000; Saadeldin *et al.*, 2015; Yamijala *et al.*, 2009).

The impacts to water distribution networks are not exclusive to just weather, soil and operational factors. Many other factors also have the potential to impact upon buried infrastructure, some of which are currently unquantifiable and in some

cases even currently unknown, such as the malicious tampering of pipes (Boxall *et al.*, 2007). If possible, an improved representation of new factors in statistical modelling may improve the accuracy of current pipeline failure predictions and help the further understanding of the risks posed upon the water distribution network. It is important to note that statistical modelling approaches cannot summarise all of the potential impacts on a network but can include (based on user expertise) variables which are most likely to be predictive of water infrastructure failure in a particular region.

1.3 Research Focus

In the UK, water distribution networks typical span across regional areas and are subject to a wide range of contrasting environmental conditions, which can impact on the failure rates of different pipeline materials. This research focusses on Anglian Water's distribution network[1], which is situated in the region of East Anglia of the UK. East Anglia has amongst the highest ground movement potential in the country due to extensive deposits of clay soils and a large intra-annual difference in the available soil water content from winter to summer seasons (Farewell *et al.*, 2017; Pritchard *et al.*, 2014). For regulation purposes, and to help Anglian Water obtain a weather-adjusted baseline for the number of bursts they should expect given the weather they experienced in a given year, a series of statistical models have been developed for Anglian Water plc. under the 'Water Infrastructure Serviceability Performance Assessment' (WISPA) project. WISPA aimed to generate a variable baseline of burst water mains, based on the observed weather and soil conditions. This provided a favourable alternative to using static "tramline" targets, which failed to represent variability in prevailing weather or environmental conditions. The motivation behind the development of statistical models for asset failure prediction was the series of cold winters (which had not been seen for the previous 15 years) which meant that Anglian Water

[1] Geographically, Anglian Water is the largest water and sewerage company in the UK, covering 20% of England and Wales' total land area. Potable drinking water is supplied to a total of 2 million households and 124,063 businesses, with over 1 billion litres of water being supplied each day. The Anglian Water network is serving one of the fastest growing parts of the UK, with a predicted 34% increase in the number of households by 2031 (Anglian Water, 2018).

struggled to meet their regulatory targets for the number of bursts mains. Anglian Water was the first utility operator in the UK to use model predictions to set targets for burst water mains in their annual reporting.

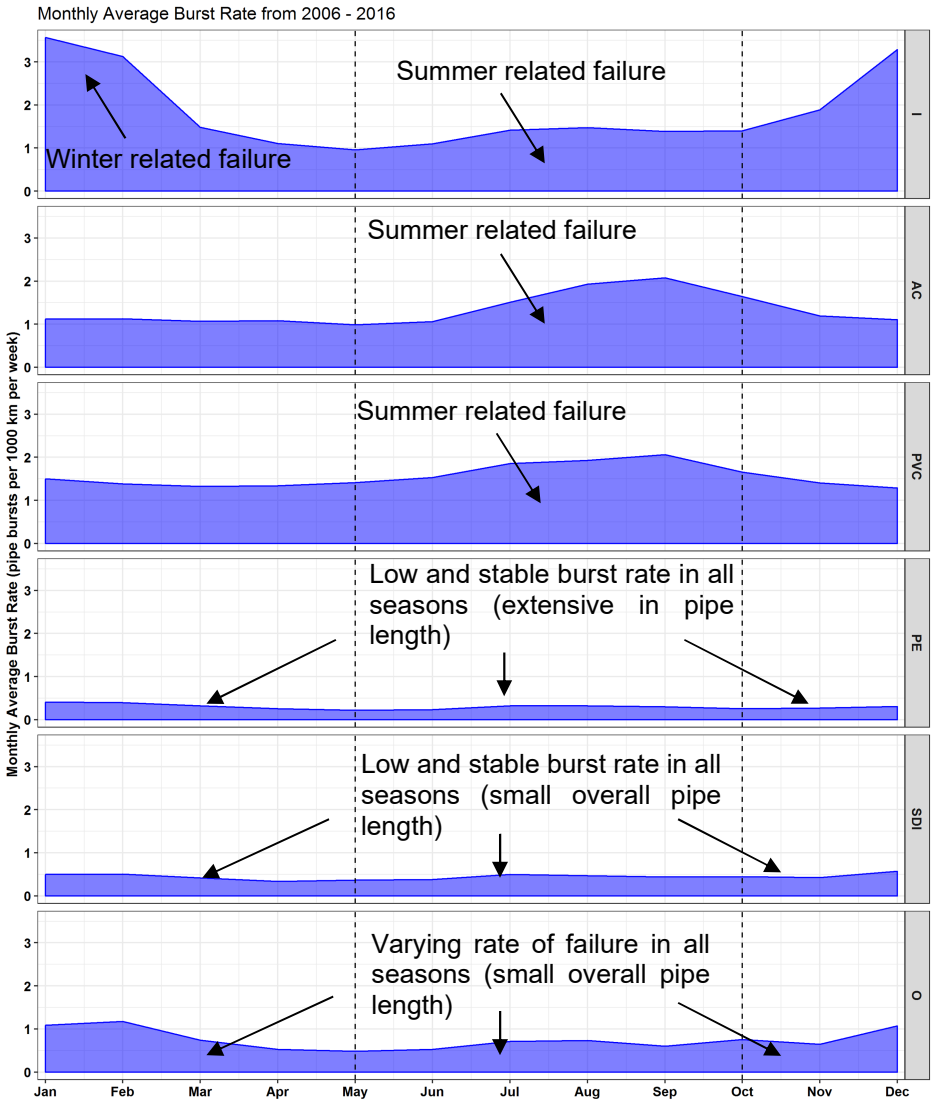

Figure 1-1: Monthly mean average burst rates (bursts per 1,000 km per week) for the 6 main pipe material groups over the Anglian Water region

Note: pipe materials are Iron (I), asbestos cement (AC), polyvinylchloride (PVC), polyethylene (PE), steel and ductile Iron (SDI) and pipes classified as "other" (O)

There are numerous factors which can lead to the failure of water pipes. Broadly, factors can be categorised into two main groups, environmental (weather, soils, trees) and operational (asset health, age, material, diameter, hydraulic pressure, previous pipeline repairs, construction methods, and previous maintenance

strategies or investment). All factors leading to pipeline failure can be further classified to be either static or dynamic, with many variables exhibiting a time dependant or seasonal relationship to burst water mains (Farmani et al., 2017). Therefore, identifying the factors which directly contribute to pipe failure is inherently difficult and has been the focus of numerous scientific investigations (Folkman, 2018; Gould et al., 2011; Kimutai et al., 2015; Martínez-Espiñeira et al., 2017; Pelletier et al., 2003). With such a wide range of contributory factors, this thesis does not aim to re-describe all known variables, except from the ones included in the developed models, but a comprehensive summary of a range of variables can be found in (Gould et al., 2011; Kleiner and Rajani, 2001; Wilson et al., 2017; Yamijala et al., 2009).

Environmental factors such as weather and soil can have either a direct or indirect impact on the failure of different pipe materials. An example of a direct weather-related impact on pipeline failure would be the successive embrittlement of iron pipes caused by prolonged periods of cold weather or air frost. However, indirect impacts of prolonged cold weather are also evident, where the mechanism of increased external pressure being exerted on the pipeline by frost-loading of the surrounding soil has been reported (Gould et al., 2011). Conversely, prolonged periods of dry weather (leading to increased evapotranspiration and high soil moisture deficit) can cause shrinkable soils, such as clay, to shrink. This can lead to increased external pressure in the surrounding soil leading to pipeline failure. This mechanism of failure has shown to increase observed circumferential failures in some pipe materials, particularly in Asbestos Cement (AC) pipes (Folkman, 2018). Weather variables influencing pipeline failure tend to be highly dynamic and can vary on day to day basis. Other environmental factors, such as soil and tree impacts tend to be less dynamic, with factors, such as soil pH, depth to underlying bedrock and the presence of trees being regarded as static where variability only occurs over a long-time period.

The operational management of water pipelines have a known and established impact on pipeline failure. Due to the often-extensive lengths of pipes in a water companies' asset-base, there are numerous operational factors which can lead to the failure of water pipes. The most common factor leading to pipe failure is

the age of the asset, where older pipes are more likely to fail. For many UK water companies, there are still small lengths of the distribution network which can be dated back to the early 1900's. Several studies statistically describe the relationship between asset age and pipeline failure probability (Kleiner and Rajani, 2001; Yamijala et al., 2009). The asset-base of individual water companies is often highly varied and encompasses a wide-range of different pipe materials of different ages, materials and diameters. In addition to this, water demand, pressure management, and internal water pipe temperatures have also been found to have a statistically significant impact on the failure of different pipe materials. Therefore, a statistical modelling approach must be capable of including a range of different variables, from both environmental and operational datasets. Independently, the factors which can lead to pipe bursts are well-known and extensively researched. However, due to the wide-range of contributory factors, the risks of pipeline failure are very heterogenous and can vary at small spatial scales. This means that water utility companies need to understand the specific risks which are prevalent in their own distribution network, as one set of specific risks may not necessarily be applicable to another geographical location.

Due to the seasonality of pipeline failures in some pipe materials (Figure 1-1), statistical models are highly suitable for the prediction of pipeline failure. For this reason, the models developed in WISPA focussed on operational variables such as pipe age and diameter and various soil and weather conditions. For the ease of investigation, approximately 30 pipeline materials have been grouped, based on their physical properties, into 6 simplified material categories. For some pipe materials, failure rates are highly influenced by the antecedent and prevailing environmental conditions typical of the different seasons (Clayton *et al.*, 2010; Gould *et al.*, 2011; Kleiner and Rajani, 2001). Figure 1-1 highlights this seasonality, where notable increases in pipeline failure exist for Iron, asbestos cement (AC) and polyvinylchloride (PVC) pipes, for the summer and winter seasons, respectively. For pipeline materials, such as polyethylene (PE) and steel and ductile Iron (SDI), the failure rates are less influenced by season, and have a near continual rate of failure. Across the Anglian Water distribution network, pipes classified as "other" (O) have a varying failure rate across the

different seasons, with no distinct trend evident for any particular season. The seasonal trends of failure rates in different pipe materials relate directly to the model accuracies gained within the WISPA project, see Figure 1-2.

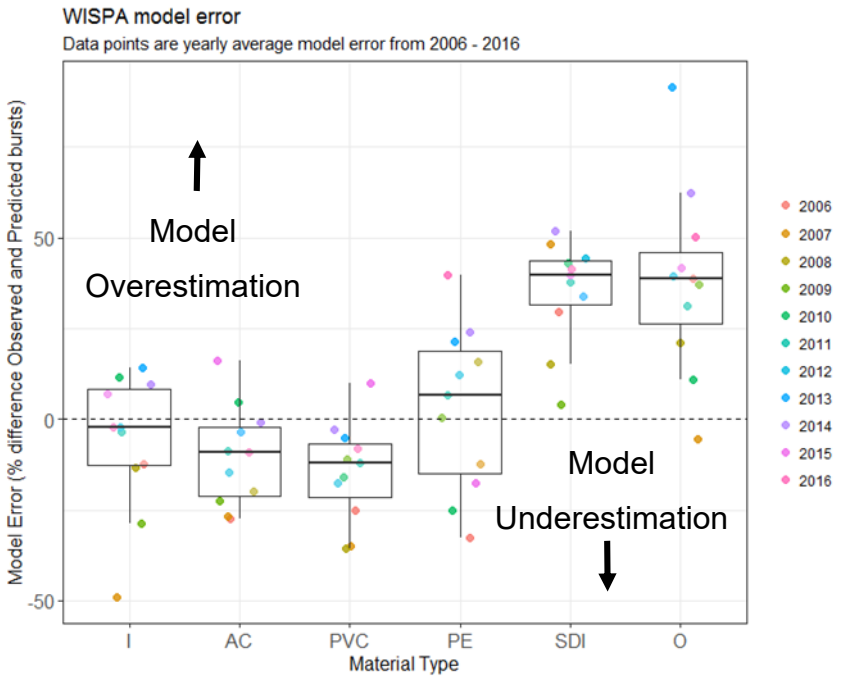

Figure 1-2: Annual WISPA model performance for individual material types. Data points represent the annual model error. Model error is calculated by the % difference in observed and predicted bursts (annually)

Note: pipe materials are Iron (I), asbestos cement (AC), polyvinylchloride (PVC), polyethylene (PE), steel and ductile Iron (SDI) and pipes classified as “other” (O)

Pipe materials which exhibit a highly seasonal rate of failure, such as Iron, AC and PVC, attained the lowest average model error in the previously developed WISPA models. Together, these three pipeline materials represent over 64% of the total network length and ~89% of all observed bursts (from 2006 – 2016). The largest range in model error from WISPA was recorded in PE pipes, with model predictions showing numerous instances of over and under-prediction. Polyethylene pipes make up over 27% of the network but account for only 7% of observed bursts. Nearly all model prediction for SDI and O pipes was over-predicted during the 11 years, to a high degree on inaccuracy. However, these

material groups only represent around 8% of the distribution network, and under 4% of observed bursts (Figure 1-2).

This research aims to improve the modelling performance of the previously developed WISPA models, by 1) rigorous data cleaning, 2) material-specific variable selection, 3) material specific model building, 4) using additional secondary data sources such as trees and satellite derived variables. An exploration of the specific environmental conditions which lead to pipeline failure is anticipated to permit an enhanced ability to predict pipe failure. The work presented in this thesis provides a critique of the previously developed WISPA models, as well as an evaluation of new methods and datasets to construct a series of enhanced statistical models for Anglian Water. These models were designed to accurately predict the total number of bursts in a given period (week, quarter or annually).

Accurate burst information is important as it can help inform important business decisions, such as determining the total capital invested into tackling pipeline failures, the accurate setting of burst targets which are agreed with the regulator, and in the long-term, helping to reduce the customer bill's by making water distribution more efficient.

On this premise, many companies are looking towards data-driven approaches to aid business operations. Statistical models can help identify pipes which are most at risk of failure based on the pipes physical condition (material, age, diameter and previous incidences of failure) and the prevailing environmental conditions (soil, weather and tree conditions). In doing so, modelled predictions can be used to identify the most at-risk parts of the distribution network.

The models developed within this thesis include both static and dynamic operational and environmental datasets (Table 1-1). Akaike's Information Criterion (AIC) was used to identify predictive variables of pipeline failure for each pipeline material. The variables which was identified as being predictive for each material type were consequently used to construct material-specific stepwise Poisson regression models. Only variables which are predictive of pipeline failure for each material type are included in the model testing, allowing for the

development of parsimonious models which include only relevant variables for each pipe material type. Stepwise Poisson regression models were selected due to their ability to include several different classes of data, which is important given the wide-range of different factors contributing to pipeline failure. The general mathematical form for the stepwise Poission regression models is detailed in Section 2.3.1.

An advantage of using statistical models when aiding burst detection is the ability to remove bias and subjectivity when deploying task forces. Statistical models, when used correctly, can succinctly synthesise a large amount of empirical information which is not always achievable by humans, particularly in times of high business pressure. Furthermore, statistical models can account for a wide-range of different scenarios and help inform critical business decisions, both retrospectively and prospectively. However, it is important to note that such models are not intended to replace human efforts in predicting burst pipes, but act as an operational tool to help inform and aid operational management decisions.

The methods and modelling approaches presented within the thesis are transferable to other geographical locations, as the techniques can be applied where relevant data exists. The evaluation of several new datasets and the inclusion of new environmental parameters has the potential to improve pipeline failure prediction and is an essential step in the evolution of different water pipeline failure models.

1.4 Motivation of research

The increased accessibility of additional environmental datasets and improvement in computing power has facilitated statistical modelling at successively finer spatial and temporal resolutions. This research investigates the potential of several data-driven techniques to create a series of material-specific water pipeline failure models for an entire drinking water distribution network. An overview of the different datasets used throughout this research is presented in Table 1-1. Through an increased understanding of the environmental and operational factors leading to water infrastructure failure, utility

operators are able to monitor the performance of the network and identify geographical areas which would benefit from increased investment or different proactive management practices.

From an industrial perspective, the previously developed WISPA models have been useful for regulation purposes and for asset planning, however, Anglian Water now wants to use statistical modelling for proactive asset management and the short-term forecasting of asset failure. On this premise, the reliability of the pipeline failure prediction needs to be improved.

No previous studies have included the representation of trees into statistical models predicting bursts in different pipeline materials. This thesis evaluates the use of a recently released national tree inventory, which presents a novel opportunity to analyse the impacts of trees to an entire water distribution network. In addition to the representation of trees in statistical models, this research also evaluates a beta version of a new high resolution (5 km tiled) weather dataset from the Met Office (MORECS) and evaluates the use recently launched satellite data (Sentinel 1) for the low-cost and regional monitoring of soil related ground deformation. A description of the datasets used in this thesis are outlined in Table 1-1.

The wider motivation of this work can be seen as seeking to conserve water resource by further understanding the environmental risks posed to buried water distribution networks helping to mitigate against further pipeline failures. Moreover, the modelling techniques described within this thesis are suitable for the prediction of other buried assets, such as sewerage and gas pipelines.

Table 1-1: Description of the datasets used throughout this thesis.

Type	Dataset	Reference	Data Type	Resolution	Details	Used in Objectives
Soils	National Soils Map	Hallett et al., 2017	GIS Vector	1:250,000	GIS map containing information relating to a wide range of soil factors, including soil texture, corrosivity to iron, depth to bedrock, shrink swell likelihood, hydrology of soils type and hydrological rock type. *Data licensing type: Shared data from the National Soils Research Institute*	2, 3, 4, 5
	National Perils Directory	Pritchard et al., 2014	GIS Vector	1:250,000	GIS map containing information relating to the flood susceptibility, climate adjusted clay hazard, soft and compressible soil hazard. *Data licensing type: Shared data from Cranfield University*	2, 3, 4, 5
Weather	MORECS	Met Office, 2018a	Tabular	40 km and 5 km tiled grid	The Met Office Rainfall and Evaporation Calculation System (MORECS) dataset, containing weekly estimates of Soil Moisture Deficit (SMD), temperature, sunshine hours and vapour pressure. *Data licensing type: Shared data from the Met Office*	2, 3, 4
	Met Office Regional Climate Summaries	Met Office, 2018b	Tabular	Monthly	Regional climate summaries including temperature, rainfall, days air frost (mean, minimum and maximum). The climate summaries provide climatic information from 1971-2018. *Data licensing type: Open data from the Met Office*	2, 3, 4
	Daily Volumetric Soil Water Content	CEH, 2017	Tabular	*In situ* point	30-minute measurements of volumetric water content taken at a depth of 10 cm at two sites. Measurements were taken with a time-domain transmission probe. *Data licensing type: Open data from the Centre of Ecology and Hydrology*	2,3,4

Tree	National Tree Map	Bluesky, 2018	GIS Vector	Centimetre	The GIS layer includes the height and canopy extent of all trees over 3.5 m in a GIS format. *Data licensing type: Shared data from Bluesky*	5
Operational	Pipe Network	Anglian Water	GIS Vector	Centimetre	GIS vector layer including information to the location, length, material, age, and diameter of pipeline. *Data licensing type: Shared data from Anglian Water*	3, 4
	Burst Dataset	Anglian Water	GIS Vector	Meter	GIS vector file which includes information upon the timing and location of a historical archive of pipeline bursts. Further information includes the type of burst (i.e. proactive or reactive). *Data licensing type: Shared data from Anglian Water*	1, 2, 3, 4
Satellite	Sentinel 1	ESA, 2015	Synthetic Aperture Radar (SAR)	3 x 20 m in Range and Azimuth	Sentinel 1a images collected in Single Look Complex (SLC) and Interferometric Wide (IW) mode. The images were collected in VV polarisation. The relevant precise orbit information was also acquired to ensure correct co-registration of image pairs. *Data licensing type: Open data from the European Space Agency (ESA)*	1, 2, 3, 4
	SRTM	NASA, 2017	Digital Elevation Model (DEM)	90 m	Provides a rasterised output of elevation at a 90 m resolution. *Data licensing type: Open data from the National Aeronautics and Space Administration (NASA)*	5

1.5 Research aim and objectives

The requirement of more accurate and fully representative models for the prediction of water pipeline failure is vital to assist the operational management of water distribution networks. There is scope to improve current methods of pipeline failure prediction by enhancing the representation of soil, weather and trees variables within operational statistical models.

The aim of this research is:

To establish the impact of soils, weather and trees on water infrastructure failure and to develop a series of material-specific drinking water pipeline failure models to predict bursts for an entire distribution network.

In order to test the aim, a series of research objectives have been developed to:

1. Determine the impact of extensive data cleaning and pre-processing on the improvement of water-pipeline failure predictive modelling.
2. Use variable selection techniques to identify predictive variables, and with these, construct a series of material-specific water infrastructure failure models.
3. Measure the impact of trees on the failure rates of buried water pipes and establish whether a national tree inventory can enhance predictive models.
4. Quantify the impact of using separate winter and summer models for the enhanced prediction of water infrastructure failure.
5. Measure and evaluate the seasonal response of soil-related road and rail infrastructure movement using the Persistent Scatterers Interferometry technique to test if this would be a useful input to statistical modelling of pipeline failure.

1.6 Research approaches and thesis format

This thesis adopts a range of different approaches in order to address the research aim and objectives. The thesis has been prepared in the style of formatted papers, with Chapters 2 to 5 representing individual research papers.

Objective 1 concerns the process of cleaning the historical pipe burst data provided from Anglian Water plc., ensuring that a fully consistent and quality-controlled dataset is available for the investigation of Objectives 2 to 4. Pre-processing of the historical dataset of pipeline failure is a fundamental step to ensure that errors are not introduced into the dataset which is used for statistical modelling.

Objective 2 uses a quantitative approach to select the predictive soil, weather and operational variables for specific pipeline material types. Material-specific predictive variables are then used for the development of individual pipeline failure models using Poisson regression, which is a form of Generalised Linear Model (GLM), as the statistical modelling technique. This technique was selected for consistency with previous Ofwat approved methods of pipeline failure prediction. Objective 2 tests the use of a beta version of a new meteorological dataset (MORECS), which offers an improved spatial resolution of key weather variables in comparison to previously available datasets. The developed models are then used as a foundation for Objectives 3 and 4. Objectives 1 and 2 are detailed within an individual research paper, which is included as Chapter 2 of the thesis.

Objective 3 investigates the use of an additional secondary dataset, which represents the presence of trees of different heights in proximity to the distribution network and evaluates the dataset's ability to enhance the models developed in Objective 2. Rate of failure analysis is used in Objective 3 to provide a quantitative investigation into the impact of trees on historical cases of pipeline failure. Using the same variable selection approach in Objective 2 (AIC), combinations of predictive tree height variables was identified and then added to the material

specific Poisson regression models. A direct comparison between the models developed in Objective 2, and the tree enhanced models developed in Objective 3 was then compared. Objective 3 is addressed as an individual research paper and is included as Chapter 3 in the thesis.

Objective 4 builds upon the most predictive models as developed through Objectives 1 to 3, and further enhances these models by increasing the representation of seasons into the model training and testing. This method is a novel approach to separately test and train statistical models using individual summer and winter datasets. Objective 4 is addressed as an individual research paper and is included as Chapter 4 in the thesis.

Having tested new weather and tree datasets for their use in the statistical modelling of pipeline failure, Objective 5 investigates if near-real time remote sensing data can also be used to generate key modelling variables. A technique named Persistent Scatterers Interferometry (PSI) (Ferretti *et al.*, 2001, 2000) is used to generate detailed regional measurements of soil-related surface deformation. Time series analysis was then undertaken to identify seasonal trends in soil-induced movement in different classes of above-ground infrastructure. Objective 5 is addressed within Chapter 5 of the PhD thesis, and is presented as an individual published research paper.

Chapter 6 provides a synthesis of the findings outlined in Chapters 2 to 5 and describes how the results contribute to the research aim as outlined in this chapter. An account of the findings for each objective is also provided in Chapter 6, discussing the novelty and intellectual contribution of the thesis. Alongside this, a series of recommendations are also made to provide guidance for future studies.

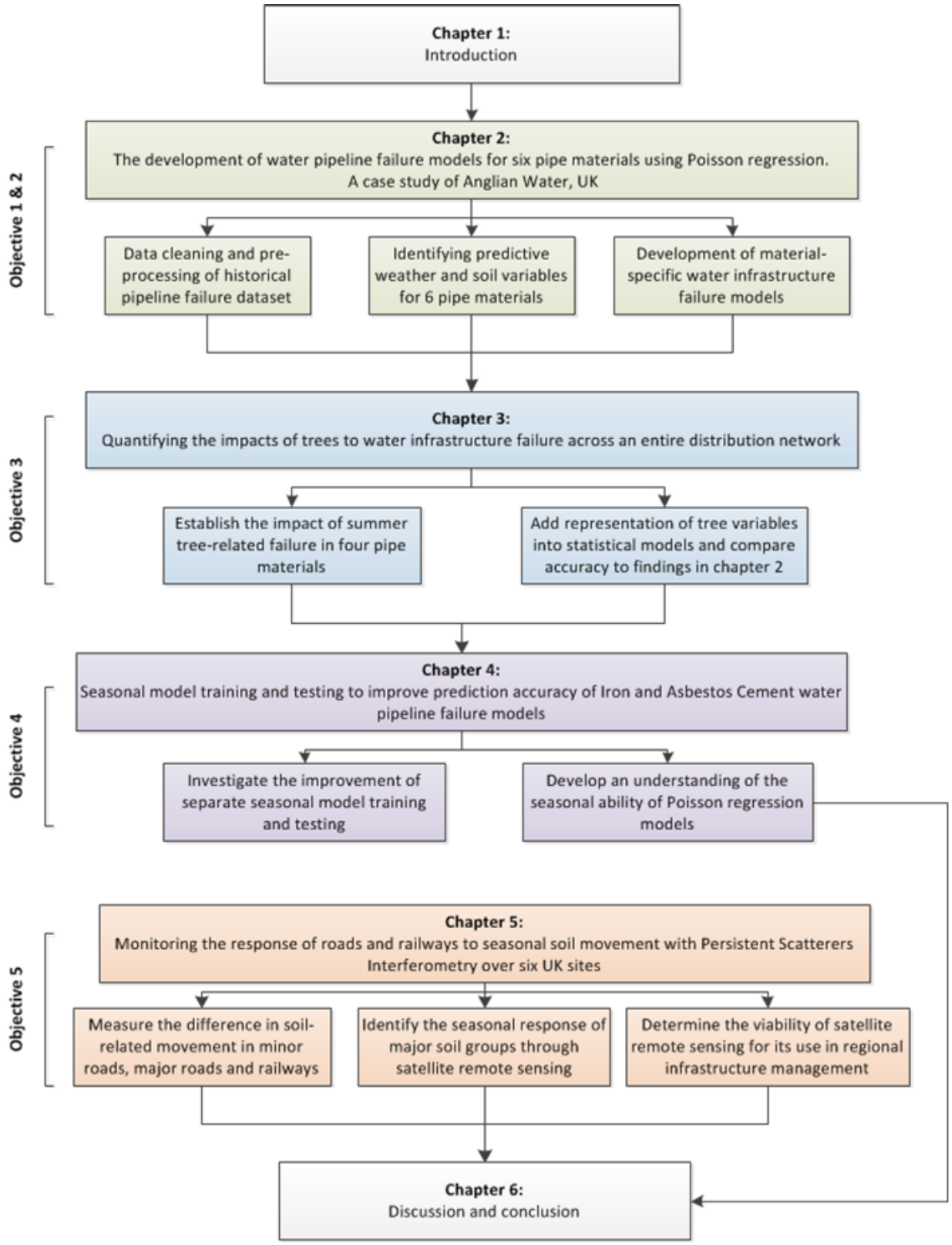

Figure 1-3: An overview of the research methodology followed. Each chapter has a series of individual objectives which have been designed to satisfy five individual objectives of the PhD thesis

1.7 Dissemination from the PhD thesis

At the time of writing this thesis, one paper has been published in an international peer-reviewed journal and 3 papers are currently being prepared for submission in a similar format to Chapters 2, 3 and 4. Matthew North has been the first author on all chapters formatted as academic papers, and has written the content, conducted the practical research and data analysis, and has drawn discussion and conclusions. Dr Timothy Farewell and Dr Stephen Hallett have also contributed by acting as academic supervisors and have also helped in the preparation in submitted and published manuscripts. Additional advice and guidance for the technical aspect of the research methodology has been provided by Dr Daniel Farewell and Miss Audrey Bertelle in Chapters 2 and 5, respectively. A specific note of the author contribution is provided at the end of each manuscript. All papers are intended for submission to international peer-reviewed journals, of which the details are provided below:

Currently un-submitted papers

North, M., Farewell, T., Farewell, D., Hallett, S. (un-dated) 'The development of water pipeline failure models for six materials using Poisson regression. A case study of Anglian Water, UK.' *Intention of publishing in "Environmental Modelling and Software"* **(Chapter 2)**

North, M., Farewell, T., Hallett, S. (un-dated) 'Quantifying the impacts of trees on water infrastructure failure across an entire distribution network: enhancing current statistical-based methods of drinking water pipeline failure prediction' *Intention of publishing in "Environmental Modelling and Software"* **(Chapter 3)**

North, M., Farewell, T., Hallett, S. (un-dated) 'Seasonal model training and testing to improve prediction accuracy of Iron and asbestos cement water pipeline failure models' *Intention of publishing in the "Proceedings of the Institute of Civil Engineers (ICE) – water management" collection* **(Chapter 4)**

Published papers

North, M., Farewell, T., Hallett, S. and Bertelle, A. (2017) 'Monitoring the Response of Roads and Railways to Seasonal Soil Movement with Persistent Scatterers Interferometry over Six UK Sites', *Remote Sensing*, 9(922). doi: 10.3390/rs9090922. **(Chapter 5)**

Note: Despite each chapter being presented as an individual research paper, references to individual chapters are made (i.e. Chapter 2) as opposed to formatted citations (i.e. North et al., 2017). This is due many of the manuscripts not yet published in their final form to date.

1.8 Bibliography

Anglian Water, 2018. Anglian Water Fast Facts [WWW Document]. URL https://www.anglianwater.co.uk/about-us/fast-facts-file.aspx (accessed 17.9.18).

Berardi, L., Giustolisi, O., Kapelan, Z., Savic, D.A., 2008. Development of pipe deterioration models for water distribution systems using EPR. J. Hydroinformatics 10, 113. https://doi.org/10.2166/hydro.2008.012

Bluesky, 2018. National Tree Map Dataset Description [WWW Document]. URL https://www.bluesky-world.com/ntm (accessed 2.8.18).

Boxall, J.B., O'hagan, A., Pooladsaz, S., Unwin, D.M., 2007. Estimation of burst rates in water distribution mains. Proc. Inst. Civ. Eng. - Water Manag. 160, 73–82. https://doi.org/doi.org/10.1680/wama.2007.160.2.73

CEH, 2017. COSMOS UK Soil Moisture Data. Nature Environment Research Council [WWW Document]. URL https://cosmos.ceh.ac.uk/data (accessed 11.04.19)

Clayton, C.R.I., Xu, M., Whiter, J.T., Ham, A., Rust, M., 2010. Stresses in cast-iron pipes due to seasonal shrink-swell of clay soils. Proc. Inst. Civ. Eng. - Water Manag. 163, 157–162. https://doi.org/10.1680/wama.2010.163.3.157

Cook, D., Husband, P., Boxall, J., 2015. Operational management of trunk main discolouration risk. Urban Water J. 13, 382–395. https://doi.org/https://doi.org/10.1080/1573062X.2014.993994

ESA, 2015. Sentinel 1 mission description [WWW Document] URL https://sentinel.esa.int/web/sentinel/missions/sentinel-1 (accessed 11.04.19)

Farewell, T.S., Jude, S., Pritchard, O., 2017. The influence of soil on the impacts of burst water mains on infrastructure and society : A mixed methods investigation. Nat. Hazards Earth Syst. Sci. in press, 1–30.

Farmani, R., Kakoudakis, K., Behzadian, K., Butler, D., 2017. Pipe Failure Prediction in Water Distribution Systems Considering Static and Dynamic Factors. Procedia Eng. 186, 117–126. https://doi.org/10.1016/j.proeng.2017.03.217

Ferretti, A., Prati, C., Rocca, F., 2001. Permanent Scatterers in SAR Interferometry. IEEE Transactions on Geoscience and Remote Sensing. 39, 8–20.

Ferretti, A., Prati, C., Rocca, F., 2000. Nonlinear Subsidence Rate Estimation Using Permanent Scatterers in Differential SAR Interferometry. IEEE Transactions on Geoscience and Remote Sensing. 38, 2202–2212.

Folkman, S., 2018. Water Main Break Rates in the USA and Canada : A Comprehensive Study. Mech. Aerosp. Eng. Fac. Publ. Paper 174.

Gould, S.J.F., Boulaire, F.A., Burn, S., Zhao, X.L., Kodikara, J.K., 2011. Seasonal factors influencing the failure of buried water reticulation pipes. Water Sci. Technol. 63, 2692–2699. https://doi.org/10.2166/wst.2011.507

Hallett, S.H., Sakrabani, R., Keay, C.A., Hannan, J.A., 2017. Developments in land information systems : examples demonstrating land resource management capabilities and options. Soil Use Manag. 33, 514–529. https://doi.org/10.1111/sum.12380

Institution of Civil Engineers, 2014. The State of the Nation Infrastructure 2014. Inst. Civ. Eng.

Kabir, G., Tesfamariam, S., Loeppky, J., Sadiq, R., 2016. Predicting water main failures : A Bayesian model updating approach. Knowledge-Based Syst. 110, 144–156. https://doi.org/10.1016/j.knosys.2016.07.024

Kimutai, E., Betrie, G., Brander, R., Sadiq, R., Tesfamariam, S., 2015. Comparison of statistical models for predicting pipe failures: Illustrative example with the city of Calgary water main failure. J. Pipeline Syst. Eng. Pract. 6. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000196

Kleiner, Y., Rajani, B., 2001. Comprehensive review of structural deterioration of water mains: statistical models. Urban Water 3, 131–150. https://doi.org/10.1016/S1462-0758(01)00033-4

Makar, J.M., Kleiner, Y., 2000. Maintaining water pipeline integrity. AWWA Infrastruct. Conf. Exhib., Balitmore, Maryland, March 12-15.

Martínez-Espiñeira, R., García-Rubio, M.A., González-Gómez, F., 2017. Which factors, and to what extent, influence the condition of urban water distribution networks. An empirical analysis of the Spanish case. Water Resour. Econ. 18, 20–33. https://doi.org/10.1016/j.wre.2017.02.002

Met Office, 2018a. UK regional climate summaries 1981 - 2018 [WWW Document]. URL https://www.metoffice.gov.uk/public/weather/climate (accessed 5.9.18).

Met Office, 2018b. Met Office Rainfall Evapotranspiration Calculation System - Specialist Datasets [WWW Document]. URL https://www.metoffice.gov.uk/services/industry/data/specialist-datasets (accessed 21.5.18).

NASA, 2017. Shuttle Radar Topography Mission. NASA Jet Propulsion Labatory [WWW Document]. URL https://www2.jpl.nasa.gov/srtm/ (accessed 11.04.2019)

OFWAT, 2018. Water loss per day UK - Discover Water [WWW Document]. URL https://discoverwater.co.uk/leaking-pipes (accessed 8.8.18).

Pelletier, G., Mailhot, A., Villeneuve, J.-P., 2003. Modeling Water Pipe Breaks—Three Case Studies. J. Water Resour. Plan. Manag. 129, 115–123. https://doi.org/10.1061/(ASCE)0733-9496(2003)129:2(115)

Pritchard, O.G., Hallett, S.H., Farewell, T.S., 2014. Soil Impacts on UK Infrastructure: current and future climate. Eng. Sustain. Proc. Inst. Civ. Eng. 167, 170–184. https://doi.org/doi:10.1680/ensu.13.00035

Rajani, B., Kleiner, Y., 2001. Comprehensive review of structural deterioration of water mains : physically based models. Urban Water 3. https://doi.org/10.1016/S1462-0758(01)00032-2

Rostum, J., 2000. Statistical Modelling of Pipe Failures in Water Networks. Nor. Univ. Sci. Technol. 1–132.

Saadeldin, R., Hu, Y., Henni, A., 2015. Numerical analysis of buried pipes under field geo-environmental conditions. Int. J. Geo-Engineering 6, 6. https://doi.org/10.1186/s40703-015-0005-4

Vale, H., Poole, N., 2011. The case for change - current and future water availabilty. Environ. Agency.

Wang, R., Dong, W., Wang, Y., Tang, K., Yao, X., 2013. Pipe Failure Prediction : A Data Mining Method 1208–1218.

Wilson, D., Filion, Y., Moore, I., 2017. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. Urban Water J. 14, 173–184. https://doi.org/10.1080/1573062X.2015.1080848

Wols, B. a., van Thienen, P., 2014. Modelling the effect of climate change induced soil settling on drinking water distribution pipes. Comput. Geotech. 55, 240–247. https://doi.org/10.1016/j.compgeo.2013.09.003

Yamijala, S., Guikema, S.D., Brumbelow, K., 2009. Statistical models for the analysis of water distribution system pipe break data. Reliab. Eng. Syst. Saf. 94, 282–293. https://doi.org/10.1016/j.ress.2008.03.011

2 The development of water pipeline failure models for six materials using Poisson regression: a case study of Anglian Water, UK

This chapter investigates Objectives 1 and 2, and is presented in the form of one unpublished research paper, intended for the journal Environmental Modelling and Software:

North, M., Farewell, T., Farewell, D., Hallett, S. (2018) The Development of water pipeline failure models for six materials using Poisson regression: a case study of Anglian Water, UK. (unpublished)

The components of this chapter can be considered in three parts: 1) the performance of data-cleaning and pre-processing to create a more reliable dataset for statistical modelling, 2) the selection of key environmental (soil and weather) and operational variables for the prediction of water infrastructure failure, and 3) building a series of material specific water infrastructure failure models for the Anglian Water distribution network. This chapter forms the basis of the models discussed in Chapters 3 and 4.

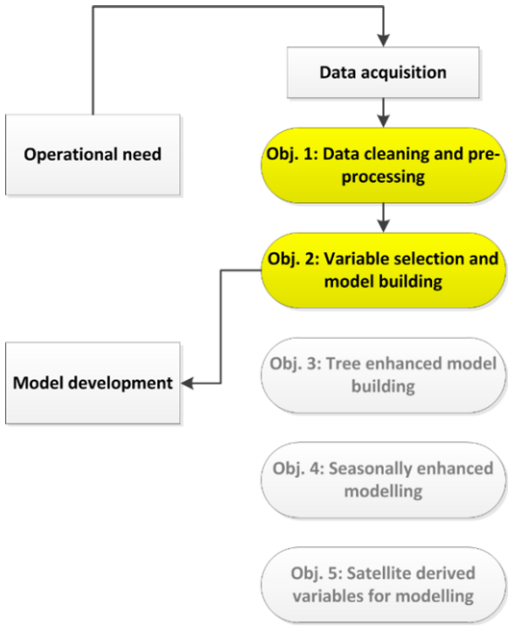

Figure 2-a: Objectives aimed to be investigated within this chapter in context of the overall thesis

The development of water pipeline failure models for six materials using Poisson regression: a case study of Anglian Water, UK.

M. North[a], T. Farewell[a], D. Farewell[b], S. Hallett[a]

[a] *School of Water, Energy and Environment, Cranfield University, Bedfordshire, MK43 0AL, United Kingdom*
[b] *Division of Population Medicine, School of Medicine, College of Biomedical and Life Sciences, Cardiff University, CF14 4YS, United Kingdom*

Corresponding Author: t.s.farewell@cranfield.ac.uk; tel.: +44 (0)1234 752978

Abstract

Statistical models that predict the location and timing of water pipeline failure allow utility companies to manage assets in a more cost-effective manner. Different pipe materials exhibit different failure mechanisms, which are controlled by variations in weather, soil and infrastructure-related factors. This study used Poisson GLMs and selected variables using Akaike's Information Criterion (AIC) to develop six material-specific water pipeline failure models (iron, asbestos cement, polyvinylchloride, polyethylene, steel and ductile Iron, pipes classified as "other"). An 8-year (2008-2016) dataset of infrastructure failure over an entire 38,000 km water distribution network was used for model training and prediction, using a 50% hold out sample. Model input datasets included soil, weather and infrastructure information. Results have highlighted 31 variables which are predictive of asset failure. The lowest error between observed and predicted bursts was attained for asbestos cement pipes (0.59%) and Iron pipes (2.64%). Findings are discussed in context of the environmental causes of pipeline failure and the applications of the proposed methods, considering the implications for the wider utility sector.

Key words: *Water Infrastructure, Pipe Burst, Weather, Soils, Generalized Linear Models, Variable Selection*

Highlights

- Identification of soil, weather and operational variables predictive of asset failure
- Water pipeline failure predictions using Poisson regression
- Variable selection and model building performed using Akaike's Information Criterion
- Development of material-specific pipeline failure models
- Prediction of pipeline failure over a whole water distribution network

2.1 Introduction

Globally, water distribution networks are ageing and are subject to a wide-range of environmental and infrastructure factors which induce pipeline failure. Understanding the factors leading to infrastructure failure is vital for utility companies seeking to maintain a resilient and efficient potable distribution network. Water companies use mathematical modelling to predict asset failure (Kleiner and Rajani, 2001; Pritchard *et al.*, 2014; Yamijala *et al.*, 2009). This is of particular importance given global climatic change, as the processes leading to asset deterioration may be accelerated or adopt different spatial or temporal patterns (Pritchard *et al.*, 2015a; Wols and van Thienen, 2014).

There is a close relationship between the environment and pipeline failures (Farmani *et al.*, 2017; Francis *et al.*, 2014; Wols and van Thienen, 2014; Yamijala *et al.*, 2009). In understanding the spatial and temporal variability of these relationships, statistical modelling can aid the prediction of the timing and location of burst water mains. Drawing upon predictions of pipeline failure, water utilities may adopt proactive management approaches to mitigate network risks, to aid reporting to regulatory bodies, to minimise disruption to water provision, and to reduce expensive reactive management practices and related financial penalties.

Weather factors, such as rainfall, high or low temperature, days of air frost, sunshine and high Soil Moisture Deficit (SMD) can exert direct or indirect stresses on the pipe network by impacting on the soil that surrounds the pipe, or the water inside it (Gould *et al.*, 2011; Wols and van Thienen, 2014). Differential soil

movements are common in many clay rich soils and increase failure rates of buried assets (Francis *et al.*, 2014; Wols *et al.*, 2014). In temperate climate countries, such as the UK, the seasonal shrinking and swelling of clay soils is controlled by soil moisture (Chapter 5). Soil movement correlates with higher rate of burst pipes in summer (periods of soil shrink) and autumn/winter (periods of soil swell) (Gould *et al.*, 2011; Wols and Van Thienen, 2014). Other factors, such as air temperature, have a more direct impact on the failure rates of some pipe materials, where the materials possess different tolerances to temperature thresholds (Wols and Thienen, 2016).

2.1.1 Approaches for modelling water pipeline failure

Modelling approaches suitable for the analysis of pipeline failure may be categorised into physical and statistical approaches (Kleiner and Rajani, 2001; Rajani and Kleiner, 2001). Physical models simulate the behaviour of pipe materials by analysing the mechanical performance of the pipeline itself, the internal loads due to operational practices, and material deterioration due to external factors (Rajani and Kleiner, 2001). These models are either probabilistic (including potentially unknown mechanisms of failure or random variation) or deterministic (including only known mechanisms of failure with no random variation). These methods are useful in identifying how pipes react under different environmental and operational conditions. However, they require a large amount of detailed input data which is often costly or difficult to obtain. Typical input data suitable for physical models was summarised by Rajani and Kleiner (2001) into three categories: operational factors (such as structure of material, material type, soil-pipe interactions and installation quality), loading on the pipe (internal pressure and external pressure) and material deterioration (such as chemical, biological and electro-chemical interactions). Conversely, statistical models provide a useful alternative due to their ability to be used with varying levels of input data (Farmani *et al.*, 2017; Rajani and Kleiner, 2001; Yamijala *et al.*, 2009). Geographical Information Systems (GIS) can attribute pipeline data with soil, weather and infrastructure data at a range of different spatial scales. Typically, this has been used to investigate the environmental conditions present at the time

of and preceding pipeline failure (Wols and van Thienen, 2014). Statistical models are often used to predict water pipe failure due to their relative low cost by comparison to physical models, and their ability to be used with a range of different input data. Several types of statistical models exist and have been applied to predict water pipe failure, such as Proportional Hazard model (Park, 2004), Artificial Neural Networks and Neuro-fuzzy Systems (Jafar *et al.*, 2010; Tabesh *et al.*, 2009), Bayesian belief networks (Francis *et al.*, 2014), and Generalised Linear Models (GLM's) (Farmani *et al.*, 2017; Yamijala *et al.*, 2009). Authoritative reviews of the various statistical methods for predicting water pipe failure are available (Kleiner and Rajani, 2001; Wilson *et al.*, 2017).

Selection of the correct model and input variables is critical for proactive asset management (Farmani *et al.*, 2017). The model should be accurate, reproducible, adaptable and practical in an industrial context. This study investigates the ability of Poisson regression generalised linear models to predict pipeline failures. This approach was selected for its simplicity, adaptability, practicality and reproducibility.

2.1.2 Variable and model selection methods for statistical models

To develop an accurate and interpretable statistical model, careful selection of input variables is vital (Li *et al.*, 2013). Several techniques have been developed which quantify the predictive ability of individual variables. Statistical tests such as Akaike's Information Criterion (AIC) and Bayesian Information Criterion have been used in many disciplines since their creation in the 1960's (Morozova *et al.*, 2015). Several studies have used AIC to evaluate the influence of covariates on the prediction of a response variable, or for final model selection, using water infrastructure failure datasets (Park, 2004; Weirich *et al.*, 2015). AIC is suitable for exploratory analysis and time series applications, such as forecasting, whereas Bayesian Information Criterion is better suited to confirmatory analysis and controlled experiments (Aho *et al.*, 2014). Both methods can be employed in a stepwise approach, where the effect of either adding variables (forwards selection), or removing variables (backwards elimination), or a hybrid mixture of the two approaches can be adopted (Zhang, 2016). Other variable selection

techniques include model averaging, such as Bayesian Model Averaging (BMA), and penalised regression methods, such as Least Absolute Shrinkage and Subset Operator (LASSO) and the Elastic Net. Model averaging methods such as BMA allow the incorporation of *a priori* knowledge to variable selection and allow for model uncertainty to be averaged across all models selected (Morozova *et al.,* 2015). Penalised regression methods such as LASSO or Elastic Net have been less widely used with water utility datasets but have been implemented for variable selection and model building in other scientific disciplines. Penalised regression methods decrease model complexity whilst keeping all explanatory variables, which is one advantage of these models. Whilst other approaches could have been used to select variables, this study has implemented AIC stepwise variable selection, as this is easily calculated from GLM models, which is the currently approved method of pipeline failure prediction by the UK water industry regulator, Ofwat. Furthermore, in comparison with other variable selection techniques, AIC stepwise methods are relatively easy to implement and communicate, an important consideration when addressing the pragmatic reproducibility and application of the work in an industrial context.

2.1.3 Overview of this study

This paper presents a statistical approach to select variables and build models to predict the location and timing of water mains bursts across an entire 38,000 km potable water distribution network, over an 8-year (2008 – 2016) period for the Anglian Water region of the UK (Figure 2-1).

Variable selection was undertaken using AIC to build, what is here termed, predictive stepwise GLM for six groups of pipe materials: Iron, Asbestos Cement (AC), Polyvinylchloride (PVC), Polyethylene (PE), Steel and Ductile Iron (SDI) and pipes classified as “Other” (O). O pipes represent either unclassified or rare pipe materials. Following a preliminary process of variable selection, a series of individual stepwise GLMs (Nelder and Wedderburn, 1972) was built for each pipe material based on the selection of the most predictive variables as indicated by the AIC.

This study improves understanding of the impact of environmental conditions on the rates of failure in common pipe materials. The development of material-specific stepwise pipe failure models highlights contributory environmental factors which directly impact specific pipe materials. Older pipe materials such as Iron and AC pipes are strongly influenced by environmental conditions such as temperature (Wols and Thienen, 2014), SMD (Gould *et al.*, 2011) and soil properties (Farmani *et al.*, 2017). Newer pipe materials, such as PE, are more resistant to environmental conditions (Davis *et al.*, 2007).

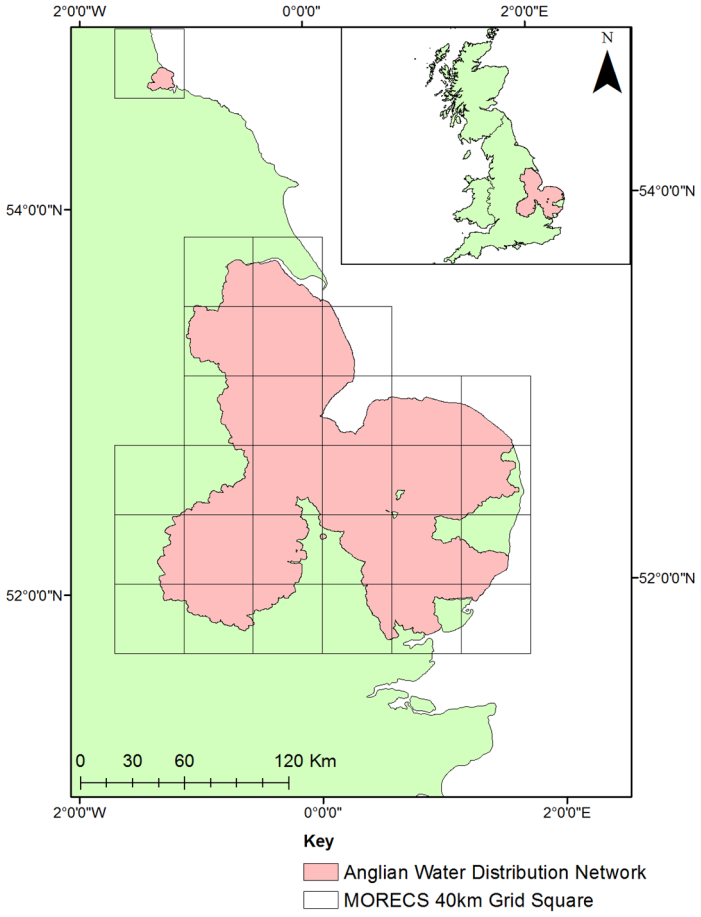

Figure 2-1: Study area extent of the Anglian Water service area. The drinking water distribution network is shown as red, and the relative position of the service area is shown as an insert map

Note: MORECS is the Met Office Rainfall and Evapotranspiration Calculation System, and represents the weather data used within this study

2.2 Materials

2.2.1 Study area

Anglian Water's drinking water distribution network covers an area of approximately 27,500 km^{-2} in East Anglia, and the town of Hartlepool (Figure 2-1). The prevailing climate of the region is temperate oceanic, with distinct seasonal variations in temperature and precipitation events throughout the year. Temperatures across East Anglia range from a typical maximum average temperature of 22.2°C in July to a minimum average temperature of 1.1°C in February. Rainfall is highest during months November, December and January, whilst the driest months are July, August and September (Met Office, 2018a).

Across East Anglia there are considerable deposits of silts, clays and peat soils which are derived from marine and riverine alluvium from previous glaciations and transgressions of the North Sea (Pritchard *et al.*, 2015b). Due to the intra-annual variation of sunshine hours, temperature and precipitation there is a strong seasonal pattern of SMD, with a high SMD during summer and autumn months, and a low SMD during winter and spring. This, combined with the abundance of shrinkable soils in East Anglia (Pritchard *et al.*, 2015a), create high ground movement potentials in this region.

Because pipe failures are influenced by infrastructure parameters (material, age and diameter), weather parameters (temperature, soil moisture deficit), and soil parameters (corrosivity, texture, ground movement potential), these are the three main families of data used in this study. Substantial data cleaning was required to ensure data consistency, veracity and validity to train and test the predictive models, which is detailed in Section 2.2.5. The following sections describe the input data and the data processing required.

2.2.2 Infrastructure data

Pipe and burst data was provided by Anglian Water plc., representing the East Anglian region of the UK, see Figure 2-1.

2.2.2.1 Pipe network data

The oldest pipes in the Anglian Water distribution network are Iron pipes, followed by AC, PVC, PE & SDI pipes. O pipes have a wide range of typical installation dates ranging from 1881 to 2001 (Table 2-1) and are generally constructed from rare materials (such as lead or glass reinforced plastic) or remain unclassified. A description of the operational variables included in the models is provided in Table 2-1, a description of the specific diameter and age bands is provided in Table 2-2, and a summary of all material types and the distribution network is provided in Table 2-3.

Table 2-1: Description of the infrastructure variables included

Variable Name	Description	Units	Variables
Material	Simplified material (Table 2-3)	Category	1
Diameter Band	Categorised bands (Table 2-2)	Category	1
Age Band	Categorised band (Table 2-2)	Category	1
Pipe age mean	Mean age of pipe	Date	1
Pipe age max	Maximum age of pipe	Date	1

Table 2-2: Age Band and Diameter Band ranges used in this study

Age Band	Range
0	Unknown
1	<1881
2	1881 to 1900
3	1901 to 1920
4	1921 to 1940
5	1941 to 1960
6	1961 to 1980
7	1981 to 2000
8	2001 to 2020
Diameter Band	**Range**
1	<165 mm
2	165 to 320 mm
3	321 to 625mm
4	>625mm

Table 2-3: Pipe installation date, length and number of bursts by pipe material **Note: pipe materials are Iron (I), Asbestos Cement (AC), Polyvinylchloride (PVC), Polyethylene (PE), Steel and Ductile Iron (SDI), and pipe materials classified as "Other" (O). Total length of pipe is based on service week (07.10.2008). The total bust number value extends temporally from 07.10.2008 to 27.09.2016**

Material	Installation range	Total length (km) (% of network)	Total bursts (% of bursts)	bursts per km of pipe
I	1881 to 1921	11,735 (30. 22%)	19,212 (48.10%)	1.63
AC	1920 to 1941	7,259 (18.69%)	9,027 (22.6%)	1.24
PVC	1960 to 2001	6,126 (15.77%)	7,395 (18.51%)	1.20
PE	1981 to present	10,538 (27.14%)	2,818 (7.05%)	0.26
SDI	1960 to 2001	1,902 (4.89%)	739 (1.85%)	0.38
O	1881 to 2001	1,267 (3.26%)	749 (1.87%)	0.59
Total		38,827	39,940	

2.2.2.2 Pipe burst data

A total of 39,940 bursts was analysed from the 7th October 2008 to the 27th September 2016 and was provided by Anglian Water. For this investigation, bursts which are classified as “reactive” was used as opposed to total number of both proactive and reactive bursts. A reactive burst is defined as a burst which was identified or reported by a third party, such as a member of public. A proactive burst is defined as a burst which was identified by a member of the company and was not reported officially by the public. Reactive bursts was selected for investigation as it is this category of pipeline failure in which Anglian Water was unable to predict the timing and location of. The largest burst rate per kilometre of pipe is seen in Iron, AC and PVC pipes. These pipe materials account for 89.21% of the overall reactive bursts during the observation period.

2.2.3 Soil data

This study used Cranfield University’s datasets: National Soil Map and Natural Perils Directory and associated soil property datasets (Hallett *et al.*, 2017; Pritchard *et al.*, 2014), to determine the influence of soil variables on the rate of water infrastructure failure. The 1:250,000 maps include information related to a wide range of physical factors relating to soil, and is a result of an extensive field-based campaign from 1939 to 1987 (Hallett *et al.*, 2017). Soil is critical due to its immediate proximity to the laid pipe materials. In total, 49 soil variables was selected for investigation. These variables were categorised into five groups relating to soil texture, corrosivity to Iron, depth to rock, shrink-swell likelihood and the hydrology of soils (see Table 2-4). Soils information was attributed to each pipeline and burst using a GIS (see Section 2.2.5). 16 variables was included to evaluate the influence of clay, silt, sand and peat soil texture on the failure of buried assets. Variables included the minimum, maximum and average percent of soil texture in each soil association (being a collection of soil types typically occurring together in a landscape), along with a variable to describe the texture of the most abundant soil type in each soil association. This allowed an evaluation as to how soil texture impacts upon the failure of buried assets in common soil properties in the UK. 6 variables was included which evaluate soil’s

corrosivity to ferrous Iron. Alongside this, 3 variables was included which evaluated the minimum, maximum and the average pH of the soil association in which the burst occurred. The soils pH can directly impact the corrosivity of the soil, and therefore the potential degradation of different pipe materials, particularly metallic pipe materials. The depth of underlying geology has the potential to control ground movement potential as it can determine the available soil water content (Chapter 5). Therefore, several variables describing the drainage and the hydrological properties of the soil, including hydrology of soils type and hydrological rock type, was included as explanatory variables as they can influence soil ground movement potential in the region. On this basis, the average, maximum and minimum depths of bedrock was also included as initial variables for this study. A full description of these variables is in Boorman, Hollis and Lilly (1995).

Table 2-4: Description of the soil variables included in this study

Attribute	Description	Variables
Data from the National Soil Map of England and Wales		
Clay	% Clay soil content. (Minimum, maximum, weighted mean and the value of the most abundant soil series).	4
Silt	% Silt soil content. (Minimum, maximum, weighted mean and the value of the most abundant soil series).	4
Sand	% Sand soil content. (Minimum, maximum, weighted mean and the value of the most abundant soil series).	4
Peat	% Peat soil content. (Minimum, maximum, weighted mean and the value of the most abundant soil series).	4
Organic Carbon	% Organic carbon soil content (minimum, maximum, weighted mean).	3
pH	The pH of the soil (measured mean, min and max) of the soil association.	3
Depth to Rock	Depth to bedrock (mean, min and max) (cm).	3

Corrosivity to Iron	5 level categorical classification of the corrosivity of the soil to ferrous Iron at 1m depth. Dominant class and % composition of individual classes.	6
Shrink Swell	6 levels categorical classification of the shrinkability of the soils. Dominant class and % composition of individual classes in the soil association.	7
Hydrology of Soil and Geology	Provides information of the water regime of the soils. Hydrological soil type (HOST) and Hydrological rock type (HYDROCK). Dominant class in the soil association provided.	2
Flood Susceptibility	Indication if the soil shows evidence of having been flooded.	1
Climate-adjusted clay hazard	A 9-class shrink swell classification indicating the likelihood of the soil to expand and contract, modified by the potential SMD (1961-1990). Calculated for the highest shrink swell class at different minimum % contributions.	6
Soft and compressible soils hazard	Soils which contain soft soils or are at risk of containing soft soil.	2
	Total variables:	49

2.2.4 Weather data

2.2.4.1 Days air frost

The total number of day's air frost has been included as a variable with a known impact on the failure rate of buried pipelines, particularly Iron (Wols and Thienen, 2016). This data was acquired, at no cost, from the Met Office climate summaries dataset for eastern England, where the average number of day's air frost has been recorded for the period from 1971 to 2018 (Met Office, 2018a). These data provided the regional monthly average number of days in which air frost occurred across the Anglian Water distribution network.

2.2.4.2 MORECS weather data (40 km and 5 km)

Weather and climate data was obtained from the Met Office Rainfall and Evaporation Calculation System (MORECS) (Met Office, 2018b). MORECS provides weekly estimates of actual evaporation, potential evaporation, SMD,

and hydrologically effective runoff at two spatial resolutions, 40 km and 5 km tiled output. The gridded output from both datasets (MORECS 40 km and 5 km) was attributed to each pipeline and burst incident using a GIS. In total, 74 weather variables was selected for investigation (Table 2-5). These variables are categorised across four themes: rainfall, SMD, sunshine, and temperature. Variables which represented the minimum, maximum and mean rainfall, SMD and temperature in a 1, 2- and 4-week period preceding the burst event was included in the model's development. The inclusion of multiple variables, over different time periods, allowed for the representation of all potential environmental conditions in the case where there was a lag between the date of the actual burst occurrence and the reported burst date.

2.2.4.2.1 Derived temperature variables

A range of different temperature variables was included in this study due to the acknowledged influence temperatures exert upon the failure rate of different pipe materials, and the many ways temperature can be investigated (Farmani *et al.*, 2017; Wols and Thienen, 2016, 2014). The inclusion of multiple temperature variables, at 1°C increments (from 0 to 10°C), allowed the effective identification of thresholds most likely to lead to infrastructure failure in different pipe materials. In following this approach, it was posited that if multiple thresholds were identified then an alternative, more continuous, modelling of temperature could be considered, but this was not expected to be the case. A summary of the methods used to create the temperature accumulation and temperature change are described in Table 2-5.

2.2.4.2.2 Soil Moisture Deficit (SMD)

Numerous SMD variables was included in this study, due to the known impact of differential soil movements on water infrastructure failure (Wols and van Thienen, 2014). Several variables, describing different SMD conditions was included to identify potential threshold where pipeline failure, in different pipe materials, occur. Due to the abundance of soils with high shrink swell potential across East Anglia (Pritchard *et al.*, 2015b), it is likely that SMD is an important variable for predicting water pipe failure across the water distribution network. MORECS 5

km includes a modelled output of SMD in different land cover types (grassland, deciduous and coniferous). The available water content, based on soil texture, are included as either low (e.g. sandy soils), medium (e.g. loamy soils) or high (e.g. peaty soils) helped to include a better representation of SMD as input variables.

2.2.4.2.3 Rainfall

Rainfall influences the amount of available soil moisture, therefore controlling ground movement potential in clay soils. On this basis, a total of 6 rainfall parameters was included in the initial variable selection. These variables were derived from both the 40 km and 5 km MORECS datasets and represent the total accumulated rainfall at 1, 2 and 4-week intervals.

2.2.4.2.4 Sunshine hours

Sunshine hours was included due to the potential impact on differential soil movements across the region. Sunshine has a direct impact on the rates of evapotranspiration therefore potential ground movement potential (Clark, 2002).

2.2.4.2.5 Atmospheric Vapour Pressure

A variable describing the atmospheric vapour pressure was also included as an initial explanatory variable. Vapour pressure is the partial pressure that water vapour exerts at any one time, its units are measured in hectopascals (hPa). Vapor pressure has been reduced in each MORECS grid square to mean sea level using a lapse rate of -0.025 hPa / 100 m respectively (Hough and Jones, 1997).

Table 2-5: Description of the weather variables included in this study

Variable Name	Description	Units	Dataset	Variables
Days Air Frost	Total number of day's air frost in a month	days	Met Office Climate Statistics	1
Rainfall	Total accumulated rainfall over 1, 2 and 4-week period preceding the reported burst date	mm	MORECS 5 and 40km	6

SMD	Weekly mean SMD under different land uses (grassland, deciduous, coniferous, real land use) and in soils with different water holding abilities (High, Medium, Low)	mm	MORECS 5 and 40 km	10
SMD Change	The absolute change in mean SMD from the week the burst was reported to 1, 2 or 4 weeks before	mm	MORECS 5 and 40km	6
Sunshine Hours	Total weekly sunshine hours preceding the reported burst date	hours	MORECS 5 km	1
Vapour Pressure	Vapour pressure is a variable which indicated the partial pressure that atmospheric water vapour exerts at any one time. Variable describes the weekly vapour pressure in the week preceding the reported burst date	hPa	MORECS 5 km	1
Temperature Max	Maximum air temperature recorded in the week the burst was reported	°C	MORECS 40 km	1
Temperature Min	Minimum air temperature recorded in the week burst was reported	°C	MORECS 40 km	1
Temperature Mean	Mean air temperature recorded in the week the burst was reported	°C	MORECS 40 km	1
Temperature Change	The absolute change in mean air temperature from the week the burst was reported to 1, 2 or 4-weeks before	°C	MORECS 5 and 40 km	6
Temperature ≤ (0 to 10°C) over 1 week	The accumulated air temperature beneath a threshold (0 - 10°c, in 1°c increments) in a 1-week period preceding the reported burst date. Air temperature accumulations are calculated in 1°c increments below the threshold value. The greater the value of this variable, the colder the temperature was, for a prolonged period of time, in the previous week to the burst date	°C	MORECS 5 and 40 km	20
Temperature ≤ (0 to 10°C) over 4 weeks	The accumulated air temperature beneath a threshold (0 - 10°c, in 1°c increments) in a 4- week period preceding the reported burst date. Air temperature accumulations are calculated in 1°c increments below the threshold value. The greater the value of this variable, the colder the air temperature was, for a prolonged period of time, in the previous weeks to the burst date	°C	MORECS 5 and 40 km	20
			Total Variables:	74

2.2.5 Data preparation

To ensure that a fully consistent and error-free dataset was used for statistical modelling, data preparation was undertaken to ensure that the successful merging pipe, burst and environmental datasets. The result was a weekly summary of bursts per operational pipes joined with the environmental variables (as described in Sections 2.2.3 and 2.2.4). Pipes cohorts was created based on the grouping of pipe material, age and diameter. This ensured the creation of homogenised sections of pipes with similar characteristics. A description of the methods used to create pipe cohorts is given in Figure 2-2. The creation of pipe cohorts is essential to reduce the computational resource needed for statistical analysis and has been previously noted in several investigations (Kleiner and Rajani, 2001; Rostum, 2000; Xu *et al.*, 2011).

Individual datasets was created for the 6 individual material groups. Further sampling was then undertaken to attain separate model train and test datasets. For this, a 50% random sample was taken from the dataset, this created two datasets of equal length which was used for model training and model testing.

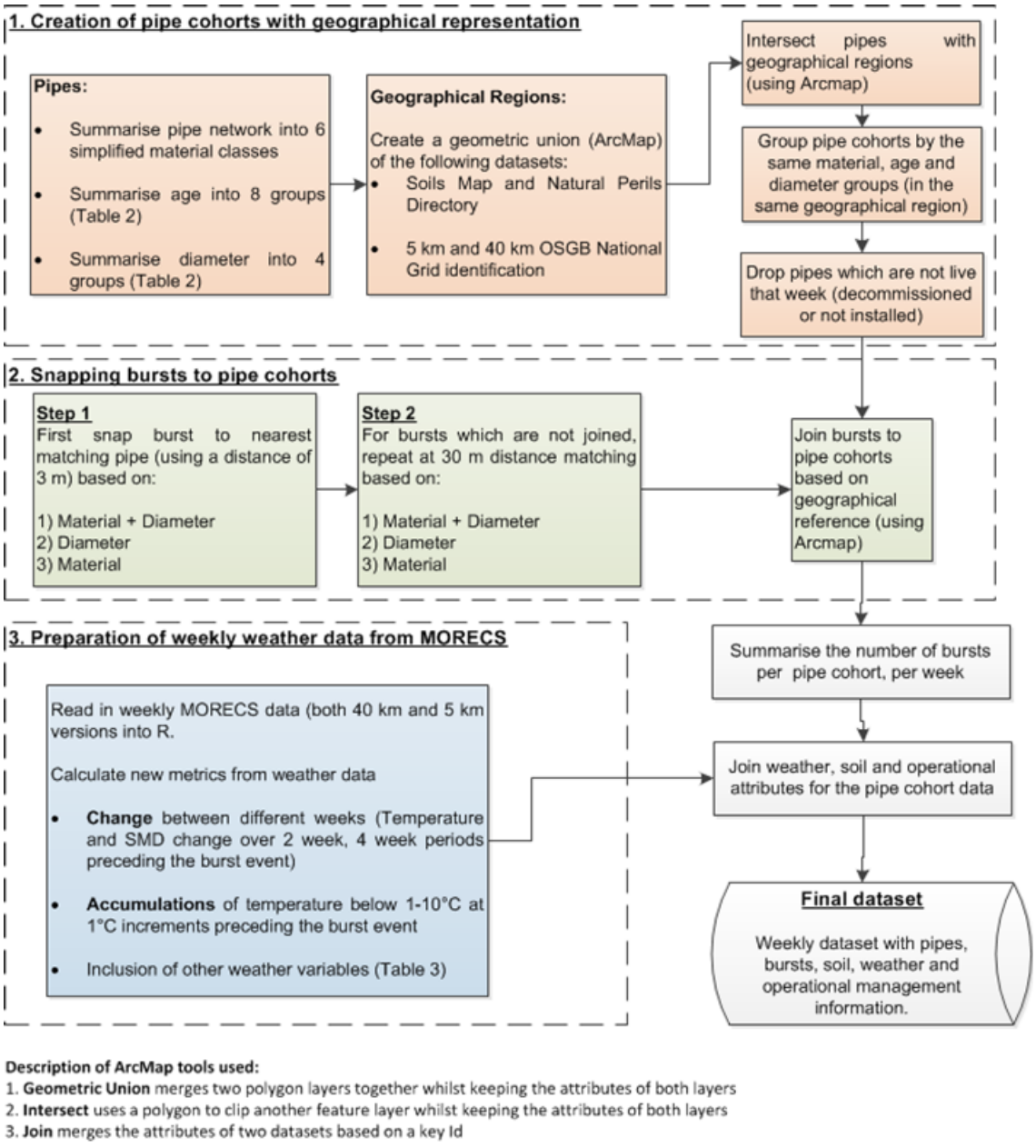

Figure 2-2: An overview of the pre-processing steps used to join the pipe network, burst and environmental data together, for the creation of a material-specific pipe cohort dataset

2.3 Methods

2.3.1 AIC variable selection method and stepwise GLM method

GLM's have been extensively used for environmental modelling due to their ability to evaluate an array of independent variables and non-normal count data, such as failures from infrastructure systems (Kleiner and Rajani, 2001; Yamijala *et al.*,

2009). GLM's generalise linear regression and link a set of explanatory variables to a response variable (Zhang, 2016). The Poisson GLM used in this study is given in equation 2-1:

$$P(Y = y|\bar{x}) = e^{-\mu}\frac{\mu^y}{y!}$$

where

$$\mu = E(Y|\bar{x})$$

and

$$\log(\mu) = \log(L) + \beta_0 + \beta_1 x1 + \cdots + \beta_k xk \qquad \text{eq. 2-1}$$

In this case, the total number of reactive bursts recorded in each pipe cohort (expressed above as y) is the response variable. The list of explanatory variables (represented as $\bar{x}$ variables) are described in detail in Tables 2-1, 2-4 and 2-5. An offset of log length, log(L), was used to provide a model of the rate of pipe breaks, so direct comparisons of failure rates can be made between different pipe materials. In this study, variable selection was performed using the Akaike Information Criterion (AIC) (Akaike, 1974). This technique was selected to perform model comparisons based on each models' ability to predict the response variable. The formula for AIC is given in equation 2-2:

$$AIC = (-2)\log(maximum\ likelihood) + 2(n\beta) \quad \text{eq. 2-2}$$

AIC was used as the basis to add explanatory variables which do, and remove those that do not, add to the model's ability to predict the response variable. This allowed the development of parsimonious stepwise models with the minimum number of required variables (Yamijala *et al.*, 2009).

AIC was used initially to select influential variables to be used to create a stepwise GLM for each pipe material. In the first instance, the full list of variables (plus a null model) was evaluated using a series of Poisson GLMs to predict the response of pipe bursts based on each variable individually. A subset of the full dataset was used in this initial step to minimise the computational resource needed to evaluate the large range of variables. For each pipe material, a sample

dataset of 5 million observations, covering the study period, was used to run the GLM. For each of the pipe materials, a null model was run first. After this, AIC values was used as a threshold to eliminate covariates which were worse performing than the null model. This created a list of variables which was taken forward to the next step. The best performing variable (as indicated by the lowest AIC value) was then removed from the explanatory variable list and added to a stepwise model. At each step, variables with a higher AIC value than the best performing variable in the previous step was excluded. This allowed the removal of variables which did not add value to the stepwise GLM. This process was continued until either there was no more variables left to evaluate, or a maximum number of 12 steps was reached. The maximum of 12 steps was selected to minimise the possible influence of overfitting or the selection of uninformative parameters from noise. This process of variable selection and GLM model building is presented below in Figure 2-3.

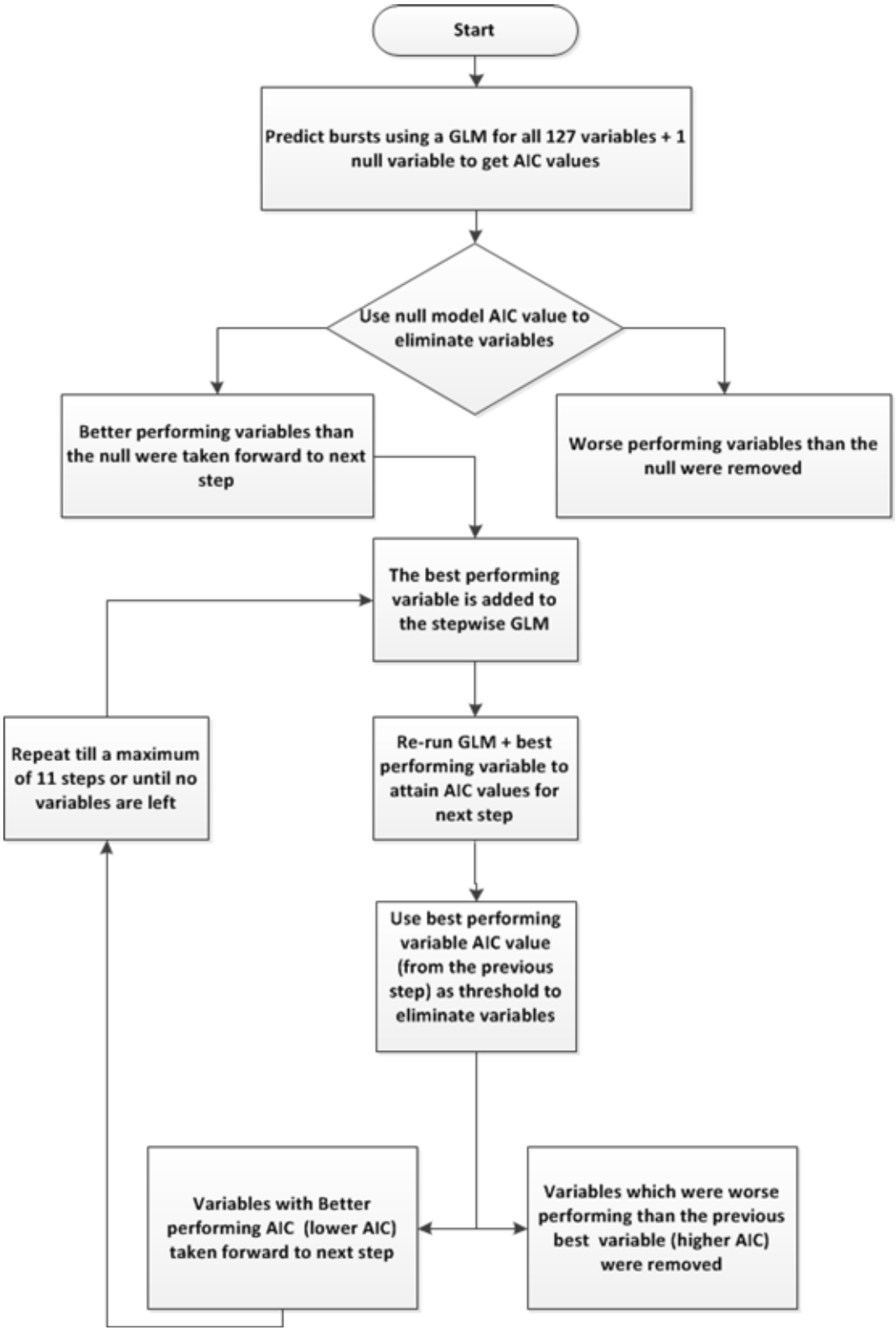

Figure 2-3: A flowchart of the methods used to select influential variables and create the stepwise GLMs

2.3.2 Model building and testing

Upon the final selection of stepwise models for each pipe material, predictions were made for each developed model's, and predictions run so the best model could be selected based on the minimum Root Mean Squared Error (RMSE) between observed and predicted bursts. GLM models were trained using 50% of the full dataset and tested using the remaining 50%, as a hold-out sample. Predictions of pipe failure was recorded for each of the model steps for each material. The residual between observed and predicted, AIC, RMSE and Mean Absolute Error (MAE) was extracted from the model fit or calculated to describe the fit of the models. RMSE and MAE are standard metrics to evaluate model performance, analysing the deviance between a set of observed and predicted outcomes (Farmani *et al.*, 2017).

All data analysis was undertaken in the open-source software R (version 3.2.3) (R Core Team, 2015) using Cranfield University's high-performance computing facility. The use of the high-performance computing facility was integral to the method, owing to the large volume of utility data and computational requirements of the methods chosen.

2.4 Results

2.4.1 Variable selection

Diameter Band and Age Band were the most frequently selected variables in this study, both being selected a total of 4 times across the final models (Table 2-6). Diameter Band was identified as being a predictive variable in Iron, AC, SDI and O, whilst Age Band was selected as being a predictive variable in Iron, AC, PVC, and O models. The inclusion of pipe age minimum, maximum and average variables was also found to be predictive in the Iron, SDI, PVC and PE models, which suggests using the absolute age of the pipes (either maximum, minimum or average) is a viable alternative to categorised age in some investigated pipeline materials. Using a dynamic age range is important as it helps the models to adapt over time, as we would expect the rate of failure of a pipe material to change as they age.

Soil properties was repeatedly identified as being predictive of pipeline failure in many of the final models. The climate-adjusted clay hazard variable was selected as an important explanatory variable for Iron, AC and SDI pipe materials. This indicates the potential impact of soil shrink and swell on the failure rate of these pipe materials. Soils which have a high corrosivity to Iron was also selected as an important explanatory variable in the AC, SDI and PVC final models. Depth to bedrock geology was identified as an influential variable for both Iron and AC pipes, whilst the hydrological rock type variable was identified as being predictive for the failure of PE pipes. The hydrology of soil type variable was also identified as a predictive within the SDI model. The identification of depth to bedrock, hydrological rock type and hydrology of soil type as predictive covariates suggest that soil drainage, periodic water-logging and soil hydrology are important factprs impacting upon the failure of these pipe materials. Soil texture was also identified as being predictive for several of the pipe materials, particularly in the AC model. Two sand texture variables was selected for the AC model, thereby indicating the influence of sand soil texture to the failure of AC pipelines. The final selected model for PVC also included silt minimum and soft and compressible soil hazard variables therefore suggesting that soft, unconsolidated alluvium is an important variable for the predicting the failure of PVC pipes.

Table 2-6: Variable selection count across all pipe materials for all models. Numbers represent the order of selection in stepwise model building

Note: grey numbers indicate that the variable was identified as being predictive but was not included in the final model

		I	AC	PVC	PE	SDI	O
Infrastructure	Age Band	7	8	3			3
	Diameter Band	2	1		4	1	1
	Pipe age minimum					3	
	Pipe age maximum	9		5			
	Pipe age average				3		

Soil	Corrosivity to Iron		4	6	4	
	Climate-adjusted clay hazard	3	5		8	
	Depth to rock	6	2			
	Hydrology of Soils Type (HOST)				2	
	Hydrological Rock Type (HYDROCK)			2		
	Shrink Swell	10				
	Sand Minimum			8		
	Sand Maximum		7			
	Sand washout hazard		9			
	Silt Minimum			2		
	Soft soil (% “at risk”)					6
	Soft and compressible soils hazard			9		
Weather	Day’s Air Frost	5				
	SMD (MORECS 40 km)		3			
	SMD change 2 weeks (MORECS 40 km)				10	
	SMD Deciduous Medium soil (MORECS 5 km)	4				4
	SMD Deciduous Low soil (MORECS 5 km)				6	
	Temperature change 1 week (MORECS 5 km)	8			7	
	Temperature ≤ 0°C 4 weeks (MORECS 5 km)				5	
	Temperature ≤ 2°C 1 week (MORECS 5 km)					5
	Temperature ≤3°C 4 weeks (MORECS 40 km)			4		
	Temperature ≤4°C 4 weeks (MORECS 40 km)		6			
	Temperature ≤5°C 1 week (MORECS 5 km)	1				
	Temperature ≤5°C 1 week (MORECS 40 km)	11				
	Temperature ≤5°C 4 weeks (MORECS 40 km)			1		2
	Rainfall (MORECS 40 km)			7		
	Vapour Pressure (MORECS 5 km)			1		

Important weather variables identified as being predictive of pipeline failure include 10 temperature and 4 SMD variables, along with the total number of day's air frost, rainfall and vapour pressure. Temperature was identified as an important variable in all pipe materials investigated, with Iron, SDI, and O pipeline failure models including two or more temperature related variables. The final Iron model includes 2 temperature related variables from MORECS 5 km dataset (Temperature change 1 week and Temperature ≤5°C 1 week) and the equivalent variable from MORECS 40 km dataset (Temperature ≤5°C 1 week). The selection of temperature variables from both MORECS 40 km and 5 km datasets suggests that spatial resolution of temperature measurements is less important for the prediction of pipeline failure. The number of day's air frost was also included in the Iron model, thereby further confirming the vulnerability of Iron pipes in temperatures under 5°C. Despite the inclusion of temperature variables at 1°C increments under 10°C, all variables selected show that pipe materials are more susceptible to failure under a 5°C threshold. Further to this, AC, PVC and PE models included temperature variables which are represent longer durations of cold temperatures over 4 weeks, whilst Iron pipes included variables which represent temperature change over a shorter 1-week period, see Table 2-6. O and SDI pipeline failure models contain temperature variables of both a 1 and 4-week duration, suggesting there is no clear response to the duration of cold weather and the failure of these pipe materials.

SMD variables were included in the final Iron, AC, SDI and O models. The most common SMD variable type to be identified as being predictive is SMD in a deciduous land cover type in either low or medium density (MORECS 5 km dataset), see Table 2-6. In all final models which contain an SMD related variable (with the exception of O) a climate-adjusted clay hazard variable is also included. This suggests that there is prominent effect of clay shrink across the Anglian region which impacts upon the failure of Iron, AC, and SDI pipeline materials.

The failure of PVC is also influenced by weather variables, rainfall and vapour pressure, and is the only pipe material for which these variables was identified as being predictive.

Table 2-7: Final model variables for the six different pipe materials

Note: material groups are Iron (I), asbestos cement (AC), polyvinylchloride (PVC), polyethylene (PE), steel and ductile Iron (SDI), and pipes classified as "other" (O)

Material	Variables
I	Temperature ≤5°C 1 week (MORECS 5 km) + diameter band + climate-adjusted clay hazard + SMD deciduous (M) + day's air frost + depth to bedrock + age band + temperature change 1 week (MORECS 5 km) + pipe age max + shrink-swell (1) + temperature ≤5°C 1 week (MORECS 40 km)
AC	Diameter band + depth to bedrock + SMD + corrosion to Iron + climate-adjusted clay hazard + temperature ≤4°C 4 weeks (MORECS 40 km) + sand maximum + age band + sand washout hazard
PVC	Vapour pressure + silt minimum + age band + temperature ≤3°C 4 weeks (MORECS 40 km) + pipe age maximum + corrosion to Iron + rainfall + sand minimum + soft and compressible soils hazard
PE	Temperature ≤5°C 4 weeks (MORECS 40km) + hydrological rock type + pipe age average
SDI	Diameter band + hydrology of soils type + pipe age minimum + corrosion to Iron + temperature ≤0°C 4 weeks (MORECS 5 km) + SMD deciduous (L) + temperature change 1 week (MORECS 5 km) + climate-adjusted clay hazard
O	Diameter band + temperature ≤5°C 4 weeks (MORECS 40 km)+ age band + SMD deciduous (M) + Temperature ≤2°C 1 week (MORECS 5 km)

2.4.2 Model prediction

When applied to the 50% hold-out sample, all final stepwise models improve the predictive ability in comparison to a null model (Figure 2-4). This is evidenced by a reduction of RMSE as additional variables are added to the stepwise models. As shown in Table 2-8, the range in percent difference between predicted and observed bursts was -3.15% to 21.85%, over the 8-year study period.

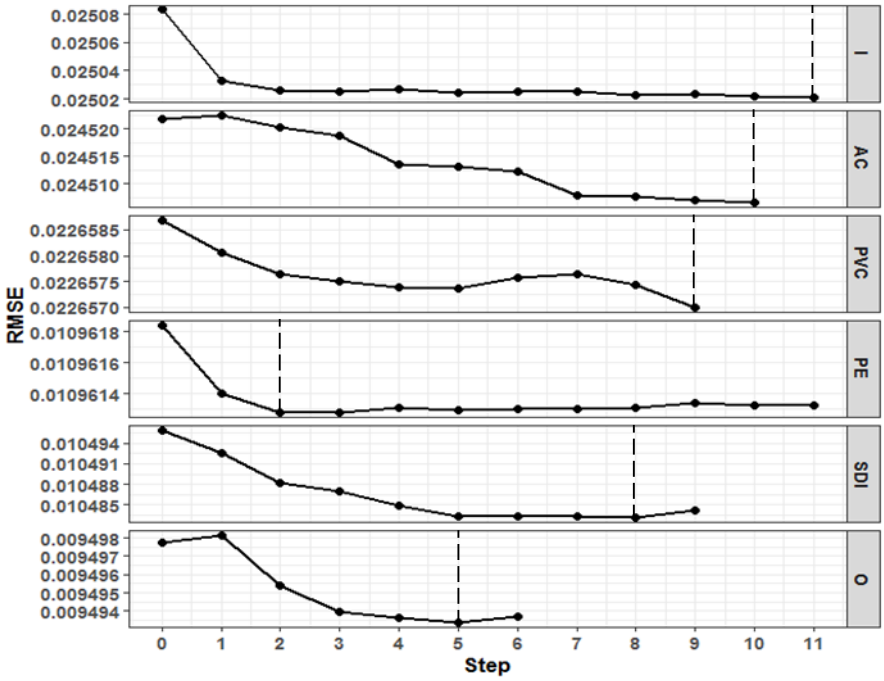

Figure 2-4: RMSE plotted against step for all models. Dashed lines represent the lowest RMSE which also represents the final model selection

The pipelines exhibit a strong seasonal pattern of bursts, with a marked increase of failures in the winter season (see Figure 2-5). The model captures this seasonal variation well. An overestimation of the model was observed for winter 2011, where ~100 bursts were over-predicted. However, the residual difference between predicted and observed bursts for Iron pipes is -112.1 bursts over the 8-year period, which is an overestimation of 2.64% of the observed amount. The predictive accuracy of the Iron model is evidenced by a low RMSE and MAE value of 0.0250 and 0.0012, which is aligned to the results reported within Yamijala et al. (2009).

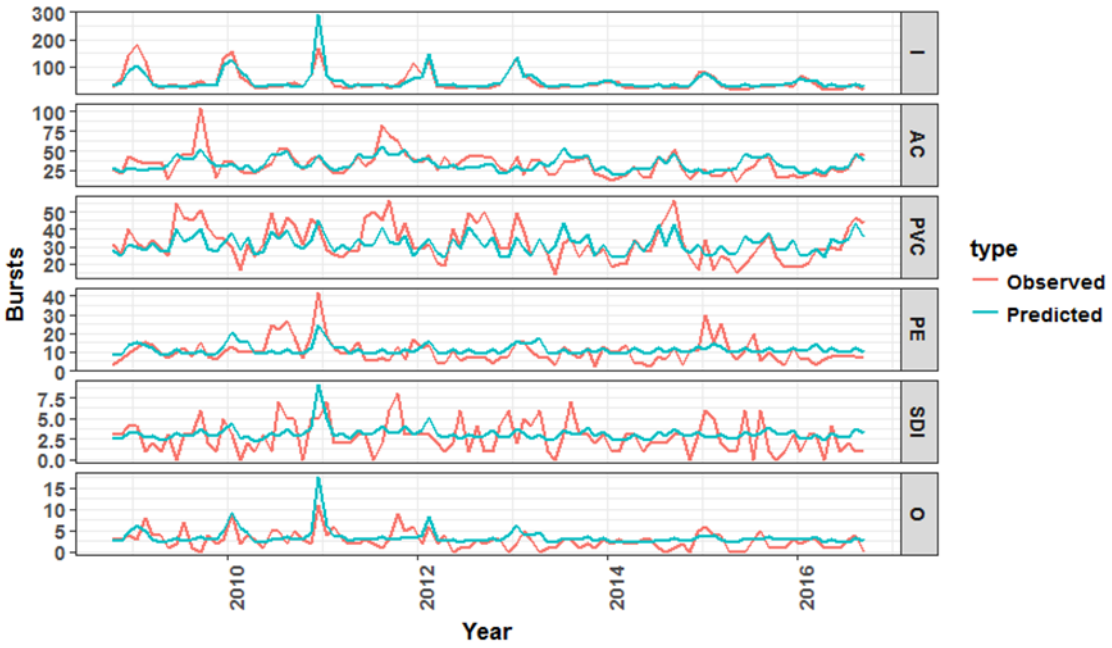

Figure 2-5: Time series of monthly model predictions vs. observations for final models in all pipe materials. Predictions were made using a 50% hold out sample

The final model selected for AC overestimated observed bursts by 0.59% over the 8-year observation period, this is the lowest percent change between observed and predicted bursts for all six material-specific models. The RMSE and MAE for AC was 0.0245 and 0.0011, respectively. The peaks of observed pipeline failures mostly occur during summer months (Figure 2-5), the model predictions largely capturing this pattern, particularly in years 2013, 2014 and 2015. An underestimation of the observed bursts is evident in the summers of 2009 and 2011. The residual between observed and predicted bursts is 18.9 bursts over the 8 years of observation.

PVC pipelines exhibit higher rates of failure in summer months, except for spring 2013. The predictions for the final PVC model appear to be relatively well aligned to the observed bursts. The residual between observed and predicted bursts for the final model is -98.1 bursts over the 8-year period, which is -3.15% of the observed bursts. No seasonal pattern in observed or predicted bursts is evident. A low RMSE and MAE value evidences the predictive ability of the PVC model, see Table 2-8.

The final PE model underestimated the observed bursts by 8.46%, leading to a residual of -84.6 bursts over the 8-year observation period. The predictions had an RMSE and MAE of 0.1096 and 0.0002, respectively. As shown in Figure 2-5, there is no distinct seasonal variation evident for PE bursts with regular bursts events occurring annually.

The final model for SDI overestimated the observed bursts by 13.1%, with a residual of 34.6 between observed and predicted bursts. There is no consistent seasonal pattern to the observed failure rate of SDI pipes, with failures occurring in all seasons (Figure 2-5). The model performs poorly at capturing the peaks in pipe failures. However, SDI pipelines comprise less than 5% of the overall length of the distribution network for Anglian Water, and just 1.85% of bursts come from this material (Table 2-3).

The largest error between predicted and observed bursts is found in O pipes, where the model overestimated bursts by 21.85%. The residual difference of O pipes is 58.8, during the 8-year observation period. It is important to note that O pipes constitute only 3.26% of the entire length of the water distribution network, and just 1.87% of bursts are classified in this material. Observed failures for O pipes follow a similar seasonal trend to Iron pipes, with distinct peaks of failures in winter. This is most likely because many of the pipes classified as “other” are metallic and have a similar failure mechanism to Iron pipes. The model captures this seasonal variation well, from years 2010 – 2012, however seems to be unresponsive to variations in pipe failures from 2013 – 2017. The final model selected for O pipes, in comparison to the other models, has the lowest RMSE and MAE of 0.0094 and 0.001, respectively.

Table 2-8: Final model selection and summary statistics

Note: observation period is the 7th October 2008 to the 27th September 2016

Material	Observed (o)	Predicted (p)	Residual difference (p – o)	% change (p – o)	RMSE	MAE
I	4242	4354.1	112.1	2.64%	0.0250	0.0012
AC	3157	3175.9	18.9	0.59%	0.0245	0.0011
PVC	3133	3034.9	-98.1	-3.15%	0.0226	0.0090
PE	999	1083.6	84.6	8.46%	0.1096	0.0002
SDI	264	298.6	34.6	13.1%	0.0104	0.0002
O	269	327.8	58.8	21.85%	0.0094	0.0001

2.5 Discussion

The literature reveals that pipe materials possess different tolerances to environmental and operational conditions. A series of stepwise GLM's were created to explore and evaluate the influence of soil type and weather on the rate of failure for 6 common pipe materials. This paper has permitted a) the selection of the most informative parameters for predicting pipeline failure for individual material types, and b) the assessment of the ability of a modelling approach to predict water pipe failure, drawing on an 8-year historic dataset from Anglian Water plc.

The results summarised in Table 2-8 suggest that all the developed models have a good ability to predict bursts. The RMSE values in this study are below the errors reported in other studies which use other modelling techniques such as Evolutionary Polynomial Regression (Farmani *et al.*, 2017) and are similar to the results published in (Tabesh *et al.*, 2009) where Artificial Neural Networks and Neuro-fuzzy Systems prediction methods were evaluated. The increase in RMSE

upon the addition of new variables (Figure 2-4) can be attributed to the increased variance within the datasets, along with the possible effect of overfitting the model to response variable. However, in using a quantified stepwise approach, the model which has attained the lowest RMSE, with the minimum number of variables allowed for the selection of a model with the highest model fit and minimum number of required variables.

The development of six individual stepwise GLM's, based on pipe material, enabled the selection of predictive covariates being related specifically to the material type. This led to the development of six parsimonious models, without the need for inclusion of un-informative parameters relating to other material types. It is noted by Wols and van Theinen (2014), that separating pipe materials into distinct classes is a critical step in determining the individual impacts of weather, soil and operational factors on the rates of failure. Moreover, analysis of individual material types is a computationally efficient way of predicting bursts over an entire distribution network.

Several repetitions in the selected predictive variables are apparent, with the most notable being pipe age and diameter. As key infrastructure factors, it was anticipated that age and diameter would prove predictive for all pipe materials. Pipes of different materials are often installed at certain time periods (Table 2-3). An improvement to this analysis would be to create material-specific age and diameter bands, to represent better the variance within the material group. However, the use of variable pipe age min, max and average was also included within this study, and was predictive for the failure Iron, SDI, PVE and PE pipes, which demonstrates its viability as an alternative to categorised age bands.

This study identified that temperatures under a 5°C weekly mean is more predictive of bursts than temperatures between 5-10°C for all investigated pipe materials. This information is important for water utilities to understand the thresholds within which pipelines are most likely to fail. The duration of temperature was also analysed, and it was found that bursts in AC, PVC and PE pipe materials is easier to predict using temperature changes over a longer 4-week period, whilst the failure of Iron pipes is easier to predict using temperature

change and temperature over 1 week. This corresponds to the findings presented by Rajani *et al.* (2012), where they noted that short-term fluctuation of extreme low temperatures (< 0°C) induced higher failure rates within Iron pipes. Conversely, prolonged dry weather in summer periods is known to impact upon soil shrinkage rates which leads to a higher failure rates in AC pipelines (Gould *et al.*, 2011). This mechanism is clearly evident in Figure 2-5, where notable peaks of AC failure occur within summer and autumn periods, where shrinkable soils, such as clay, are most likely to be shrunk. For materials, Iron, AC and SDI the finalised selected models incorporate both an SMD along with a clay soil variable. This indicates that these pipe materials are subjected to clay shrink-swell processes, as the literature suggests (Gould *et al.*, 2011).

Variables from both the MORECS 40 km and 5 km data products was identified as being predictive in all six pipe material types investigated. An equal number of temperature variables was selected from both datasets for predicting asset failure. Spatially, temperature variations across the Anglian region will remain relatively uniform, except for coastal areas. Therefore, it can be assumed that the representation of temperature at either a 5 km or 40 km spatial resolution is sufficient for predicting water pipe failure. The MORECS 5 km SMD variables was found to be better at predicting asset failure in comparison to the 40 km spatial resolution dataset. SMD varies at a much finer resolution than temperature, owing to the spatial heterogeneity of soil types. Therefore, the increased spatial representation of SMD at a 5 km scale, with representation on land cover type and density, proved to be more predictive of asset failure. Smaller scale micro-meteorological factors such as shading, relief and topography have not been included in this study due to a lack of data representing such conditions. Furthermore, in order to represent micro-meteorological variables within the model would require a further splitting the distribution network into smaller cohorts which represent such factors, leading to an increasingly complex dataset.

It is unclear from the existing literature whether temperature or shrink-swell related risks are important to the failure of PE and PVC pipelines (Rajani and Kleiner, 2001). It is noted by Davis *et al.* (2007) that PVC (and PE) pipelines are

newer pipe materials, and the failures attributed to these material types can sometimes be explained by poor manufacturing processes, or issues in installation, and not always to the prevailing environmental conditions.

The selection of depth to bedrock, hydrological rock type and hydrology of soil type variables in the final models of Iron, SDI and PE suggests that there is an influence of soil hydrology and soil drainage on the failure rate of these materials. There is a well-established relationship between the shrink-swell likelihood and soil moisture availability (Gould *et al.*, 2011; Wols and van Thienen, 2014). However, periodic water-logging and soil drainage is an important consideration missing from many of the current statistical models. An increase of soil saturation around metallic pipelines (such as Iron and SDI) has previously been noted to lead to higher rates of failure (Park, 2004; Wasim *et al.*, 2017). However, increased soil saturation can also create a seasonal soil movement pattern in certain soils (Chapter 5), investigating the effects of soil drainage and period water-logging on observed infrastructure movement using a satellite-based method. In shallow, freely-draining soils, minimal infrastructure movement occurred. However, in soils which are periodically water-logged a distinct seasonal variation in observed above-ground infrastructure movement occurred. These infrastructure movements corresponded to seasonal SMD. This seasonal movement in soil types might help further explain failures caused by differential soil settlement, along with the impact of waterlogged soil conditions causing corrosion or pitting failures to pipelines (Kleiner and Rajani, 2001). Soils which have impeded drainage across the Anglian region include loamy and clayey soils with naturally high ground water, loamy and sandy soils with naturally high ground water and a peaty surface, slowly permeable seasonally wet loamy and clayey soils, and slightly acid loam and clayey soils. Combined, these soil groups constitute 31.3% of soils in England and Wales, therefore evidencing the potential further impact of these soil groups on infrastructure failure extending beyond the study area selected in this study (Hallett *et al.*, 2017).

Several studies have considered soil parameters in predicting water pipe failure (Park, 2004; Wols *et al.*, 2014; Yamijala *et al.*, 2009). Soil texture is a key variable

in determining soil moisture availability, movement potential and drainage (Boorman *et al.*, 1995). This study highlighted two sand texture variables (the maximum percentage of sand in the soil association and sand washout hazard) which are predictive of asset failure in AC pipes. There is a low likelihood that sand is having a direct impact on the pipe material itself (with the exception of sand abrasion). However, sandy soils can give rise to cascading infrastructure failures, which is the successive failure of other buried assets which are co-located to the water distribution network. In a recent investigation (in the same study area), cascading infrastructure events are more than three times as likely to occur within a sandy soil (Farewell *et al.*, 2017). This has the potential to lead to a higher incidence of reactive bursts being recorded within sandy soils, therefore making the sand texture class variable predictive.

Silt minimum and soft and compressive soils hazard was identified as being predictive for PVC pipelines. Davis *et al.* (2008) used a physical probabilistic modelling technique to investigate the impact of failures within PVC pipelines. The authors found that pipeline fractures developed from internal cracks which existed at the time of manufacture. With increasing external pressure, the manufacturing defects are exploited and a rise in pipeline failure was observed to occur. Silty soils are highly compressible (Pritchard *et al.*, 2014), therefore this mechanism of soil compression inducing higher failure rates for PVC pipelines is highly likely for this pipe material.

The application of the work discussed within can be used anywhere detailed soil, weather and a historical archive of pipeline failure data exists. National soil datasets such as Statsgo in North America (US Department of Agriculture, 2008) and Australian Soil Resource Information System (ASRIS, 2018) provide a comprehensive inventory of soil-related variables, like the ones used in this study. Weather datasets such as Agri4cast in Europe (Joint Research Centre, 2018) are also freely-available and provide information upon key variables similar to the ones used in this study.

2.6 Conclusions

The combination of the environmental datasets used, and the models developed, offer the potential to help water utilities undertake proactive management of assets based on distinct environmental conditions. There is a strong seasonal pattern of failure of Iron, SDI, O and AC pipe materials across the Anglian Water network. Therefore, it is possible to predict bursts for these pipe materials with a high statistical accuracy. Key findings of this work include the prominent effect of clay shrink / swell, the seasonal waterlogging and soil hydrology impacts on pipeline failure and the identification of < 5°C temperature threshold for asset failure. An inclusion of a variable representative of soil temperature at pipe depth (c. 1 m) would be expected to improve the model's predictive ability, as discrepancies are likely to exist using air temperature as a proxy. The availability of such data is currently lacking but would be possible through the deployment of an *in-situ* network of smart sensors placed in the soil around the pipe in strategic locations.

The modelling approach of variable selection using AIC and Poisson GLMs has permitted the development of informative, yet parsimonious models for six common pipe materials. The results obtained are analogous to previous studies, confirming the ability of these models to predict water pipe failure. A limitation of this current study is the lack of information relating to the type of burst. Information concerning the type of failure observed (i.e. longitudinal, circumferential, pitting etc.) could be incorporated in future studies where such data are available.

Further understanding of the mechanisms leading to PE and PVC pipeline failure is still required. For the newer pipe materials types, e.g. PVC and PE, as the installed assets age, a better understanding of the factors leading to their deterioration will be gained, as the historical archive of data grows (Davis *et al.*, 2008). However, clear relationships between Iron, AC, SDI and O pipeline to environmental conditions exist. Headway has been made in furthering the understanding of PVC failure with the identification of soft, unconsolidated soil variables being predictive of failure.

The findings within have contributed to the general understanding of the environmental pressures which lead to asset failure across the Anglian Water network. These methods have the potential to be applied in other locations where data are available. Such information will lead to the development of more resilient and efficient water distribution networks, a key concern in the light of global climatic change. The models described within this study are now in operational use by Anglian Water plc as part of their Water Infrastructure and Serviceability Performance Assessment (WISPA) project.

Acknowledgements

This work was supported by the UK Natural Environment Research Council [NERC Ref: NE/M009009/1]. All authors made significant contribution to this work. M.N, T.F, and D.F designed the research methodology, M.N and T.F conducted the research and all authors contributed to the writing of the manuscript. We acknowledge the advice of Anglian Water plc. staff, and the provision of the asset data described here.

2.7 Bibliography

Aho, K., Derryberry, D., Peterson, T., 2014. Model selection for ecologists : the worldviews of AIC and BIC. Ecology 95, 631–636. https://doi.org/10.1890/13-1452.1

Akaike, H., 1974. A New Look at the Statistical Model Identification. IEEE Trans. Automat. Contr. 19.

ASRIS, 2018. ASRIS - Australian Soil Resource Information System [WWW Document]. URL http://www.asris.csiro.au (accessed 21.5.18).

Boorman, D., Hollis, J., Lilly, A., 1995. Hydrology of soil types: a hydrologically based classification of soils in the United Kingdom. Wallingford, UK.

Clark, C., 2002. Measured and estimated evaporation and soil moisture deficit for growers and the water industry. Meteorol. Appl. 9, 85–93. https://doi.org/10.1017/S1350482702001093

Davis, P., Marlow, D., Moglia, M., Davis, P., Silva, D. De, Marlow, D., Moglia, M., Gould, S., Burn, S., 2008. Failure prediction and optimal scheduling of replacements in asbestos cement water pipes. J. Water Supply Res. Technol. - AQUA 57, 240. https://doi.org/10.2166/aqua.2008.035

Davis, P.Ã., Burn, S., Moglia, M., Gould, S., 2007. A physical probabilistic model to predict failure rates in buried PVC pipelines. Reliab. Eng. Syst. Saf. 92, 1258–1266. https://doi.org/10.1016/j.ress.2006.08.001

Farewell, T.S., Jude, S., Pritchard, O., 2017. The influence of soil on the impacts of burst water mains on infrastructure and society : A mixed methods investigation. Nat. Hazards Earth Syst. Sci. in press, 1–30. http://dx.doi.org/10.5194/nhess-2017-433

Farmani, R., Kakoudakis, K., Behzadian, K., Butler, D., 2017. Pipe Failure Prediction in Water Distribution Systems Considering Static and Dynamic Factors. Procedia Eng. 186, 117–126. https://doi.org/10.1016/j.proeng.2017.03.217

Francis, R.A., Guikema, S.D., Henneman, L., 2014. Bayesian Belief Networks for predicting drinking water distribution system pipe breaks. Reliab. Eng. Syst. Saf. 130, 1–11. https://doi.org/10.1016/j.ress.2014.04.024

Gould, S.J.F., Boulaire, F.A., Burn, S., Zhao, X.L., Kodikara, J.K., 2011. Seasonal factors influencing the failure of buried water reticulation pipes. Water Sci. Technol. 63, 2692–2699. https://doi.org/10.2166/wst.2011.507

Hallett, S.H., Sakrabani, R., Keay, C.A., Hannan, J.A., 2017. Developments in land information systems : examples demonstrating land resource management capabilities and options. Soil Use Manag. 33, 514–529. https://doi.org/10.1111/sum.12380

Hough, M.N, Jones, R.J.A, 1997. The United Kingdom Meteorological Office rainfall and evaporation calculation system: MORECS Version 2.0 – an overview. Hydrol. Earth Syst. Sci. 1, 227-239. https://doi.org/10.5194/hess-1-227-1997

Jafar, R., Shahrour, I., Juran, I., 2010. Application of Artificial Neural Networks (ANN) to model the failure of urban water mains. Math. Comput. Model. 51, 1170–1180. https://doi.org/10.1016/j.mcm.2009.12.033

Joint Research Centre, 2018. Agri4Cast Resources Portal [WWW Document]. URL agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx (accessed 21.5.18).

Kabir, G., Tesfamariam, S., Loeppky, J., Sadiq, R., 2016. Predicting water main failures : A Bayesian model updating approach. Knowledge-Based Syst. 110, 144–156. https://doi.org/10.1016/j.knosys.2016.07.024

Kleiner, Y., Rajani, B., 2001. Comprehensive review of structural deterioration of water mains : statistical models. Urban Water 3, 131–150.

Li, S., Wang, S., Lin, X., 2013. Variable selection and estimation in generalized linear models with the seamless L0 penalty. Can. J. Stat. 40, 745–769. https://doi.org/10.1002/cjs.11165.Variable

Met Office, 2018a. UK regional climate summaries 1981 - 2018 [WWW Document]. URL https://www.metoffice.gov.uk/public/weather/climate (accessed 5.9.18).

Met Office, 2018b. Met Office Rainfall Evapotranspiration Calculation System - Specialist Datasets [WWW Document]. URL https://www.metoffice.gov.uk/services/industry/data/specialist-datasets (accessed 21.5.18).

Morozova, O., Levina, O., Uusküla, A., Heimer, R., 2015. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. BMC Med. Res. Methodol. 15, 1–17. https://doi.org/10.1186/s12874-015-0066-2

Nelder, A.J.A., Wedderburn, R.W.M., 1972. Generalized Linear Models. J. R. Stat. Soc. A 135, 370–384.

North, M., Farewell, T., Hallett, S., Bertelle, A., 2017. Monitoring the Response of Roads and Railways to Seasonal Soil Movement with Persistent Scatterers Interferometry over Six UK Sites. Remote Sens. 9. https://doi.org/10.3390/rs9090922

Park, S., 2004. Identifying the Hazard Characteristics of Pipes in Water Distribution Systems by using the Proportional Hazards Model : 2 . Applications. KSCE J. Civ. Eng. 8, 669–677. https://doi.org/10.1007/BF02823558

Pritchard, O.G., Hallett, S.H., Farewell, T.S., 2015a. Probabilistic soil moisture projections to assess Great Britain's future clay-related subsidence hazard. Clim. Change 133, 635–650. https://doi.org/10.1007/s10584-015-1486-z

Pritchard, O.G., Hallett, S.H., Farewell, T.S., 2015b. Soil geohazard mapping for improved asset management of UK local roads. Nat. Hazards Earth Syst. Sci. 15, 2079–2090. https://doi.org/10.5194/nhess-15-2079-2015

Pritchard, O.G., Hallett, S.H., Farewell, T.S., 2014. Soil Impacts on UK Infrastructure: current and future climate. Eng. Sustain. Proc. Inst. Civ. Eng. 167, 170–184. https://doi.org/doi:10.1680/ensu.13.00035

R Core Team, 2015. R: A language and environment for statistical computing [WWW Document]. URL https://www.r-project.org (accessed 21.5.18).

Rajani, B., Kleiner, Y., 2001. Comprehensive review of structural deterioration of water mains : physically based models. Urban Water 3. https://doi.org/10.1016/S1462-0758(01)00032-2

Rajani, B., Kleiner, Y., Sink, J., 2012. Exploration of the relationship between water main breaks and temperature covariates. Urban Water 9006, 67–84. https://doi.org/10.1080/1573062X.2011.630093

Rostum, J., 2000. Statistical Modelling of Pipe Failures in Water Networks. Nor. Univ. Sci. Technol. 1–132.

Tabesh, M., Soltani, J., Farmani, R., Savic, D., 2009. Assessing pipe failure rate and mechanical reliability of water distribution networks using data-driven modeling. J. hydroinformatics 11, 1–17. https://doi.org/10.2166/hydro.2009.008

US Department of Agriculture, 2008. Natural Resources Conservation Service - Web Soil Survey [WWW Document]. URL http://websoilsurvey.nrcs.usda.gov/ (accessed 20.5.18).

Wasim, M., Mojtba, M., Robert, D., 2017. Corrosion behaviour of pipes in soils and in simulated soil solution, in: Corrosion and Previention. Australian Corrosion Association, Blackburn, Australia.

Weirich, S., Silverstein, J., Rajagopalan, B., 2015. Resilience of Secondary Wastewater Treatment Plants : Prior Performance Is Predictive of Future Process Failure and Recovery Time. Environ. Eng. Sci. 32. https://doi.org/10.1089/ees.2014.0406

Wilson, D., Filion, Y., Moore, I., 2017. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. Urban Water J. 14, 173–184. https://doi.org/10.1080/1573062X.2015.1080848

Wols, B. a., van Thienen, P., 2014. Modelling the effect of climate change induced soil settling on drinking water distribution pipes. Comput. Geotech. 55, 240–247. https://doi.org/10.1016/j.compgeo.2013.09.003

Wols, B.A., Thienen, P. Van, 2016. Impact of climate on pipe failure: predictions of failures for drinking water distribution systems. Eur. J. Transp. Infrastruct. Res. 16, 240–253.

Wols, B.A., Thienen, P. Van, 2014. Impact of weather conditions on pipe failure : a statistical analysis. J. Water Supply Res. Technol. - AQUA 212–223. https://doi.org/10.2166/aqua.2013.088

Wols, B.A., Van Daal, K., Van Thienen, P., 2014. Effects of climate change on drinking water distribution network integrity: Predicting pipe failure resulting from differential soil settlement. Procedia Eng. 70, 1726–1734. https://doi.org/10.1016/j.proeng.2014.02.190

Wols, B.A., Van Thienen, P., 2014. Modelling the effect of climate change induced soil settling on drinking water distribution pipes. Comput. Geotech. 55, 240–247. https://doi.org/10.1016/j.compgeo.2013.09.003

Xu, Q., Chen, Q., Li, W., Ma, J., 2011. Pipe break prediction based on evolutionary data-driven methods with brief recorded data. Reliab. Eng. Syst. Saf. 96, 942–948. https://doi.org/10.1016/j.ress.2011.03.010

Yamijala, S., Guikema, S.D., Brumbelow, K., 2009. Statistical models for the analysis of water distribution system pipe break data. Reliab. Eng. Syst. Saf. 94, 282–293. https://doi.org/10.1016/j.ress.2008.03.011

Zhang, Z., 2016. Variable selection with stepwise and best subset approaches. Ann. Transl Med. 4, 136. https://doi.org/10.21037/atm.2016.03.35

3 Quantifying the impact of trees on water infrastructure failure across an entire distribution network

This chapter investigates Objective 3, and is presented in the form of one unpublished research paper, intended for the journal Environmental Modelling and Software:

North, M., Farewell, T., Hallett, S. (2018) Quantifying the impact of trees on water infrastructure failure across an entire distribution network (unpublished)

The two components of this chapter comprise: 1) an identification as to the rates of failure in different pipe cohorts under low (0-5%), medium (5-30%), and high (30-100%) tree density and soil shrink swell conditions, and 2) establishing how a series of statistical models can be further improved through the inclusion of tree-related data variables. The chapter describes and evaluates the Poisson regression models developed in Chapter 2, considering methods of enhancing current methods of prediction by evaluating the use of a national tree inventory.

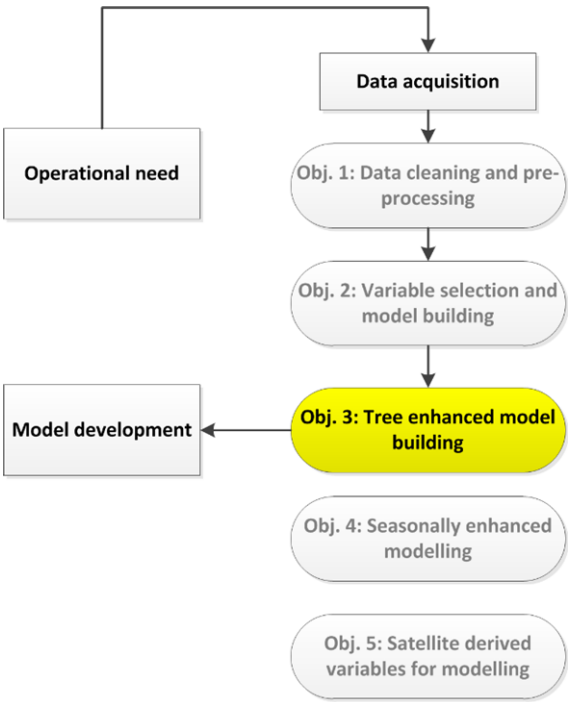

Figure 3-a: Objectives aimed to be investigated within this chapter in context of the overall thesis

Quantifying the impact of trees on the failure rates of different pipeline materials across an entire distribution network

M. North[a], T. Farewell[a], S. Hallett[a]

[a] *School of Water, Energy and Environment, Cranfield University, Bedfordshire, MK43 0AL, United Kingdom*

Corresponding Author: t.s.farewell@cranfield.ac.uk; tel.: +44 (0)1234 752978

Abstract:

Trees and large vegetation influence localised soil movements by increasing water loss through their evapotranspiration, causing proximal clay-rich soils to shrink. Such conditions can increase failure rates in drinking water pipelines. This study measured the failure rates of four pipe materials under varying tree density and soil conditions, investigating whether a series of material-specific pipeline failure models could be improved using data from a national tree inventory. Failure rates in Iron and polyvinylchloride pipes was found to be 6.4 and 3.5 times higher respectively in shrinkable clay soils in areas of high tree densities in comparison to pipes with the same tree density in non-shrinkable soils. Polyethylene pipes were found to be stable under all tree density and soil conditions, whereas asbestos cement pipes had varying failure rates. Predictive tree height variables was selected using Akaike's Information Criterion, with these being then added to a series of Poisson regression models. However, the inclusion of tree height did not significantly improve model accuracy, except by a limited improvement in the prediction of Iron pipeline failure.

Keywords: *Trees, Water Infrastructure, Soil, Weather, Failure Rate, Akaike's Information Criterion, Poisson Regression*

Highlights

- Tree density, height and distance impact on asset failure
- Variable selection using Akaike's Information Criterion (AIC)
- Tree-induced summer soil shrinkage and associated water asset failure
- Enhancement of predictive models incorporating relevant secondary datasets

3.1 Introduction

Trees are integral to sustainable urban environments due to their ability to reduce air temperature and pollution (Nowak *et al.*, 2006), to sequester carbon (Nowak and Crane, 2002), and to exert a positive impact on the health and well-being of people (Chiesura, 2004; Yannas, 2001). However, the detrimental impact of trees on the built environment has long been recognised in the scientific literature, particularly the impacts of trees on above-ground infrastructure such as buildings, pavements and roads (Mercer and Reeves, 2011; Randrup *et al.*, 2001; Watson *et al.*, 2014). The impacts of trees on above-ground infrastructure is often visible, resulting in the cracking, heaving or subsidence of structures (Jones and Jefferson, 2012; Watson *et al.*, 2014). However, establishing the impact of trees on below-ground assets, such as water, sewerage and gas pipelines, is more difficult due to the associated cost and disruption of inspecting these buried assets (Torres *et al.*, 2017).

3.1.1 Tree impacts to water infrastructure failure

Trees can impact on infrastructure either directly or indirectly. Direct interaction of trees on buried infrastructure can include processes such as root penetration, where tree roots grow along the path of least resistance and have the potential to exploit cracks or installation defects in pipelines. Tree roots have also been observed to proliferate in areas close to water sources such as leaking water pipes or sewer systems, making this type of buried utility particularly vulnerable (Gasson and Cutler, 1998; Jones and Jefferson, 2012; Torres *et al.*, 2017). However, tree root intrusions are unlikely to generate bursts in drinking water mains, due to the high internal pressure within the pipe. More commonly the

impact of trees on high-pressured water mains are indirect and include processes such as exacerbated clay-related shrink and swell and changes to the available soil water content within the 'active zone' (Day *et al.*, 2010; Navarro *et al.*, 2009). The active zone is the area which is influenced by the extent and depth of tree-related impacts on the soil (Jones and Jefferson, 2012).

Several factors converge to influence the depth and extent of the active zone, which may be summarised in 3 categories: tree, soil and external factors. Tree influences can include the species of tree, root depth and structure, height, age, gross primary productivity, asymmetric root growth, and the influence of a stand of multiple trees together (Biddle, 1979; Guo, 2017; Mercer and Reeves, 2011). A comprehensive summary of the physical interactions of trees and the active zone can be found within Biddle (1979). Numerous soil properties can influence the depth and extent of the active zone, including soil texture, drainage, hydrological regime, soil hydraulic conductivity, soil suction, and shrink swell potential (Guo, 2017). External factors influencing both soil and trees can include the presence of impervious surfaces such as pavements and roads, increased localised drainage, and also weather influences, particularly the balance between rainfall and evapotranspiration (Watson *et al.*, 2014). Based on unique combinations of tree, soil, weather, and external factors, the zone of influence a tree can exert is highly variable in depth and extent, as it is controlled by numerous interacting factors.

Typically the active zone depth is 1.5 to 2 m (Biddle, 1979), however, other studies have shown that this can extend to 4 m in extreme cases in the UK, and up to 6 m in other countries (Jones and Jefferson, 2012). Some further studies have also suggested that the zone of influence of trees can be extended by 1.5 times the original depth and extent where multiple trees are located together (Guo, 2017), further highlighting the potential impact of trees on pipes assets which are typically buried at 0.75 to 2 m depth.

In shrinkable clay soils, trees can intensify the extent of soil shrinkage due to increased water loss through evapotranspiration. This process is controlled by the prevailing weather, most notably the balance in water availability arising from

the rates of precipitation and evapotranspiration. A common index for measuring the soil water balance is Soil Moisture Deficit (SMD), which is a measurement used to describe the amount of water needed (in mm) to return the soil to field capacity (Pritchard *et al.*, 2015a). In temperate climate countries, the associated pattern of clay shrink swell is highly seasonal, with a maximum soil shrinkage in summer and early autumn (high SMD), and soil heave during winter and late spring (low SMD). This seasonal variation in SMD has been shown to impact upon above ground infrastructure such as roads and railways (Chapter 5), but also on the failure rates of different pipeline materials (Chapter 2).

Due to the known impacts of trees on subsidence, the UK's National Joint Utilities Group (NJUG) has provided a set of regulations for the guidance of siting utilities near trees (NJUG, 2007). These regulations also consider the ground movement potential of clay and peat soils and propose three zones of influence where trees are likely to cause tree-related damage to utility networks. Due to the relative infancy of the NJUG guidelines, in comparison to the age of water distribution networks, these regulations do not provide a solution to the impact of trees on pre-existing infrastructure networks but do act as a guidance for the installation of new utilities.

3.1.2 Methods to measure and predict tree related damage

Several *in situ* techniques exist to investigate tree-induced ground movement, such as electrical resistivity tomography (Jones *et al.*, 2009), and levelling and site surveys of tree-induced ground movement (Gasson and Cutler, 1998; Guo, 2017). These techniques provide an accurate method to measure tree-related ground movement directly. However, they are not practical for the analysis of tree-impacts on buried utility networks across a regional area, due to the limited spatial extent of observation and high associated costs. Therefore, mathematical modelling using a range of different environmental datasets and infrastructure information provides a favourable alternative, or complement, to physical measurements. Such techniques can summarise the impacts and interactions on a range of different environmental pressures across a whole distribution network, in a manner aligned to the needs of utility operators.

Water companies seek to become proactive in the way they manage their assets (Kleiner and Rajani, 2001; Yamijala *et al.*, 2009). Statistical models have the ability to predict water pipe failure by attributing a range of different environmental (typically weather and soils) and operational factors to a historical record of previous bursts. Following the selection of an appropriate statistical model, historical locations of previous pipe failures and associated environmental data can be represented using a Geographical Information System (GIS), and the relationships between operational and environmental factors can be statistically modelled.

Accurate and detailed tree inventories are available at the local, regional or national scale (Östberg, 2013). With such data becoming increasingly accessible, there is now greater potential to statistically model the interactions between trees and the built environment. This study presents a measurement of the impact of tree-related ground movement on the failure of buried water pipes to establish whether variables created from a national tree map can enhance the predictive ability of four material-specific water pipeline failure models. The tree-related properties considered represent the location, height and density of trees proximal to drinking water pipes. Using both rate of failure analysis and Akakie's information Criterion (AIC) (Akaike, 1974), influential tree variables have been selected which are predictive of pipeline failure for four common pipe materials, Iron, Asbestos Cement (AC), Polyvinylchloride (PVC) and Polyethylene (PE). The predictive variables identified was added to a series of material-specific pipe failure models, described and developed in Chapter 2. A direct comparison between the tree-enhanced and previously developed models is further provided.

3.2 Study area, model and dataset description

3.2.1 Study area

Anglian Water drinking water distribution network covers an area of c. 27,500 km^{-2} in East Anglia, England (Figure 3-1). The prevailing climate of the region is temperate oceanic, with distinct seasonal variations in temperature and regular precipitation events throughout the year. Temperatures across the East Anglian region range from a typical maximum average daily temperature of 22.2°C in July

to a minimum average temperature of 1.1°C in February. Rainfall is highest during the months November, December and January, whilst the driest months are July, August and September (Met Office, 2018a).

Across East Anglia, England, there are considerable deposits of silts, clays and peat soils which are derived from marine and riverine alluvium from previous glaciations and transgressions of the North Sea (Pritchard *et al.*, 2015b). Due to the intra-annual variation of sunshine hours, temperature and precipitation there is a strong seasonal pattern of SMD. High SMD is typically recorded during summer and autumn months, and a low SMD typically recorded during winter and spring. On this basis, East Anglia has amongst the highest ground movement potential in the UK due to the abundance of shrink swell prone soils and prevailing weather conditions (Pritchard *et al.*, 2015a). Trees are widespread across the East Anglia region, with large variations in the density, height and species of trees which are located against infrastructure networks, often varying with changes in urban and rural landscapes.

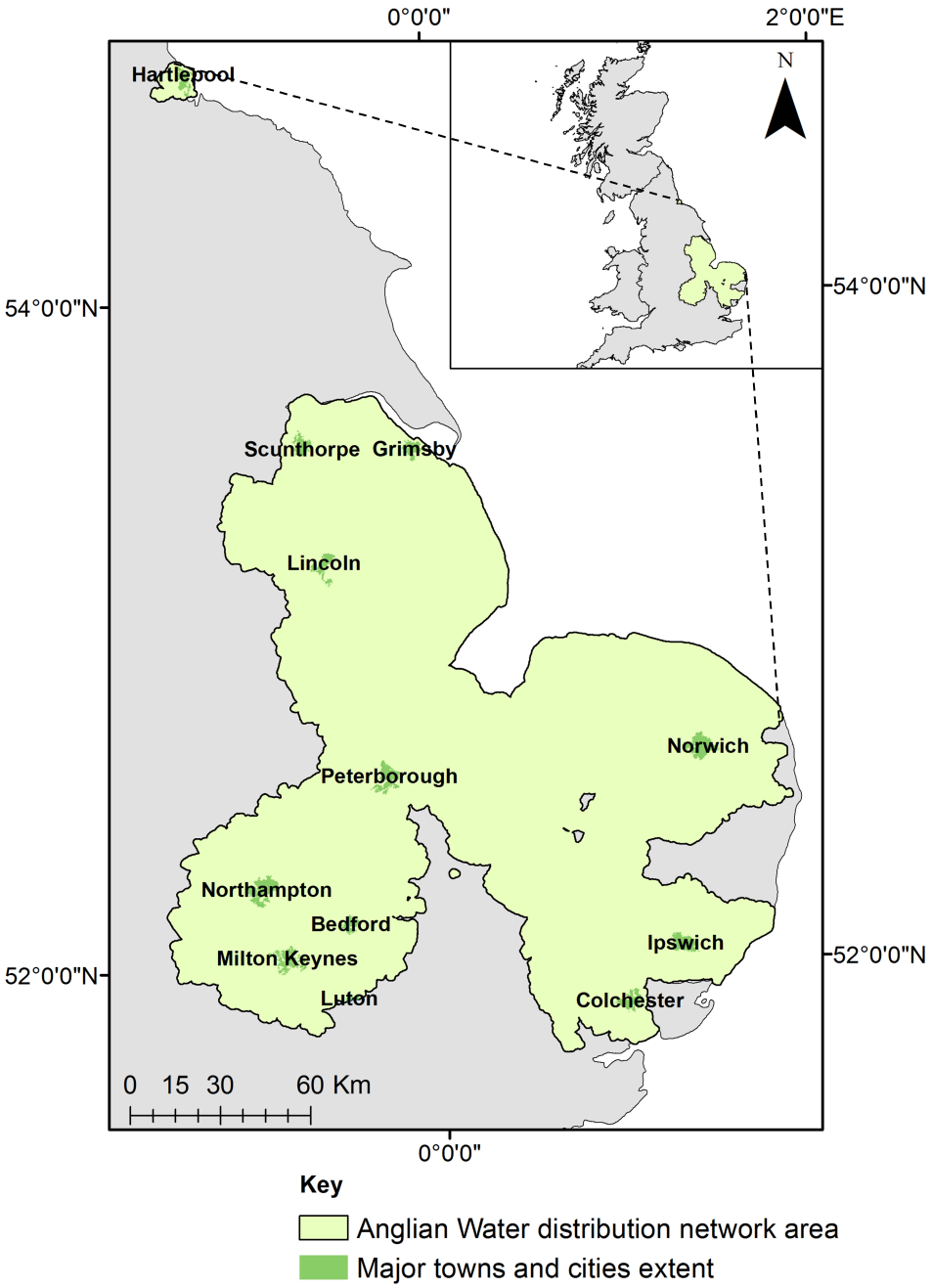

Figure 3-1: Study area extent of the Anglian Water service area; (insert map shows the relative position of the distribution network in context of the UK)

3.2.2 Tree dataset description

The tree inventory used for this investigation is provided by commercial provider Bluesky Ltd. (Bluesky, 2018). The map, known as the "National Tree Map" ™, uses aerial imagery sources to map every tree over 3 m, covering the full extent of England and Wales. The accuracy of the map is >90 % for the full extent of England and Wales and >95 % for trees within 50 m from buildings (Bluesky, 2018). The output is in distributed in a GIS format, and provides the point location and canopy extent of individual trees, with attributes including the individual tree height. No information is included pertaining to the tree species.

3.2.3 Water Infrastructure data description

Water infrastructure data was provided for this investigation by Anglian Water plc, comprising a dataset with historical incidences of pipeline failure from 28th December 2005 to 27th December 2016. This data has been cleaned and pre-processed to ensure that a consistent, complete and accurate dataset is used for statistical modelling. A description of the pre-processing and data cleaning steps to create the pipe cohorts used for statistical analysis are detailed in Chapter 2.

A total of 4 main pipe material groups have been selected for investigation, representing over 96 % of the total length of the Anglian Water pipe network. These pipe material groups comprise Iron, AC, PVC and PE. Two other material categories, steel and ductile Iron pipes are classified as "other", was excluded from this study due to their small sample size. The total length of pipe analysed in this study is 38,438 km, being the mean network length over the period of investigation (Table 3-1). The total number of reactive bursts across the 4 materials analysed was 42,623 over the 10-year observation period, where a reactive burst is defined as a burst reported to the Anglian Water leakage team by a third party, rather than one "proactively" identified by the company's leakage inspector teams.

Table 3-1: Summary of the failure rates of the investigated pipe materials across the Anglian Water distribution network. Burst rates (pipe bursts per 1,000 km per week) are calculated for reactive bursts only

Material	Typical Installation Range	Total Length (km)	Total Reactive Bursts	Burst Rate (Bursts/ 1000km/ per week	Total Bursts (May-Oct)	Burst Rate (May-Oct)	Total Bursts (Oct-May)	Burst Rate (Oct-May)
AC	1920 to 1941	7,000	9,282	2.31	4,809	2.86	4,473	1.91
I	1881 to 1921	11,226	20,737	3.21	3,172	1.17	17,565	4.68
PVC	1960 to 2001	5,951	9,275	2.71	2,220	1.55	3,731	1.87
PE	1981 to present	10,761	3,229	0.52	650	0.25	2,579	0.71
Total:		39,438	42,623					

3.2.4 Pipeline failure model description

The water pipeline failure models used within this study are developed using Poisson regression, a form of generalised linear model (Coxe *et al.*, 2009; Nelder and Wedderburn, 1972). These models have been built using a step-wise model building approach using Akaike's Information Criterion (AIC) as the variable selection method (Chapter 2). A full description of the model building and variable selection process is provided in Chapter 2. This Chapter provides a full description of the datasets and the variables included in the models. Table 3-2 describes the variables included in the material-specific models. For a specific description of the weather, soil and operational variables, details are given in subsequent Tables 3-2 to 3-6.

Table 3-2: Variables included in the original material-specific water pipeline failure models (Chapter 2)

Material	Model Variables
AC	Diameter band + SMD + corrosion to Iron + climate-adjusted clay hazard + sand min + depth to bedrock + age band + temperature ≤4°C over 4 weeks
I	Days air frost + temperature ≤5°C over 1 week + diameter band + shrink swell + climate-adjusted clay hazard + SMD + age band + temperature change over 1 week + depth to bedrock + pipe age maximum
PVC	Silt minimum + vapour pressure + age band + corrosion to Iron + soft and compressive soils hazard + sand minimum + pipe age maximum + rainfall + temperature ≤3°C over 4 weeks
PE	Temperature ≤4°C over 1 week + hydrological rock type + pipe age average

3.2.5 Weather Variable Description

A series of variables which represent the weather conditions at the time of and preceding failure are included as being predictive of water pipeline failure. Weather data has been obtained from the Met Office Rainfall Evapotranspiration Calculation System (MORECS) dataset (Met Office, 2018b). This provides a gridded output of modelled meteorological data with a spatial resolution of 5 km. The 5 km MORECS dataset is a beta version of the dataset and has been provided to the authors for testing to determine its suitability for statistically modelling water infrastructure failure. The number of day's air frost data has been acquired from the Met Office regional climate statistics, and represents the number of days per month when air frost was observed in East Anglia (Met Office, 2018a). The dynamic weather variables included in the models represent weekly temperature, SMD, rainfall and vapour pressure. A description of the included weather variables is given in Table 3-3.

Table 3-3: A description of the weather variables included in the original material-specific water pipeline failure models

Variable	Description	Included in models
Days air frost	Total number of days air frost in a month	I
Rainfall	Total accumulated rainfall in the 1 week preceding the reported burst date	PVC
SMD	Weekly mean SMD under different land uses (grassland, deciduous, coniferous, real land use) and in soils with different water holding abilities (High, Medium, Low) in the week preceeding the reported burst date	AC, I
Temperature ≤ (0 to 10°C) over 1 weeks	The accumulated air temperature beneath a threshold (0 - 10°c, in 1°c increments) in a 1-week period preceding the reported burst date. Air temperature accumulations are calculated in 1°c increments below the threshold value. The greater the value of this variable, the colder the temperature was, for a prolonged period of time, in the previous week to the burst date	I, PE
Temperature ≤ (0 to 10°C) over 4 weeks	The accumulated air temperature beneath a threshold (0 - 10°c, in 1°c increments) in a 4- week period preceding the reported burst date. Air temperature accumulations are calculated in 1°c increments below the threshold value. The greater the value of this variable, the colder the temperature was, for a prolonged period of time, in the previous weeks to the burst date	AC, PVC
Temperature Change	The absolute change in mean air temperature from the week the burst was reported to 1, 2 or 4-weeks before	I
Vapour pressure	Vapour pressure is a variable which indicated the partial pressure that atmospheric water vapour exerts at any one time	PVC

3.2.6 Soils Data Description

Soils information has been taken from the National Soil Map of England and Wales and associated soil property datasets (Hallett *et al.*, 2017). The soil map used in this study is produced at a 1:250,000 scale and contains detailed information relating to a wide range of different physical and chemical factors, including soil texture, corrosivity, depth to rock, shrink and swell capacity and soil hydrological regime. For a full description of the soil variables included in the final models, refer to Table 3-4.

Table 3-4: A description of the soil variables included in the original material-specific water pipeline failure models

Variable	Description	Included in models
Climate-adjusted clay hazard	% of clay in the most abundant soil series	AC, I
Corrosion to Iron	6 level categorical classification of the corrosivity of the soil to ferrous Iron at 1m depth. Dominant class and % composition of individual classes	AC, PVC
Depth to bedrock	Depth to bedrock (mean, min and max) (cm)	AC, I
Hydrological rock type	Provides information of the hydrological rock type of the soil. Dominant class in the soil association provided. A full description of this variable is provided by (Boorman *et al.*, 1995)	PE
Sand minimum	Minimum % of sand content	AC
Shrink swell	A 6 level categorical classification of the shrinkability of the soil	I
Silt minimum	Minimum % of silt soil content	PVC
Soft and compressible soils hazard	Soils at risk of including soft soil or compressible soils	PVC

3.2.7 Operational Variable Description

Due to the known importance of operational variables in the prediction of water pipeline failure (Kleiner and Rajani, 2001), a series of key variables representing the age and diameter of pipeline was included (Table 3-5). Several age-related variables was included accommodating the various ways age can be represented. The variables included were found, using AIC in a forward's stepwise approach, to be the most predictive variables for each specific material type, as described in Chapter 2. The measurements of the pipeline used to define the diameter and age bands are given in Table 3-6.

Table 3-5: A description of the operational variables included in the original material-specific water pipeline failure models

Variable	Description	Included in models
Diameter Band	Categorised bands (Table 3-6)	AC, I
Age Band	Categorised bands (Table 3-6)	AC, I, PVC
Pipe age (maximum)	Maximum age of pipe	I, PVC
Pipe age (average)	Mean age of pipe	PE

Table 3-6: Age Band and Diameter Band ranges used in this study

Age Band	Range
0	Unknown
1	<1881
2	1881 to 1900
3	1901 to 1920
4	1921 to 1940
5	1941 to 1960
6	1961 to 1980
7	1981 to 2000
8	2001 to 2020

Diameter Band	Range
1	<165 mm
2	165 to 320 mm
3	321 to 625 mm
4	>625mm

3.2.8 Tree variables

A series of variables was derived from the National Tree Map, representing the height, location and density of trees proximal to pipelines across the Anglian Water distribution network. Trees have different water consumptions, based on their height, diameter, age, foliage, species and gross primary productivity (Biddle, 1972). Directly, the water consumption of trees can have a measurable impact on ground movement in shrinkable soils, such as clay, as trees can change the localised water balance often creating a deficit of water in the immediate proximity of the tree, causing clay soils to shrink (Clayton et al., 2010). Trees of younger age, increased leaf area, with a larger basal height, typically have a higher water consumption which increases the risk of ground movement and the failure of adjacent water pipes (Biddle, 1972, Clayton *et al.*, 2010). These characteristics change with individual tree species, however, a disadvantage of using the National Tree Map is that it includes no classification of individual tree species. Alternative methods, such as image classification from optical sensors (spaceborne and airborne) can be used to identify individual tree species based on their spectral properties (Walton *et al.,* 2008). Individual tree classification methods were not used in this study due to the additional complexity and computational resource needed to classify the species of all trees adjacent to the water distribution network.

To reduce the computational demand of the data processing in the GIS tool ArcMap (version 10.3.1), the National Tree Map was processed using 9 individual British National Grid 100 km^{-2} Ordinance Survey (OS) grid squares. A further pre-processing step included removing all trees located >40 m from the pipe network.

Table 3-7: Tree height band and associated measurements

Tree Height Band	Measured Height
1	< 5 m
2	>5 – 10 m
3	>10 – 20 m
4	>20 – 30 m
5	>30 m

The Anglian Water pipe network was split into cohorts representing homogenous pipe materials, ages and diameters located in the same soil types. The creation of pipe cohorts permitted the analysis of tree impacts on homogenised cohorts of pipes, to increase computational efficiency. The average length of each pipe cohort is 0.633 km long (with a standard deviation of 1.139 km). For each pipe cohort, multiple buffers of 10, 20 and 40 m was created and the total area of tree canopy (as a percentage) within each buffer zone was recorded. For each pipe cohort's buffer zone, the total area of trees in each height category (Table 3-7) was recorded enabling the total percentage of canopy coverage for each tree height band to be calculated per pipe cohort (Figure 3-2). This led to the formation of variables which indicated the percentage of the total area of trees per height band in each individual pipe cohorts. The variables created were descriptive of tree height, the proximity of trees to the pipe network (at 10, 20 and 40 m distances) and the density of tree cover surrounding each pipe cohort (as a percentage cover). Another variable was created which evaluated all tree canopy cover, irrespective of height, for each pipe cohort. This variable was used for the investigation of the rate of failure for different pipe materials, described in Section 3.3.1.

Due to the potential annual changes to the pipe distribution network, the processing of the pipe network and National Tree Map data was repeated yearly, to ensure that all pipes over the 10-year observation period had the correctly associated tree-related variables for pipes which were in service that year.

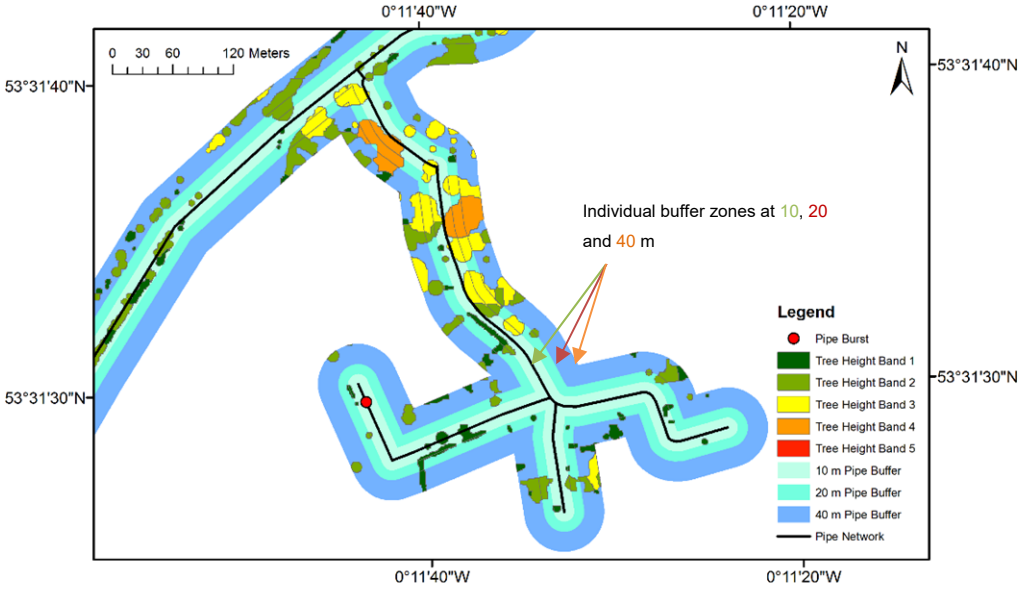

Figure 3-2: A visualisation of the GIS approach taken to create tree variables. 10, 20 and 40m buffer zones are indicated as shades of blue adjacent to the pipeline which is coloured in black.

3.3 Methods

3.3.1 Rate of failure analysis

Initial analysis investigated the mean average failure rates (bursts per 1,000 km of pipe per week) for AC, Iron, PVC and PE pipe materials per month. This investigation allowed the analysis of the trends of water pipe failure rates over the period of observation, to identify the changes in failure rate, per pipe material, between the summer and winter seasons. No representation of tree influence was used, but the results provided a baseline for the average failure rates for different pipeline materials.

Further investigations included calculating the failure rates (bursts per 1,000 km per week) of different pipeline materials under different tree density and soil shrink and swell conditions, at 10, 20 and 40 m buffer distances. A total of 3 categories of tree densities (percentage of pipe cohort covered by tree canopy) was used, which represents 0-5 % (low density tree cover), 5-30 % (medium density tree cover) and 30-100 % (high density tree cover). These categories

were used to ensure that a representative length of pipe and number of bursts are recorded in each category so that a sufficient sample size was maintained for the calculation of burst rates.

For this investigation, a 6-level categorical classification of the soil shrink and swell potential was used, which was acquired from the National Soil Map of England and Wales (Hallett *et al.*, 2017). The shrink swell variable used describes the shrink swell potential of the most dominant soil type within the soil association. The 6-level categorical classification was further classified into 4 bands, representing low, medium, high and very high shrink swell potential (Table 3-8). This reclassification was done to ensure that a sufficient sample size of pipes was included in analysis of each shrink swell category. Shrink swell class 6 is an alluvial clay or peat, which has very high shrink swell potential that is not realised unless effective drainage is installed to at least 2 m depth, hence its inclusion in the “Medium” shrink swell potential category.

Table 3-8: Shrink swell categories used in this present study. Shrink swell class is the original banding from the National Soils Map of England and Wales. Shrink swell potential category is how they have been grouped in the present study

Shrink swell potential category	Shrink swell class	Volumetric shrinkage
“Low”	1, 2	<3%
“Medium”	6, 3	5 – 12 %
“High”	4	12 – 15 %
“Very High”	5	>15%

The rates of failure investigating tree density and shrink swell potential was calculated for the summer period only (May to October) which is the season most likely to show a stronger pattern of failure rates in shrink-swell soils (Mercer and Reeves, 2011). For the calculation of seasonal failure rates, no burst data was removed, and the failure rates were calculated using 100% of the burst and pipe data from the respective season being investigated.

3.3.2 Akaike's Information Criterion (AIC) analysis

Akaike's Information Criterion (AIC) has been widely used as a variable selection method in scientific disciplines (Aho *et al.*, 2014). AIC provides a test that quantifies how predictive a set of covariates are to a response variable in a given set of data. To this extent, AIC can be used for the creation of statistical models (Zhang, 2016), but also for the identification of predictive covariates. A full mathematical description of AIC is given in (Akaike, 1974). A specific example of the use of AIC for building stepwise Poisson regression models for water infrastructure failure is provided (Chapter 2).

To identify the most influential tree variables, relating to tree height and the proximity to buried assets, a series of individual AIC analyses was undertaken. This helped establish the relative predictive ability of all newly developed tree covariates in comparison to the weather and soil covariates already included into pipeline failure models.

As a second step, the predictive ability of different combinations of different tree heights was investigated. This analysis was undertaken with a soil shrink swell variable, due to the known impacts of trees on shrinkable soils (Guo, 2017; Mcpherson and Peper, 1996; Mercer and Reeves, 2011). The different combinations of tree height bands were evaluated sequentially using AIC from the smallest tree height band (T1) to the largest tree height band (T5). This resulted in the testing of 6 combinations of tree height and shrink swell variables.

3.3.3 Poisson Regression – Tree Enhanced Modelling

For each material type, the most predictive combination of tree height variables, as indicated by the lowest AIC, was then added to the water infrastructure failure models described in Table 3-2. A direct comparison between the models without and those with the inclusion of tree factors was undertaken. Both models were trained and tested on identical data (using a 50% hold-out sample). Model performance was evaluated using the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) between observed and predicted bursts. RMSE and MAE are standard statistical metrics used to evaluate model performance and

can be used to compare different models directly. All data analysis and predictive modelling was undertaken in the open-source software R (version 3.2.3) (R Core Team, 2015).

3.4 Results

3.4.1 Seasonal failure rates of pipe materials

There are clear seasonal trends observed in the failure rates (bursts per 1,000 km of pipe per week) in AC, Iron and PVC pipes (Figure 3-3). Conversely, the failure rate in PE pipes shows no seasonal trend and exhibits a near constant monthly average failure rate of 0.25 to 0.5 bursts per 1,000 km of pipe per week. The highest monthly average failure rate is recorded in Iron pipes, where there are > 3 bursts per 1,000 km of pipe per week observed during the winter season between December and February. A notable rise in monthly failure rates during the summer season is also observed in AC, PVC and Iron pipes, from June to September. AC pipes show the largest increase of ~1 burst per 1,000 km per week from June to September, whilst PVC and Iron pipes show a smaller increase of ~0.5 bursts per 1,000 km or pipe per week from June to July, and then a slower increase towards September (Figure 3-3).

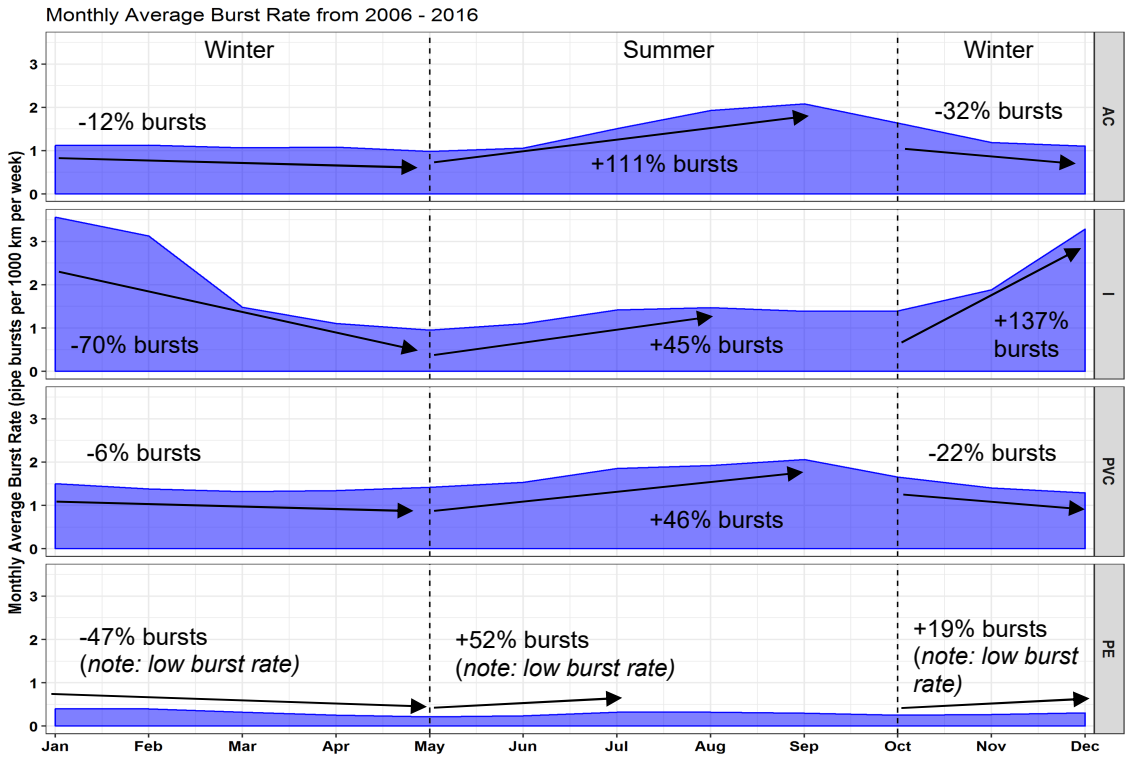

Figure 3-3: The average monthly rate of failure (pipe bursts per 1,000 km of pipe per week) for Asbestos Cement (AC), Iron (I), Polyvinylchloride (PVC) and Polyethylene (PE) pipelines

Note: percentages indicate the relative change in bursts rates (representing the largest difference) from either January to May, May to October, or October to December. Dashed lines indicate the summer and winter periods investigated in this study

3.4.2 The impact of tree density and proximity on the failure rates of pipe materials

The analysis into the impact of tree density and soil shrink swell potential on water pipeline failure was undertaken for the summer period only. This has been chosen as the risk of tree-related ground movement to the water utility network is greatest in summer, as there is a higher water demand causing proximal clay-rich soils to shrink. The summer period used, as highlighted in Figure 3-3, captures the peak of AC, Iron and PVC pipeline failures, and includes a total of 5 months (May to October). Upon the final grouping of pipe cohorts into tree density and shrink swell categories, entries with no bursts recorded and <1 km of pipe were excluded from analysis, ensuring that a fully representative sample size is maintained and that meaningful burst rates are reported.

The failure rates in Iron and PVC pipes, at a 40 m distance, was 6.4 and 3.5 times higher respectively than the failure rates observed under the same tree density in a low shrink swell soil (Figure 3-4).

AC pipes were found to have varying failure rates under all tree densities and soil shrink swell potentials. The lowest failure rate recorded in AC was observed in pipes situated in low shrink swell potential soils with low tree density (0-5 % coverage) in a 40 m buffer distance, resulting in a failure rate of 1.59 bursts per 1,000 km of pipe per week. Conversely, the highest failure rates in AC pipes was recorded in a high shrink swell soil under medium tree density conditions (5-30 % coverage), being 4.09 bursts per 1,000 km of pipe per week. The failure rate in AC pipes was found to increase proportionately with ground movement potential and higher tree densities.

The summer failure rates in Iron pipes was found to increase with ground movement potential and tree density. This pattern of failure is replicated in the 10, 20 and 40 m analysis (Figure 3-4). The failure rate in Iron pipes, buried in a low shrink swell potential soil and low tree density (0-5 %) was 1.90, 1.84 and 1.55 pipe bursts per 1,000 km per week in 10, 20 and 40 m distance, respectively. The failure rates increase in both a medium and high shrink swell soil, with the highest failure rates in Iron pipelines reaching 7.73 bursts per 1,000 km per week under very high shrink swell soils and high tree density (30-100 %) at a 40 m distance.

Despite the clear increase in failure rates under high tree density and very high ground movement potential (in the 40 m analysis), the failure rates in PVC pipes remain consistent in low, medium and high shrink swell potential and tree densities. No clear trends in the variation of failure rates are therefore apparent for PVC material, with failure rates ranging from 2.29 under medium ground movement potential and high tree density (30-100 %) at a 20 m distance, to 3.94 bursts per 1,000 km per week under very high ground movement potential and high tree density (30-100 %) in a 20 m buffer distance. Failure rates under the same very high ground movement potential and tree density in 40 m was 2.4 times greater than the failure rates in a 20 m distance, at 9.7 bursts per 1,000 km

per week, therefore indicating the influence of high tree density and ground movement potential to the failure of PVC pipes.

The lowest rates of pipe failure, in all shrink swell and tree density conditions, was observed in PE pipes, aligned with the results in Figure 3-3. Under all shrink swell and tree density conditions the failure rates of PE pipe material are mostly consistent at < 1 bursts per 1,000 km of pipe per week. There is a slightly larger failure rate in PE pipes in very high shrink swell soils, under medium tree density (5-30 % canopy coverage) at 10 m, where the failure rates of PE reach 1.4 bursts per 1,000 km per week.

3.4.3 AIC analysis of tree height variables vs. operational, weather and soil variables

Individual analysis of the predictive ability of each model variable was assessed using AIC. Following the results in Section 3.3.2, only variables representing the percentage of tree canopy coverage in different tree heights bands (1 – 5) up to a 40 m distance was included for this investigation. A comparison of the predictive ability of tree height categories (bands 1 to 5) and the previously selected operational, weather and soil model variables (Chapter 2) are highlighted in Figure 3-5. A null variable was also included so a direct comparison to a dummy variable could be determined. Lower AIC values indicate a greater contribution of the individual variable in predicting reactive bursts. As individual models were developed for each material type, the AIC values obtained are not comparable across the material specific models.

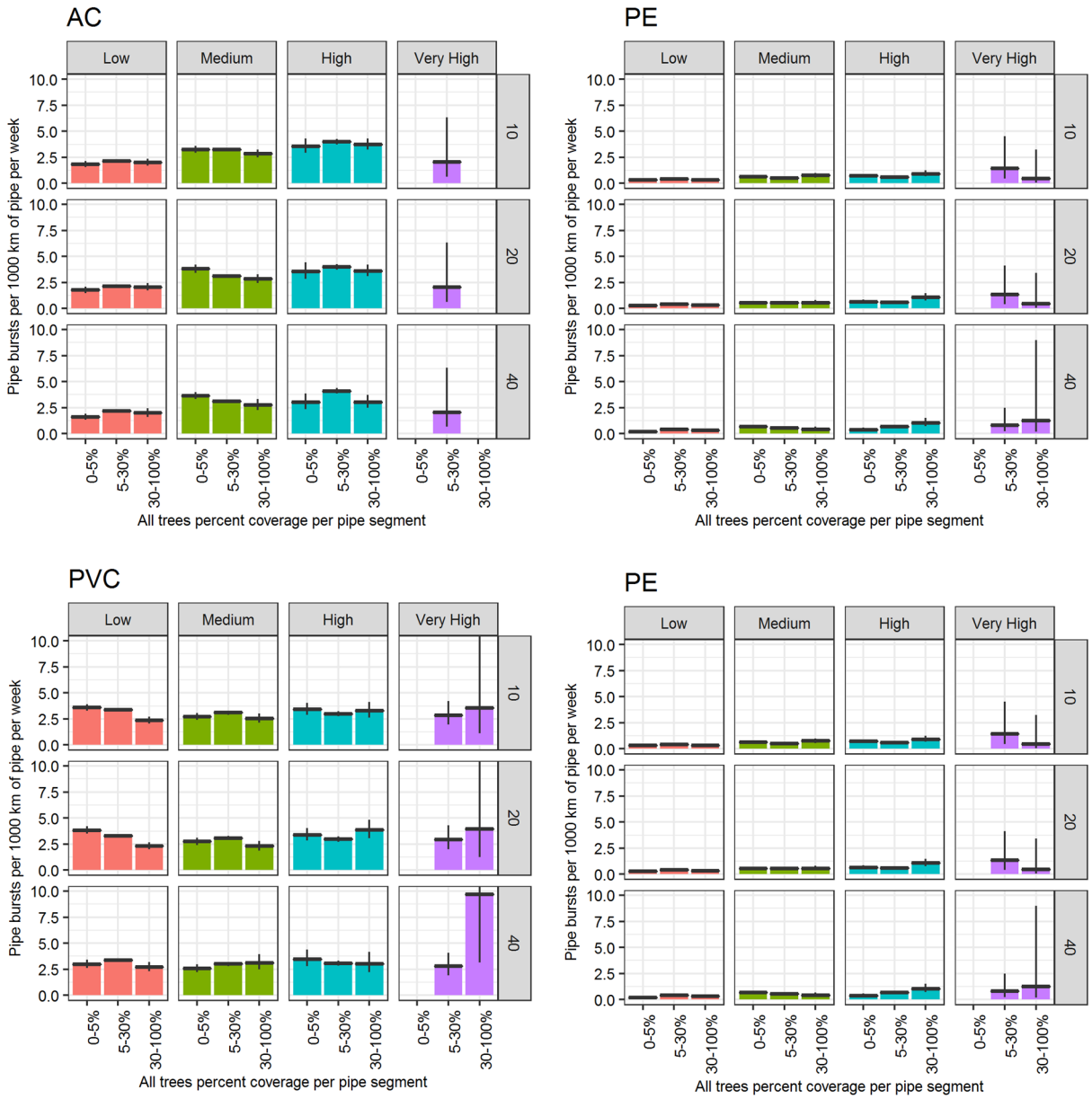

Figure 3-4: Summer (1st June – 31st October) water pipeline failure rates for 4 materials Asbestos Cement (AC), Iron, Polyethylene (PE) and Polyvinylchloride (PVC) under different tree densities and shrink swell conditions. For each material, results are also shown for the analysis of tree density at 10 m, 20 m, and 40 m distances. Error bars represent the 95% confidence interval for the Poisson mean estimate

Note: Low, Medium, High, and Very High refer to the soil shrink swell category described in Section 3.2.6

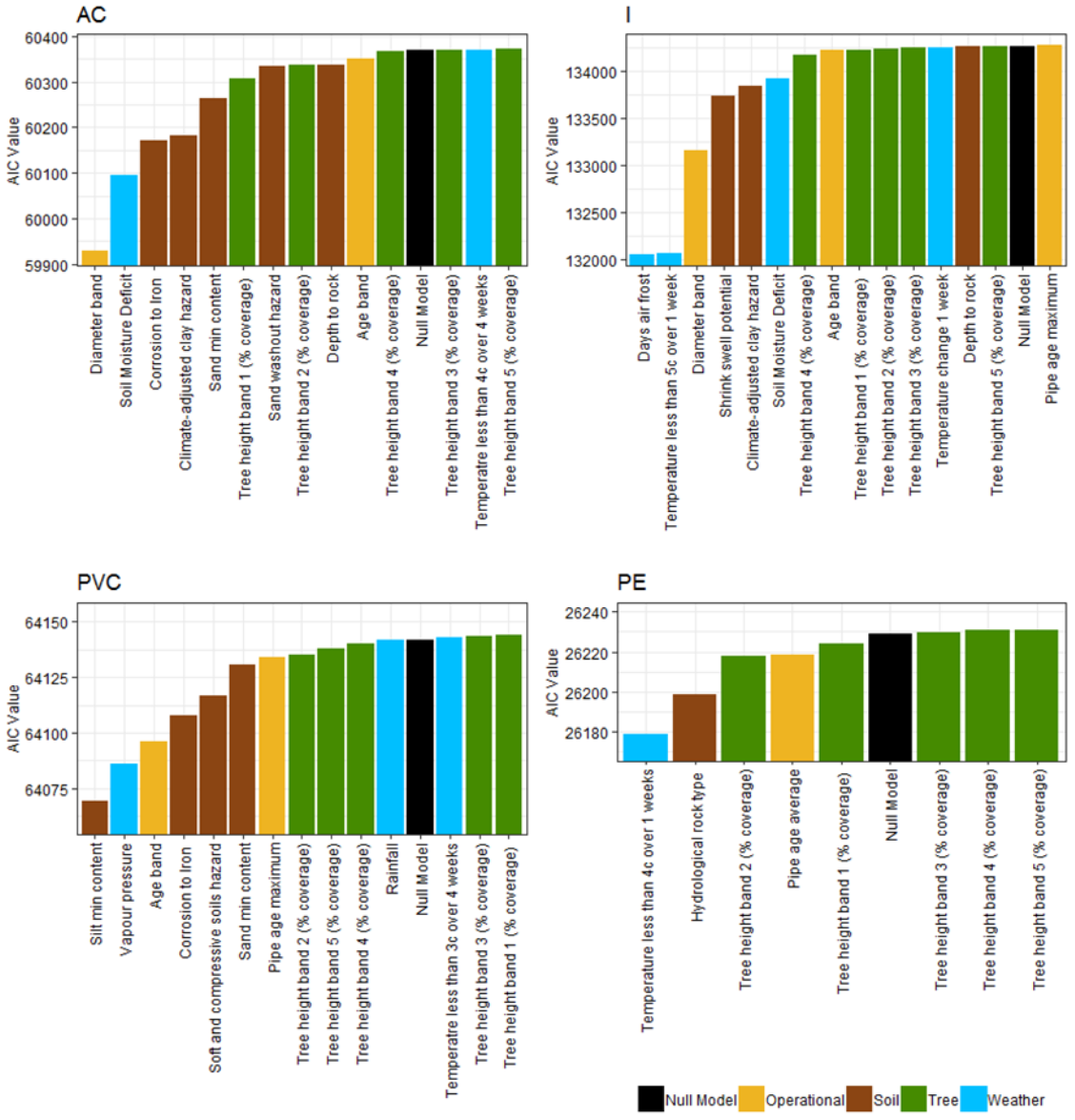

Figure 3-5: Individual AIC analysis comparing the predictive ability of tree height variables in comparison to the operational, weather and soil variables included in the material specific models

In all pipe materials investigated, no variables representing tree height was found to improve the quality of model fit, in comparison to operational, weather and soil variables (Figure 3-5). For AC pipes, variables representing trees of small height (i.e. height band 1 and 2) was identified as being the most predictive, with

variables representing taller trees (band 5) being identified as the least predictive of pipeline failure. For AC pipes, operational variables (Diameter Band), weather variables (SMD) and soil variables (corrosivity, climate-adjusted clay hazard, sand minimum content and sand washout potential) was highlighted being predictive of pipeline failure, see Figure 3-5. Tree height bands 3 and 5, and also Temperature ≤ 4°C over 4 weeks was less predictive than the null model, suggesting that these variables hold no ability to improve model fit, despite their initial selection in Chapter 2.

In Iron pipes, the most influential covariates are the previously selected weather variables (total number of days air frost, temperatures ≤5°C over 1 week and SMD), operational variables (diameter band) and soil variables (shrink swell potential, corrosivity and climate-adjusted clay hazard). All tree variables have a very similar AIC, suggesting minimal contribution to the quality of model fit in different tree height variables.

In PVC pipes, a mixture of soil, weather and operational variables was identified as being more predictive than tree height variables. There is greater difference in AIC value between the best performing tree variable (tree height band 2) and the least predictive tree variable (tree height band 1). A total of 3 variables was identified as being less predictive than the null model, which included 2 tree height variables (bands 3 and 1) and a weather variable (temperature ≤3°C over 4 weeks.

In PE pipes, smaller trees (tree height band 2 and 1) was selected as being the most predictive of pipeline failure, whilst taller trees (height band 3, 4 and 5) was selected as being less predictive than the null model. In total, only 3 other variables are included in the PE pipeline failure model (temperature ≤4°C over 1 week, hydrological rock type and pipe age average), and only tree height bands 1 and 2 variables was selected as being more predictive than the null model.

3.4.4 The impact of combining tree height variables for the prediction of pipeline failure

For all pipe materials there is no consistent selection of a combination of tree height variables which are predictive of pipeline failure, see Figure 3-6. In Iron, AC and PE pipes, the largest reduction in AIC, is observed with the addition of the tree height band 1 variable. In PVC pipes, the largest reduction in AIC is observed with the addition of tree height band 1 and 2 variable.

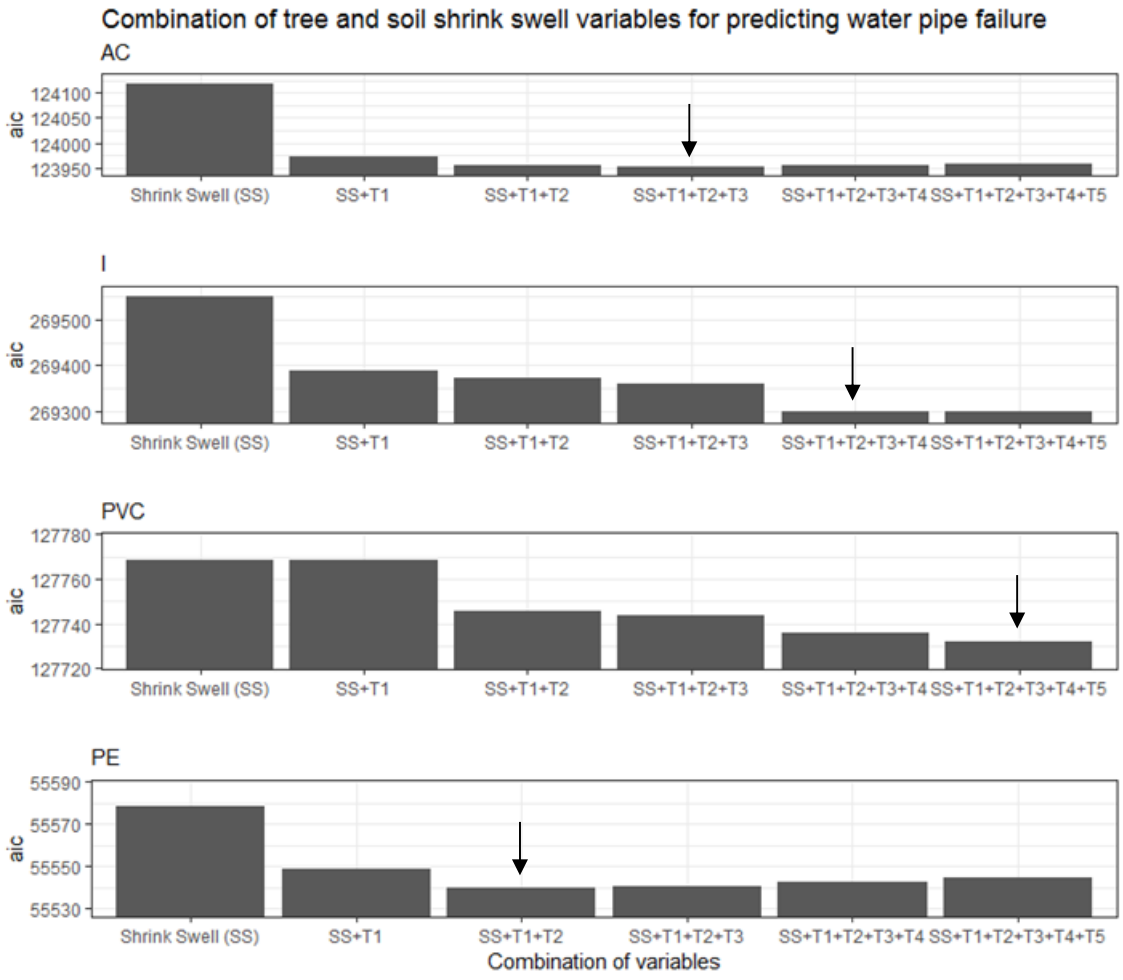

Figure 3-6: AIC analysis for the different combinations of tree height variables and shrink swell in predicting water pipe failure for six pipe materials. T1 to T5 represents tree height bands 1 to 5. Black arrows indicate the combinations of tree height variables with the lowest AIC value

3.4.5 Enhancing current water failure prediction models

Upon the selection of the most predictive combination of tree height variables for each material types (Figure 3-6), the tree height variables were added to the models previously described. A description of the variables included in the tree-enhanced models are shown below in Table 3-9.

Table 3-9: A list of the final model variables for Asbestos Cement (AC), Iron (I), Polyvinylchloride (PVC) and Polyethylene (PE) tree-enhanced drinking water pipe failure models

Material	Model Variables
AC	Diameter band + SMD + corrosion to Iron + climate-adjusted clay hazard + sand min + depth to bedrock + age band + temperature ≤4°C over 4 weeks + shrink swell + tree height 1 + tree height 2 + tree height 3
I	Days air frost + temperature ≤5°C over 1 week + diameter band + shrink swell + climate-adjusted clay hazard + SMD + age band + temperature change over 1 week + depth to bedrock + pipe age maximum + tree height 1 + tree height 2 + tree height 3 + tree height 4
PVC	Silt minimum + vapour pressure + age band + corrosion to Iron + soft and compressible soils hazard + sand minimum + pipe age maximum + rainfall + temperature ≤3°C over 4 weeks + shrink swell + tree height 1 + tree height 2 + tree height 3 + tree height 4 + tree height 5
PE	Temperature ≤4°C over 1 week + hydrological rock type + pipe age average + shrink swell + tree height 1 + tree height 2

The final results of the tree-enhanced model predictions showed no improvement in prediction accuracy using tree-related variables, apart from a very slight improvement in model error (-0.007 %) in the Iron model, see Table 3-10.

Table 3-10: Summary statistics of models with and without the representation of trees. RMSE is the Root Mean Squared Error and MAE is the Mean Absolute Error

Note: 'Mat' is material type referring to either asbestos cement (AC), Iron (I), polyvinylchloride (PVC) or polyethylene (PE)

Model	Mat	Observed (o)	Predicted (p)	Residual difference (p – o)	Model Error % difference (p - o)	RMSE	MAE
No trees	AC	4579	4698.21	119.21	2.201%	51.043	40.334
With trees	AC	4579	4698.57	119.57	2.209%	51.215	40.528
No trees	I	10385	10356.00	-28.97	-0.279%	147.577	112.003
With trees	I	10385	10356.70	-28.34	-0.272%	147.783	112.308
No trees	PVC	4651	4609.93	-41.07	-0.883%	32.038	26.057
With trees	PVC	4651	4609.41	-41.59	-0.894%	32.017	26.061
No trees	PE	1553	1677.87	124.87	8.040%	33.032	29.543
With trees	PE	1553	1678.93	125.93	8.108%	32.388	29.324

For all models, the inclusion of the tree variables accounted for a change in burst prediction which was <1 burst over the 10-year period, suggesting poor model sensitivity to the addition of new covariates. The overall model error is low, with the most accurate model predicting bursts to -0.2 % of the observed bursts over a 10-year period. The least performing model was PE, which over-predicted bursts by 8 % during the 10-year period.

3.5 Discussion

The seasonal differences in pipe failure rates has been widely reported in the scientific literature (Clayton *et al.*, 2010; Gould *et al.*, 2011; Kleiner and Rajani, 2001; Wols and Thienen, 2014). Understanding the material-specific and seasonal trends in failure rates is integral for effective infrastructure management, as it helps utility operators to improve the life cycle costing of pipe materials in a quantitative manner (Davis *et al.*, 2007). Moreover, an improved

understanding into the failure rates of different pipeline materials helps utility operators to proactively manage the distribution network, by prioritising the rehabilitation of the most at-risk assets. By using relevant environmental datasets, this paper has demonstrated how to gain a greater understanding of the environmental impacts on the water distribution network using relevant secondary datasets. This paper has a) quantified summer time failure rates for four common pipe materials under varying soil and tree density conditions, b) investigated whether a series of water infrastructure failure models can be enhanced using tree-related variables which are representative of tree height.

3.5.1 Pipeline failure rate analysis under tree density and soil shrink swell conditions

Investigating tree impacts up to a distance of 40 m of the pipe network is aligned with previous studies (Mercer and Reeves, 2011), where tree-related infrastructure damage has been noted at similar distances owing to the water consumption of different tree species and associated soil movements. Gasson and Cutler (1998) noted the importance of recognising soil type when investigating the impacts of trees on infrastructure, suggesting that trees in a non-shrinkable soil can be situated, with less risk, at closer proximity to infrastructure. When analysing the impacts of trees on a regional network of buried infrastructure, assuming a “worse-case” scenario of the potential maximum extent of tree impact on infrastructure is a suitable approach. This ensures that the potential impacts of trees, irrespective of soil type or tree species, are considered.

The highest failure rates observed in this study was recorded in Iron and PVC pipelines, with a tree density of 30-100 % and a very high shrink swell potential soil. This finding confirms the impact of a stand of trees have on the failure rates of Iron and PVC pipe materials in summer. Typically, the failure rates in Iron, PVC and AC pipe materials increase with higher tree density and soil shrink swell potential. The increase in summer failure rates of Iron and AC has been previously documented and has been attributed in the main to soil shrinkage. PE pipes exhibited no variability in failure rate under all soil and tree density

conditions. A lack of information is available in the scientific literature relating to the failure rate of PE pipe materials. One potential reason for this being the relative infancy of this pipe material in comparison to other materials making it more stable and tolerant to the environmental conditions being discussed (Pietrucha-Urbanik, 2015).

The impact of multiple trees located together on the failure rates of different water pipeline materials has not been previously investigated, despite the known impact of individual trees on differential soil movements (Biddle, 1979; Clayton *et al.*, 2010; Day *et al.*, 2010). The increase in failure rates under very high ground movement potential and high tree density conditions was expected, given the widely reported impact of individual trees causing differential soil movement in clay rich soils. The failure rates of Iron and PVC at 40m reported under very high shrink swell potential and 30-100 % tree density cover was 7.73 and 9.70 bursts per 1,000 km of pipe per week, respectively. By comparison, the failure rates reported in the same tree density conditions (30-100 % at a 40 m distance) in a low shrink swell soil was just 1.19 and 2.72 bursts per 1,000 km per week for Iron and PVC pipes respectively. No relationships were found for AC and PE pipes under these conditions. However, there was a small sample size of pipe recorded within very high shrink swell soil and high tree density (30-100 %) which did not allow the generation of a failure rate for these pipe materials.

3.5.2 AIC analysis of tree height and combination of tree height variable

This study reveals that pipe materials are impacted differently by different tree heights, with a different combination of tree variables selected as being predictive for each pipe material. Only in PVC pipe material a combination of all tree heights was identified as being predictive (tree height bands 1 to 5), whilst in PE only small trees (tree height bands 1 and 2) was identified as predictive. Trees of low to medium height was selected as being most predictive of failure in AC and Iron pipes.

A potential reason for the inconsistent selection of tree height variables being predictive of pipeline failure are the numerous other factors which contribute to the failure of buried assets. The AIC analysis verified that operational, weather and soil factors are more predictive than tree variables in all material types investigated. In addition, above-ground indicators such as tree height are not always descriptive of below-ground properties of tree root architecture (Burgess *et al.*, 2001). For example, asymmetric root growth, root depth, and tree root intrusion are all known factors impacting upon infrastructure but have not been represented in the National Tree Map and cannot be described by using tree height data alone. An inclusion of information which represents the below-ground properties of trees, such as rooting information, rooting depth, and rooting extent, has the potential to further develop the models described within, and help better represent trees in the models described.

3.5.3 Tree-enhanced water infrastructure failure models

The inclusion of additional tree-related variables led to the increase of model error for AC, PVC and PE pipe failure models, with only Iron pipes achieving a slight reduction in model error (Table 3-10). One potential reason for the increase in error is the inclusion of more variables leading to overfitting of the GLM and increased complexity within the model. Furthermore, the AIC analysis shown in Figure 3-5 highlights that the most dominant and predictive variables are not tree related, and consist of a range of different soil, weather and operational variables. For Iron and AC pipes in particular, weather and soil variables such as SMD, temperature and soil texture are highly predictive of asset failure and was already represented in the material-specific models. Therefore, the lack of sensitivity of the models to the addition of new variables, such as tree height and density, might be explained by the most dominant variables being already included, and the addition of less predictive variables not improving overall model accuracy.

This study used pipe cohorts for the investigation of tree density, where pipes of the same material, age and diameter were grouped together. Tree density was calculated for each individual pipe cohort and represented as a percentage of canopy coverage. The mean average length of the pipe cohorts analysed is 0.663

km long, however, the standard deviation of all pipe cohort lengths is large at 1.139 km. An alternative approach for modelling tree impacts and density to the distribution network would be to split the network into regular divisions of pipe length (i.e. 500 m), then investigating the impact of tree density of different tree heights on regular lengths of pipe. It is expected that in using regular lengths of pipe, a clearer and more comparable representation of tree density and height would be gained across the network. A potential problem with this approach would be the representation of soil heterogeneity, as soil varies at an irregular scale. Therefore, setting a maximum distance of 500 m, which includes smaller lengths of pipe cohort, would overcome this. However, the creation of smaller pipe cohorts may increase the computational resource needed for analysis, particularly when investigating the impact of trees across a regional network of buried infrastructure.

The tree inventory used within this investigation has no representation of tree species, which is a known major contributing factor to ground movement in shrinkable soils (Mercer and Reeves, 2011; Östberg, 2013; Randrup *et al.*, 2001). To this extent, a classification of tree species using high resolution satellite or airborne multispectral imagery may provide more detailed information for predicting the impacts of tree-related ground movement and provide means of generating a key variable for statistical modelling (Ruiliang, 2013). Other studies have also described the impact of the removal of trees on infrastructure damage (Rawlins *et al.*, 2015). In using a static tree inventory, no representation of the potential removal of trees is included. Other studies have described change detection methods to identify the removal of trees and large vegetation using a time series of either satellite or aerial images (Walton *et al.*, 2008). The development and integration of such information within the statistical models may increase the representation of trees in the included modelling covariates and help further improve current methods of prediction.

The methods developed within this study can be used anywhere detailed soil, weather and tree inventories exist. The value of the work discussed is particularly relevant to urban geographical locations which contain highly shrinkable soils and

vegetation co-located against buried infrastructure. Soils with high ground movement potential are found throughout the world (Mokhtari and Dehghani, 2012). However, ground movement potential is only realised in locations where there is a sufficient SMD to permit the soils to shrink and swell in accordance with fluxes in the available soil water content. Typically, these geographical locations are in temperate climate countries with distinct wet and dry seasons, or also in semi-arid or arid environments where shrinkable soils can respond to short period of rainfall (Jones and Jefferson, 2012). This research has highlighted that high densities of trees situated in shrinkable soils can lead to increased failure rates in some pipe materials. Such information is important for the successful pro-active management and planned rehabilitation of buried assets.

3.6 Conclusions

Unique relationships between tree densities, soil shrink swell potential and the rates of failure in four common pipe materials have been evaluated across an entire water distribution network. Clear increases in failure rates was observed for Iron and PVC pipe materials which are buried in highly shrinkable soils with 30-100 % of tree canopy coverage in the pipe cohort (Figure 3-4). AC pipes show a varied response in failure rates, typically increasing with soil shrink swell potential and increased tree density. Modern PE pipes showed a stable rate of failure under all soil shrink swell conditions. Analysis of tree density up to a distance of 40 m from the pipe network provided an increased representation of pipeline failure in Iron and PVC pipe materials and is a recommendation for future studies wishing to undertake a similar analysis.

Despite the tree variables having a clear impact upon the rates of failure in Iron and PVC pipes, the use of tree height and density variables within a series of material specific predictive models did not improve model accuracy. However, future research could include adding a representation of tree species as a model variable and identifying recent felling or planting of trees which are co-located to the network. Furthermore, the use of a regular length of pipe cohort to measure tree canopy coverage has the potential to increase the representation of tree density for individual sections of pipeline. Another recommendation of further

research would be to undertake statistical analysis of the residuals between predicted and observed bursts to determine whether any cyclical patterns exist in the model predictions. This may help to further identify any repetitive weaknesses exist in the model, such as poor prediction accuracy in a given season, or month.

The findings have contributed to the wider general understanding of the impact of trees on four common pipe materials (AC, Iron, PVC and PE), across an entire distribution network. These methods have the potential to be applied to other locations where the data permits. Such methods are critical for the planning and pro-active management of water infrastructure, particularly given increasing pressure on global water resources and expected increase in water demand.

Despite the known impacts of trees on utility networks, as identified in this paper, it is important to prevent tree loss due to the numerous positive impact's trees have on the environment and society. Several methods exist for the successful co-location of trees and buried assets. For example, the artificial control of soil moisture regime in high risk soils, planting of deep-rooting tree species, planting of non-water intensive tree species and also installation of physical barriers have all been found to mitigate against the impact of trees on infrastructure (Randrup *et al.*, 2001). These management techniques, combined with the findings presented in this paper, provide water utility companies the opportunity to pro-actively manage the most at-risk assets based on unique soil and tree conditions.

Acknowledgements

This work was supported by the UK Natural Environment Research Council [NERC Ref: NE/M009009/1]. All authors made significant contribution to this work. M.N and T.F designed the research methodology, M.N conducted the research and all authors contributed to the writing of the manuscript. We acknowledge the advice of Anglian Water plc. staff, and the provision of the asset data described here.

3.7 Bibliography

Aho, K., Derryberry, D., Peterson, T., 2014. Model selection for ecologists : the worldviews of AIC and BIC. Ecology 95, 631–636. https://doi.org/10.1890/13-1452.1

Akaike, H., 1974. A New Look at the Statistical Model Identification. IEEE Trans. Automat. Contr. 19.

Biddle, P.G., 1979. 2: Tree Root Damage to Buildings - An Arboriculturist's Experience. Arboric. J. 3, 397–412.

Bluesky, 2018. National Tree Map Dataset Description [WWW Document]. URL https://www.bluesky-world.com/ntm (accessed 2.8.18).

Boorman, D., Hollis, J., Lilly, A., 1995. Hydrology of soil types: a hydrologically based classification of soils in the United Kingdom. Wallingford, UK.

Burgess, S.S.O., Adams, M.A., Turner, N.C., White, D.A., Ong, C.K., 2001. Tree roots: Conduits for deep recharge of soil water. Oecologia 126, 158–165. https://doi.org/10.1007/s004420000501

Chiesura, A., 2004. The role of urban parks for the sustainable city. Landsc. Urban Plan. 68, 129–138. https://doi.org/10.1016/j.landurbplan.2003.08.003

Clayton, C.R.I., Xu, M., Whiter, J.T., Ham, A., Rust, M., 2010. Stresses in cast-iron pipes due to seasonal shrink-swell of clay soils. Proc. Inst. Civ. Eng. - Water Manag. 163, 157–162. https://doi.org/10.1680/wama.2010.163.3.157

Coxe, S., West, S.G., Aiken, L.S., 2009. The analysis of count data: A gentle introduction to poisson regression and its alternatives. J. Pers. Assess. 91, 121–136. https://doi.org/10.1080/00223890802634175

Davis, P.Ã., Burn, S., Moglia, M., Gould, S., 2007. A physical probabilistic model to predict failure rates in buried PVC pipelines. Reliab. Eng. Syst. Saf. 92, 1258–1266. https://doi.org/10.1016/j.ress.2006.08.001

Day, S.D., Wiseman, P.E., Dickinson, S.B., Harris, J.R., 2010. Tree Root Ecology in the Urban Environment and Implications for a Sustainable Rhizosphere. Arboric. Urban For. 36, 193–205.

Defra, 2011. Water for Life. https://doi.org/10.1007/s13398-014-0173-7.2

Gasson, P.E., Cutler, D.F., 1998. Can we live with trees in our towns and cities? Arboric. J. 22, 1–9. https://doi.org/10.1080/03071375.1998.9747188

Gould, S.J.F., Boulaire, F.A., Burn, S., Zhao, X.L., Kodikara, J.K., 2011. Seasonal factors influencing the failure of buried water reticulation pipes. Water Sci. Technol. 63, 2692–2699. https://doi.org/10.2166/wst.2011.507

Guo, L., 2017. Field Monitoring and Numerical Analysis of the Influence of Trees on Soil Moisture and Ground Movement in an Urban Environment. RMIT University.

Hallett, S.H., Sakrabani, R., Keay, C.A., Hannan, J.A., 2017. Developments in land information systems : examples demonstrating land resource management capabilities and options. Soil Use Manag. 33, 514–529. https://doi.org/10.1111/sum.12380

Jones, G.M., Cassidy, N.J., Thomas, P.A., Plante, S., Pringle, J.K., 2009. Imaging and monitoring tree-induced subsidence using electrical resistivity imaging. Near Surf. Geophys. 7, 191–206. https://doi.org/10.3997/1873-0604.2009017

Jones, L.D., Jefferson, I., 2012. Chapter 33 Expansive Soils, in: ICE Manual of Geotechnical Engineering, Volume 1 - Geotechnical Engineering Principles, Problematic Soils and Site Investigation. pp. 413–441. https://doi.org/10.680/moge.57074.0463

Kleiner, Y., Rajani, B., 2001. Comprehensive review of structural deterioration of water mains : statistical models. Urban Water 3, 131–150. https://doi.org/10.1016/S1462-0758(01)00033-4

Mcpherson, E.G., Peper, P.P., 1996. Costs of street tree damage to infrastructure. Arboric. J. 20, 143–160.

Mercer, G., Reeves, A., 2011. The Relationship between Trees , Distance to Buildings and Subsidence Events on Shrinkable Clay Soil. Arboric. J. 33, 229–245.

Met Office, 2018a. UK regional climate summaries 1981 - 2018 [WWW Document]. URL https://www.metoffice.gov.uk/public/weather/climate (accessed 9.5.18).

Met Office, 2018b. Met Office Rainfall Evapotranspiration Calculation System - Specialist Datasets [WWW Document]. URL https://www.metoffice.gov.uk/services/industry/data/specialist-datasets (accessed 21.5.18).

Mokhtari, M., Dehghani, M., 2012. Swell-Shrink Behavior of Expansive Soils, Damage and Control. Electron. J. Geotech. Eng. i, 2673–2682.

Navarro, V., Candel, M., Yustres, A., Sanchez, J., Alonso, J., 2009. Trees, soil moisture and foundation movements. Comput. Geotech. 36, 810–818. https://doi.org/10.1016/j.compgeo.2009.01.008

Nelder, A.J.A., Wedderburn, R.W.M., 1972. Generalized Linear Models. J. R. Stat. Soc. A 135, 370–384.

Nowak, D.J., Crane, D.E., 2002. Carbon storage and sequestration by urban trees in the USA. Environ. Pollut. 116, 381–389. https://doi.org/10.1016/S0269-7491(01)00214-7

Nowak, D.J., Crane, D.E., Stevens, J.C., 2006. Air pollution removal by urban trees and shrubs in the United States. Urban For. Urban Green. 4, 115–123. https://doi.org/10.1016/j.ufug.2006.01.007

OFWAT, 2018. Water loss per day UK - Discover Water [WWW Document]. URL https://discoverwater.co.uk/leaking-pipes (accessed 8.8.18).

Östberg, J., 2013. Tree Inventories in the Urban Environment. Methodological Development and New Applications. Landsc. Planning, Hortic. Agric. Sci. Swedish University of Agricultural Sciences.

Pietrucha-Urbanik, K., 2015. Failure analysis and assessment on the exemplary water supply network. Eng. Fail. Anal. 57, 137–142. https://doi.org/10.1016/j.engfailanal.2015.07.036

Pritchard, O.G., Hallett, S.H., Farewell, T.S., 2015a. Probabilistic soil moisture projections to assess Great Britain's future clay-related subsidence hazard. Clim. Change 133, 635–650. https://doi.org/10.1007/s10584-015-1486-z

Pritchard, O.G., Hallett, S.H., Farewell, T.S., 2015b. Soil geohazard mapping for improved asset management of UK local roads. Nat. Hazards Earth Syst. Sci. 15, 2079–2090. https://doi.org/10.5194/nhess-15-2079-2015

R Core Team, 2015. R: A language and environment for statistical computing [WWW Document]. URL https://www.r-project.org (accessed 21.5.18).

Randrup, T.B., McPherson, E.G., Costello, L.R., 2001. A review of tree root conflicts with sidewalks , curbs , and roads. Urban Ecosyst. 5, 209–225. https://doi.org/10.1023/a:1024046004731

Rawlins, B.G., Harris, J., Price, S., Bartlett, M., 2015. A review of climate change impacts on urban soil functions with examples and policy insights from England, UK. Soil Use Manag. 31, 46–61. https://doi.org/10.1111/sum.12079

Ruiliang, P., 2013. Tree Species Classification. Remote Sens. Nat. Resour. 239–258. https://doi.org/doi:10.1201/b15159-19

Torres, M.N., Rodríguez, J.P., Leitão, J.P., 2017. Geostatistical analysis to identify characteristics involved in sewer pipes and urban tree interactions. Urban For. Urban Green. 25, 36–42. https://doi.org/10.1016/j.ufug.2017.04.013

Walton, J.T., Nowak, D.J., Greenfield, E.J., 2008. Assessing Urban Forest Canopy Cover Using Airborne or Satellite Imagery. Arboric. Urban For. 34, 334–340.

Watson, G.W., Hewitt, A.M., Custic, M., Lo, M., 2014. The management of tree root systems in urban and suburban settings: A review of soil influence on root growth. Arboric. Urban For. 40 (4). 193 – 217.

Wols, B.A., Thienen, P. Van, 2014. Impact of weather conditions on pipe failure : a statistical analysis. J. Water Supply Res. Technol. - AQUA 212–223. https://doi.org/10.2166/aqua.2013.088

Yamijala, S., Guikema, S.D., Brumbelow, K., 2009. Statistical models for the analysis of water distribution system pipe break data. Reliab. Eng. Syst. Saf. 94, 282–293. https://doi.org/10.1016/j.ress.2008.03.011

Yannas, S., 2001. Toward more sustainable cities. Sol. Energy 70, 281–294. https://doi.org/10.1016/S0038-092X(00)00091-8

Zhang, Z., 2016. Variable selection with stepwise and best subset approaches. Ann. Transl Med. 4, 136. https://doi.org/10.21037/atm.2016.03.35

4 Seasonal model testing and training to improve the prediction accuracy of Iron and asbestos cement water pipeline failure models

This chapter investigates Objective 4 and is presented in the form of one unpublished research paper, intended for the journal Proceedings of the Institution of Civil Engineers – Water Management.

North, M., Farewell, T., Hallett, S. (2018) Seasonal modelling to improve the prediction accuracy of Iron and asbestos cement water pipeline failure models. (unpublished)

This chapter investigates the continued development of the models described in Chapter 3. The impact of using separate seasonal and non-seasonal model training and testing datasets is evaluated within. For this investigation, focus has been made only to the development of Iron and asbestos cement pipes due to their highly seasonal pattern of pipeline failure.

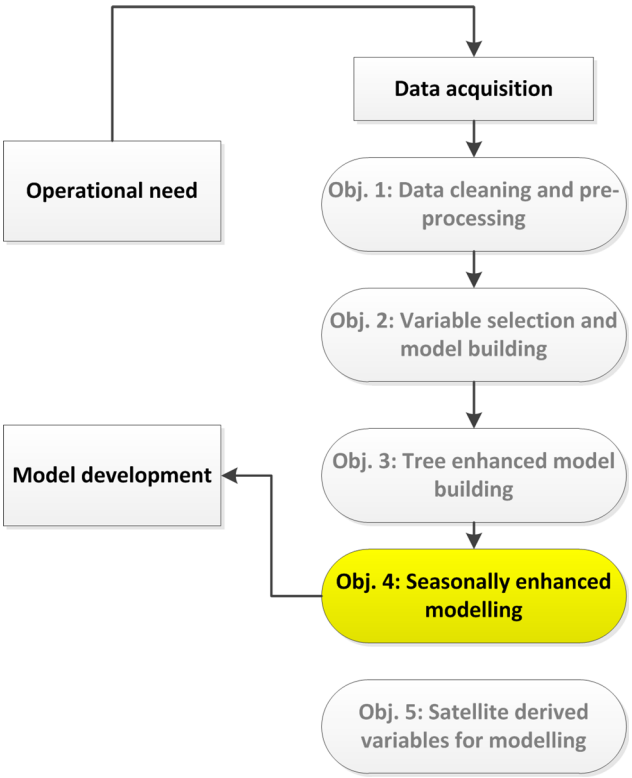

Figure 4-a: Objectives aimed to be investigated within this chapter in context of the overall thesis

Seasonal modelling to improve the prediction accuracy of Iron and asbestos cement water pipeline failure models

M. North[a], T. Farewell[a], S. Hallett[a]

[a] *School of Water, Energy and Environment, Cranfield University, Bedfordshire, MK43 0AL, United Kingdom*

Corresponding Author: t.s.farewell@cranfield.ac.uk; tel.: +44 (0)1234 752978

Abstract:

The failure rates of Iron and Asbestos Cement (AC) pipelines are strongly controlled by environmental conditions typical of the different seasons. This study evaluates the impact of seasonally training two material-specific Poisson regression models using data representative of either the summer or winter season only. It is expected that an increased representation of using only seasonal data, representative of either the summer or winter period only, has the potential to improve current methods of pipeline failure prediction. The model covariates include detailed weather, soils and tree variables, and an 11-year dataset of historical pipe failure across an entire water distribution company in the UK. Using identical data, direct comparisons between seasonally trained and non-seasonally trained models was made using a series of statistical measures. An improvement in the prediction of Iron pipes was obtained when trained on separate seasonal data, with a reduction of annual average model error (% difference between observed and predicted burst values) of -0.36%. There was no improvement in the predictive ability of AC pipes when trained on seasonal data, with a reduction of annual model error of -0.08%, and an increase in RMSE and MAE.

Keywords: *Seasonal, Soil, Weather, Trees, Poisson regression, Model training, Model testing*

4.1 Introduction

In temperate climate countries, the failure rates of Iron and Asbestos Cement (AC) pipes are strongly influenced by the weather conditions typical of the different seasons (Clayton *et al.*, 2010; Davis *et al.*, 2008; Gould *et al.*, 2011). For the case of the Anglian Water distribution network (UK), a clear increase of Iron pipeline failure rates occurs during the winter months, whilst a clear increase in AC pipeline failures occur during the summer months (Figure 4-1). This seasonal pattern of failure lends itself well to statistical modelling of predicting infrastructure failure, and has been the subject of numerous previous investigations (Farmani *et al.*, 2017; Kimutai *et al.*, 2015; Rostum, 2000; Wilson *et al.*, 2017; Yamijala *et al.*, 2009; Chapter 2; Chapter 3).

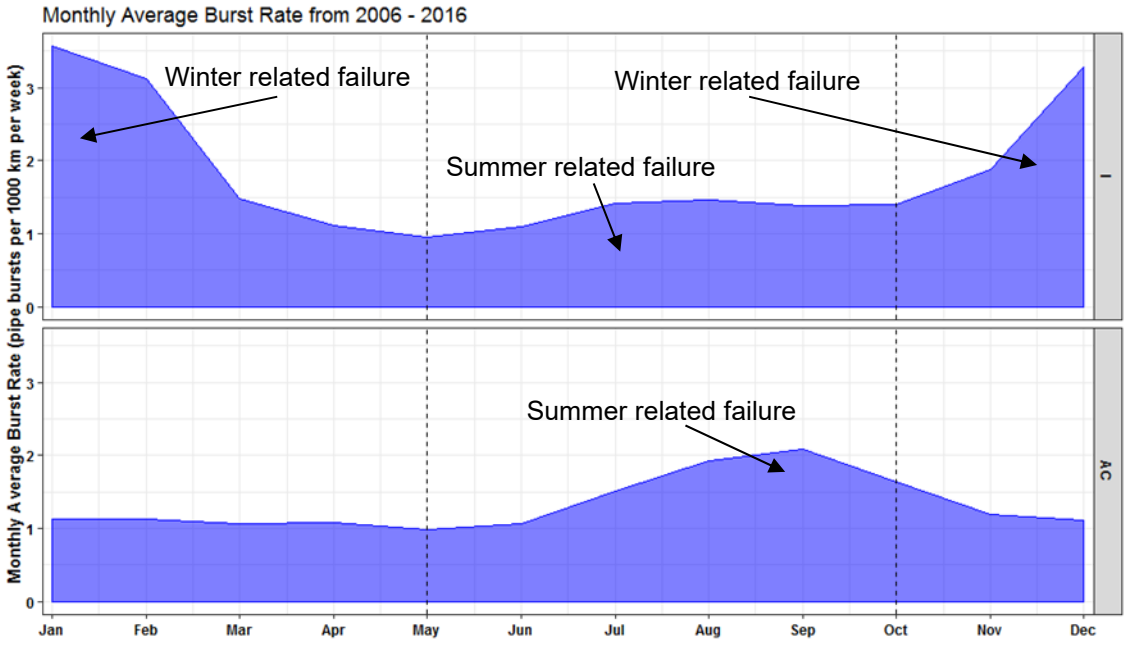

Figure 4-1: Monthly average burst rate (pipe bursts per 1,000 km of pipe per week) for Iron (I) and Asbestos Cement (AC) pipes for the Anglian Water distribution network, UK. Dashed lines represent the marked boundaries of summer and winter seasons used in this study. Tick marks on the x axis indicate the middle of the month

The development of statistical methods for pipeline failure prediction has been of key research interest since the 1980's, as statistical approaches allow water utility operators to understand the stresses to the network in a cost-effective and non-

disruptive manner. Statistical modelling allows utility companies to summarise a wide range of operational and environmental pressures upon a network, without the need for expensive *in situ* observations.

4.1.1 Seasonal influences of pipe failure

Previous investigations have revealed a wide range of environmental and operational causes of failures within Iron and AC pipes (Clayton *et al.*, 2010; Davis *et al.*, 2008; Chapter 2; Chapter 3). Iron pipes are typically the oldest pipe materials used in water distribution networks in the UK, with some operational pipes dating back to the late 1800's. Iron pipeline failure is induced by cold weather, where pipes fail due to the successive embrittlement and long-term weakening of the pipeline material, often propagating from pipe manufacturing or installation defects (Farrow *et al.*, 2017). Due to the cold weather mechanism of failure (such as successive embrittlement and thermal forces between the internal and external temperatures within the pipe), weather variables such as accumulated low temperatures, number of day's air frost and temperature changes have been previously found to be highly effective for modelling Iron pipeline failure (Chapter 2).

In the case of AC pipes, the environmental factors leading to pipeline failure are typical of the summer season within the UK (Wols and van Thienen, 2014). AC pipes were widely installed between 1920 and 1940, many of which are still currently operational. Several studies have noted the rise of AC pipe failures during dry summer months. This increase in failure has largely been attributed to the increased bending stresses of the pipe, caused by differential soil movement (in shrinkable clay-rich soils), and has led to an increased frequency of observed circumferential failures (Farrow *et al.*, 2017). To this extent, weather variables such as Soil Moisture Deficit (SMD) and soil variables such as the indication of clay soils or high shrink swell potential have been found to be effective for the prediction of AC failures (Chapter 2).

Iron pipes also have a small rise in failures from June to September (Figure 4-1). The increase of failures in Iron pipes during summer can also be attributed to differential soil movements, in a similar mechanism to the common pipeline

failures in AC, and has been previously investigated (Clayton *et al.*, 2010). To this extent, modelling variables such as SMD, trees, and soil shrink swell potential are also effective for the modelling of Iron pipeline failures, often combined with the cold weather variables indicative of winter pipeline failure (Chapter 2).

4.1.2 Current status of prediction accuracy

A series of material-specific water infrastructure failure models have been developed for the Anglian Water distribution network, which covers a large region of approximately 27,500 km^{-2} in the east of the UK (Chapter 3). Material specific variable selection has identified a series of key operational, soil, weather and tree variables which are predictive of failure for individual material types, a description of the models and included variables are provided in Section 4.2.2. For a full description of model development, variable selection and datasets used refer to (Chapter 2; Chapter 3).

For the models developed in Chapter 3, the mean yearly model error (% difference between observed and predicted bursts) for both Iron and AC models is -1.62% and 2.44%, respectively. Figure 4-2 highlights the consistent underestimation of bursts during the winter season (as indicated by the blue circle) in Iron pipes, and summer bursts (as indicated by the orange circle) in AC pipes. For both Iron and AC, model performance during the “off-peak” season of bursts appears to be satisfactory, with model error increasing with the number of observed bursts.

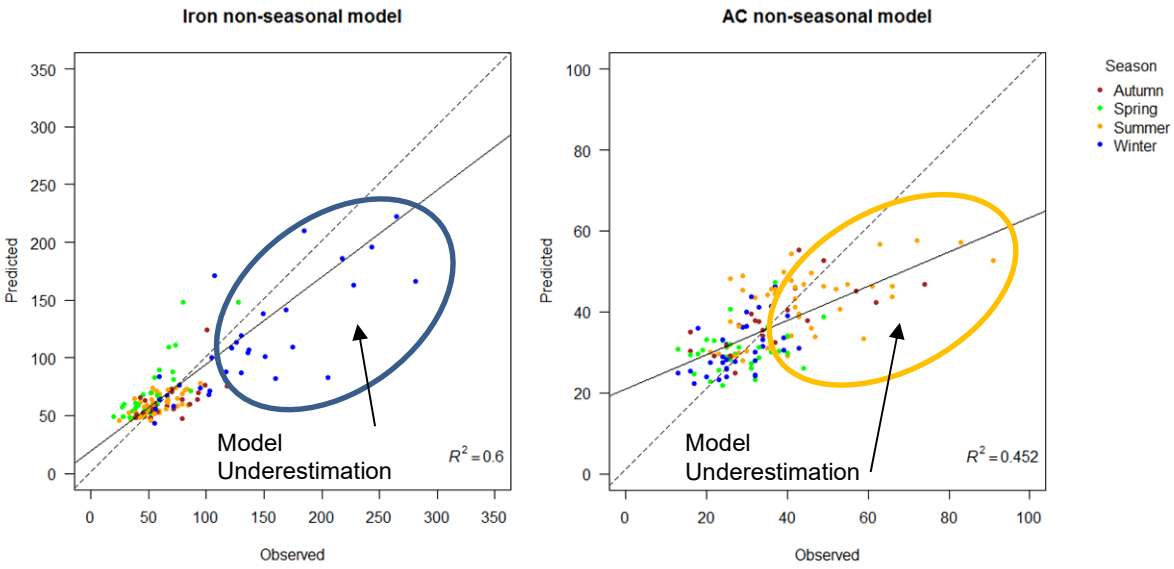

Figure 4-2: Scatter plot showing the seasonal variations in prediction accuracy for Iron (I) and Asbestos Cement (AC). Dashed line is a 1:1 linear line, whilst the black line notes the line of best fit from the linear model. Data points represent model predictions per month

4.1.3 Research aim and motivation

This study aims to establish whether using separate “winter” and “summer” training and testing datasets can improve the prediction accuracy for Iron and AC pipeline failure models. It is expected, that with an increased representation of the distinct seasonal characteristics in specific training datasets, the Poisson regression model may achieve improved prediction accuracy. This study aims to achieve a further understanding in the ability of Poisson regression models to represent the seasonal trends in pipeline failure, establishing whether the use of separate summer and winter only training and testing datasets can improve the performance of the models.

4.2 Materials and Methods

4.2.1 Anglian Water datasets description

For this investigation, a historical dataset of pipe bursts for the period from 28th December 2005 to 27th December 2016 was provided for model testing and training by Anglian Water plc.

Only Iron and AC pipe materials were used due to the highly seasonal pattern of failure rates and notable underestimation of predicted bursts achieved in previous investigations, Figure 4-2 (Chapter 3). Iron and AC pipes have a combined length of over 18,226 km across the distribution network, which is 49% of the total length of the Anglian Water distribution network. A total of 30,019 observed bursts was recorded in Iron and AC pipes, which is 70% of the total observed bursts during the observation period. A summary of pipe length, total number of bursts and associated seasonal burst rates (bursts per 1,000 km of pipe per week) in both summer and winter seasons is provided in Table 4-1. Summer has been defined from the 1st May – 1st October, and winter has been defined as the 1st October – 30st April, to capture both the winter and summer peak failures in Iron and AC pipes respectively (Chapter 3).

Table 4-1: Summary of the failure rates of Iron and AC pipe materials during the observation period. Burst rates (bursts per 1,000 km of pipe per week) are calculated for reactive bursts only

Note: summer failure rates are calculated from the 1st May to the 30th September, and winter failure rates are calculated from the 1st October to the 30th April

Material	Typical installation range	Total length (km)	Total bursts	Burst rate	Total bursts (May-Sep)	Burst rate (May-Sep)	Total bursts (Oct-Apr)	Burst rate (Oct-Apr)
I	1881 to 1921	11,226	20,737	3.21	3,172	1.17	17,565	4.68
AC	1920 to 1941	7,000	9,282	2.31	4,809	2.86	4,473	1.91
Total:		18,226	30,019					

Separate datasets were created for Iron and AC pipe materials from a cleaned and quality checked dataset. Details of the data pre-processing and data quality checks are detailed within (Chapter 2). Measures were made to ensure that only pipes which are operational at the time of the burst was included in the analysis, and that decommissioned and out of service pipes were excluded. This ensured the reliable estimation of pipe network length.

4.2.2 Environmental dataset description

Detailed information relating to the prevailing characteristics of soil, weather and trees was used as model variables in this study. Soils data was taken from the National Soils Map of England and Wales and associated soil property datasets (Hallett *et al.*, 2017). The Met Office Rainfall and Evaporation Calculation System (MORECS) dataset, with a spatial resolution of 5 km, was used for the generation of weather variables for this investigation (Met Office, 2018b). Detailed information regarding the density of trees aligning the distribution network, in a 40 m distance, was also calculated from Bluesky's "National Tree Map", which represents the heights and locations of every tree >3 m in the UK (Bluesky, 2018). For brevity, a full description of the datasets and the formation of specific variables included within the models has not been provided within but is extensively described in Chapters 2 and 3.

4.2.3 Model description

Poisson regression, which is a form of Generalised Linear Model (GLM), works by linking a response variable to a series of explanatory variables. For this study, the number of reactive bursts is the response variable, and the explanatory variables are a series of environmental variables describing the soil, weather and tree conditions. A reactive burst is defined as a burst which has been discovered by a third party, and not one by the leakage detection teams. Only reactive bursts have been used as the response variable, as these are the bursts which have been reported by the public and pose the largest threat of fines by the regulatory body, Ofwat. Due to the scheduled maintenance and reporting of proactive bursts within the distribution network, the dates recorded for timing of proactive burst often does not correlate to the actual time of burst occurrence, leading to

difficulties when trying to predict proactive bursts, hence their exclusion from this present study. Reactive bursts are often reported in a timely manner, therefore can be linked more closely to the environmental conditions at the time of pipeline failure.

Explanatory weather variables are dynamic and vary over time. Poisson regression links the impact of both dynamic (such as weather) and static variables (such as soils, trees and operational factors) to the response variable, which in this case is reactive bursts. In this study, a series of material-specific Poisson regression models which are described and developed in (Chapter 3) have been used. The variables included in the material-specific models are shown in Table 4-2. For brevity, a full description of the model variables is not provided within but can be found in (Chapter 3).

Table 4-2: Model variables included in the final model selection. Tree height band measurement are 1 (< 5 m), band 2 (5 – 10 m), band 3 (10 – 20 m), band 4 (20 – 30 m)

Variable type	Iron	Asbestos Cement (AC)
Operational	Diameter Band	Diameter Band
	Age Band	Age Band
	Maximum Pipe Age	
Soils	Climate-adjusted clay hazard	Climate-adjusted clay hazard
	Depth to bedrock	Depth to bedrock
	Shrink Swell Potential	Corrosivity to Iron
		Minimum Sand Content
		Maximum Sand Content
Weather	Number of days of air frost	Weekly mean SMD (real land use) in the week of the reported burst date
	The absolute change in mean SMD from the week the burst was reported to 1, 2 or 4-weeks before	The accumulated air temperature 4°C in a 4- week period preceding the reported burst date. Air temperature accumulations are calculated in 1°c increments below the

		threshold value. The greater the value of this variable, the colder the temperature was, for a prolonged period of times, in the previous weeks to the burst date
	The accumulated air temperature beneath 5°C in a 1-week period preceding the reported burst date. Air temperature accumulations are calculated in 1°c increments below the threshold value. The greater the value of this variable, the colder the temperature was, for a prolonged period of time, in the previous week to the burst date	
	Weekly mean SMD (real land use) in the week of the reported burst date	
Trees	% area coverage of trees classified as height band 1 up to a 40 m distance from the pipeline	% area coverage of trees classified as height band 1 up to a 40 m distance from the pipeline
	% area coverage of trees classified as height band 2 up to a 40 m distance from the pipeline	% area coverage of trees classified as height band 2 up to a 40 m distance from the pipeline
	% area coverage of trees classified as height band 3 up to a 40 m distance from the pipeline	% area coverage of trees classified as height band 3 up to a 40 m distance from the pipeline
	% area coverage of trees classified as height band 4 up to a 40 m distance from the pipeline	

4.2.4 Model performance evaluation

Identical datasets and material specific model variables was used for the evaluation of seasonally and non-seasonal model training and testing to ensure that results could be directly compared. To create the seasonal model, the training and testing data was further split, in accordance to the observations date based on a summer (1st May to 30th September) or winter (1st October to 30th April) split. This resulted in 4 datasets, summer train and test datasets, and winter

train and test datasets. For the evaluation of the non-seasonal model, a total of 2 datasets (train and test) was used which included burst observations throughout all seasons. Upon model testing of the seasonal models, the datasets were re-joined to ensure that a directly comparable dataset was re-created. For a full overview of the data preparation and sampling of the data, see Figure 4-3.

Model performance was evaluated using the statistical measures Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), correlation coefficient of determination (R^2), and also the percentage difference between observed and predicted values. These statistical metrics are common for the evaluation of predicted versus observed values and have been widely used in the scientific literature (Farmani *et al.*, 2017; Kabir *et al.*, 2016; Kimutai *et al.*, 2015; Yamijala *et al.*, 2009).

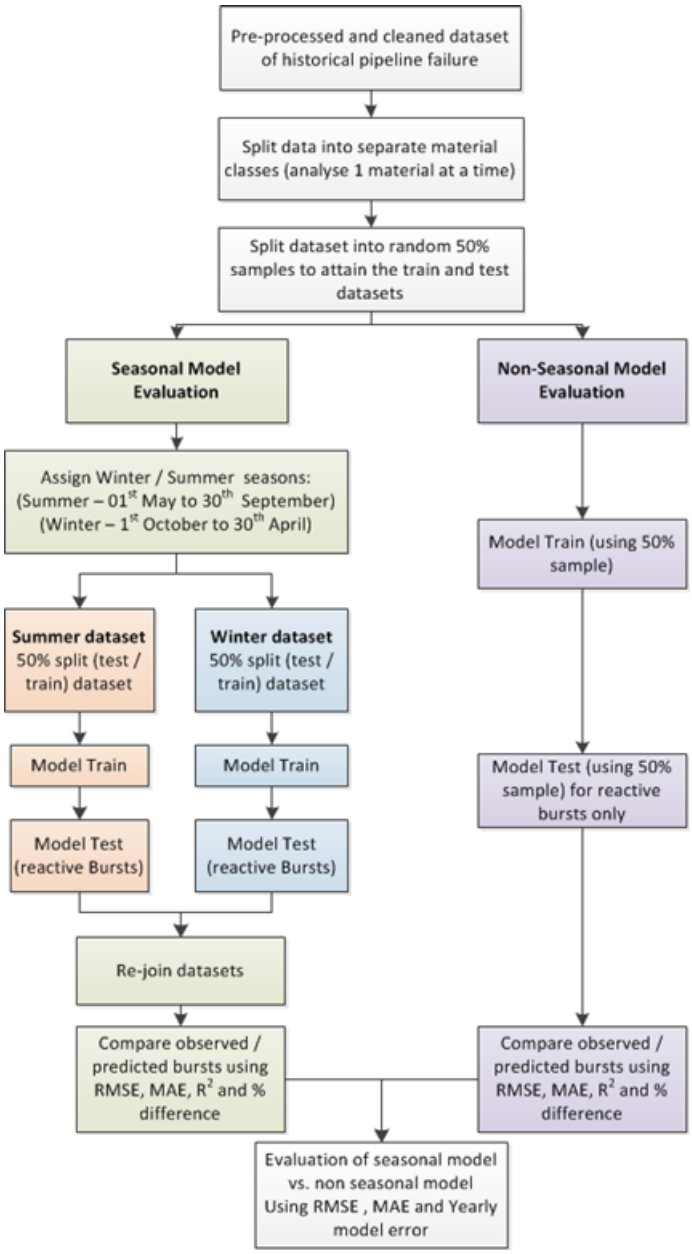

Figure 4-3: An overview of the methodology followed in the seasonal vs. non seasonal model predictions

Note: RMSE is the Root Mean Squared Error, MAE is the Mean Absolute Error, R^2 is the correlation coefficient of determination between observed and predicted values, and model error is the percentage difference between observed and predicted bursts

4.3 Results

Only the Iron model achieved a slight improvement in performance when it was trained and tested on a separate winter and summer dataset. This improvement

in performance is evidenced by a reduction of RMSE and MAE across the observation period (Table 4-3).

When trained on the seasonal data, AC increased the 11-year RMSE by 0.65 and MAE by 0.05 from the non-seasonally trained data, therefore suggesting that there was no value in seasonally training this model (Table 4-4). The original models, which were not trained on the separate seasonal data, obtained an average yearly model error average of -1.62% and 2.44% for Iron and AC, respectively. For both materials, the models which were trained using seasonal data had a lower yearly model error average, with a reduction in -0.36% and 0.08% model error for Iron and AC respectively.

Figure 4-4 shows the distribution of monthly summarised observed and predicted bursts for both the seasonally trained and non-seasonally trained models. Aligned with the improvement of model error, RMSE and MAE, Iron shows a small increase in R^2, from 0.6 to 0.635, when the model is trained on separate seasonal datasets. AC showed a very small increase in R^2, of just 0.01, therefore suggesting there is no improvement when seasonally training AC pipe material.

For both Iron and AC models, very minor changes are observed when assessing the model error each year between the seasonally trained and non-seasonally trained models, see Table 4-3 and 4-4, respectively. By using seasonal training data, a reduction in yearly model error was observed in Iron for 8 continuous years (2006 to 2013), and an increase in model error was observed from 2014 to 2016. For AC, the results are mixed, and a total of 5 years showed a reduction in model error and a total of 6 years showed an increase in model error, with no consistent years of model over or underestimation. The largest reduction in error between the seasonally and non-seasonally trained models for Iron pipes was achieved in 2011, where the models annual average error reduced by 3.47%. The largest reduction in error between the seasonally and non-seasonally trained models for AC pipes was achieved in 2012, with an annual reduction in error of 1.43%.

Table 4-3: Statistical description of the difference between non-seasonal and seasonally trained models for Iron

Note: RMSE is Root Mean Squared Error and MAE is Mean Absolute Error. RMSE and MAE are analysis of the deviance between yearly Observed (Obs) and Predicted (Pred) as shown in the table. Model Error is the percentage difference between yearly Obs and Pred burst values. Green text indicates an improvement in model error, and red text indicates a reduction in model error

Non Seasonal Model

Year	RMSE	MAE	Obs	Pred	residual	RMSE	MAE	Model Error
2006	0.04599	0.00395	1053	1018.00	-35.00			-3.44%
2007	0.04534	0.00355	1026	835.79	-190.21			-22.76%
2008	0.04180	0.00344	906	929.48	23.48			2.53%
2009	0.05054	0.00447	1279	1065.20	-213.80			-20.07%
2010	0.04939	0.00512	1194	1529.98	335.98			21.96%
2011	0.04305	0.00336	950	857.81	-92.19	147.78	112.31	-10.75%
2012	0.04372	0.00350	949	909.51	-39.49			-4.34%
2013	0.04130	0.00356	889	1047.17	158.17			15.10%
2014	0.03491	0.00246	621	688.27	67.27			9.77%
2015	0.03894	0.00274	722	740.62	18.62			2.51%
2016	0.03939	0.00287	796	734.82	-61.18			-8.33%
		Total :	*10385*	*10356.65*			*Average:*	*-1.62%*

Seasonal Model

Year	RMSE	MAE	Obs	Pred	residual	RMSE	MAE	Model Error
2006	0.04599	0.00397	1053	1031.98	-21.02			-2.04%
2007	0.04533	0.00357	1026	842.39	-183.61			-21.80%
2008	0.04180	0.00343	906	922.65	16.65			1.80%
2009	0.05053	0.00448	1279	1072.98	-206.02			-19.20%
2010	0.04937	0.00503	1194	1482.13	288.13			19.44%
2011	0.04304	0.00341	950	885.57	-64.43	133.03	101.59	-7.28%
2012	0.04372	0.00351	949	916.93	-32.07			-3.50%
2013	0.04129	0.00352	889	1027.12	138.12			13.45%
2014	0.03492	0.00247	621	693.98	72.98			10.52%
2015	0.03894	0.00276	722	750.40	28.40			3.78%
2016	0.03938	0.00286	796	729.91	-66.09			-9.06%
		Total:	*10385*	*10356.04*			*Average:*	*-1.26%*

Table 4-4: Statistical description of the difference between non-seasonal and seasonally trained models for AC

Note: RMSE is Root Mean Squared Error and MAE is Mean Absolute Error. RMSE and MAE are analysis of the deviance between yearly Observed (Obs) and Predicted (Pred) as shown in the table. Model Error is the percentage difference between yearly Obs and Pred burst values. Green text indicates an improvement in model error, and red text indicates a reduction in model error

Non-seasonal model

Year	RMSE	MAE	Obs	Pred	Residual	RMSE	MAE	Model error
2006	0.04592	0.00374	522	457.46	-64.54			-14.11%
2007	0.04075	0.00301	413	367.31	-45.69			-12.44%
2008	0.04124	0.00304	407	398.80	-8.20			-2.06%
2009	0.04536	0.00362	493	448.31	-44.69			-9.97%
2010	0.04050	0.00331	404	456.31	52.31			11.46%
2011	0.04414	0.00373	488	508.37	20.37	51.22	40.53	4.01%
2012	0.03982	0.00299	388	394.77	6.77			1.71%
2013	0.04144	0.00327	425	451.75	26.75			5.92%
2014	0.03721	0.00280	347	383.24	36.24			9.46%
2015	0.03708	0.00284	309	431.44	122.44			28.38%
2016	0.03952	0.00301	383	400.81	17.81			4.44%
		Total:	*4579*	*4698.57*			*Average:*	*2.44%*

Seasonal Model

Year	RMSE	MAE	Obs	Pred	Residual	RMSE	MAE	Model error
2006	0.04591	0.00375	522	459.31	-62.69			-13.65%
2007	0.04075	0.00299	413	364.04	-48.96			-13.45%
2008	0.04124	0.00304	407	398.48	-8.52			-2.14%
2009	0.04536	0.00362	493	449.52	-43.48			-9.67%
2010	0.04050	0.00331	404	457.94	53.94			11.78%
2011	0.04413	0.00373	488	509.79	21.79	51.87	40.58	4.27%
2012	0.03981	0.00297	388	389.09	1.09			0.28%
2013	0.04145	0.00328	425	454.47	29.47			6.49%
2014	0.03722	0.00279	347	382.74	35.74			9.34%
2015	0.03709	0.00285	309	433.56	124.56			28.73%
2016	0.03953	0.00300	383	399.10	16.10			4.03%
		Total:	*4579*	*4698.05*			*Average:*	*2.36%*

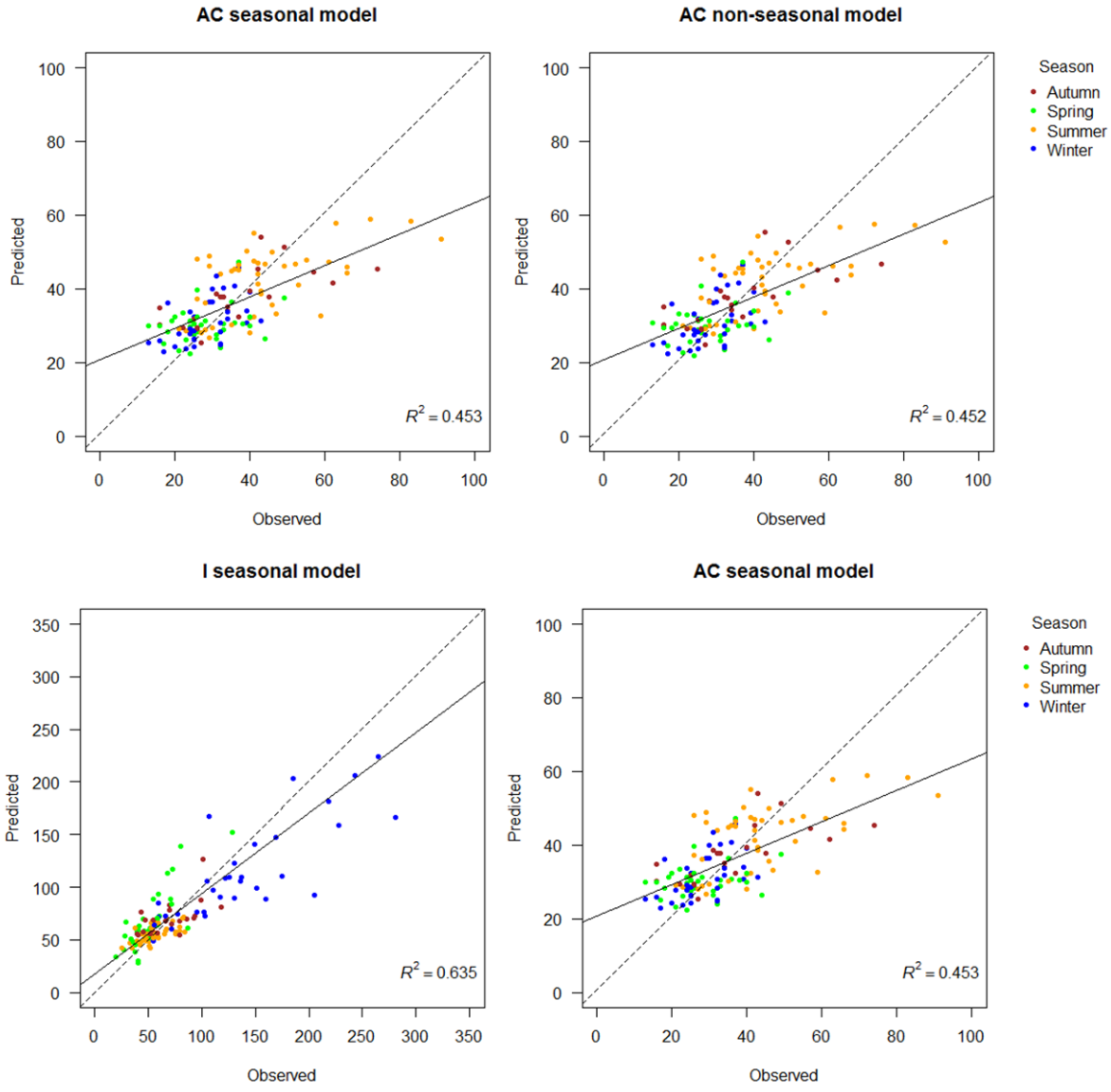

Figure 4-4: Scatter plots comparing the predicted vs. observed yearly number of bursts in Iron (I) and Asbestos Cement (AC) pipelines for seasonal and non-seasonal models

Note: Model outputs which were not trained using the seasonal split in data are on the left, whilst the seasonally trained models are shown on the right. The dashed line represents a 1:1 linear trend line, whilst the black line represents the line of best fit through the data points from the linear model

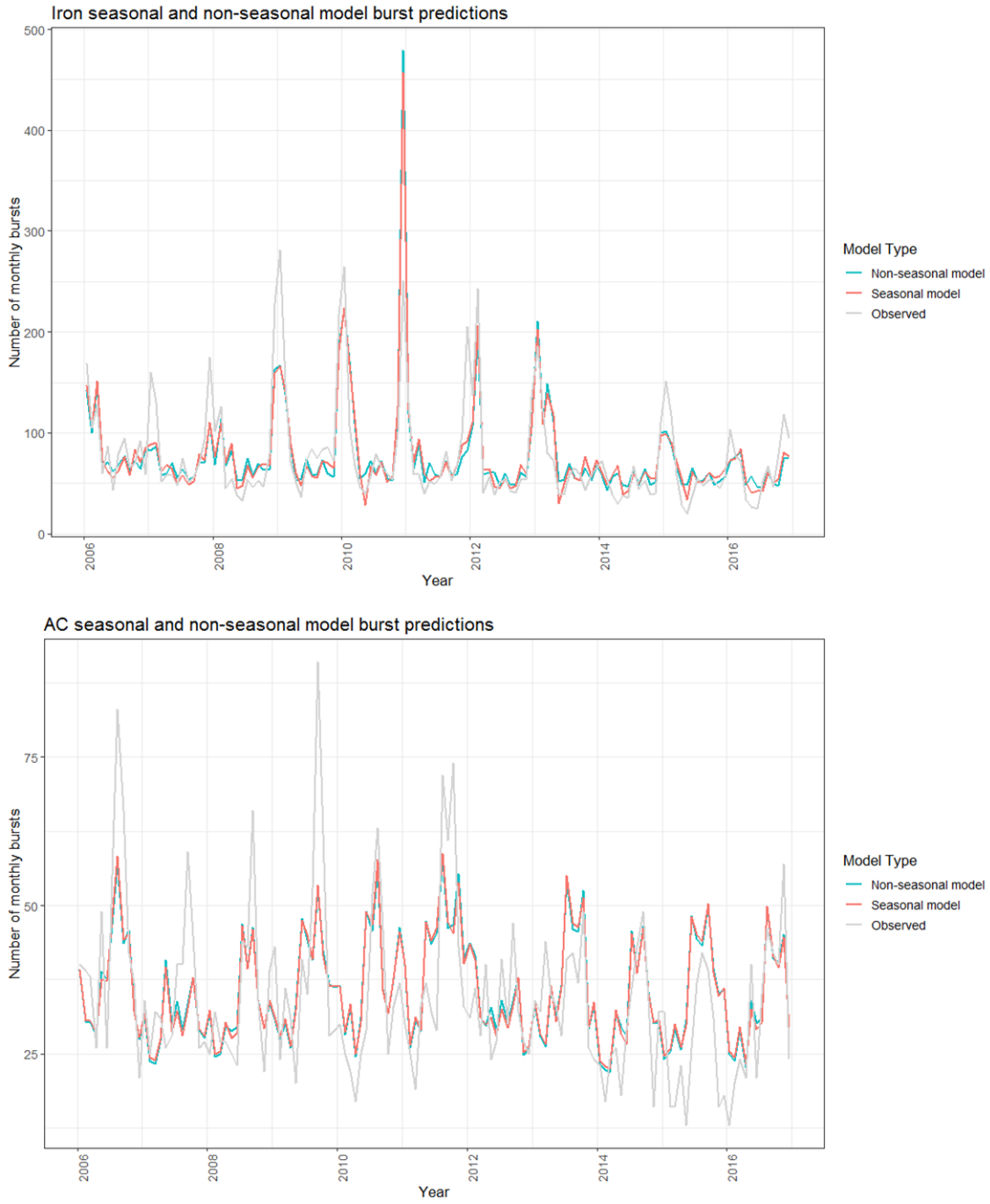

Figure 4-5: Time series plot of non-seasonally trained model, seasonally trained model, and observed bursts for Iron and AC pipes. Bursts represent the number of monthly bursts

4.4 Discussion

The overall performance in predicting bursts in Iron pipes was improved slightly when using a seasonal training and test dataset. This is evidenced by a consistent reduction in annual model error from 2006 to 2013. The largest reduction in model error from the non-seasonal to the seasonal model was gained in 2011 (-3.47%), however, a notable reduction in error (2.52%) was also observed in 2010. Iron pipes follow a strong seasonal pattern of failure and have a uniform and consistent trend of bursts with distinct peaks during winter months (Figure 4-5). A defined peak of bursts is seen in the observed values during every winter season, with the exception of 2014, when there was a lower number of observed bursts (621) in comparison to other years (mean yearly observed bursts is 944). Despite this, both the seasonal and non-seasonal models did well in predicting Iron pipeline failure for this year, with model RMSE alike to previous investigations predicting pipeline failure (Farmani *et al.*, 2017; Tabesh *et al.*, 2009). This suggests that the conditions leading to a lower rate of observed bursts are fully represented in the explanatory variables, as both the seasonal and non-seasonal models captured the reduction of observed bursts for 2014, see Figure 4-5.

Observed failures in AC pipelines show varying annual trends, with peak burst observations occurring during summer and autumn time, and other peaks of bursts also occur regularly throughout other parts of the year. The seasonally trained model did not show an improvement over the non-seasonally trained model, with only a slight reduction in percent change between predicted and observed values over the 11-years of analysis. RMSE and MAE increased with the seasonally trained data test, suggesting there is little increased benefit of using a seasonally trained and tested model for AC pipeline failure prediction. With a highly fluctuating annual pattern of observed pipeline failures, splitting the original dataset into separate summer and winter was not as effective at capturing the peak burst values for AC pipe material. Despite this, the RMSE of both the seasonal and non-seasonal models is aligned to previous investigations using alternative methods for the prediction of pipeline failure (Farmani *et al.*, 2017; Tabesh *et al.*, 2009).

Upon comparison of the seasonally and non-seasonally trained models, a very small difference in enhanced predictive ability was observed. As described in (Chapter 2; Chapter 3), several of the final model variables in Iron and AC pipeline failure models already have a prior representation of seasonality within them. For example, weather variables included in the Iron model (SMD, number of days air frost, and temperature ≤5°C over 1 week) and AC model (SMD, and temperature ≤4°C over 4 weeks) can sufficiently describe the conditions typical of either the summer or winter seasons. The representation of dynamic weather variables already included within the Poisson regression models may part explain the lack of improvement when using separate seasonal training datasets for Iron and AC pipeline failure prediction. This is because the changes in weather typical of the different seasons are already represented, via the included variables, within the model.

No directly comparable studies can be found which investigates the change in model performance upon using a separate winter and summer training and test datasets for the prediction of water pipeline failure. This is despite the known seasonal causes of pipeline failure and widespread application statistical models (Clayton *et al.*, 2010; Gould *et al.*, 2011; Kleiner and Rajani, 2001). A recommendation for future research would be to undertake separate variable selection and model building, using techniques outlined in (Chapter 2), to build a series of summer and winter specific models. This may help to further improve the representation of seasonal factors leading to pipeline failure. Undertaking separate seasonal variable selection and model building would allow the development of a parsimonious model which is fully representative of the season it is modelling, as opposed to using a model which has been built upon variables which are representative of pipeline failure throughout the different seasons.

One disadvantage of using separate winter and summer train/test datasets is the reduction in sample size used for predictive modelling. Statistical approaches rely on a sufficient sample size to generate reliable and accurate predictions (Rostum, 2000). Upon further sub-sampling of the data to create separate winter and summer train and test datasets, the overall sample size used for model training

and testing is reduced but becomes more representative. This study has overcome this by using a sufficiently long time series of observations, spanning 11-years of analysis, to maintain a sufficient sample size in each dataset. However, for smaller scale investigations, the approach of creating separate winter and summer train and test datasets could become problematic if an insufficient number of burst observations are included in the seasonal train and test datasets.

This study evaluates the seasonal representation of environmental factors (soils, weather and trees) and operational factors (age and diameter) to the failure of Iron and AC pipes. Upon the evaluation of numerous weather, soil and tree datasets, and now with the impact of seasonal model training and testing, the representation of meaningful and predictive environmental factors within the developed models are nearly at a maximum. Therefore, other variables such as key operational factors, which have not been represented within the models, provide the most potential for further improving the predictive ability. Factors such as network pressure management, previous incidences of bursts, installation and manufacturing factors, water temperature and water source provide the most promise, in this regard. The integration of data through citizen science, new satellite sensors, acoustic loggers, and smart-water monitoring systems provide promise for the generation of new data which may improve predictions of pipeline failure in statistical models.

4.5 Conclusions

This paper has demonstrated an alternative approach for model training and testing suitable for a series of previously developed water infrastructure failure models (Chapter 3). The aim of this study was to establish whether using separate winter and summer training and testing datasets could improve water pipeline prediction accuracy for Iron and AC pipes. The results demonstrated that a small improvement in the predictive ability of the failure of Iron pipes was achieved upon using separate winter and summer train and test datasets, and little to no improvement was achieved for the predictive ability of the failure of AC pipelines. A small improvement in performance was discussed to be largely

attributed to the representation of seasons already being included within several weather variables, such as SMD and Temperature.

This work has helped establish whether currently operational methods for water infrastructure failure prediction can be improved using separate seasonal datasets. Further recommendations for this work are to undertake separate variable selection and model building for separate winter and summer pipeline failure models. The development of separate winter and summer models, with the potential inclusion of network pressure and water demand may help fully capture the intra-annual variability of water supply between summer and winter seasons (Wengström, 1993), and is a promising area of statistical model development. The continued development of operational predictive models is essential for water utility companies to build network resilience, by proactively identifying bursts and improving the accuracy of burst targets, which help utility companies to avoid fines by regulators. Despite the methods in this paper being applied to only one utility company in the UK, the techniques are transferable to different water utility companies globally where similar data permits.

Acknowledgements

This work was supported by the UK Natural Environment Research Council [NERC Ref: NE/M009009/1]. All authors made significant contribution to this work. M.N and T.F designed the research methodology, M.N conducted the research and all authors contributed to the writing of the manuscript. We acknowledge the advice of Anglian Water plc. staff, and the provision of the asset data described here.

4.6 Bibliography

Bluesky, 2018. National Tree Map Dataset Description [WWW Document]. URL https://www.bluesky-world.com/ntm (accessed 2.8.18).

Clayton, C.R.I., Xu, M., Whiter, J.T., Ham, A., Rust, M., 2010. Stresses in cast-iron pipes due to seasonal shrink-swell of clay soils. Proc. Inst. Civ. Eng. - Water Manag. 163, 157–162. https://doi.org/10.1680/wama.2010.163.3.157

Davis, P., Marlow, D., Moglia, M., Davis, P., Silva, D. De, Marlow, D., Moglia, M., Gould, S., Burn, S., 2008. Failure prediction and optimal scheduling of replacements in asbestos cement water pipes. J. Water Supply Res. Technol. - AQUA 57, 240. https://doi.org/10.2166/aqua.2008.035

Farmani, R., Kakoudakis, K., Behzadian, K., Butler, D., 2017. Pipe Failure Prediction in Water Distribution Systems Considering Static and Dynamic Factors. Procedia Eng. 186, 117–126. https://doi.org/10.1016/j.proeng.2017.03.217

Farrow, J., Jesson, D., Mulheron, M., Nensi, T., Smith, P., 2017. Achieving zero leakage by 2050: the basic mechanisms of bursts and leakage. UKWIR ref 17/WM/08/60.

Gould, S.J.F., Boulaire, F.A., Burn, S., Zhao, X.L., Kodikara, J.K., 2011. Seasonal factors influencing the failure of buried water reticulation pipes. Water Sci. Technol. 63, 2692–2699. https://doi.org/10.2166/wst.2011.507

Hallett, S.H., Sakrabani, R., Keay, C.A., Hannan, J.A., 2017. Developments in land information systems : examples demonstrating land resource management capabilities and options. Soil Use Manag. 33, 514–529. https://doi.org/10.1111/sum.12380

Kabir, G., Tesfamariam, S., Loeppky, J., Sadiq, R., 2016. Predicting water main failures : A Bayesian model updating approach. Knowledge-Based Syst. 110, 144–156. https://doi.org/10.1016/j.knosys.2016.07.024

Kimutai, E., Betrie, G., Brander, R., Sadiq, R., Tesfamariam, S., 2015. Comparison of statistical models for predicting pipe failures: Illustrative example with the city of calgary water main failure. J. Pipeline Syst. Eng. Pract. 6. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000196

Kleiner, Y., Rajani, B., 2001. Comprehensive review of structural deterioration of water mains : statistical models. Urban Water 3, 131–150. https://doi.org/10.1016/S1462-0758(01)00033-4

Met Office, 2018. Met Office Rainfall Evapotranspiration Calculation System - Specialist Datasets [WWW Document]. URL https://www.metoffice.gov.uk/services/industry/data/specialist-datasets (accessed 21.5.18).

Rostum, J., 2000. Statistical Modelling of Pipe Failures in Water Networks. Nor. Univ. Sci. Technol. 1–132.

Tabesh, M., Soltani, J., Farmani, R., Savic, D., 2009. Assessing pipe failure rate and mechanical reliability of water distribution networks using data-driven modeling. J. hydroinformatics 11, 1–17. https://doi.org/10.2166/hydro.2009.008

Wengström, T., 1993. Comparative analysis of Pipe Break Rates A Literature review. Chalmers Univ. Technol.

Wilson, D., Filion, Y., Moore, I., 2017. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. Urban Water J. 14, 173–184. https://doi.org/10.1080/1573062X.2015.1080848

Wols, B. a., van Thienen, P., 2014. Modelling the effect of climate change induced soil settling on drinking water distribution pipes. Comput. Geotech. 55, 240–247. https://doi.org/10.1016/j.compgeo.2013.09.003

Yamijala, S., Guikema, S.D., Brumbelow, K., 2009. Statistical models for the analysis of water distribution system pipe break data. Reliab. Eng. Syst. Saf. 94, 282–293. https://doi.org/10.1016/j.ress.2008.03.011

5 Monitoring the response of roads and railways to seasonal soil movement with Persistent Scatterers Interferometry over six UK sites.

This chapter investigates Objective 5, and is presented in the form of one published research paper.

North, M., Farewell, T., Hallett, S., Bertelle, A. (2017) Monitoring the response of roads and railways to seasonal soil movement with Persistent Scatterers Interferometry over six UK sites, Remote Sensing, 9(922). doi:10.3390/rs9090922.

The main components of this chapter can be broken down into 1), establishing the effectiveness of the Persistent Scatterers Interferometry (PSI) technique to measure soil-related deformation in different types of above-ground infrastructure, 2) identify the seasonal response of major soil groups through satellite remote sensing. A discussion of the PSI technique and its use in statistical based modelling of water pipeline failure is made in Chapter 6.

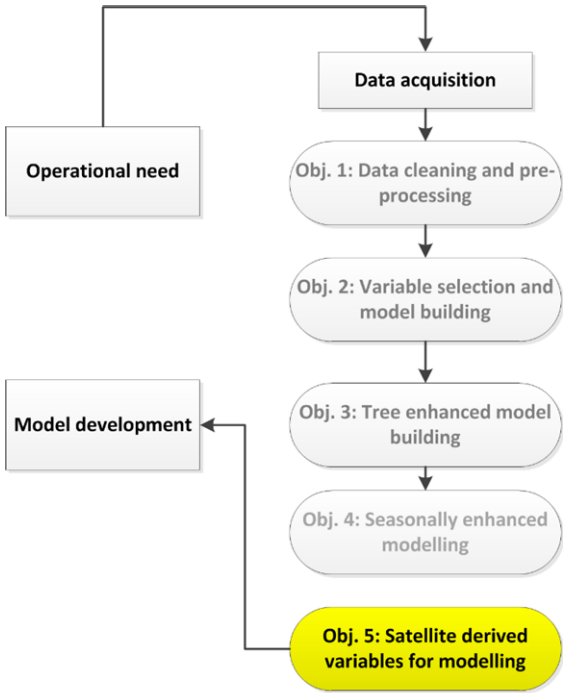

Figure 5-a: Objectives aimed to be investigated within this chapter in context of the overall thesis

Monitoring the response of roads and railways to seasonal soil movement with Persistent Scatterers Interferometry over six UK sites

Matthew North[a], Timothy Farewell[a], Stephen Hallett[a] and Audrey Bertelle[a]

[a] *School of Water, Energy and Environment, Cranfield University, Bedfordshire, UK, MK43 0AL*

Corresponding Author: t.s.farewell@cranfield.ac.uk; tel.: +44 (0)1234 752978

Abstract:

Road and rail networks provide critical support for society, yet they can be degraded by seasonal soil movements. Currently, few transport network operators monitor small-scale soil movement, but understanding the conditions contributing to infrastructure failure can improve network resilience. Persistent Scatterers Interferometry (PSI) is a remote sensing technique offering the potential for near real-time ground movement monitoring over wide areas. This study tests the use of PSI for monitoring the response of major roads, minor roads and railways to ground movement, across six study sites in England, using Sentinel 1 data in VV polarisation in ascending orbit. Some soils are more stable than others; a national soil map was used to quantify the relationships between infrastructure movement and major soil groups. Vertical movement of transport infrastructure is a function of engineering design, soil properties and traffic loading. Roads and railways built on soil groups prone to seasonal water-logging (Ground-water gley soils, Surface-water gley soils, Pelosols, and Brown soils) demonstrated seasonal subsidence and heave, associated with an increased risk of infrastructure degradation. Roads and railways over Podzolic soils demonstrated relative stability. Railways on Peat soils exhibited the most extreme continual subsidence of up to 7.5 mm y^{-1}. Limitations of this study include the short observation period (~13 months, due to satellite data availability) and the regional scale of the soil map, (mapping units contain multiple soil types with different ground movement potentials). Future use of a higher resolution soil map over a longer period will advance this research. Nevertheless, this study

demonstrates the viability of PSI as a technique for measuring both seasonal soil-related ground movement, and the associated impacts on road and rail infrastructure.

Keywords: Persistent Scatterers Interferometry; Sentinel 1; Synthetic Aperture Radar; Infrastructure monitoring; Soil movement; Soil compression; Shrink swell; Environmental risk; Road; Railway

5.1 Introduction

Ground movement is the soil-related geohazard most damaging to infrastructure in the UK (Institution of Civil Engineers, 2014). As such, the ability to measure the impact of soil movement on infrastructure networks, in a cost-effective manner, offers great value to utilities, insurance companies and governments. Infrastructure resilience can be compromised by infrastructure pressures (ageing assets, embrittlement, thinning, increasing demand and loading), environmental pressures (changing climate, soil movement), and financial pressures. However, a fully functioning and fault-resistant system of infrastructure is important for the critical operation of healthcare, transport, trade and commerce. The UK government is set to invest £100 billion in infrastructure by 2021 to ensure that the UK's infrastructure needs are met for future generations (HM Infrastructure and Projects Authority, 2016). Monitoring the structural condition of infrastructure networks is essential to ensure its long-term resilience. Developing suitable methods to monitor infrastructure networks, at the regional to national scale, is also a key challenge. (HM Infrastructure and Projects Authority, 2016).

Several *in situ* techniques are available for the local monitoring of infrastructure assets, including manual visual inspection, levelling, total station surveying and GPS technologies (Lan *et al.*, 2012). These approaches provide highly accurate measurements of deformation at a single point, yet they require significant investment of human resources and equipment to obtain a high density of measurements suitable for wide-scale infrastructure monitoring. On this premise, satellite remote sensing, particularly with the onset of new generation high resolution Synthetic Aperture Radar (SAR) sensors, can measure deformation

over areas several tens of km wide, whilst retaining high precision and accuracy. This allows the cost-effective monitoring of surface deformation (Crosetto and Monserrat, 2009).

Advanced Differential Synthetic Aperture Radar Interferometry (D-InSAR) techniques, such as Persistent Scatterers Interferometry (PSI), offer promise for the monitoring of large scale soil movement and long-term deformation of infrastructure networks. PSI requires at least 20 SAR images to measure surface deformation over months or years, removing the effects of atmosphere, topography and signal noise (Ferretti *et al.*, 2001, 2000). Several applications of PSI have been previously investigated. Some examples include, monitoring either natural or anthropogenic urban subsidence (Lan *et al.*, 2012; Stramondo *et al.*, 2008; Zhao *et al.*, 2009), measuring the intra-annual variability of soil related ground movement (Culshaw et al., 2006; Huang and Lee, 2006; Lan et al., 2012), the detection of natural hazards such as landslides (Lauknes *et al.*, 2010; Tofani *et al.*, 2013), as well as observing the structural condition of infrastructure (Chen *et al.*, 2012; Ventisette *et al.*, 2011).

The PSI technique is of great potential value to utility companies wishing to monitor assets in near real-time and could ultimately lead to observations of an entire infrastructure network with a high spatial and temporal resolution. In a recent study of Mexico City, a density of 575 PS targets per km^{-2} was achieved using the European Space Agency's Sentinel 1 data (Crosetto et al., 2015). With the advent of very high resolution SAR data, such as TerraSAR-X, a PS density of up to 5,201 targets per km^{-2} have been recorded (Yu *et al.*, 2013). Lower resolution datasets such as the ERS, and the ENVISAT satellite constellations have an archive of data extending back to 1992, enabling historical evaluations of surface deformation, if required (Terrafirma, 2010). However, new satellite missions such as the Sentinel 1 offer an increased spatial resolution with a quicker revisit period of 6 days drawing on a combination of Sentinel 1a and 1b, making it a promising new tool for environmental monitoring. On this basis, PSI applied to Sentinel 1 data could provide high density measurements of ground

movement on a regional scale and over a range of different land cover types (Lan *et al.*, 2012). Furthermore, PSI outputs can be integrated into a Geographical Information System (GIS) which allows for a greater understanding of the relationships between surface deformation and the natural environment (Aldiss et al., 2014; Meisina and Zucca, 2006). Additional data such as soil maps, geological maps and climate and meteorological data, can all provide information to further explain observed deformation phenomena (Aldiss *et al.*, 2014). Therefore, remote infrastructure observation with PSI has the potential to prompt proactive asset maintenance, increase network resilience and reduce the need for expensive *in situ* monitoring.

One example of a PSI investigation is Aldiss *et al.*'s study (Aldiss *et al.*, 2014), where a large time series of 60 ERS and ENVISAT images in descending orbit were collected between March 1997 – December 2005, with ground movement being quantified over a 95 x 55 km scene over London, UK. The study found that PSI targets in large parts of the Thames estuary subsided between 0.9 and 1.5 mm y^{-1} with the fastest soil subsidence rate being, on average 2.1 mm y^{-1}. By combining PSI outputs with geological data in a GIS, the authors noted that regional patterns of uplift and subsidence are controlled by both deep geological features, such as the relative mass of underlying geology, and shallow geological features such as fault lines. Boyle *et al.*'s study (2000), noted that the shrink swell behavior of the London clays can give rise to 50 mm y^{-1} of vertical movement (cumulative shrink and swell) over wide areas. This rate of movement was attributed largely to dry summers and wet winters. Both studies (Aldiss et al., 2014; Boyle et al., 2000) noted the importance of external factors such as extreme meteorological conditions, local topography, urban fabric, vegetation, and anthropogenically-induced subsidence such as tunneling and water abstraction.

National soil maps, such as the one used in this study (Keay *et al.*, 2009), often estimate ground movements based on the laboratory assessment of clay soils' volumetric change potential. Such categorical maps are widely used by utilities,

with classes such as “high” and “low” ground movement potential (Pritchard *et al.*, 2015a). Quantification of the vertical movement of soils is difficult due to numerous external factors (vegetation, agriculture, erosion etc.). Therefore, applying PSI over engineered surfaces which maintain high coherence can act as one possible method to quantify soil movement.

Previous studies have investigated the effects of engineering design on the rates of observed surface movement (Lan *et al.*, 2012; Yu *et al.*, 2013). The importance of foundation depth is consistently noted as a predominant control over surface deformation of structures. This paper seeks to measure directly how road and rail networks respond to the movement of different soils. By investigating the use of PSI over 6 different UK urban areas (Bristol, Bath, Bournemouth, Kings Lynn, Peterborough and Grantham), a wide range of environmental conditions, soil types and asset responses can be included in the analysis. This work will improve the current knowledge of how different types of infrastructure respond to soil movement, particularly in the UK, and will help determine the suitability of the PSI technique to monitor critical infrastructure networks nationally.

5.2 Materials and methods

This study uses Sentinel 1a C-band SAR data to produce a time-series of interferograms used in the PSI process. Sentinel-1 Single Look Complex (SLC) data has a swath width of 250 km in Interferometric Wide (IW) mode, with a spatial resolution (pixel size) of 3 x 20 m in range and azimuth, respectively. The tandem satellites (Sentinel 1a and 1b) cycle over 175 different orbits and have a full-global repeat cycle of 6 days. Many areas, such as the UK, receive more frequent data acquisition.

Sentinel 1 images were selected based on the same orbit number to ensure full spatial coverage of the study areas. Precision orbit ephemerides files were downloaded for each SAR image to ensure that no errors occurred in the co-registration of the data, this process, known as de-bursting, ensures phase continuity over the burst limits. If no precise orbit files are applied, then linear

features can appear in some interferograms (Crosetto et al., 2015). Images were also selected as per the minimum separation between two orbits for each pair (perpendicular baseline) at an interval of at least 12 days, so limiting the temporal and spatial decorrelation of the interferometric pairs. Data were collected in two separate areas (see Figure 5-1) so comparison between infrastructure responses over a large area could be undertaken. For the Western study area (Bristol, Bath and Bournemouth), data were downloaded from the 8th September 2015 to the 31st December 2016, with a polarization of VV (Vertical transmit and Vertical receive) in ascending orbit. In total, 23 images were downloaded for processing, including one master image which was selected as the 16th July 2016. For the Eastern study area (Grantham, Peterborough and Kings Lynn), 23 images were collected from the 13th January 2016 to the 8th March 2017, also collected in VV polarisation in ascending orbit. The master image for all Eastern England sites was selected as the 16th August 2016. See Table 5-1 for a description of the image dates, perpendicular and temporal baselines of the Sentinel 1a images used in this study.

All PSI processing was conducted using Harris Corporation's ENVI SARscape software v5.3. After the SAR images were connected (spatially and temporally), a series of interferometric pairs were created. These image pairs were then consequently co-registered to the master images' slant geometry. Each of the slave images were then co-registered to the master to calculate the phase difference between each image pair. A Doppler filter was applied to reduce the temporal decorrelation of the image pairs. In this step, a SRTM 3 Digital Elevation Model (DEM) was downloaded with a spatial resolution of 90 m to correct errors related to phase and atmospheric effects between the interferometric pairs. For each study site Ground Control Point (GCP) files were created to correct the images geometrically, after this step a Goldstein filter was applied to decrease the signal-to-noise ratio by filtering the differential phase (Goldstein and Werner, 1998). Values of 5 in range and 1 in azimuth were used in the multi-looking phase to obtain square pixels for the image pairs. This step is used to avoid the effects of over and under sampling on geocoded images.

The interferometric pairs were then processed by the first and second inversion steps in the ENVI PSI workflow to locate the potential Persistent Scatterers (PS). The first inversion step estimates the residual height and displacement velocity of the study area, and is applied to flatten the interferograms. Reference points are used in this step (one or more) and are automatically selected (based on high coherence and minimal deformation) for the removal of the offset phase from all interferograms. For the processing of large areas, the scene is split into sub-areas where individual processing is undertaken. A mosaicking operation is then conducted to merge all smaller subset areas to produce the final velocity map, and in this instance several reference points are selected. The second inversion step removes the atmospheric phase effects from the interferograms using the low and high pass filters, which correct the spatial and temporal distributions of atmospheric effects. The low pass filter accounts for the spatial distribution of atmospheric variations. The high pass filter accounts for the temporal distribution of atmospheric variations. The second model inversion creates the date-by-date displacements, which is then used to create the geocoded output, which is displacement along the Line of Sight (LOS). All deformations are relative to the reference points which were previously identified as being stable, through detailed visual checks of these data. Visual checks included making sure that the reference points had a stable deformation profile and that they were a suitable target (i.e. a highly engineered structure or exposed bedrock). The interferometric workflow is based on Ferretti *et al.*'s methodology, and for a mathematical account of the algorithms used see (Ferretti *et al.*, 2001, 2000).

Table 5-1: Dates of Sentinel 1 images and their perpendicular (Bperp) and temporal (Btemp) baselines for Bristol, Bath, Bournemouth, Grantham, Kings Lynn and Peterborough

	Bristol		Bath		Bournemouth	
Date	***Bperp***	***Btemp***	***Bperp***	***Btemp***	***Bperp***	***Btemp***
19-Nov-2015	-51.04 m	-240	-49.76 m	-240	-48.22 m	-240
01-Dec-2015	-26.69 m	-228	-25.62 m	-228	-23.93 m	-228
13-Dec-2015	36.48 m	-216	37.01 m	-216	36.28 m	-216
06-Jan-2016	79.81 m	-193	80.28 m	-193	80.69 m	-193
18-Jan-2016	-13.55 m	-180	-12.22 m	-180	-10.56 m	-180
30-Jan-2016	-13.55 m	-168	34.51 m	-168	36.29 m	-168
11-Feb-2016	61.99 m	-156	62.75 m	-156	63.24 m	-156
30-Mar-2016	-78.46 m	-108	-77.46 m	-108	-75.60 m	-108
23-Apr-2016	29.07 m	-84	28.88 m	-84	28.15 m	-84
10-Jun-2016	-51.98 m	-36	-51.34 m	-36	-50.14 m	-36
04-Jul-2016	31.98 m	-12	31.74 m	-12	30.64 m	-12
16-Jul-2016*	0.00 m	0	0.00m	0	0.00m	0
28-Jul-2016	-17.18 m	12	-16.62 m	12	-18.17 m	12
09-Aug-2016	-41.68 m	24	-41.10 m	24	-39.72 m	24
21-Aug-2016	-23.03 m	36	-22.90 m	36	-23.06 m	36
02-Sep-2016	118.61 m	48	117.87 m	48	116.71 m	48
14-Sep-2016	32.81 m	60	32.81 m	60	32.63 m	60
26-Sep-2016	-24.01 m	72	-23.57 m	72	-23.30 m	72
01-Nov-2016	69.37 m	108	69.65 m	108	69.59 m	108
13-Nov-2016	74.92 m	120	75.22 m	120	74.93 m	120
25-Nov-2016	32.35 m	132	33.09 m	132	34.33 m	132
19-Dec-2016	-37.74 m	156	-36.24 m	156	-34.43 m	156
31-Dec-2016	35.72 m	168	36.40 m	168	36.39 m	168

	Grantham		Kings Lynn		Peterborough	
Date	***Bperp***	***Btemp***	***Bperp***	***Btemp***	***Bperp***	***Btemp***
13-Jan-2016	29.43 m	-216	33.8966 m	-216	31.4909 m	-216
06-Feb-2016	72.79 m	-192	73.777 m	-192	72.6985 m	-192
01-Mar-2016	-89.76 m	-168	-84.46 m	-168	-88.76 m	-168
18-Apr-2016	-19.41 m	-120	-18.63 m	-120	-19.28 m	-120
12-May-2016	-105.47 m	-96	-101.63 m	-96	-104.62 m	-96
05-Jun-2016	31.86 m	-72	31.38 m	-72	30.18 m	-72
29-Jun-2016	-62.16 m	-48	-60.12 m	-48	-61.79 m	-48
23-Jul-2016	8.829 m	-24	9.02 m	-24	9.028 m	-24
04-Aug-2016	12.73 m	-12	12.8 m	-12	12.63 m	-12
16-Aug-2016*	0.00 m	0	0.00 m	0	0.00 m	0
09-Sep-2016	-72.56 m	24	-69.2 m	24	-71.73 m	24
21-Sep-2016	-79.30 m	36	-75.10 m	36	-77.98 m	36
03-Oct-2016	46.36 m	48	46.49 m	48	46.15 m	48
15-Oct-2016	71.74 m	60	72.11 m	60	71.76 m	60
27-Oct-2016	21.81 m	72	22.85 m	72	20.59 m	72
08-Nov-2016	-42.22 m	84	-36.16 m	84	-41.21 m	84
26-Dec-2016	75.52 m	132	76.92 m	132	75.13 m	132
07-Jan-2017	26.39 m	144	26.50 m	144	23.90 m	144
19-Jan-2017	-21.13 m	156	-24.98 m	156	-19.71 m	156
31-Jan-2017	-27.96 m	168	-22.01 m	168	-26.86 m	168
12-Feb-2017	53.10 m	180	55.20 m	180	53.38 m	180
24-Feb-2017	58.00 m	192	60.06 m	192	58.65 m	192
08-Mar-2017	37.62 m	204	38.51 m	204	38.01 m	204

Transport infrastructure data were freely sourced from the UK Ordnance Survey Meridian 2 dataset, which provides a frequently updated and accurate network of principle infrastructure types across Great Britain. The data provides information on the location and extent of motorways, A-roads, B-roads, Unclassified roads and Railways at the nominal viewing scale of 1: 50,000. Data are provided in a GIS ready format, with road classification provided as attributes. Road types were categorised into minor roads (Unclassified and B-roads) and major roads (A-roads and Motorways) to simplify the analysis. From these categorisations, a buffer distance of 2 meter's (from the center line of infrastructure) was created for minor roads, 3 meter's for major roads and railways, and 4 meter's for motorways. This buffer zone was then used to subset the PSI output, ensuring that only those PSI points in close proximity to the infrastructure were analysed, reducing the number of PS points returned from pavements (sidewalks), surrounding buildings and vegetation and adjacent unconsolidated ground.

By analysing the PSI output in a GIS environment, each PS point was attributed with information pertaining to the major soil group, which permitted contrasting the information about ground deformation with major soil groups. This study uses Cranfield University's National Soil Map (Keay *et al.*, 2009). The 1:250,000 scale soil map contains information related to soil properties including ground movement potential (based on the laboratory assessment of volume change potential of soils, categorized from "low" to "high"), texture, drainage, fertility, land cover and habitats. The major soil groups included in this study were Brown soils, Ground-water Gley soils, Lithomorphic soils, Pelosols, Podozolic soils, Surface-water Gley soils, and Peat soils, see Table 5-2 below for their descriptions (FAO 2015).

Validation of the PSI outputs was undertaken over the closed Meldon quarry site (50.716084 N, -4.026326 W), see Figure 5-1. Meldon Quarry has been disused since 2007 (Dartmoor National Park Authority, 2017), and comprises a mixture of metamorphic and igneous rocks. To this extent, the geology is highly stable and so presents itself as an area of high coherence, excepting natural erosion and

deposition. The PSI processing over Meldon Quarry was undertaken with an identical model set up as for the 6 other test sites, so direct comparison upon the reliability of the results could be determined.

Table 5-2: Short descriptions of the major soil groups discussed in this study along with their associated classification in the World Reference Base (WRB)

Major Soil Group Type	Description	Soil Movement Potential	WRB Classification
Brown soils	Widespread soils with predominantly brownish or reddish sub surface. They have no gleying above 40 cm depth and are mainly associated with agriculture land use.	Moderate-High	Arenosols Cambisols Luvisols Regosols
Ground-water Gley soils	Soils normally developed over permeable materials which appear uniformly gleyed. These soils are subject to periodic waterlogging by fluctuating groundwater-tables.	High	Gleysols
Lithomorphic soils	Often shallow soils which have been formed over bedrock, or soft material at 30 cm depth.	Moderate-Low	Arenosols Histosols Leptosols Phaeozems
Pelosols	Slowly permeable clay soils with no gleyed subsurface horizon above 40 cm depth. These soils can show significant desiccation in dry seasons.	Moderate-High	Cambisols Luvisols
Podzolic soils	Soils usually formed as a result of acid weathering conditions, and have an unincorporated acid layer at their surface.	Moderate-Low	Podzols Umbrisols
Surface-water Gley soils	Seasonally waterlogged, and slowly permeable soils, which appear predominantly mottled above 40 cm depth.	High	Planosols Stagnosols
Peat soils	Organic soils derived from partially decomposed plant remains that accumulated under waterlogged conditions.	High	Histosols

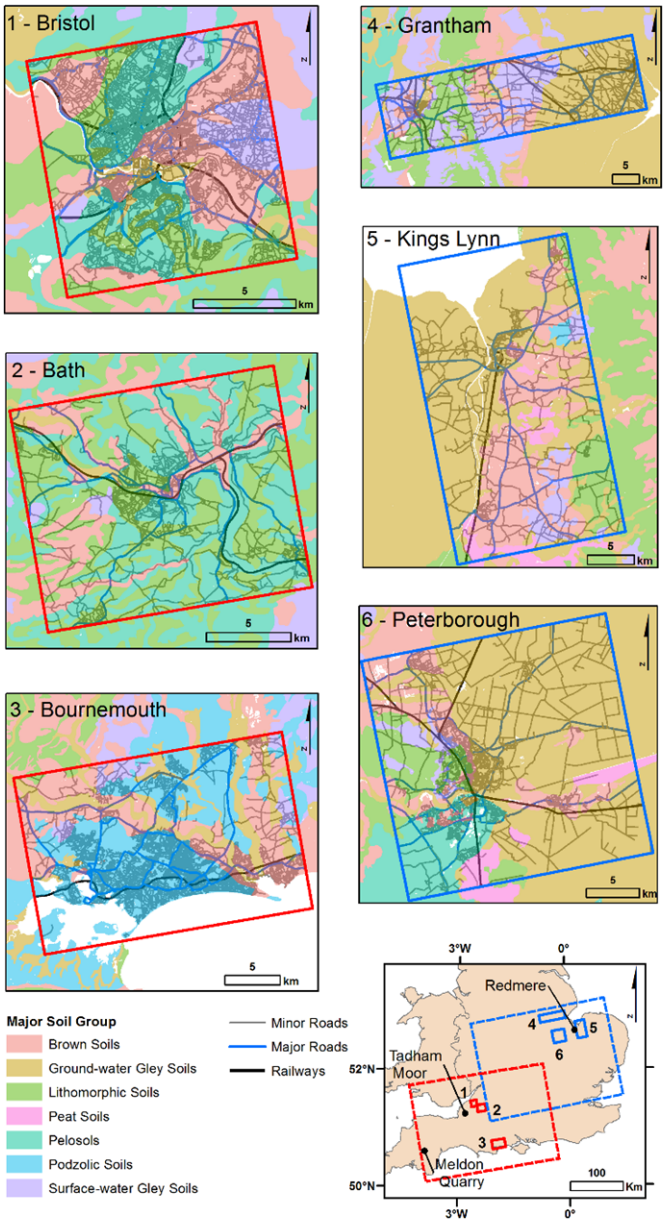

Figure 5-1: Study area extents shown with major soil group and infrastructure extent. Western sites (1: Bristol; 2: Bath, 3: Bournemouth, shown in red) Eastern sites (4: Grantham; 5: Kings Lynn; 6: Peterborough, shown in blue). The insert map locates the sites within the UK. Dashed red and blue outlines represent the area extent of the Sentinel 1 data frames (Western England – relative orbit number 30, Eastern England – relative orbit number 132). The locations of Meldon Quarry validation test site, Tadham Moor and Redmere meteorological stations, have been shown in the insert map

In total, six sites were selected for this study, separated into two distinct study areas, see Figure 5-1 and Table 5-3. The prevailing climate across all sites is classified as temperate oceanic, with all areas receiving regular precipitation events throughout the year. Sites situated in the West of England receive on average 195 mm more precipitation annually than sites in Eastern England. Of the six sites, Bournemouth receives the highest annual precipitation, 835 mm, whilst Grantham receives the least, 608 mm. Temperature ranges exhibit very little variation between all six sites, with typical maximum average summer temperatures of ~21° C and minimum average winter temperatures of ~1° C. Bournemouth receives the highest range in monthly total precipitation, from 47.8 mm in July to 100.7 mm in December. Grantham receives highest monthly precipitation in October, 59.3 mm, and 36.8 mm in February. Bristol and Bath show similar precipitation trends to Bournemouth, whilst Kings Lynn and Peterborough show similar trends to Grantham. Due to the seasonal changes in temperature and precipitation all study sites have annual profiles of soil moisture which gives rise to the conditions required for the shrinking and swelling of clay soils. Moreover, soils which are prone to seasonal waterlogging (i.e. Ground-water Gley soils, Brown soils, Pelosols and Surface-water Gley soils) may also show shrink and swell cycles in accordance to available soil moisture. *In situ* soil moisture data were obtained from two representative meteorological stations (data provided by the Centre for Ecology and Hydrology). The data provided in Figure 5-2 shows the daily averaged (from 30 minute measurements) soil Volumetric Water Content (VWC) taken at a depth of 10 cm using a Time-domain Transmission probe. As expected, the Eastern site (Redmere) held the lowest VWC value 14.8% (2nd September 2016), with a range of 26.6%, whilst the Western site (Tadham Moor) had a significantly larger VWC, 49.5% (1st August 2016), and a range of 33.9%, Figure 5-2. Therefore, some degree of surface deformation was expected in all study sites where clay soils, or soils which have significant volume change potential, are present.

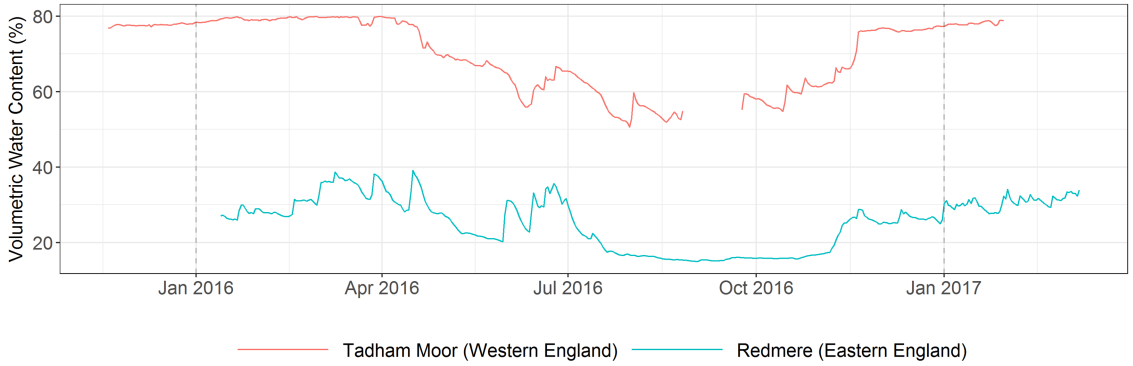

Figure 5-2: Daily Volumetric Water Content (VWC - %) for the corresponding periods of PSI investigation for Western and Eastern England. VWC is measured at 10cm depth from two *in situ* meteorological stations: Tadham Moor (Western England, 51.207099, -2.828639) in red and Redmere (Eastern England, 52.443551, 0.433083) in blue

The Bristol and Bath study sites both have had an extensive history of coal mining dating back nearly 800 years (Terrafirma, 2010), as well as mineral mining which can be dated back to 2000 years. Due to high likelihood of surface deformation, a previous PSI investigation using ERS and ENVISAT data have been undertaken over Bristol and Bath (Terrafirma, 2010), and despite the historic mining activity, compressible alluvium and shrink swell soils being present, no surface deformation was observed using PSI in this study. This present study measures recent surface deformations over Bristol and Bath using Sentinel 1 data. However, our results can be compared with previous Differential Synthetic Aperture Radar Interferometry (DinSAR) investigations in the area, offering continuity of surface deformation data between previous and current satellite missions.

The other sites have been selected based on the high potential for infrastructure responses to shrink swell cycles of soil. Sites selected in Eastern England have the highest shrink-swell potential of soils in England, based on the largest annual Soil Moisture Deficit (SMD) (Pritchard *et al.*, 2015b). It is noted (Pritchard *et al.*, 2015b), that Eastern England has the highest SMD rates in the country, and is calculated from the daily precipitation, evapotranspiration and drainage of the soil, and describes the amount of water needed for the soil to return to field

capacity (in mm). Furthermore, the response of railways to peat soils deformation in the Peterborough area was included, which is of importance given the significant peat wastage rates described in (Holman and Kechavarzi, 2011). The predominant land use in Eastern England sites is agriculture. Bournemouth has been selected as a control site, with freely and quickly draining bedrock lithology (of chalk and gravels) and Podzolic soils which are not prone to soil deformation. To our knowledge, Bournemouth, Kings Lynn, Peterborough and Grantham have not been included in any previous PSI investigations. The sites selected all have different urban densities, which help in understanding the effectiveness of PSI for monitoring infrastructure across a range of landscapes. The study areas of Bristol and Bournemouth have the most-dense urban fabric, whilst Bath and Peterborough contain a mix of rural and urban areas. Grantham and Kings Lynn are mostly a rural (agricultural) landscape, with pockets of dense settlements allowing for the PSI process to be undertaken. A description of the lengths of infrastructure, study area extent and a description of urban coverage is provided in Table 5-3.

Table 5-3: Study area characteristics and infrastructure lengths for the study sites

Study Site (Western or Eastern Area)	Area (km^{-2})	Length of minor roads (km)	Length of major roads (km)	Length of railways (km)	Historic mining present?	Urban coverage density
Bristol (W)	118	817	108	34	Yes	Mostly Urban
Bath (W)	165	265	76	31	Yes	Urban/Rural
Bournemouth (W)	284	1207	126	22	No	Mostly Urban
Grantham (E)	688	837	149	71	No	Semi-Rural
Peterborough (E)	539	907	144	55	No	Urban/Rural
Kings Lynn (E)	585	1261	190	130	No	Semi-Rural

5.3 Results

5.3.1 Minor road, major road and railways infrastructure movement

A total of 70,406 PS points was analysed for minor roads, major roads and railways across the six study areas. Western sites returned higher PS densities than Eastern sites in all infrastructure classes analysed, due to the study areas focusing predominantly on urban centres. Returned PS densities were highest for railways in all six study sites, with an average density of 18.23 PS targets returned per km of track, suggesting suitability of this infrastructure class for PSI analysis. Due to their relative widths, major roads (motorways buffered at 4 m, A-roads buffered at 3 m) returned a higher PS density than minor roads (B-roads and Minor-roads buffered at 2 m), returning 15.34 PS targets compared to 9.28 PS, respectively, per km of infrastructure. The resulting PSI velocity maps are shown in Figure 5-3. PS densities varied across the six study sites, with urban areas such as Bristol and Bournemouth achieving a very high PS density of 3799 PS km^{-2} and 2058 PS km^{-2} respectively. Bath and Peterborough, which are an urban / rural mix attained only 795 and 730 PS km^{-2}, respectively. As expected, the semi-rural study areas of Grantham and Kings Lynn achieved significantly less at 414 PS km^{-2} and 343 PS km^{-2}, respectively.

Railways exhibited the largest range of movement for all infrastructure types investigated, see Table 5-4. This is particularly so for Bath, where the average change in railway track level was -2.59 mm y^{-1} (standard deviation 5.38 mm y^{-1}), with Brown soils and Pelosols showing the highest rates of deformation. On average, railway tracks subside -0.88 mm y^{-1}. However, cyclical patterns of infrastructure movement of greater velocities that 0.88 mm y^{-1} were observed for Brown soils, Pelosols, Ground-water Gley soils and Surface-water Gley soils in Bristol, Grantham, Kings Lynn and Peterborough. Railways heave up to 2.5 mm for these major soil groups during winter months (December to April), and subside up to 5 mm during the summer months (April to September). Railways over Peat soils in Peterborough, exhibited the highest subsidence rate observed in this study, -7.5 mm y^{-1}. As Table 5-5 highlights, a very small sample size of 113 PS points were analysed for railway tracks over Peat soils, so further investigation of

this would be necessary to comment on these findings with confidence. Deformation of railways in Bath and Bournemouth present a linear pattern of deformation. Railway infrastructure in Bath is observed to subside when over Brown soils and Pelosols, whilst Bournemouth shows no deformation over Podzolic soils, see Figure 5-4. No subsidence was detected for railways over Lithomorphic soils in Bath and Peterborough.

Due to the extensive coverage of major and minor roads, more major soil groups have been analysed in the Western and Eastern study areas. Patterns and deformation trends remain mostly similar between minor roads, major roads and railways, with minimal differences between them, see Table 5-4. Cyclical patterns of deformation are observed in Bristol, Grantham, Kings Lynn and Peterborough, with Brown soils, Surface-water Gley soils, Lithomorphic soils, Pelosols, and Ground-water Gley soils all showing subsidence in summer months, and corresponding heave during winter months. Bristol presents the most uniform deformation profile of shrink and swell, with subsidence during the drying period (spring into summer), and heave during the wetting period (autumn into winter). The Eastern study sites show a larger variation of surface deformation between the major soil groups, with each major soil group showing different surface deformation characteristics. Ground-water Gley and Brown soils in Peterborough show a lagged response to the general shrink swell sequence in all infrastructure classes, with the maximum subsidence observed during December 2016. This lag corresponds to the Eastern study areas soil moisture profile (Figure 5-2), where a low soil moisture content was recorded until late November 2016.

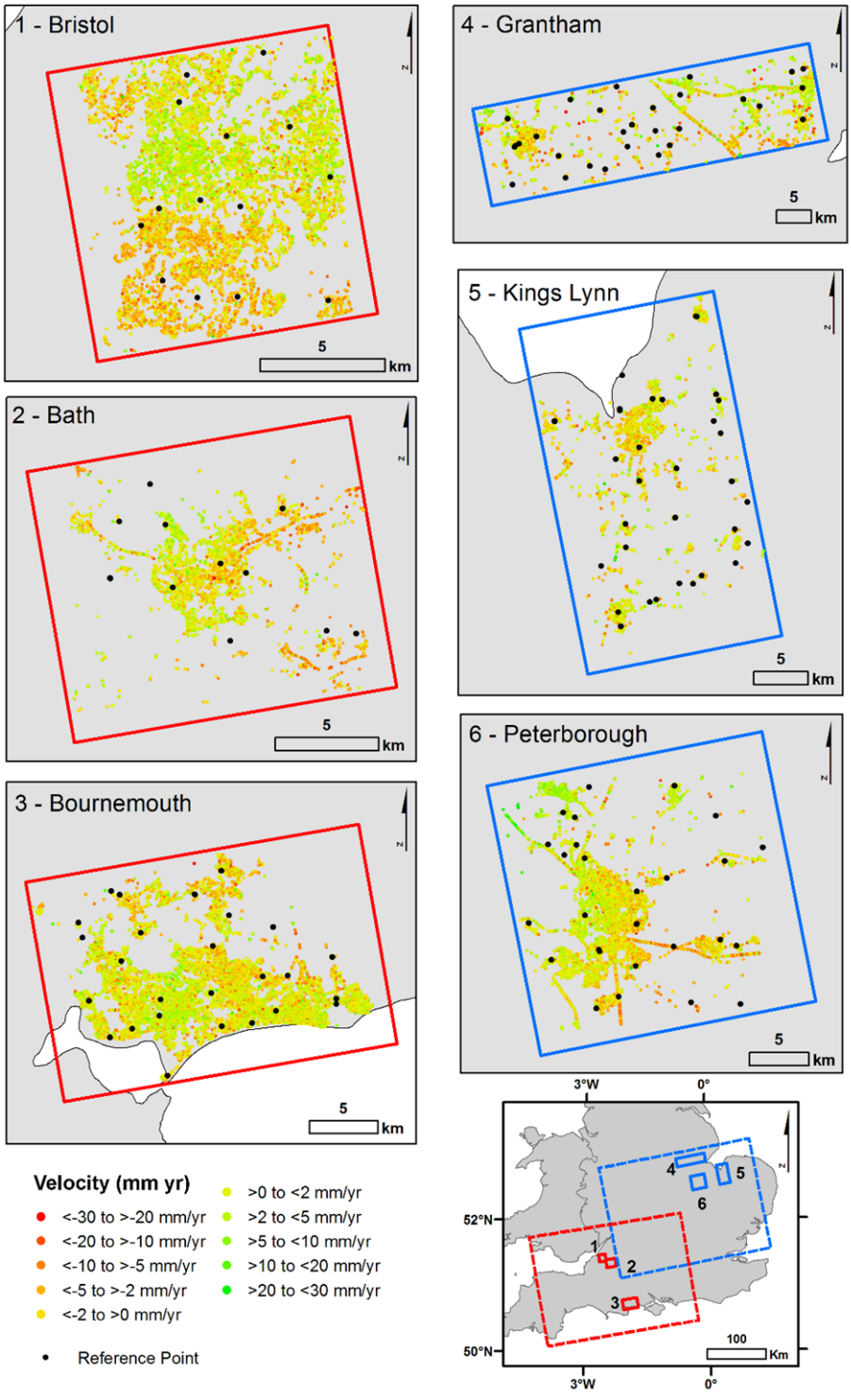

Figure 5-3: Surface deformation map showing PSI values expressed as millimetres per year for the six study areas. Dark red indicates subsidence of up to 30mm y^{-1} whilst bright green indicates uplift of up to 30mm y^{-1}. Insert map shows the location of the sites in the UK

Table 5-4: Summary description of PS targets analysed in Bristol, Bath and Bournemouth

Type	Study Area	Average PS deformation rate mm y^{-1}	Standard deviation	Maximum uplift value mm y^{-1}	Minimum subsidence value mm y^{-1}	Number of PS points	Density of PS (PS km^{-2})
Minor road (West)	Bristol	-0.24	3.18	25.05	-21.35	16557	13.16
	Bath	-0.60	3.00	11.90	-23.52	4005	7.71
	Bournemouth	0.26	3.09	23.16	-25.34	16986	13.11
	Mean	*-0.19*	*3.09*	*20.03*	*-23.40*	*12516*	*11.32*
Major road (West)	Bristol	0.04	2.83	16.13	-21.91	2920	25.54
	Bath	-0.83	2.91	17.29	-18.28	1211	14.08
	Bournemouth	0.30	2.87	15.44	-22.78	1554	10.64
	Mean	*-0.16*	*2.87*	*16.28*	*-20.99*	*1891*	*16.75*
Railway (West)	Bristol	-0.84	4.30	21.99	-18.92	688	22.93
	Bath	-2.59	5.38	14.85	-24.45	484	15.61
	Bournemouth	0.14	4.07	19.22	-15.28	487	18.03
	Mean	*-1.09*	*4.58*	*18.68*	*-19.55*	*553*	*18.85*
Minor road (East)	Grantham	-0.02	3.69	22.62	-25.06	5005	5.46
	Peterborough	0.38	3.32	24.96	-20.36	9504	10.39
	Kings Lynn	-0.25	2.91	23.47	-18.39	3882	5.89
	Mean	*0.03*	*3.30*	*23.68*	*-21.27*	*6130*	*7.24*
Major road (East)	Grantham	0.03	3.87	20.76	-24.89	1048	6.55
	Peterborough	0.24	3.08	25.56	-14.38	2328	16.16
	Kings Lynn	-0.20	3.03	13.16	-13.45	736	19.08
	Mean	*0.02*	*3.32*	*19.82*	*-17.57*	*1370*	*13.93*
Railway (East)	Grantham	-0.01	4.27	18.05	-19.14	1220	15.64
	Peterborough	-1.59	5.57	23.40	-20.67	1607	29.21
	Kings Lynn	-0.38	4.97	20.64	-16.00	184	8.00
	Mean	*-0.66*	*4.93*	*20.69*	*-18.60*	*1003*	*17.61*

Table 5-5: Sample size of PS points analysed for minor roads, major roads and railways by major soil group. Grey shading represents very low sample sizes (<100) which have been consequently removed from analysis

		Brown soils	Ground-water Gley soils	Lithomorphic soils	Pelosols	Pozolic soils	Surface-Water Gley soils	Peat soils
West	Minor roads	9276	1134	3925	7409	13490	2210	0
	Major roads	1990	440	768	882	1165	385	0
	Railways	457	204	108	469	333	10	0
East	Minor roads	5028	7885	1549	2426	16	1428	0
	Major roads	887	1717	386	838	0	279	0
	Railways	524	1284	291	354	0	445	113

Figure 5-4: Trends in vertical movement (mm) by major soil group and infrastructure type. Points show median value for all PSI points on an infrastructure type and major soil group. Solid line show a loess-smoothed trend through the plotted medians. Dashed lines identify 1st January. To ensure validity, classes with less than 100 PSI points have been removed from this plot (Table 5-5)

5.3.2 Validation test site

Using identical parameterisation as per the analysis described, a PSI investigation over the disused Meldon quarry (Figure 5-1) was undertaken to assess the general performance of the technique over a known area of no deformation. The results suggest that there was a deformation with a range of 2.24 mm between the minimum and maximum median PSI values within the quarry. Generally, there is 1 to 1.5 mm of error attributed each month to the median of all data values, represented by grey shading in Figure 5-5. The annual profile of surface deformation demonstrates a smooth, stable annual trend, with fluctuations of ± 1.5 mm during March and April. Surface deformation has been calculated by the median of 1,579 PS points which have been returned from bare rock surfaces within the quarry itself. It is expected that there is minimal movement in this quarry, apart from the natural weathering processes affecting the quarry's exposed outcrops. Therefore, it can be assumed that an error of approximately ± 1.5 mm can be attributed to the PSI measurements obtained in this study. The full area of observation contained 186,749 PS points, with an average range of 11.64 mm. This further confirms the stability of the PSI points over an area of known stability (i.e. Meldon Quarry) and gives confidence in the accuracy of the outputs in the six study areas included in this study.

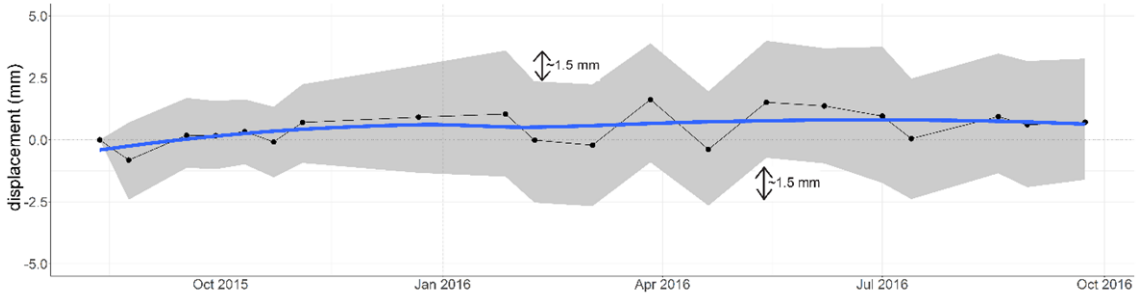

Figure 5-5: Median ground deformation (mm) for 1579 PSI points over Meldon Quarry, Dartmoor National Park (50.716084, -4.026326). Solid line: loess-smoothed trend through the plotted median PSI values. Grey ribbon indicates the inter-quartile range of Median values

5.4 Discussion

Sentinel 1 is highly effective to monitor infrastructures over the selected study sites. The response to seasonal soil moisture changes has been detected in three types of infrastructure investigated. Observed surface deformation has been discussed as a function of major soil group type, soil moisture change and the infrastructure's engineered design. Observed deformation across all sites reveal unique spatial and temporal patterns of measured movement, which have been discussed as functions of regional water use and the shrink swell potential of the different soil types studied. Sentinel 1 has proved itself to be effective at capturing these patterns across relatively wide study areas. This holds operational promise for long-term monitoring of infrastructure using Sentinel 1 with PSI.

It is evident that different soil types have an influence on observed infrastructure movement. This is particularly so for soils which are seasonally waterlogged and exhibit cycles of drying and wetting. Surface deformation of infrastructure has been observed to deform (along LOS) in all infrastructure classes analysed, mostly coinciding with the expected periods of soil shrink (summer months), and soil heave (autumn and winter months). The main soil groups included in this study which are prone to seasonal volume change are Ground-water Gley soils, Brown soils, Pelosols, and Surface-water Gley soils. All these major soil groups condition the heave and shrink of infrastructure, see Figure 5-4. Several surface deformation trends do not appear to follow seasonal pattern of shrink and swell. These unexpected results may be attributed to regional factors associated to the regional water regime, such as intense agriculture, water abstraction, irrigation and the presence of rivers and lakes. Direct analysis of these factors is out of scope for this paper, yet for Eastern study sites it should be noted that this region is under highly intensive agricultural land use, with significant drainage and irrigation, so it may not follow the expected seasonal response to soil moisture change. In such areas, variations from the expected deformation patterns are particularly evident for Brown soils (common agricultural soils) and Ground-water Gley soils in Peterborough, where a lag in infrastructure shrink is present,

demonstrating a maximum subsidence in December 2016. This lag might be the result of active water management, such as irrigation and drainage from agricultural practices, which stabilises water content during summer, and drains land during winter. These effects are particularly prevalent in the Fenlands (Eastern study sites) where the artificial pumping of water manages peat drainage (Holman and Kechavarzi, 2011). Moreover, upon analysis of the soil moisture content for Redmere (Figure 5-2), a low soil moisture was recorded until November that year, which further explains the lagged result seen in surface deformation over soils directly affected by the, often anthropogenically controlled, soil moisture content of this area. Remotely sensed soil moisture measurements have the potential to enhance the understanding of the relationships between surface soil moisture change and observed ground movement using PSI. Several limitations to this method would exist such as ensuring that the acquisition dates of the soil moisture data were aligned to the dates of the SAR images. This is particularly important given that surface soil moisture has the potential to change on a daily time scale, given the preceding weather conditions. On this basis, this method was not applied during this investigation, as *in situ* measurements provided a more continuous description of the soil moisture conditions throughout the study.

By comparison, Bournemouth shows relative stability (Figure 5-4) in infrastructures. The main soil groups are Podzolic soils, Brown soils, Ground-water Gley soils and Surface-water Gley soils. The observed stability can be attributed to the freely draining soils over permeable bedrock in Bournemouth, consisting predominantly of chalk, limestone, sands and gravels.

Variations in deformation rates are evident for Lithomorphic soils (FAO, 2015), in Bath, Bristol, Grantham, Kings Lynn and Peterborough. For example, Bath, Kings Lynn and Peterborough show minimal deformation over Lithomorphic soils whilst Grantham and Bristol show significant deformation profiles. As a major soil group, these soils are mostly very shallow and consist of seven sub-groups (Clayden and Hollis, 1984). Therefore, analysis of the major soil group in this case may not

fully represent the area's soil characteristics and thus the resultant infrastructure movement being observed. Further analysis of more detailed soil classes is required to fully characterise the interactions between Lithomorphic Soil types and infrastructure movement.

Despite its low sample size of just 113 PS points, infrastructure underlain by Peat soils showed subsidence rate of -7.5mm y^{-1}. Peterborough is the only site to have included analysis of Peat soils, but the findings of this remain important due to the high potential impact subsidence can have on the railway network operation and safety. Rates of deformation observed in this study are slightly less than (Holman and Kechavarzi, 2011) assessment of the East Anglia Fenland wastage, where it is noted that mean average wastage (over open fenland) can vary between 19 mm yr^{-1} and 127 mm y^{-1} for thin and thick deposits of Peat soils respectively (Holman and Kechavarzi, 2011). However, (Holman and Kechavarzi, 2011; Thomson *et al.*, 2007) make no direct mention of the impact of infrastructure on peat wastage. The railway observed within this study is situated on a raised, highly-engineered embankment just to the west of the Holme fen nature reserve. The impact of the railway embankment compressing the peat may lead to compaction of the Peat soil, which is suggestive as to the cause of rates of subsidence observed within this study.

A validation of the PSI output was undertaken over the disused Meldon Quarry, in Dartmoor National Park (Figure 5-5). This test site was selected for its highly stable bedrock and coherence. No vertical movement was expected during this observation period, apart from some deposition from the weathering of the quarry walls. Figure 5-5 shows minimal movement from the 1579 PS points on the quarry. However, in between March and May 2016, uplift and subsidence of approximately ±1.5 mm is observed. As this movement is consistent across the inter-quartile range of the points, and more sudden and extreme than expected, it is suggested that this movement is due to issues of the data, data cleansing or PSI technique. This level of uncertainty must be taken into consideration when reporting these results. Given the scale of the movements considered in this

study and their seasonal patterns, observations of the trend, rather than individual movement, is suggested as being appropriate. Obtaining levelling or GPS data during this study's observation period for a number of locations in all study sites would improve the validation of the results.

The rates of deformation in this study are aligned to previous investigations which have used PSI to monitor earth surface movement in the UK (Aldiss *et al.*, 2014). Apart from (Boyle et al., 2000; Pritchard et al., 2015b), there is limited work reporting investigation of the relationships between major soil group type and infrastructure movement using PSI, particularly in the UK, so limited direct comparisons can be made. Several papers have investigated infrastructure deformation using the PSI technique (Chen *et al.*, 2012; Lan *et al.*, 2012; Zhao *et al.*, 2009), but these studies use PSI to assess the structural condition of infrastructure assets in areas of known continual ground deformation. These studies note the importance of regional deformation phenomena, such as ground-water abstraction, legacy mining works or seismic activities as key considerations to make when investigating infrastructure deformation at a network level. The rates of deformation from these studies (Chen *et al.*, 2012; Lan *et al.*, 2012; Zhao *et al.*, 2009) are highly variable, but often much higher than those reported in this study (up to -73.3 mm y^{-1}) in the case of (Lan *et al.*, 2012).

Peduto *et al.* (2017) used PS InSAR to monitor the impacts of soil movement upon building damage in Rotterdam, Netherlands. Observed deformations of >10 mm y^{-1} were attributed to high risk soil types such as clay and peat. Despite the observed surface deformation velocities being higher than in this current study, the high-risk soils identified correlate to this present study, therefore increasing confidence in the findings presented. Spatial correlations of deformation were analysed, and PS observations categorized into moving / non-moving, using a 2 mm y^{-1} threshold. The approach of categorizing buildings into moving / non-moving classes would help better identify at-risk areas in linear infrastructure assets which might help to prompt proactive management by utility operators.

Previous PSI investigations of Bristol and Bath have revealed no evidence that a legacy of mining in this area has induced any long-term subsidence (Terrafirma, 2010). This paper found no notable trends in mining-induced subsidence when analysing the PSI output with historic mining extents. This is aligned to the results in (Terrafirma, 2010), however the limited observation period of this study should be taken into consideration. The Terrafirma study (2010) also noted that differential compaction was evident across both Bristol and Bath, which is confirmed in this current study, which is seen to reflect the diverse range of soil types found across the Bristol and Bath basin. To our knowledge, this is the first study of its kind investigating Bournemouth, Peterborough, Kings Lynn and Grantham using PSI, so no direct comparison of results can be made.

5.5 Conclusions

This study has applied a surface deformation investigation using PSI in six UK study sites to determine the impact that different soils have on the LOS deformation of roads and railways. Sentinel 1 has been shown to be effective at measuring surface deformation of these thin, linear infrastructure assets across a range of environmental settings, thus demonstrating PSI with Sentinel 1's potential for wide scale infrastructure monitoring. By combining PSI outputs with the 1:250,000 scale National Soil Map, this study has identified, in these areas, four major soil groups which pose a ground movement risk to infrastructure networks (Brown soils, Ground-water Gley soils, Pelosols, and Surface-water Gley soils). These soils are characterised by seasonal waterlogging. The change in soil moisture is associated with volume change in clay soils. Minor and major roads showed a similar response to soil movement, whilst railways appeared to act independently, particularly when over Peat soils. Podozolic soils remained stable in this study for all infrastructure classes investigated, due to the permeable nature of these soils over freely draining bedrock. Further investigations are required to determine the response of Lithomorphic soils, as the results were inconclusive due to the broad taxonomic nature of this group. Major soil groups in this study can contain numerous soil series, with many different soil textures and properties. This can lead to differential ground

movement even within a classified major soil group. The use of higher resolution soils data will help better explain observed deformation trends. Re-running the PSI analysis with a higher resolution soil map would allow better characterization of the observed movements to individual soil series. Furthermore, inclusion of a higher resolution DEM (such as Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global) would also help achieve better correction in the geocoding stage. The inclusion of new Sentinel 1 data to increase the length of time series would also be highly beneficial to reveal long term trends soil-related infrastructure deformation.

Furthermore, a comparison between PSInSAR and other techniques such as the Small BAseline Subset (SBAS) (Berardino *et al.*, 2002) and Intermittent SBAS (ISBAS) (Sowter *et al.*, 2013) would help to further establish the suitability of these techniques for wide scale monitoring of infrastructure deformation. SBAS and ISBAS generate near continuous coverage of ground deformation measurements, even over vegetated areas. This could help to better identify surface deformation phenomena occurring adjacent to linear infrastructure assets, which have the potential to cause a long-term risk or embankment failure.

In the absence of *in situ* levelling data, a validation site was established within a disused quarry in Dartmoor National Park, England. The results from this exercise returned the expected result of an associated error of ±1.5 mm, which suggests a good performance of the Persistent Scatterers Interferometry technique investigating soil impacts on infrastructure.

This work contributes to an increased understanding of the risks posed by soil movement to infrastructure networks in the UK. It has helped to quantify relationships between infrastructure movement and major soil groups. This is critical for the planning and monitoring of large-scale infrastructure projects, such as the planned High-Speed Rail 2 project in the UK, where an understanding of the soil related risk is key to the success of the project. This paper has demonstrated the potential of PSI and Sentinel 1 to be used as a tool for remotely monitoring environmental risks to transport infrastructure networks.

Acknowledgments: This work was supported by the UK Natural Environment Research Council [NERC Ref: NE/M009009/1]. Costs for open access publication are provided through Cranfield University via the RCUK block grant. The authors would also like to acknowledge the Centre for Ecology and Hydrology (NERC) for providing meteorological data.

Author Contributions: All authors made significant contributions to this work. M.N, T.F, A.B, designed the methodology and implemented the experiments. M.N, T.F and S.H wrote the paper, and revised the manuscript to its final format.

5.6 Bibliography

Aldiss, D., Burke, H., Chacksfield, B., Bingley, R., Teferle, N., Williams, S., Blackman, D., Burren, R., Press, N., 2014. Proceedings of the Geologists ' Association Geological interpretation of current subsidence and uplift in the London area , UK , as shown by high precision satellite-based surveying. Proc. Geol. Assoc. 125, 1–13. https://doi.org/10.1016/j.pgeola.2013.07.003

Berardino, P., Fornaro, G., Lanari, R., Member, S., Sansosti, E., Member, S., 2002. A New Algorithm for Surface Deformation Monitoring Based on Small Baseline Differential SAR Interferograms. IEEE Geosci. Remote Sens. 40, 2375–2383.

Boyle, J., Stow, R., Wright, P., 2000. In-SAR Imaging of London Surface Movement for Structural Damage Management and Water Resource Conservation; Report for BNSC Link Programme, Project 4; National Remote Sensing Centre: Farnborough, UK.

Capes, R., Marsh, S. 2009. The Terrafirma Atlas - The Terrain Motion Information Service for Europe, GMES-ESA 2009. Available Online: http://esamultimedia.esa.int/multimedia/publications/TerrafirmaAtlas/pageflip.html (accessed 06.08.2017)

Chen, F., Lin, H., Li, Z., Chen, Q., Zhou, J., 2012. Interaction between permafrost and infrastructure along the Qinghai–Tibet Railway detected via jointly analysis of C- and L-band small baseline SAR interferometry. Remote Sens.

Environ. 123, 532–540. https://doi.org/http://dx.doi.org/10.1016/j.rse.2012.04.020

Clayden, B., Hollis, J., 1984. Criteria for differential soil series; Soil Series Technical Monograph No. 17; Rothamsted Experimental Station: Harpenden, UK.

Crosetto, M., Devanthéry, N., Monserrat, O., Crippa, B., 2015. Exploitation of the full potential of PSI data for subsidence monitoring. Proc. IAHS 372, 311–314. https://doi.org/10.5194/piahs-372-311-2015

Crosetto, M., Monserrat, O., 2009. Persistent scatterer interferometry: Potentials and limits. In Proceedings of the TSPRS Hannover Workshop 2009: High Resolution Imaging for Geospatial Information, Hannover, Germany, 2-5 June 2009.

Culshaw, M., Tragheim, D., Donnelly, L.B., 2006. Measurement of ground movements in Stoke-on-Trent (UK) using radar interferometry. In Proceedings of the 10th Congress of the International Association for Engineering Geology and the Environment (IAEG 2006), Geological Society, London, UK, 6-10th September 2006, pp. 1–10.

Dartmoor National Park Authority, 2017. Meldon Aplite Quarry - Educational Register of Geological Sites, Available Online http://www.devon.gov.uk/geo-meldon-aplite-quarry.pdf (accessed 8.07.2017).

FAO, 2015. World reference base for soil resources 2014: International soil classification system for naming soils and creating legends for soil maps (Update 2015), World Soil Resources Reports World Soil Resources Report 106. ISSN 0532-0488. pp192. http://www.fao.org/3/i3794en/I3794en.pdf. (Accessed: 20 April 2019).

Ferretti, A., Prati, C., Rocca, F., 2001. Permanent Scatterers in SAR Interferometry. IEEE Geosci. Remote Sens., 39, 8–20.

Ferretti, A., Prati, C., Rocca, F., 2000. Nonlinear Subsidence Rate Estimation Using Permanent Scatterers in Differential SAR Interferometry, IEEE Geosci. Remote Sens., 38, 2202–2212.

Goldstein, R.M., Werner, L., 1998. Radar interferogram filtering for geophysical applications, Geophys. Res. Lett., 25, 4035–4038.

HM Infrastructure and Projects Authority, 2016. National Infrastructure Delivery Plan 2016-2021, HM Government, London.

Holman, I., Kechavarzi, C., 2011. A revised estimate of peat reserves and loss in the East Anglian Fens Commissioned by the RSPB, Department of Natural Resources, Cranfield University, Cranfield, UK. pp. 44.

Huang, Y., Lee, C., 2006. Detecting Ground Surface Movements With Differential Insar Techniques. In Proceedings of the ASPRS Annual Conference, Reno, Nevade, 1-5th May 2006.

Institution of Civil Engineers, 2014. The State of The Nation Infrastructure 2014. Inst. Civ. Eng.

Keay, C.A., Hallett, S.H., Farewell, T.S., Rayner, A.P., Jones, R.J.A., 2009. Moving the National Soil Database for England and Wales (LandIS) towards INSPIRE Compliance. Int. J. Spat. Data Infrastructures Res. 4, 134–155. https://doi.org/10.2902/1725-0463.2009.04.art8

Lan, H., Li, L., Liu, H., Yang, Z., 2012. Complex Urban Infrastructure Deformation Monitoring Using High Resolution PSI. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5, 643–651. https://doi.org/10.1109/JSTARS.2011.2181490

Lauknes, T.R., Piyush Shanker, a., Dehls, J.F., Zebker, H. a., Henderson, I.H.C., Larsen, Y., 2010. Detailed rockslide mapping in northern Norway with small baseline and persistent scatterer interferometric SAR time series methods. Remote Sens. Environ. 114, 2097–2109. https://doi.org/10.1016/j.rse.2010.04.015

Meisina, C., Zucca, F., 2006. PS InSAR Integrated with Geotechnical GIS : Some Examples from Southern Lombardia. Geod. Deform. Monit. Geophys. Eng. Roles 131, 65–72.

Peduto, D., Nicodemo, G., Maccabiani, J., Ferlisi, S., 2017. Multi-scale analysis of settlement-induced building damage using damage surveys and DInSAR data : A case study in The Netherlands. Eng. Geol. 218, 117–133. https://doi.org/10.1016/j.enggeo.2016.12.018

Pritchard, O.G., Hallett, S.H., Farewell, T.S., 2015a. Probabilistic soil moisture projections to assess Great Britain's future clay-related subsidence hazard. Clim. Change 133, 635–650. https://doi.org/10.1007/s10584-015-1486-z

Pritchard, O.G., Hallett, S.H., Farewell, T.S., 2015b. Soil geohazard mapping for improved asset management of UK local roads. Nat. Hazards Earth Syst. Sci. 15, 2079–2090. https://doi.org/10.5194/nhess-15-2079-2015

Sowter, A., Bateson, L., Strange, P., Ambrose, K., Fifik, M., 2013. DInSAR estimation of land motion using intermittent coherence with application to the South Derbyshire and Leicestershire coalfields. Remote Sens. Lett. 4, 979–987. https://doi.org/10.1080/2150704X.2013.823673

Stramondo, S., Bozzano, F., Marra, F., Wegmuller, U., Cinti, F.R., Moro, M., Saroli, M., 2008. Subsidence induced by urbanisation in the city of Rome detected by advanced InSAR technique and geotechnical investigations. Remote Sens. Environ. 112, 3160–3172. https://doi.org/10.1016/j.rse.2008.03.008

Thomson, A.M., Mobbs, D.C., Milne, R., Skiba, U., Levy, P.E., Jones, S.K., Billett, M.F., Van Oijen, M., Ostle, N., Foereid, B., Smith, P., Randle, T., Matthews, R.W., Gilbert, J., Halsall, L., Brewer, A., Baldwin, M., Mackie, E., Bellamy, P., Rivas-Casado, M., Bradley, R.I., Grace, J., Lewis, P., Quaife, T., Jordan, C., Tomlinson, R.W., 2007. Report Inventory and projections of UK emissions by sources and removals by sinks due to land use, land use change and forestry Annual Report, June 2007. Database 200.

Tofani, V., Raspini, F., Catani, F., Casagli, N., 2013. Persistent Scatterer Interferometry (PSI) Technique for Landslide Characterization and Monitoring. Remote Sens. 5, 1045–1065. https://doi.org/10.3390/rs5031045

Ventisette, C. Del, Intrieri, E., Luzi, G., Casagli, N., Fanti, R., Leva, D., 2011. Using ground based radar interferometry during emergency : the case of the A3 motorway (Calabria Region , Italy) threatened by a landslide. Nat Hazards Earth Syst., 11, pp. 2483–2495. https://doi.org/10.5194/nhess-11-2483-2011

Yu, B., Liu, G., Zhang, R., 2013. Monitoring subsidence rates along road network by persistent scatterer SAR interferometry with high-resolution TerraSAR-X imagery. J. Mod. Transp., 21, 236–246. https://doi.org/10.1007/s40534-013-0030-y

Zhao, Q., Lin, H., Jiang, L., Chen, F., Cheng, S., 2009. A study of ground deformation in the guangzhou urban area with persistent scatterer interferometry. Sensors 9, 503–518. https://doi.org/10.3390/s90100503

6 Discussion and conclusion

This research has investigated and implemented several quantitative techniques to enhance current methods of pipeline failure prediction. In doing so, a number of complex interactions between environmental (such as soils, weather and trees) and operational factors to the failure rates of different pipeline materials have been revealed. This chapter aims to discuss the research presented in Chapters 2 to 5 and provide a general discussion into how the methods used have helped to answer the specific research objectives and aim. The limitations of the study, the key findings from each objective, the consequent contributions to knowledge, the identification of promising future research and concluding remarks are given within this chapter.

6.1 Implications of research for the UK's water sector

The regulatory body for the UK's water distribution network (Ofwat) has allowed Anglian Water plc to adopt statistical predictions in the setting of annual burst main targets, which are adjusted to the prevailing environmental conditions. This has allowed the migration away from static burst targets, with no environmental representation, to dynamic burst targets which are based on environmental conditions and previous incidences of water pipeline failures.

This acceptance of data-driven approaches by the regulator underlines the necessity for the development of robust and accurate water infrastructure failure models. Anglian Water is the first water utility company in the UK to have Ofwat approved statistical models for the setting of annual leakage targets, through the Water Infrastructure Serviceability Performance Assessment (WISPA) initiative. As a result, this present research has focused only on the development of Poisson regression models which are the currently approved statistical methods of prediction.

This is despite the known alternatives modelling approaches as described in (Farmani *et al.*, 2017; Kleiner and Rajani, 2001; Tabesh *et al.*, 2009; Wilson *et al.*, 2017; Yamijala *et al.*, 2009). This thesis focusses on Poisson regression models to ensure that the methods developed can be applied directly into an approved industrial context. Moreover, Poisson regression models were initially selected for their relative ease of implementation in comparison to other methods, the applicability to a wide range of data types and ability to predict bursts accurately aligned with other modelling techniques (Kimutai *et al.*, 2015).

This thesis has presented a framework which water companies can adopt to identify key environmental and operational factors leading to the failure of different pipe materials and allow then to construct a series of predictive pipeline failure models from the appropriate datasets. Water companies have a corporate responsibility to conserve water and employ best practice when trading in a natural resource, therefore the reliability and efficiency of water distribution is a key priority for water companies. Considering this, water companies are striving to become innovative and effective in reducing pipeline failures and become proactive in the way they manage their assets. Therefore, statistical models, such as the ones developed within this thesis, have numerous associated positive impacts.

Water companies receive financial penalties for pipeline bursts which are found by the public (i.e. reactive bursts). Bursts which are found by the company's leakage team (i.e. proactive bursts) receive no financial penalties. On this basis, statistical models which help predict the number of bursts in a given period, under different environmental conditions, can help water companies determine what is the required operational resource needed to pro-actively detect burst water mains. Moreover, an increased understanding of the factors leading to asset failure, which is gained through variable selection in the model development, can increase the company's resilience to future pipeline failure. Statistical models and failure rate analysis can also help companies to quantify the most at-risk assets in the network, which is important for asset rehabilitation strategies and the replacement of old pipelines. Water companies have an obligation, set by the

regulator, to ensure that water pipes maintain appropriate levels of asset health to ensure the reliable distribution of water. Therefore, quantifying failure rates under different environmental conditions can provide a measure of asset health in a non-invasive and cost-effective manner, whilst minimising the need to physically inspect buried assets.

6.2 Discussion of key findings from Objective 1

There is a strong consensus in the scientific literature that data preparation, data cleaning and the creation of homogenous pipe cohorts is vital to the success of statistical modelling of water pipeline failure (Gould, 2011; Kleiner and Rajani, 2001; Rostum, 2000; Xu *et al.*, 2011). A pipe cohort is defined as a group of pipes with similar characteristics. Objective 1 was to determine the impact of data cleaning, pre-processing and the creation of pipe cohorts to the improvement of water-pipeline failure prediction. To achieve this objective, a methodological approach to data preparation was applied on a historical archive of Anglian Water utility data. Pipe cohorts were grouped by pipe material, age, diameter, which are buried in the same district metered area and soil type for the entire Anglian Water drinking water distribution network (Chapter 2).

In order to quantify the impact of data cleaning on the improvement of water-pipeline failure predictions, a series of comparisons of model testing was made using the newly developed pipe cohort data and previously developed WISPA models (Figure 6-1). Initially, the WISPA models trained on previous “un-cleaned” data, from 2004 – 2012, was re-tested using a 50% hold-out sample of the newly prepared pipe cohort data, representing bursts from 2006-2016, and the results recorded. As a second step, the WISPA models were then re-trained using the remaining 50% hold-out sample of the newly developed pipe cohort data (2006 – 2016) and re-tested using the same dataset as before. This allowed a direct comparison of the impacts of data cleaning to the prediction accuracy of the previously developed WISPA models.

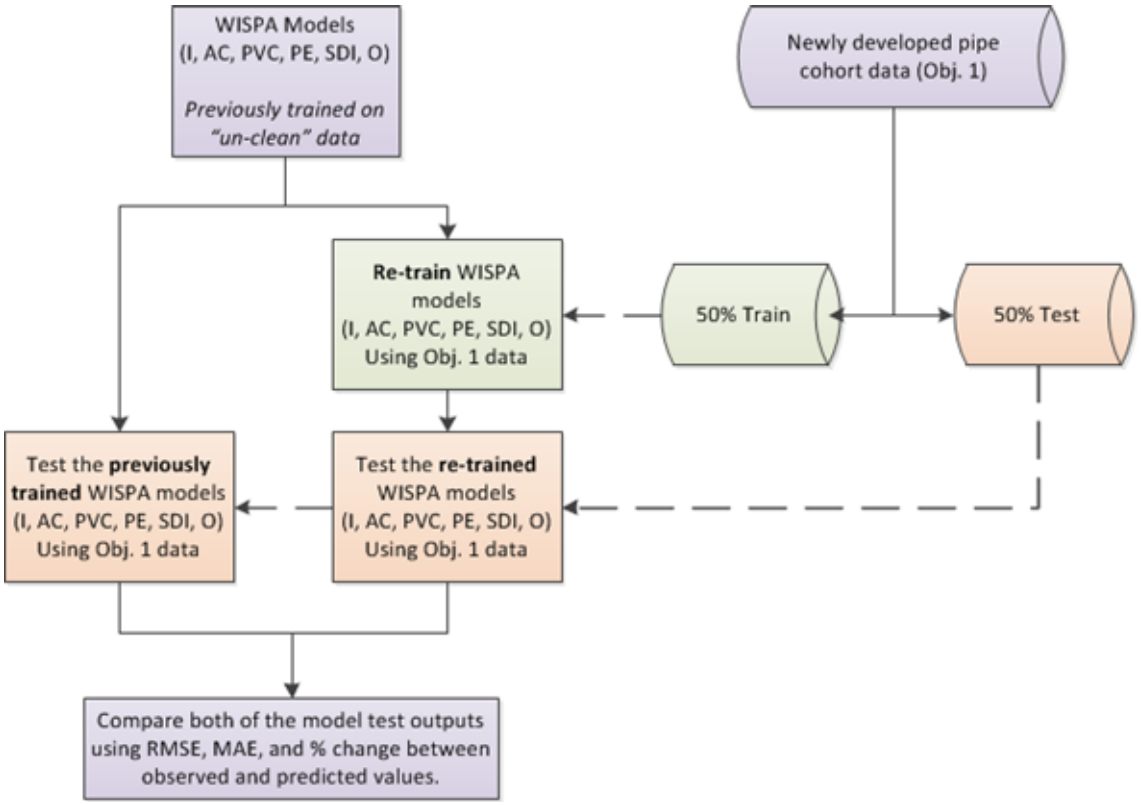

Figure 6-1: Flow diagram of the methods used to evaluate the value of data cleaning and the creation of pipe cohorts to WISPA model prediction accuracy

It was found that after data cleaning and the creation of pipe cohorts, a reduction in model error was found for all pipe materials, with the exception of polyethylene pipelines (Table 6-1). In comparison to the implementation of the other objectives (Objectives 2, 3 and 4), Objective 1 resulted in the largest reduction of model error, with the other objectives showing further small improvements in model error upon the implementation of data-cleaning on the datasets used for model testing and training.

The largest improvement in model performance was recorded in Steel and Ductile Iron (SDI) pipes and pipes classified as "other" (O), where a reduction of 31.42% and 17.41% total model error over the 10 years of analysis was achieved, respectively. The 10-year average WISPA model error for Asbestos Cement (AC), Iron and Polyvinylchloride (PVC) was reduced by 12.49%, 4.61%, 13.67%, respectively. Unexpectedly, the average 10-year model error for Polyethylene (PE) pipes increased by 4.49%. Despite the increase in model error for PE, the

mean average annual residual between predicted and observed bursts is just 4 bursts per year. In comparison to other materials, the burst rate of PE is generally low, given the total overall length of PE pipe in the distribution network (c. 10,000 km, representing 27% of the overall network length). The total number of observed bursts in 10 years (2006-2016) for PE pipes is 1,553 bursts, for which the WISPA model, trained on the new pipe cohort data, over predicted reactive bursts with a mean average of 12.5 per year. The WISPA PE model, trained on previously data, overestimated bursts with a mean average of 3.8 bursts per year. However, with marked improvements in the other 5 material groups, the benefits of data cleaning and the formation of pipe cohort data was seen to outweigh the increase in model error for just 1 pipeline material. On this basis, the newly developed pipe cohort data was used in the subsequent investigation of Objectives 2, 3, and 4.

Table 6-1: Change in model error in WISPA model after data cleaning and the creation of pipe cohorts (Objective 1)

Note: Obs is observed reactive bursts and Pred is predicted reactive bursts. Material labels are Asbestos Cement (AC), Iron (I), Polyvinylchloride (PVC), Polyethylene (PE), Steel and Ductile Iron (SDI) and pipes classified as "other" (O). Model error is the % difference between observed and predicted bursts from 2006 - 2016. RMSE is Root Mean Squared Error and MAE is Mean Absolute Error. Green indicates a reduction in model error, whilst red indicates a reduction an increase in model error

	Material	Obs bursts	Pred bursts	Model error	RMSE	MAE
WISPA	AC	4579	4161.31	-10.04%	0.04128	0.00303
WISPA data cleaned	AC	4579	4694.07	2.45%	0.04128	0.00321
WISPA	I	10385	9895.97	-4.94%	0.04330	0.00347
WISPA data cleaned	I	10385	10350.78	-0.33%	0.04329	0.00354
WISPA	PVC	4651	4061.57	-14.51%	0.03833	0.00267
WISPA data cleaned	PVC	4651	4612.25	-0.84%	0.03833	0.00284
WISPA	PE	1553	1599.66	2.92%	0.02041	0.00080
WISPA data cleaned	PE	1553	1678.16	7.46%	0.02041	0.00082
WISPA	SDI	368	569.20	35.35%	0.01461	0.00053

WISPA data cleaned	SDI	368	383.04	3.93%	0.01460	0.00042
WISPA	O	413	552.49	25.25%	0.01441	0.00046
WISPA data cleaned	O	413	448.12	7.84%	0.01441	0.00041

It is important to note the potential sources of error from either the pipe or burst datasets upon the formation of the pipe cohort data. Potential sources of error from the utility company can include mistakes being introduced to the datasets upon the incorrect recording of burst incidences, misclassification of either pipes or bursts, and errors introduced to the dataset by manipulation and transformation of data formats (Boxall *et al.*, 2007). These uncertainties within the primary datasets must be correctly identified and handled to ensure that errors do not persist within the datasets used for model prediction. Adding to this complexity, the removal of misclassified bursts or pipes is not a viable option due to the potential loss of important information within the datasets, and a statutory need to report all relevant information to the regulatory body. Therefore, the formation of a robust data preparation strategy, as like the one outlined in Chapter 2, can help minimise the potential loss of operational information, and join unknown or misclassified bursts and pipes using a methodological approach. Visual checks of the data was important throughout all data preparation steps to ensure the correct and logical joining of datasets.

6.3 Discussion of key findings from Objectives 2, 3 and 4

Objectives 2, 3 and 4 are associated with creating and enhancing Poisson regression models to improve the current Ofwat approved methods of prediction (WISPA). In doing so, an increased representation of environmental factors such as soil, weather and trees into a series of material-specific pipeline failure models was undertaken. Key topics which are relevant to all three objectives are discussed within this current section.

6.3.1 Importance of domain knowledge

Statistical modelling relies on domain knowledge for the appropriate selection and creation of variables which are likely to be predictive of pipeline failure. Throughout Chapters 2, 3 and 4, several preliminary studies were undertaken to better understand the environmental processes leading to pipeline failure.

For the work presented in Chapter 2, a preliminary investigation of the impacts into different soil and weather conditions on the failure rates of different pipeline materials was undertaken, but for brevity this was omitted within the research paper. For instances where variables needed to be created, such as weekly accumulated temperature, weekly temperature change, soil moisture deficit change, understanding the dynamics of environmental conditions and their impact on the rates of pipeline failure was of key importance for the creation of informative modelling covariates. A similar approach was adopted within Chapter 3, results from which was included in the research paper, where the pipeline failure rates were calculated under different tree density, soil shrink swell potential, pipe material and distance of analysis. This approach allowed the effective identification of the environmental conditions which led to higher rates of pipeline failure and created a sufficient understanding of the environmental processes occurring to allow the creation of informative variables used in statistical modelling. For Chapter 4, a preliminary investigation was undertaken to identify the average monthly failure rates of Iron and AC pipes. This helped to highlight the distinct intra-annual patterns of pipeline failure and was used to create the “summer” and “winter” train and test datasets for the Anglian Water region.

Unless previous investigations have been undertaken directly over the same study area, the use of exploratory data analysis is important for any future studies wishing to replicate the research presented within. This is because information cannot always be inferred from previous studies which was undertaken in different geographical areas, due to the variations in weather, soil and tree dynamics which create spatially varying environmental risks.

6.3.2 Choice of variable selection technique

Variable selection has been a key component of this research, as it has facilitated the non-bias and quantitative identification of predictive weather, soil and tree variables which were used to form the material-specific pipeline failure models. Akaike's Information Criterion (AIC), used in a forwards selection stepwise approach, was used in Chapters 2 and 3 as the variable selection method (Akaike, 1974). The rationale behind the choice of the AIC technique is multi-fold. Firstly, AIC is a metric which is easily generated from Poisson regression models. Several studies, across numerous scientific disciplines have used AIC as a method to choose variables which are used within stepwise regression models (Li et al., 2013; Morozova et al., 2015; Williams et al., 2015; Zhang, 2016). Secondly, to develop an operationally viable method for water utility companies wishing to adopt statistical models, using metrics such as AIC, which are easily generated from the Ofwat approved models, is highly appropriate to maintain ease of implementation.

In some cases, the use of AIC has selected variables which are previously known to impact water pipeline failure, such as the climate-adjusted clay hazard variable, cold weather temperatures and total number of days with air frost (Chapter 2). However, AIC also proved effective at identifying new, and previously un-explored variables, such as hydrology of soils, depth to bedrock, and combinations of tree height and density variables (Chapter 2; Chapter 3).

6.3.3 Stepwise model building approach

The previously developed WISPA models included similar variables for all material types, despite numerous previous investigations highlighting the material-specific interactions with different environmental processes (Clayton *et al.*, 2010; Davis *et al.*, 2008, 2007; Gould, 2011; Kleiner and Rajani, 2001; Makar and Kleiner, 2000; Rajani and Kleiner, 2001). Objective 2 aimed to build a series of six material-specific water infrastructure models, which have an increased representation of the failure mechanisms common across different pipeline materials. This removed the need to include other un-informative parameters relating to other material types, leading to the development of a parsimonious

model, which is a key priority in the formation of statistical models (Aho *et al.*, 2014; Farmani *et al.*, 2017; Rajani *et al.*, 2012; Zhang, 2016). Due to the size of a typical water distribution network, splitting the network into distinct material groups is a computationally-effective way of analysing the large volume of operational and environmental data. Therefore, the employment of material specific models over an entire network is not as troublesome as it might first appear. This being said, the use of high-performance computing was integral for the stepwise model building approach, owing to the large volume of operational and environmental datasets which was represented weekly.

Previous investigations have noted the issue of multicollinearity within variable selection and model building processes using an AIC step-wise approach (Morozova *et al.*, 2015). This issue was controlled in Chapter 2 by limiting the size of the developed model to a maximum of 12 steps. The use of Root Mean Squared Error (RMSE) was used as a statistical metric to evaluate the model's performance at every step. This helped to ensure that the addition of new variables had a positive impact on the model's predictive ability, and that the smallest size model (in terms of model covariates), with the highest predictive accuracy, was selected as the most appropriate model. The maximum use of 12 steps was used in this study as no material specific model showed an improvement in predicting bursts past this step. For other studies wishing to implement a step-wise modelling approach, setting the number of steps used for model creation should be flexible to accommodate for different modelling covariates and datasets. Users should look to stop the forward stepwise approach when there is no additional value in adding variables.

6.3.4 Verification of model predictions

Throughout Chapters 2, 3 and 4, a verification of the model outputs has been undertaken using a range of statistical measures using a 50% hold out sample. Statistical evaluation has been key in identifying the most accurate Poisson regression model, based on the residuals between predicted and observed bursts, RMSE, MAE, R^2, and the percent difference between predicted and observed bursts.

By statistically evaluating model performance, the modelled outputs can be compared with other studies, or to results from other modelling techniques. This helps to position the accuracy of the models developed against other studies, raising confidence in the model predictions. Moreover, a quantification of the improvement to the model upon the addition or removal of covariates can help improve the understanding of the model's sensitivity to included model variables.

Further verification of model accuracy could include the implementation of other statistical evaluations, such as Receiver Operating Characteristic (ROC). Such statistical measures help to identify the fraction of correct predictions in comparison to total predictions in individual classes, and help to develop an understanding of true/false positive rates (Wilson *et al.*, 2017; Winkler *et al.*, 2018).

An understanding into the sensitivity of the developed statistical models was gained in Chapter 2, where RMSE was used as an indicator to identify the influence of adding new variables to the model's ability to predict bursts. A detailed sensitivity analysis investigation, such as Monte Carlo Analysis, might be beneficial as it could determine the effect of model variance over numerous test runs using different samples of model train and test datasets. This would help build a more robust understanding of the model's reliability to consistently predict water pipeline bursts, which would help build upon the method implemented.

6.3.5 Discussion of tree enhanced model

Chapter 3 investigated the impact of tree density, tree height and the proximity of trees to pipes and the failure rates of four common pipeline materials across an entire drinking water distribution network. Previous investigations into the impact of trees on the failure of buried assets have been lacking to-date, especially at a regional scale. However, several studies have investigated the impact of individual trees on above-ground infrastructure, such as buildings and roads, (Mercer and Reeves, 2011; Navarro *et al.*, 2009) or at the local scale to buried infrastructure networks, such as sewerage networks (Östberg *et al.*, 2012; Torres *et al.*, 2017). Using a national tree inventory, Chapter 3 developed an appropriate method for utility companies wishing to understand the impacts of trees, soil and

weather to buried drinking water pipes across a regional area, which is aligned to the scale of observation required by utility operators.

Despite the identification of increased failure rates in pipe materials under very high shrink swell soils and high tree densities, there was almost no improvement in the model's predictive ability upon the inclusion of variables representing tree height and density. An alternative approach for modelling tree impacts would be to split the pipe network into a regular division of pipe lengths (i.e. 500 m intervals) and investigate the impact of trees on regular lengths of pipe. Currently, the analysis is undertaken using pipe cohorts, where sections of pipe of homogenous material, age and diameter are grouped together. Therefore, there is a large variation in pipe cohort length. In having a set length of pipe, tree density will become more comparable across the distribution network, and it is expected that a clearer representation of tree density will be gained within the model.

The dataset used for the creation of tree variables has no representation of tree species or any indication of the recent removal of trees. The National Tree Map was acquired in 2015 and represents the location and heights of the tree during that year. The additional representation of these factors could enhance statistical methods prediction, due to the known impact of these variables have upon differential ground movement (Mercer and Reeves, 2011; Navarro *et al.*, 2009; Östberg *et al.*, 2012). In this regard, either satellite or aerial remote sensing provides promise for the creation of additional tree variables which can be used for predictive modelling. For example, several studies have highlighted methods which enable the classification of tree species, at the individual tree canopy level, using optical sensors and the spectral signatures of different tree species (Ruiliang, 2013; Sheeren *et al.*, 2015; Zhang and Qiu, 2012). Other studies have also described change detection techniques to identify the recent removal of trees and large vegetation in urban environments using a time series of satellite or aerial images (Walton *et al.*, 2008). The inclusion of such information, combined with the national tree inventory, have the potential to increase the representation of the influences of trees on buried assets, and potentially improve the current prediction accuracy within the developed models.

6.3.6 Discussion of seasonally trained and tested model

The failure rates of some pipeline materials are controlled by the environmental conditions typical of the different seasons. The impact of training and testing both the Iron and AC Poisson regression models seasonally was detailed in Chapter 4.

Only a small improvement to the prediction accuracy of Iron pipelines was found upon using separate seasonal training and testing datasets. This was, in part, due to the prior representation of seasons within the weather variables included. On this basis, the use of dynamic variables, such as weather data, is important for the prediction of pipeline failure so that the annual trends in changing environmental conditions are fully represented within the model covariates.

This study did not evaluate the effectiveness of undertaking seasonal variable selection and model building to create separate summer and winter models. From an operational viewpoint, there is less value in the development of separate models, due to the need to deploy different models in accordance with the different seasons. This adds unnecessary complexity to the utility operations and deployment of statistical methods. However, it could be expected that the separate development of summer and winter models would increase predictive ability for each respective season, as it would lead to the development of a parsimonious model which is fully representative of environmental conditions typical of the different seasons. One disadvantage of using separate seasonal models is the reduction in the overall sample size which is available to train and test the models. A reduction in sample size has been noted to impact the accuracy of statistical methods in previous investigations (Rostum, 2000). Furthermore, “unseasonal” environmental conditions, particularly in the light of global climatic change, are likely to become more common. Developing models on data which are typical of either winter or summer seasons only, the models will have less representation of potential “unseasonal” weather, which will lead to a reduction of model error during these periods.

6.3.7 Handling uncertainty within environmental datasets and model predictions

Throughout Chapters 2, 3 and 4, relevant secondary datasets have been used to model the environmental conditions which led to the failure of drinking water pipelines. Understanding the uncertainties within the model input data are important as it can help minimise the risk of compounding error effects upon merging datasets together and using them for model training and testing.

The environmental data used for the representation of weather, soils and trees has been acquired from reputable sources, such as the Met Office, Bluesky and the National Soil Resources Institute (NSRI) of Cranfield University. It is important to understand the progeny of these datasets to increase confidence in the data being used to develop the models. Both the Met Office and Bluesky have an ISO 9001 accreditation, which means that the distributed data has been managed and quality controlled to ensure it meets customer and regulatory requirements. The NSRI have ISO 19119 accreditation, meaning that Geographical Information Services (GIS) data are quality checked and conformant to international standards. These certifications of standards affirm the quality of the data used within the thesis, but it is important to note that such certification does not replace the need of other quality checks which need to be implemented by the user. Such quality checks include understanding the type of data (i.e. categorical, continuous, and discrete) which is being used in modelling, the variance of values, the inclusion of coerced null variables, and processes such as gap-filling.

The Poisson regression models used for Chapters 2 to 4 estimate the probability of a burst occurring, based on the unique operational and environmental conditions. Where necessary, the high and low confidence limits, which are also generated upon the formation of the burst estimate, have been included in all reported results to help communicate the associated uncertainty. Other statistical modelling techniques, such as Bayesian networks, also allow the generation of probability estimates, and are highly suitable for the modelling of uncertain and complex processes, such as those found in the environment (Uusitalo, 2007). Several studies have used Bayesian networks for the modelling of water pipeline

failure (Francis *et al.*, 2014; Kabir *et al.*, 2016; Ogutu *et al.*, 2017). However, one disadvantage of Bayesian networks, over Poisson regression, is the need for the input data to be discretised. This may cause difficulties in the preparation of some datasets, such as handling categorical soils data or operational variables such as age and diameter bands, which have been developed to increase the simplicity of analysis (Uusitalo, 2007). Therefore, the ability of Poisson regression models, to include a range of variables (continuous, discrete or categorical) is a significant advantage over Bayesian methods, as it allows water utility companies to predict pipeline failure in a pragmatic and reproducible manner.

6.4 Discussion of key findings from Objective 5

Chapter 5 investigated the use of Persistent Scatterers Interferometry (PSI) combined with Sentinel 1 to measure soil-related deformation of above-ground infrastructure (Ferretti *et al.*, 2001). Sentinel 1a was launched in 2014 as part of the European Space Agency's Copernicus programme, and has an improved spatial and temporal resolution in comparison to other freely available Synthetic Aperture Radar (SAR) sensors, such as the European Remote-sensing Satellite (ERS) and Environmental Satellite (ENVISAT). Capitalising on the improved spatial resolution of Sentinel 1, and a sufficient archive of satellite data, Chapter 5 investigated the combination of PSI and Sentinel 1 for its use in measuring soil-related deformation to above ground infrastructure. The rationale of Chapter 5 was to determine the feasibility of using PSI and Sentinel 1 to generate regional land-surface movement measurements which could then consequently be used for statistical modelling of water infrastructure failure. It was hypothesised that for pipes which are buried in soils where measured above-vertical ground infrastructure movement was occurring, there would be a higher incidence of bursts in buried water pipes. Several issues, which are discussed within this section, inhibited the use of PSI measurements to be used as model covariates. However, Chapter 5 revealed the value in using PSI measurements, combined with a 1:250,000 scale detailed soils map, to measure the soil-related impacts directly against different above-ground infrastructure types. Such information is

important for transport operators and for the planning and monitoring of large-scale infrastructure projects such as the UK's High Speed Rail 2 (HS2) project.

One of the primary issues which inhibited the use of the PSI output being adopted as a modelling covariate was the consistency and density of PSI measurements. The PSI method is sensitive to a range of different factors, such as the presence of vegetation and maintenance work, such as road resurfacing, which can impact upon the density of the PSI output. Despite the overall density of PSI measurements being high in urban areas, for the purpose of statistical modelling, a high density of measurements would also have to be included which represent rural pipe cohorts, which was not always achievable due to the aforementioned reasons. As an alternative technique, the Intermittent Small Baseline Subset (ISBAS) technique (Sowter *et al.*, 2013) can provide near-continuous coverage of surface deformation measurements in both rural and urban environments, and might provide a favourable alternative to PSI for the generation of regional surface deformation measurements.

PSI data uses an archive of satellite data for the generation of surface deformation measurements. Therefore, the PSI technique can only evaluate previous deformation phenomena (up to the most recent image date), over a relatively long time series (20+ interferometric pairs). However, PSI is a potentially valuable tool to monitor known areas of high ground movement and pipeline failure within a distribution network, to understand the current status of annual shrink and swell cycles, and to further investigate surface deformation dynamics in known areas where pipeline failures are common.

Chapter 5 was the first study to combine and analyse the results of the surface deformation measurements recorded from PSI and Sentinel 1 to the National Soil map of England and Wales. This helped to contribute to the wider understanding of how major soil groups impact upon measured surface deformation. In the lack of available *in situ* levelling data, a novel approach for the validation of PSI outputs was undertaken using a disused quarry as a reference site. This approach provides a useful alternative methodology for other studies with a lack of ground measurements for validation. Owing to the computational resource and

user-expertise needed to implement PSI, it is impractical to suggest this technique as an operationally viable tool to generate measurements which are suitable for the statistical modelling of water pipeline failure. However, Objective 5 has been successful as a proof of concept for the use of satellites to measure soil-related deformation of above ground infrastructure through the unique combination of Sentinel 1, PSI and a detailed soils inventory.

6.5 Contributions to knowledge

A summary of the novel aspects of this work and the consequent contributions to knowledge are summarised in the following points:

- The creation of six material-specific water failure models which incorporate a range of different soil, weather and tree variables predictive of the failure of common pipeline materials, leading to the improvement of an existing modelling design. The methods developed in this thesis demonstrate a data-driven approach in which water companies can adopt to identify key variables which can be used to predict pipeline failure. Statistical predictions of pipeline failure can be used in the operational management of assets where information regarding the expected number of bursts under different environmental conditions can be used to aid proactive burst detection. A shift towards the proactive management of assets is important given the growing pressure on water companies to reduce total leakage from the distribution network. Operational tools, such as models developed within can help companies reduce total leakage, and therefore achieve a more sustainable business model. Minimising water loss is particularly important given global climate change, where there is a growing awareness of water conservation in the public, and a tighter governmental regulation on burst water mains.

- The generation of new knowledge upon how tree density and soil conditions lead to the differential failure rates in common pipeline materials over an entire distribution network in the UK. The methodology devised also has applicability to other water utility operators, where the data permits, and are scalable to

other buried assets such as gas pipelines and sewerage networks. Increased knowledge upon tree-soil-weather interactions to buried pipes is important given the current increase of green urban spaces. Urban greening is occurring globally to help sequestrate CO^2 and other greenhouse gasses, whilst reducing latent heat in urban heat islands. With a drive to increase urban green spaces, an understanding of the impacts of trees on buried assets is needed to achieve long-term resilience in buried infrastructure networks. Therefore, the work developed within Chapter 3 has contributed towards an improved understanding of water pipeline failure rates under different tree, soil and weather conditions. Such information is useful for utility owners and local authorities whose assets are directly impacted on by trees and are looking to prioritise the replacement of the most at-risk assets in the network. Furthermore, an improved understanding into the failure rates in different pipe materials help companies to have more accurate life-cycle costings of pipe materials, which is important for long-term business strategy.

- The development of a methodology to measure the seasonal soil-related movement in different types of above-ground infrastructure using relevant, satellite data, secondary geographical datasets and the Persistent Scatterers Interferometry technique. Such methods are important as they provide infrastructure operators the means to monitor wide-scale infrastructure networks in a cost-effective and accurate manner. This minimises the need for extensive ground measurements of levelling campaigns which are labour intensive, expensive and disruptive. The work presented within Chapter 5 also details a method for the validation of the PSI output when there is a lack of ground-truth data. An area of known stability, i.e. a disused quarry, was used to validate PSI measurements against, which provides a useful alternative for future studies wishing to undertake PSI investigations retrospectively where ground-truth data are not available.

- The testing of 3 new datasets (MORECS 5 km, National Tree Map and Sentinel 1), to determine their suitability for predicting pipeline failure:
    - The MORECS 5 km dataset provided an increased representation of meteorological variables in the models developed in Chapter 2 and led to an improved model prediction accuracy in comparison to the previously used MORECS dataset, which has a coarser resolution. Testing the beta version of MORECS 5 km dataset within this thesis has contributed to the development of the dataset in terms of providing an un-biased critique of the datasets value for burst detection in the water industry. Such information is important as the MORECS 40 km dataset is the current industry standard of meteorological data, therefore the release of MORECS 5 km is highly anticipated.

    - The National Tree Map was used to calculate, for the first time, the varying failure rates of different pipeline materials under different tree density and soil shrink-swell potential. However, the variables created from the National Tree Map did not improve statistical methods of pipeline failure prediction as expected.

    - Upon combining Sentinel 1 with the National Soils Map of England and Wales, results highlighted that minor roads, major roads and railways which are built on 4 (out of 7) major soil groups, showed a seasonal pattern of infrastructure movement. The use of Sentinel 1 data, combined with the PSI technique, was evaluated for its potential use in providing modelling covariates which describe soil related ground movement. It was concluded that the results generated by this method were not suitable to be included into the developed models, however several alternative approaches such as the use of ISBAS and higher resolution data (such as TerraSAR-X) may have better promise for water pipeline failure modelling.

6.6 Recommendations for future research

Considering the scope and the constraints of this present research, several aspects have been highlighted for future research and development, namely that:

- **An increased representation of new network management and operational factors may lead to improved model accuracy:**
  The inclusion of detailed operational and network management factors may increase the predictive ability of the models developed. This study has increased the representation of environmental factors (soils, weather and trees) within the current methods of prediction, however, there is further scope to increase the representation of wider operational and network management factors within models described. For example, a representation of pipeline pressure, or changes to the way pressure is managed within the distribution network has a known impact on the failure of buried assets, but is currently lacking in the developed models (Kimutai *et al.*, 2015). Other factors can include information upon previous pipeline failures, transient, cyclical and surge pressures within the distribution network, and external loading impacts to pipeline failure (Rezaei *et al.*, 2015).

- **Integrating data sources from new technologies into the statistical models may improve representation of factors currently missing:**
  The data generated from new emergent technologies, such as new satellites, citizen science and novel sensors, such as acoustic loggers and smart water metering systems, may provide valuable input data for water pipeline failure prediction. Assimilation of such data into the developed models may provide an increased representation of factors which have not been currently considered.

- **Directly comparing other statistical models may help position the accuracy of the currently Ofwat approved methods to other modelling techniques:**

A comparison of numerous statistical modelling techniques, such as decision tree analysis, Bayesian Belief Networks, and Proportional Hazards model would allow the direct comparison of the Poisson regression modelling technique to other methods. A critique of different models, applied over the same distribution network, may permit the change of approved methods by the regulator, Ofwat. However, sufficient reason and evidence would have to be gained in order to change the current approved methods.

- **Working towards the development of a short-term forecasting model for pipeline failure prediction using real-time weather forecasts:**
  The integration of short term (2 week) forecasts of pipeline failure is a key development area for utility companies wishing to understand network risks to aid proactive management practices. The development of a forecasting model, based on the models described combined with real-time weather forecasts, has the potential to improve the current understanding of operational risks over the distribution network, or at the district metered area level.

- **A clearer representation of trees in the statistical modelling may lead to an improvement of burst prediction:**
  The development of an alternative methodology to create variables representative of trees height and density into the developed statistical models has the potential to further improve prediction accuracy. The current methodology of creating variables which represented the percentage of tree canopy coverage (in different tree height bands up to a 40 m distance of the pipe) within individual pipe cohorts did not improve the current methods of pipeline failure prediction. This was discussed to be because of the highly varying length of pipe cohorts which were created across the distribution network. In separating the pipe network into regular divisions of length (i.e. 500 m), it could be expected that a more comparable dataset of tree-related risks will be gained and will be better represented within the statistical models. Further information relating to tree species would also be expected to enhance the predictive ability of the developed models and could be acquired through

image classification using optical remote sensing (either airborne or spaceborne) platforms.

6.7 Concluding remarks

This thesis has analysed a range of environmental factors which can lead to the increased failure of common pipeline materials over a regional distribution area in the UK. The results discussed are applicable to other temperate climate countries, with the methods developed being appropriate where similar data exists. Moreover, the environmental conditions discussed and represented within the models developed are suitable to other types of buried infrastructure, such as gas or sewerage pipelines.

A quantitative investigation into the impact of data cleaning and the creation of pipe cohorts was undertaken, using a series of previously developed water infrastructure failure models (Objective 1). This analysis highlighted the value and benefits of a methodological approach for the preparation of data used for statistical modelling and led to the reduction of model error for the previously developed WISPA models in 5 out of 6 material types. The creation of a cleaned and pre-processed dataset is not only important for the prediction of pipeline failure, but is important given the statutory need to report accurate burst and pipeline information to the industry regulator.

The data developed in Objective 1 was used subsequently for the material-specific variable selection and model building to develop new series of Poisson regression pipeline failure prediction models. Several datasets representing operational, soil and weather conditions (Chapter 2) and tree-related variables (Chapter 3) was evaluated. As a result, six material-specific pipeline failure models have been developed, which are representative of the environmental conditions which led to infrastructure failure for each individual material type. Chapter 2 evaluated a beta version of MORECS dataset, with an improved 5 km spatial resolution (Met Office, 2018), and Chapter 3 evaluated the use of the National Tree Map (Bluesky, 2018) for its use in statistical modelling of pipeline failure for the first time. Chapter 4 investigated the impact of seasonally training and testing the developed models, to increase the representation of the

seasonality of pipeline failure predictions. Chapter 5 described the acquisition of satellite-derived measurements to quantify the impacts of soils on above-ground infrastructure movement. Despite the usefulness of results in understanding soil-related risks to above ground infrastructure, the results from Chapter 5 was discussed to not be suitable for the inclusion of statistical modelling of water pipeline failure, due to the inconsistency of the measured output, interference from vegetation, and a lack of rural PSI measurements, see Section 6.4. However, Chapter 5 combined satellite-derived surface deformation measurements to the National Soil Map of England and Wales (Hallett *et al.*, 2017) for the first time, and revealed the influence of seasonal soil movement to different types of above-ground infrastructure. Such information is useful for the inference of deformation occurring to buried in assets beneath above-ground infrastructure, such as water pipes.

This thesis has evaluated a range of different environmental pressures leading to the failure of drinking water pipes. As a framework, it has described suitable approaches for utility companies, infrastructure operators, and academics to develop material specific pipeline failure prediction models using the relevant secondary environmental datasets, in a cost-effective, reproducible and pragmatic approach. Such research is timely as statistical models are now being used by Anglian Water to aid proactive management of assets and set annual burst targets. Therefore, the development of robust and accurate methods of pipeline failure prediction is of great importance.

The use of the approaches developed within this study are encouraged for all UK water utility companies, as it enables them to better understand the environmental and operational risks to the distribution network. Direct measurements of pipeline failure rates, under a range of environmental conditions, facilities the generation of accurate annual burst targets, which can help minimise the risk of large fines by the regulator. Accurate and reliable predictions of pipeline failure can also lead to the increased ability to proactively manage the network, by prioritising the rehabilitation of the most at-risk assets, based on the unique environmental and operational conditions.

By reducing water loss within the distribution network, several positive impacts can be gained. An example of a direct impact of reducing water loss, is the reduction in energy that is required to treat water pre-distribution. With fewer pipeline failures, more water can be reliably distributed to the consumer, which saves the utility company money and energy in treating potable water which would have otherwise been lost. By saving money and energy in the water treatment phase, an increase in potential investment can be given to rehabilitate the most at-risk assets in the network, which would help to reduce further water loss. Moreover, utility companies could also use saved money to fund vital research and development projects giving the utility operator a competitive edge. An example of an indirect impact of proactive leak detection is allowing business and industry to function without any disturbances in their water supply, which has numerous benefits for the nation's economy and trade. Other organisations such as schools and hospitals also require an uninterrupted supply of water and would benefit from increased resilience against pipeline failure. Therefore, even with just the two examples given, it is evident that the positive impacts of proactive management in water supply extends much further than the distribution network itself.

Ultimately, proactive leak detection, through the employment of statistical models, can lead to greater resilience within the distribution network which consequently leads to an increased operational performance. On this basis, the development of methods to proactively detect burst water pipes is a critical step for water utility companies and is essential for achieving wide-spread sustainability within the UK's water industry.

6.8 Bibliography

Aho, K., Derryberry, D., Peterson, T., 2014. Model selection for ecologists : the worldviews of AIC and BIC. Ecology 95, 631–636. https://doi.org/10.1890/13-1452.1

Akaike, H., 1974. A New Look at the Statistical Model Identification. IEEE Trans. Automat. Contr. 19.

Bluesky, 2018. National Tree Map Dataset Description [WWW Document]. URL https://www.bluesky-world.com/ntm (accessed 2.8.18).

Boxall, J.B., O'hagan, A., Pooladsaz, S., Unwin, D.M., 2007. Estimation of burst rates in water distribution mains. Proc. Inst. Civ. Eng. - Water Manag. 160, 73–82. https://doi.org/doi.org/10.1680/wama.2007.160.2.73

Clayton, C.R.I., Xu, M., Whiter, J.T., Ham, A., Rust, M., 2010. Stresses in cast-iron pipes due to seasonal shrink-swell of clay soils. Proc. Inst. Civ. Eng. - Water Manag. 163, 157–162. https://doi.org/10.1680/wama.2010.163.3.157

Davis, P., Marlow, D., Moglia, M., Davis, P., Silva, D. De, Marlow, D., Moglia, M., Gould, S., Burn, S., 2008. Failure prediction and optimal scheduling of replacements in asbestos cement water pipes. J. Water Supply Res. Technol. - AQUA 57, 240. https://doi.org/10.2166/aqua.2008.035

Davis, P.Ã., Burn, S., Moglia, M., Gould, S., 2007. A physical probabilistic model to predict failure rates in buried PVC pipelines. Reliab. Eng. Syst. Saf. 92, 1258–1266. https://doi.org/10.1016/j.ress.2006.08.001

Farmani, R., Kakoudakis, K., Behzadian, K., Butler, D., 2017. Pipe Failure Prediction in Water Distribution Systems Considering Static and Dynamic Factors. Procedia Eng. 186, 117–126. https://doi.org/10.1016/j.proeng.2017.03.217

Ferretti, A., Prati, C., Rocca, F., 2001. Permanent Scatterers in SAR Interferometry. IEEE Transactions on Geoscience and Remote Sensing. 39, 8–20. DOI: 10.1109/36.898661

Francis, R.A., Guikema, S.D., Henneman, L., 2014. Bayesian Belief Networks for predicting drinking water distribution system pipe breaks. Reliab. Eng. Syst. Saf. 130, 1–11. https://doi.org/10.1016/j.ress.2014.04.024

Gould, S.J.F., 2011. A Study of the Failure of Buried Reticulation Pipes in Reactive Soils. Monash University. PhD Thesis.

Hallett, S.H., Sakrabani, R., Keay, C.A., Hannan, J.A., 2017. Developments in land information systems : examples demonstrating land resource management capabilities and options. Soil Use Manag. 33, 514–529. https://doi.org/10.1111/sum.12380

Kabir, G., Tesfamariam, S., Loeppky, J., Sadiq, R., 2016. Predicting water main failures : A Bayesian model updating approach. Knowledge-Based Syst. 110, 144–156. https://doi.org/10.1016/j.knosys.2016.07.024

Kimutai, E., Betrie, G., Brander, R., Sadiq, R., Tesfamariam, S., 2015. Comparison of statistical models for predicting pipe failures: Illustrative example with the city of Calgary water main failure. J. Pipeline Syst. Eng. Pract. 6. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000196

Kleiner, Y., Rajani, B., 2001. Comprehensive review of structural deterioration of water mains : statistical models. Urban Water 3, 131–150. https://doi.org/10.1016/S1462-0758(01)00033-4

Li, S., Wang, S., Lin, X., 2013. Variable selection and estimation in generalized linear models with the seamless L0 penalty. Can. J. Stat. 40, 745–769. https://doi.org/10.1002/cjs.11165.Variable

Makar, J.M., Kleiner, Y., 2000. Maintaining water pipeline integrity. AWWA Infrastruct. Conf. Exhib.

Mercer, G., Reeves, A., 2011. The Relationship between Trees , Distance to Buildings and Subsidence Events on Shrinkable Clay Soil. Arboric. J. 33, 229–245.

Met Office, 2018. Met Office Rainfall Evapotranspiration Calculation System - Specialist Datasets [WWW Document]. URL https://www.metoffice.gov.uk/services/industry/data/specialist-datasets (accessed 21.5.18).

Morozova, O., Levina, O., Uusküla, A., Heimer, R., 2015. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. BMC Med. Res. Methodol. 15, 1–17. https://doi.org/10.1186/s12874-015-0066-2

Navarro, V., Candel, M., Yustres, A., S??nchez, J., Alonso, J., 2009. Trees, soil moisture and foundation movements. Comput. Geotech. 36, 810–818. https://doi.org/10.1016/j.compgeo.2009.01.008

Ogutu, G., Okuthe, P., Lall, M., 2017. A review of probabilistic modelling of pipeline leakage using bayesian networks. J. Eng. Appl. Sci. 12, 3163–3173. DOI: 10.3923/jeasci.2017.3163.3173

Östberg, J., Martinsson, M., Stål, Ö., Fransson, A.M., 2012. Risk of root intrusion by tree and shrub species into sewer pipes in Swedish urban areas. Urban For. Urban Green. 11, 65–71. https://doi.org/10.1016/j.ufug.2011.11.001

Rajani, B., Kleiner, Y., 2001. Comprehensive review of structural deterioration of water mains : physically based models. Urban Water 3. https://doi.org/10.1016/S1462-0758(01)00032-2

Rajani, B., Kleiner, Y., Sink, J., 2012. Exploration of the relationship between water main breaks and temperature covariates. Urban Water 9006, 67–84. https://doi.org/10.1080/1573062X.2011.630093

Rezaei, H., Ryan, B., Stoianov, I., 2015. Pipe failure analysis and impact of dynamic hydraulic conditions in water supply networks. Procedia Eng. 119, 253–262. https://doi.org/10.1016/j.proeng.2015.08.883

Rostum, J., 2000. Statistical Modelling of Pipe Failures in Water Networks. Nor. Univ. Sci. Technol. 1–132.

Ruiliang, P., 2013. Tree Species Classification. Remote Sens. Nat. Resour. 239–258. https://doi.org/doi:10.1201/b15159-19

Sheeren, D., Fauvel, M., Planque, C., Willm, J., Dejoux, J.F., 2015. Tree species discrimination in temperate woodland using high spatial resolution

Formosat-2 time series 1–4. IEEE Explore. Annecy, France. 10.1109/Multi-Temp.2015.7245792

Sowter, A., Bateson, L., Strange, P., Ambrose, K., Fifik, M., 2013. DInSAR estimation of land motion using intermittent coherence with application to the South Derbyshire and Leicestershire coalfields. Remote Sens. Lett. 4, 979–987. https://doi.org/10.1080/2150704X.2013.823673

Tabesh, M., Soltani, J., Farmani, R., Savic, D., 2009. Assessing pipe failure rate and mechanical reliability of water distribution networks using data-driven modeling. J. hydroinformatics 11, 1–17. https://doi.org/10.2166/hydro.2009.008

Torres, M.N., Rodríguez, J.P., Leitão, J.P., 2017. Geostatistical analysis to identify characteristics involved in sewer pipes and urban tree interactions. Urban For. Urban Green. 25, 36–42. https://doi.org/10.1016/j.ufug.2017.04.013

Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecol. Modell. 203, 312–318. https://doi.org/10.1016/j.ecolmodel.2006.11.033

Walton, J.T., Nowak, D.J., Greenfield, E.J., 2008. Assessing Urban Forest Canopy Cover Using Airborne or Satellite Imagery. Arboric. Urban For. 34, 334–340.

Williams, B., Hansen, G., Baraban, A., Santoni, A., 2015. A Practical Approach to Variable Selection — A Comparison of Various Techniques, Casualty Actuarial Society E-Forum. pp. 1–20.

Wilson, D., Filion, Y., Moore, I., 2017. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. Urban Water J. 14, 173–184. https://doi.org/10.1080/1573062X.2015.1080848

Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W., Tscheikner-Gratl, F., 2018. Pipe failure modelling for water distribution networks using boosted decision

trees. Struct. Infrastruct. Eng. 2479, 1–10. https://doi.org/10.1080/15732479.2018.1443145

Xu, Q., Chen, Q., Li, W., Ma, J., 2011. Pipe break prediction based on evolutionary data-driven methods with brief recorded data. Reliab. Eng. Syst. Saf. 96, 942–948. https://doi.org/10.1016/j.ress.2011.03.010

Yamijala, S., Guikema, S.D., Brumbelow, K., 2009. Statistical models for the analysis of water distribution system pipe break data. Reliab. Eng. Syst. Saf. 94, 282–293. https://doi.org/10.1016/j.ress.2008.03.011

Zhang, C., Qiu, F., 2012. Mapping individual tree species in an urban forest using airborne lidar data and hyperspectral imagery. Photogramm. Eng. Remote Sens. 78, 1079–1087. https://doi.org/10.14358/PERS.78.10.1079

Zhang, Z., 2016. Variable selection with stepwise and best subset approaches. Ann. Transl Med. 4, 136. https://doi.org/10.21037/atm.2016.03.35